Statistical Genetics

Summer Institute in Statistical Genetics
University of Washington
July 17-19, 2019

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/SISG19/

SpeakUp
Tutorial

What is it?

You're in class and no one dares to ask a question or, the opposite, everyone wants to participate and there's no time. Does it sound familiar to you? To improve these interactions we present the **free SpeakUp app**.

With no login, SpeakUp allows you to **create** a chat room. People around you will see it and will be able to **join** it. Once inside, they can **interact** in it by writing and rating anonymous messages. Finally it is possible to create **multiple-choice questions** to poll the audience.

SpeakUp is developed by the non-profit Seance association in collaboration with the University of Lausanne and the Swiss Federal Institute of Technology in Lausanne.

UNIL | Université de Lausanne

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Download SpeakUp from the AppStore or Google Play. You can also access it through any browser on speakup.info. For further information, contact adrian.holzer@epfl.ch

Google play

Available on the App Store

Q1: I currently live in:

- **A** 43 North America.
- **B** 1South America.
- **C** 0 Africa.
- **D** 4 Asia.
- **E** 1 Europe.
- **F** 1 Rest of the world.

Q2: I am a:

- **A** 29 Student in biological sciences.
- **B** 13Student in mathematical sciences.
- **C** 5 Postdoc or faculty.
- **D** 1 Private sector scientist.
- **E** 3 Public sector scientist.
- **F** 4 None of the above.

Q3: I know most about:

- **A** 14 Mathematics and statistics.
- **B** 2 Computer science.
- **C** 22 Genetics.
- **D** 11 Other biological sciences.
- **E** 3 Something else.

Q4: I study or work on:

- **A** 28 Humans.
- **B** 7 Non-human animals other than fish.
- **C** 2 Fish.
- **D** 6 Plants.
- **E** 5 Micro organisms.
- **F** 4 I do not study or work on biological material.

Q5: The organisms I work with are:

- **A** 37 Diploid.
- **B** 4 Haploid.
- **C** 3 Polyploid.
- **D** 8 I don't work with organisms.

Q6: The data I work with are:

- **A** 7 Non-genetic.
- **B** 4 Microsatellite.
- **C** 26 DNA sequence.
- **D** 13 Other omic data.
- **E** I don't work with data.

Q7: About R, I:

- **A** 4 Have no experience with R.
- **B** 4 Have run an R program someone else gave me.
- **C** 15 Have downloaded and run an R package.
- **D** 28 Have written and run an R program.
- **E** Have written and distributed an R package.

Q8: I have:

- **A** 6 Performed a test for Hardy-Weinberg equilibrium.
- **B** 5 Estimated F_{ST}.
- **C** Estimated kinship.
- **D** 7 Tested for association between a marker and a trait.
- Two or more of **A**, **B**, **C** or **D**.
- 33 None of the above.

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)

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

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

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
For data freeze 5 of this study:

Sequence analysis identified 410,323,831 genetic variants (381,343,078 SNVs and 28,980,753 indels), corresponding to an average of one variant per 7 bp throughout the reference genome. Among all variant alleles, 46.0% were observed once across all samples (i.e. singletons).

There is an average of 3.78 million variants in each studied genome. Among these, an average of 30,207 were novel (0.8%) and 3,510 were singletons (0.1%). Thus while there are vast numbers of rare variants in humans, only a few of these are present in each genome.

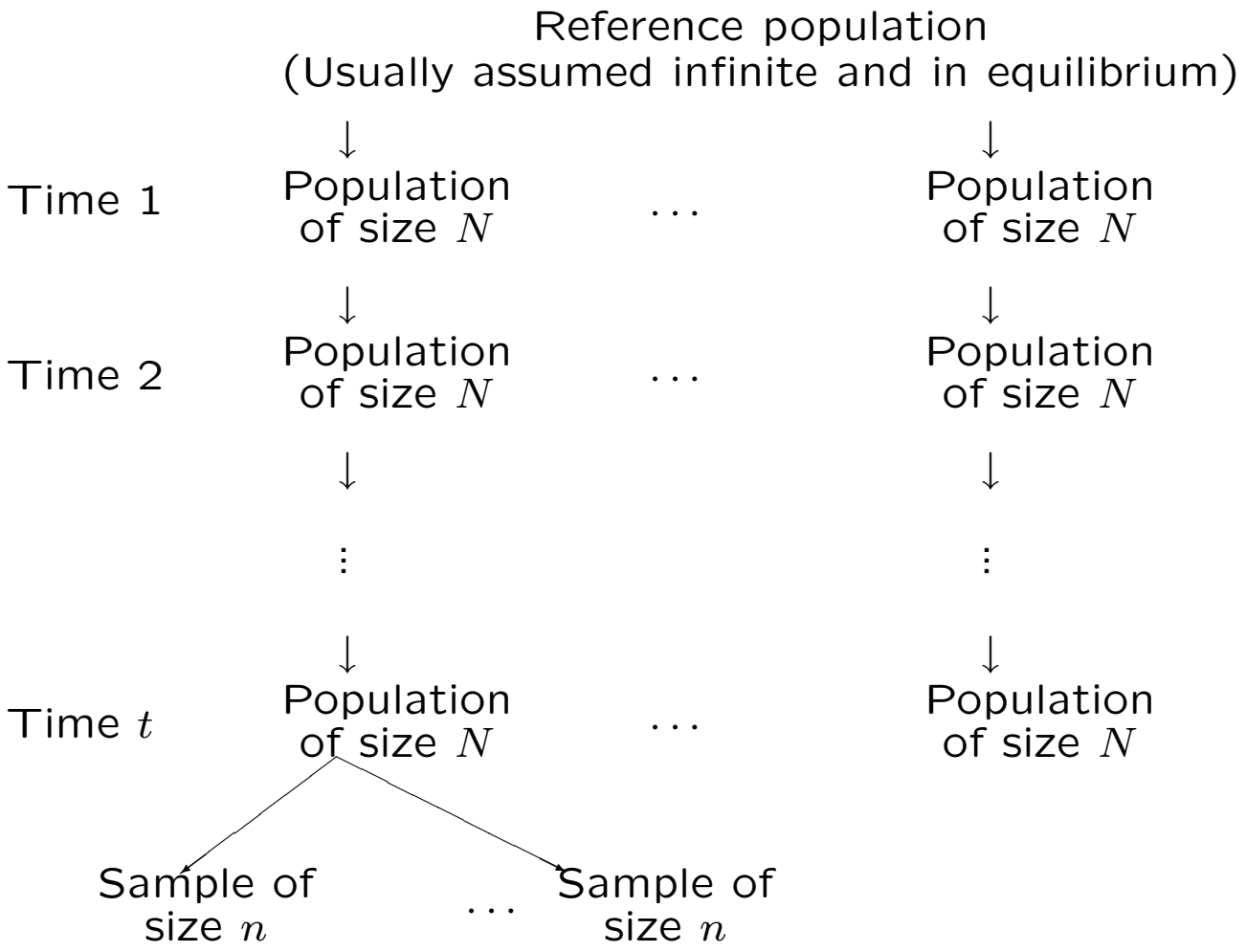
Currently over 1 billion variants found from 140,000 whole-genome sequences.

Sampling

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

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

Classical Model



Coalescent Theory

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

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

Probability

Probability provides the language of data analysis.

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

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

Subjective probability:
Probability is a measure of belief.

First Law of Probability

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

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

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

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

Second Law of Probability

If G and H are mutually exclusive events, then:

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

For example,

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

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

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

Complementary Probability

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

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

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

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

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

Third Law of Probability

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

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

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

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

For example

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

Independent Events

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

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

and

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

Events G, H are independent.

Law of Total Probability

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

$$\Pr(E) = \Pr(E|G)\Pr(G) + \Pr(E|\bar{G})\Pr(\bar{G})$$

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

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

For example

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

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:

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

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

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

F_2 generation

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

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

Seed Characters

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

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

Plant Characters

Model for stem length is

Genotype	Phenotype
GG	Long
Gg	Long
gg	Short

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

F_2 Plant Character

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

Fisher's 1936 Criticism

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

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

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

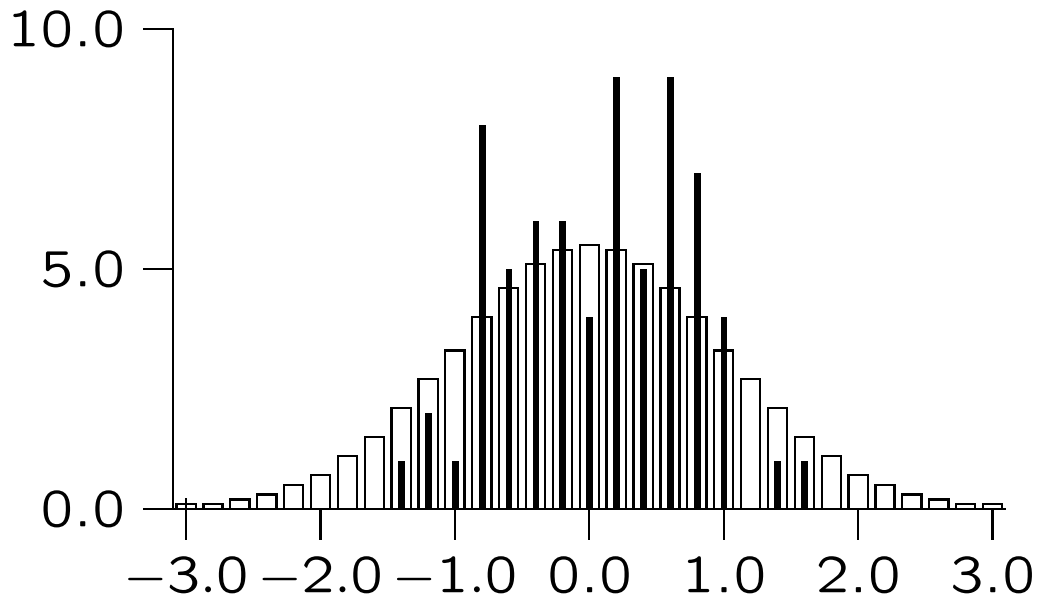
Mendel had 69 comparisons where the expected ratios were correct. Each set of data can be tested with a chi-square test:

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

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

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. 2008. "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."

2018 paper

"According to Fisher (1959), if the null hypothesis is rejected, 'The force with which such a conclusion is supported is that of the simple disjunction: Either an exceptionally rare chance has occurred, or the theory of random distribution is not true' (p. 39). Fisher's theory does not permit one to say which of the two possibilities is the case, nor to give a probability for it. Furthermore, if significance is not achieved, nothing can be concluded. In order for the probability distribution that forms the basis of a chi-square test to be valid, the hypothesis to be tested must be declared before the data are examined.

(continued on next slide)

2018 paper

Viewed in this light, there are several gaps between Fisher's calculations and his conclusion. Fisher is rejecting the multinomial null hypothesis if the chi-square is too small, which would be legitimate if the hypothesis test were declared before Weldon pointed the way, or if Fisher routinely used a two-tailed chi-square test. Neither is the case. And one still has Fisher's disjunction to contend with. Nonetheless, Fisher is a superb data-analyst, and we should not be interpreted as challenging his conclusion."

Kadane JB, Wang Z. 2018. Sums of possibly associated multivariate indicator functions; the Conway-Maxwell-Multinomial distribution. Brazilian Journal of Probability and Statistics 32:583-596.

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 of independent tosses is

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

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

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

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

The binomial probability of x successes in n trials is

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

Binomial Likelihood

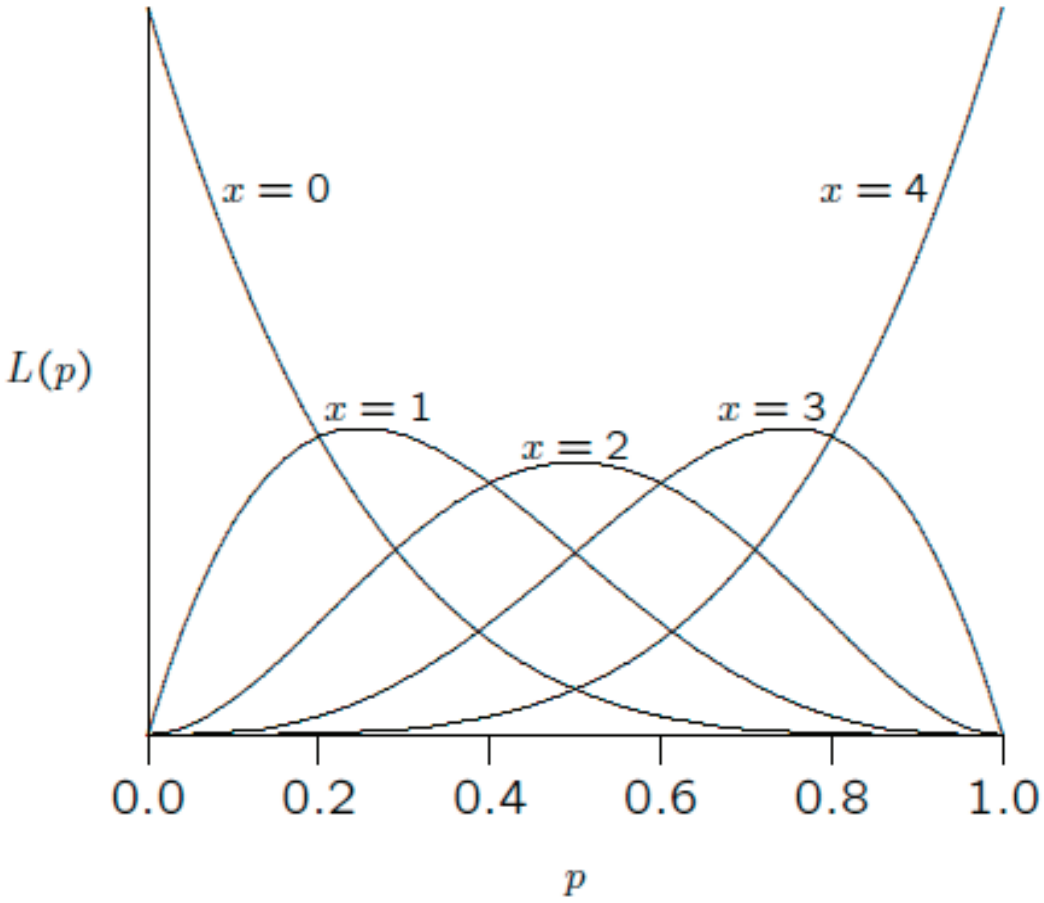
The quantity $\Pr(x|p)$ is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

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

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

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

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



Binomial Mean

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

The *sample proportion* of successes is

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

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

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

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

Binomial Variance

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

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

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

Normal Approximation

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

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

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

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

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

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

Confidence Intervals

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

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,

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

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

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

Exact Confidence Intervals: Two-sided

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

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

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

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 $n = 100$.

Exact CIs for $n = 10$

One-sided			Two-sided			
x	$\tilde{p}$	p_U	x	p_L	$\tilde{p}$	p_U
0	0.00	0.26	0	0.00	0.00	0.31
1	0.10	0.39	1	0.00	0.10	0.45
2	0.20	0.51	2	0.03	0.20	0.56
3	0.30	0.61	3	0.07	0.30	0.65
4	0.40	0.70	4	0.12	0.40	0.74
5	0.50	0.78	5	0.19	0.50	0.81
6	0.60	0.85	6	0.26	0.60	0.88
7	0.70	0.91	7	0.35	0.70	0.93
8	0.80	0.96	8	0.44	0.80	0.97
9	0.90	0.99	9	0.55	0.90	1.00
10	1.00	1.00	10	0.69	1.00	1.00

The two-sided CI is not symmetrical around $\tilde{p}$.

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

For a SNP with alleles A, B there are three genotypes:

AA	P_{AA}
AB or BA	P_{AB}
BB	P_{BB}

The probability of x lots of AA is $(P_{AA})^x$, 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 AA, and y of AB or BA and z of BB in n trials is:

$$\Pr(x, y, z) = \frac{n!}{x!y!z!}(P_{AA})^x(P_{AB})^y(P_{BB})^z$$

Multinomial Variances and Covariances

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

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

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

For the sample counts:

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

Allele Frequency Sampling Distribution

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

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

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

Within-population Variance

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

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

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

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

Within-population Variance

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

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

Then the variance can be written as

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

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

Bounds on f

Since

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

there are bounds on f:

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

or

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

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

An aside: 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.

An aside: 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:

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

and

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

so

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

Allele Dosage

The dosage X of allele A for an individual is the number of copies of A (0,1,2) that individual carries (the sum of its two allele indicators).

The probabilities for X are

$$\Pr(X = 0) = P_{aa}, \Pr(X = 1) = P_{Aa}, \Pr(X = 2) = P_{AA}$$

so the expected value of X is $2P_{AA} + P_{Aa} = 2p_A$.

The expected value of X^2 is $4P_{AA} + P_{Aa} = 2(p_A + P_{AA})$ and this leads to a variance the dosage for an individual of

$$\text{Var}(X) \;=\; 2P_{AA} + 2p_a - 4p_A^2 = 2p_A(1 - p_A)(1 + f)$$

We will come back to this result, but note here that the f term is usually not included in genetic data analysis packages.

Maximum Likelihood Estimation: Binomial

For a sample of n alleles, the likelihood of p_A when there are n_A alleles of type A is

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

and this 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 $p_A = n_A/n$. The MLE of p_A is its sample value: $\widehat{p}_A = \tilde{p}_A$.

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?

Maximum Likelihood Estimation

Because

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

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, algebraically, 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 genotype proportions based on the estimates of p_A and f:

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

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):

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

$$\hat{f} = 1 - \frac{n_{Aa}}{2n\tilde{p}_A(1-\tilde{p}_A)} = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

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. For example

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

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 regardless of degrees of freedom. 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 A_u, two possible MoM estimates are (for large sample sizes)

$$\hat{f}_W = \frac{\sum_{u=1}^{m}(\tilde{P}_{uu} - \tilde{p}_u^2)}{\sum_{u=1}^{m}\tilde{p}_u(1-\tilde{p}_u)}$$

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

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, and a sample of n individuals has n_{aa} recessive homozygotes. The genotypes of the other $(n - n_{Aa})$ individuals can be AA or Aa.If there is Hardy-Weinberg equilibrium, the likelihood for the two phenotypes is

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

Differentiating wrt p_a:

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

Setting this to zero leads to an equation that can be solved explicitly: $p_a = \sqrt{n_{aa}/n}$.

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 counts of the AA and Aa genotypes. Only the joint count $(n - n_{aa})$ of $AA + Aa$ is observed.

Estimate the missing genotype counts (assuming independence of alleles) as proportions of the total count of dominant phenotypes:

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

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

EM Algorithm for Recessive Alleles

Maximize the likelihood (using Bailey's method):

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

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

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

EM Algorithm for Two Loci

A more interesting applicatio of the EM algorithm is the estimation of two-locus gamete frequencies from unphased genotype data. For two loci with two alleles each, the ten two-locus frequencies are:

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

EM Algorithm for Two Loci

Gamete frequencies are marginal sums:

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

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

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

EM Algorithm for Two Loci

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

Assume initial value of p_{AB} and *Estimate* the missing counts as proportions of the total count n_{AaBb} 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

$$\begin{aligned} p_{AB} &= \frac{1}{2n}\left(2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB}\right) \\ &\text{or} \\ n_{AB} &= 2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB} \end{aligned}$$

Example

As an example, consider the data for SNPs rs7546189 and rs1128474 on slide 20:

	BB	Bb	bb	Total
AA	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

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

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = 12 - x$	$n_A = 12$
a	$n_{aB} = 6 - x$	$n_{ab} = x + 12$	$n_a = 18$
Total	$n_B = 6$	$n_b = 24$	$2n = 30$

$$0 \leq x \leq 6$$

Example

EM iterative equation:

$$
\begin{aligned}
x' &= 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB/ab} \\
&= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\
&= 0 + 0 + 1 + 3 \times \frac{2x(x+12)}{2x(x+12) + 2(12-x)(6-x)} \\
&= 1 + \frac{3x(x+12)}{x(x+12) + (12-x)(6-x)}
\end{aligned}
$$

Example

A good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (30 \times 12/30 \times 6/30) = 2.4$. Successive iterates are:

Iterate	x	$x/2n$
1	2.4000	0.0800
2	2.5000	0.0833
3	2.5647	0.0855
4	2.6063	0.0869
5	2.6327	0.0878
6	2.6494	0.0883
7	2.6600	0.0887
8	2.6667	0.0889
9	2.6709	0.0890
10	2.6736	0.0891
11	2.6752	0.0892
12	2.6763	0.0892
13	2.6769	0.0892
14	2.6773	0.0892
15	2.6776	0.0893
16	2.6778	0.0893
...	...	...

ALLELIC ASSOCIATION

Hardy-Weinberg Law

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

For a locus with two alleles, A, a:

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

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

For a locus with several alleles A_i:

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

Why would HWE not hold?

- Natural selection.
- LD with trait in trait-only sample.
- Population Structure/Admixture.
- Problems with data.
- etc.

Problems with Data

A SNP with genotype counts $40, 0, 60$ for AA, AB, BB is likely to cause HW rejection. What about $4, 0, 96$?

Typing systems may report heterozygotes as homozygotes, as was the likely explanation for
"To justify applying the classical formulas of population genetics in the Castro case, the Hispanic population must be in Hardy-Weinberg equilibrium. In fact, Lifecodes' own data show that it is not. ... Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium: 17 per cent observed homozygotes at D2S44 and 13 per cent observed homozygotes at D17S79 compared with only 4 per cent expected at each locus, indicating, perhaps not surprisingly, the presence of genetically distinct subgroups within the Hispanic sample."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339:501-505.

Population Structure

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 1 has 51% heterozygotes instead of the expected 49.8%.

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

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

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

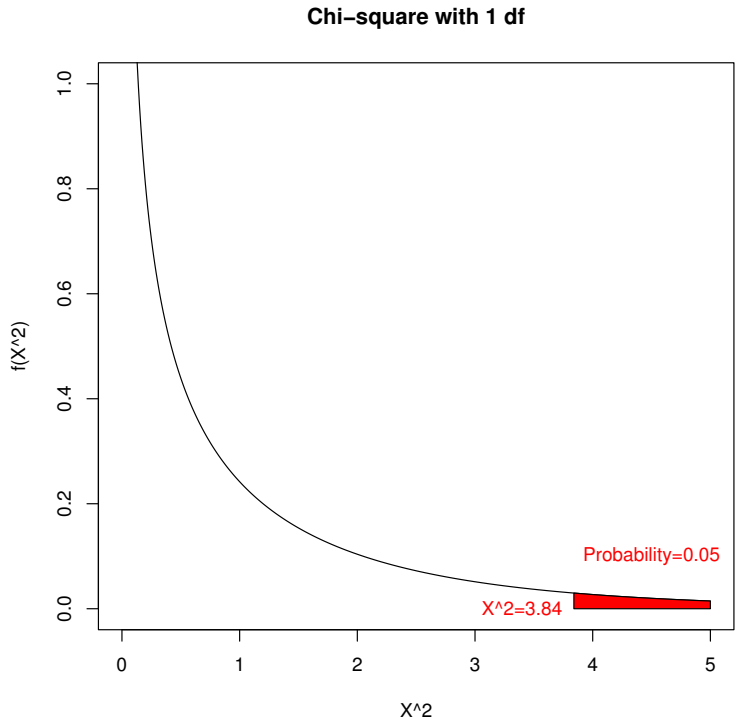
Inference about HWE

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:

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

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

Then form "expected" counts:

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

Goodness-of-fit Test

Perform the chi-square test:

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

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

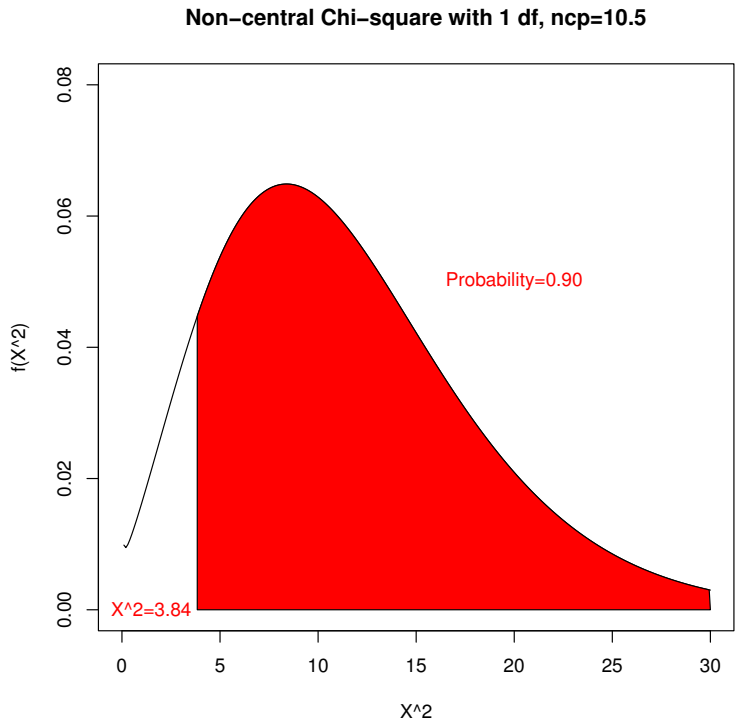
Sample size determination

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

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 $\lambda \geq 10.51$.

```
> pchisq(3.84,1,0)
[1] 0.9499565
> pchisq(3.84,1,10.51)
[1] 0.09986489
> qchisq(0.95,1,0)
[1] 3.841459
> qchisq(0.10,1,10.51)
[1] 3.843019
```

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 $\lambda = 10.5$.

Sample size determination

To achieve 90% power to reject HWE at the 5% significance level when the true inbreeding coefficient is f, need sample size n to make $nf^2 \geq 10.51$.

For $f = 0.01$, need $n \geq 10.51/(0.01)^2 = 105,100$.

For $f = 0.05$, need $n \geq 10.51/(0.05)^2 = 4,204$.

For $f = 0.10$, need $n \geq 10.51/(0.10)^2 = 1,051$.

Significance Levels and p-values

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

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

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

Exact HWE Test

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

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

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

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

Exact HWE Test

Putting these together gives the conditional probability

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

Reject the Hardy-Weinberg hypothesis if this quantity, the probability of the genotypic array conditional on the allelic array, is considered too small to allow that outcome if HWE holds. Is the probability for the data 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	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

The p-value is 0.01 and HWE is rejected at the 5% level.

Exact HWE Test Example

In this example, $\widehat{f} = 0$ and the chi-square test statistic is $X^2 = 50$. The resulting p-value is 1.54×10^{-12}, substantially different from the exact value of 0.01.

```
> 1-pchisq(50,1,0)
[1] 1.537437e-12
```

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	$\frac{10!}{7!1!2!}\frac{2^1 15!5!}{20!} = \frac{15}{323} = 0.047$

The p-value is $0.433 + 0.047 = 0.480$. Compare this to the 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 probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true and are conditional on the observed allele frequencies. 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

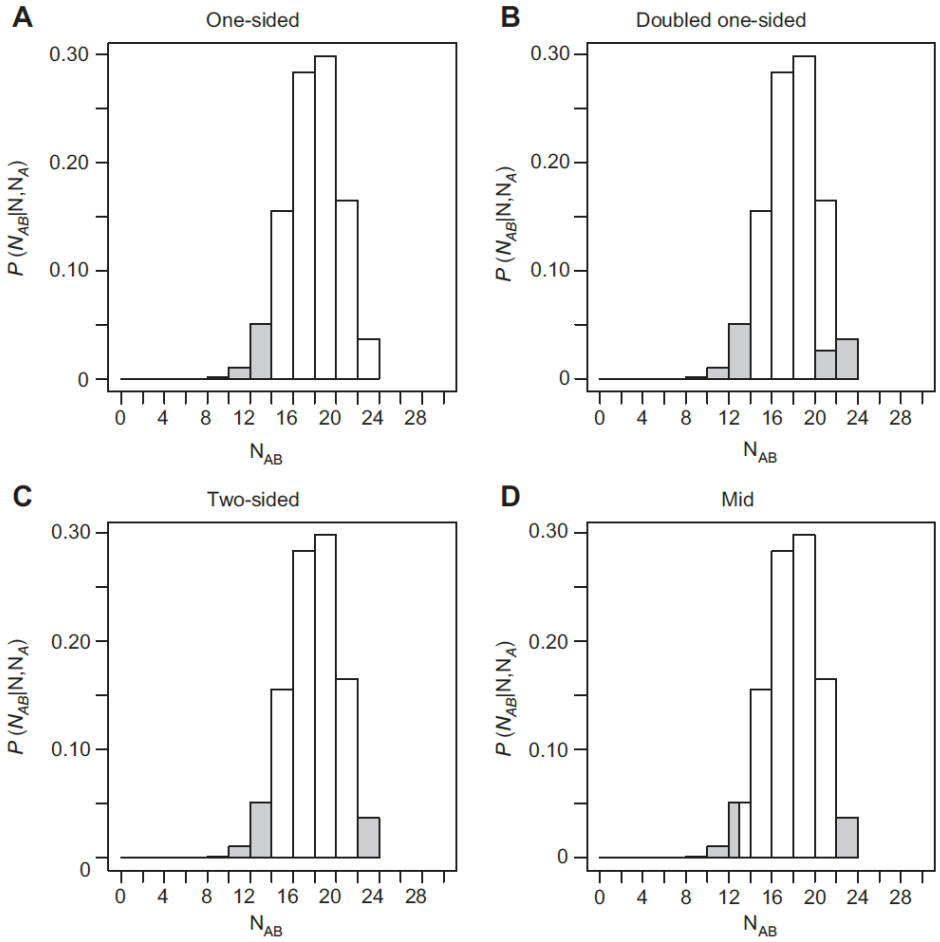


Figure 1 Computation of the *p*-value in an exact test for HWP, for a sample of 50 individuals with a minor allele count of 23, for which 13 heterozygotes were observed. (A) One-sided *p*-value in a test for heterozygote dearth. (B) *p*-value obtained by doubling the one-sided tail. (C) Standard two-sided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

Usual vs Mid p values

AA	Aa	aa	$\Pr(n_{Aa}\|n, n_A)$	p value	
				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

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

Even though the nominal significance level for a HWE test may be set at 0.05, for example, the actual significance level can be quite different. (e.g. 0.0353 vs 0.05 on the previous slide.)

The difference between nominal and actual values depends on the sample size and the minor allele frequency, as shown on the next slide.

Graffelman and Moreno, 2013

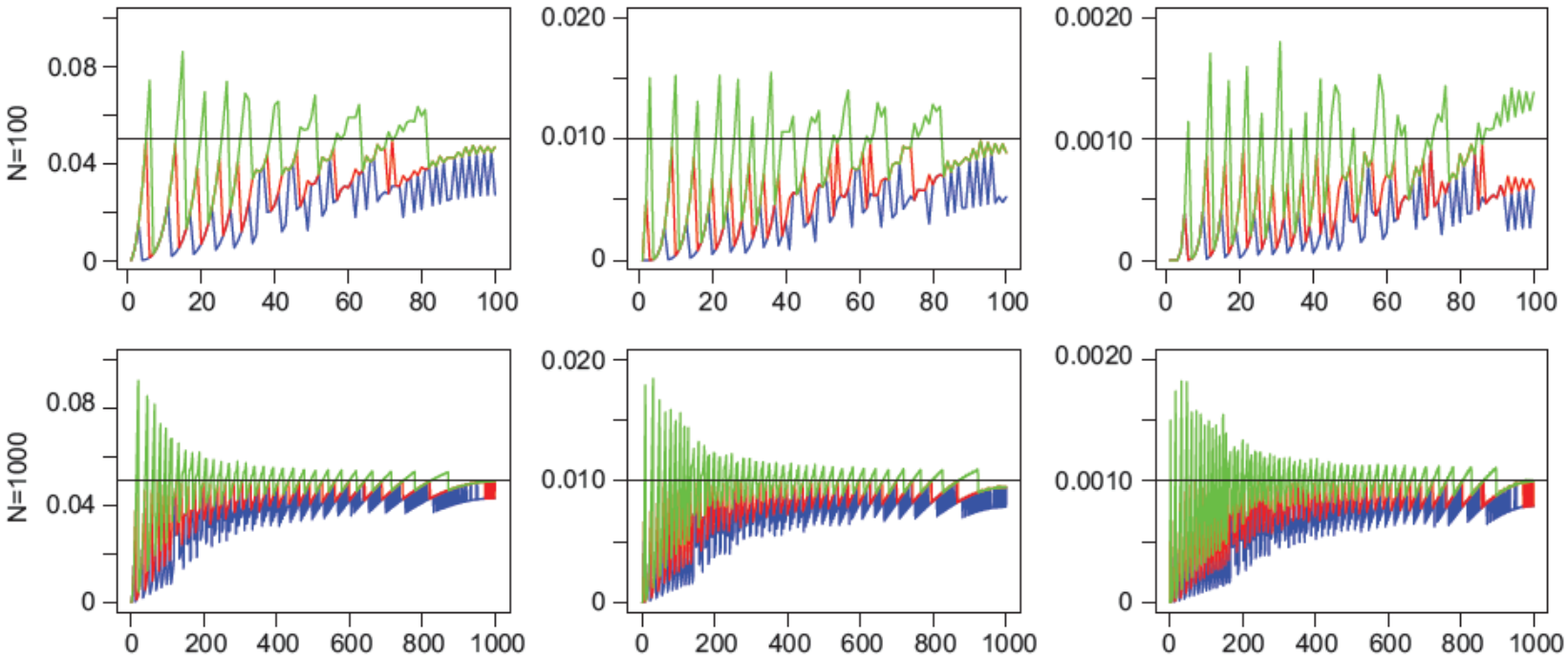


Figure 2 Type I error rate against minor allele count for different sample sizes (25, 50, 100 and 1000) and significance levels (0.05, 0.01, and 0.001) for exact tests with standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

Power of Exact Test

Calculating the power of an HWE test is easy for the chi-square test statistic as it follows from the non-central chi-square distribution.

It is more complicated for the exact test. If there is not HWE:

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

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

Power of Exact Test

Once the rejection region has been determined, the power of the test (the probability of rejecting) can be found by adding these probabilities for all sets of genotype counts in the region. HWE corresponds to $\psi = 2$. What is the power to detect HWE when $\psi = 1 (f > 0)$, the sample size is $n = 10$ and the sample allele frequencies are $\tilde{p}_A = 0.75, \tilde{p}_a = 0.25$? Note that $C = 1/[1/(5!5!0!) + 1/(6!3!1!) + 1/(7!1!2!)]$.

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

The $\psi = 2$ column shows that the rejection region is $n_{Aa} = 1$. The $\psi = 1$ column shows that the power (the probability $n_{Aa} = 1$ when $\psi = 1$) is 37.4%.

Power Examples

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

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

Graffelman and Moreno, 2013

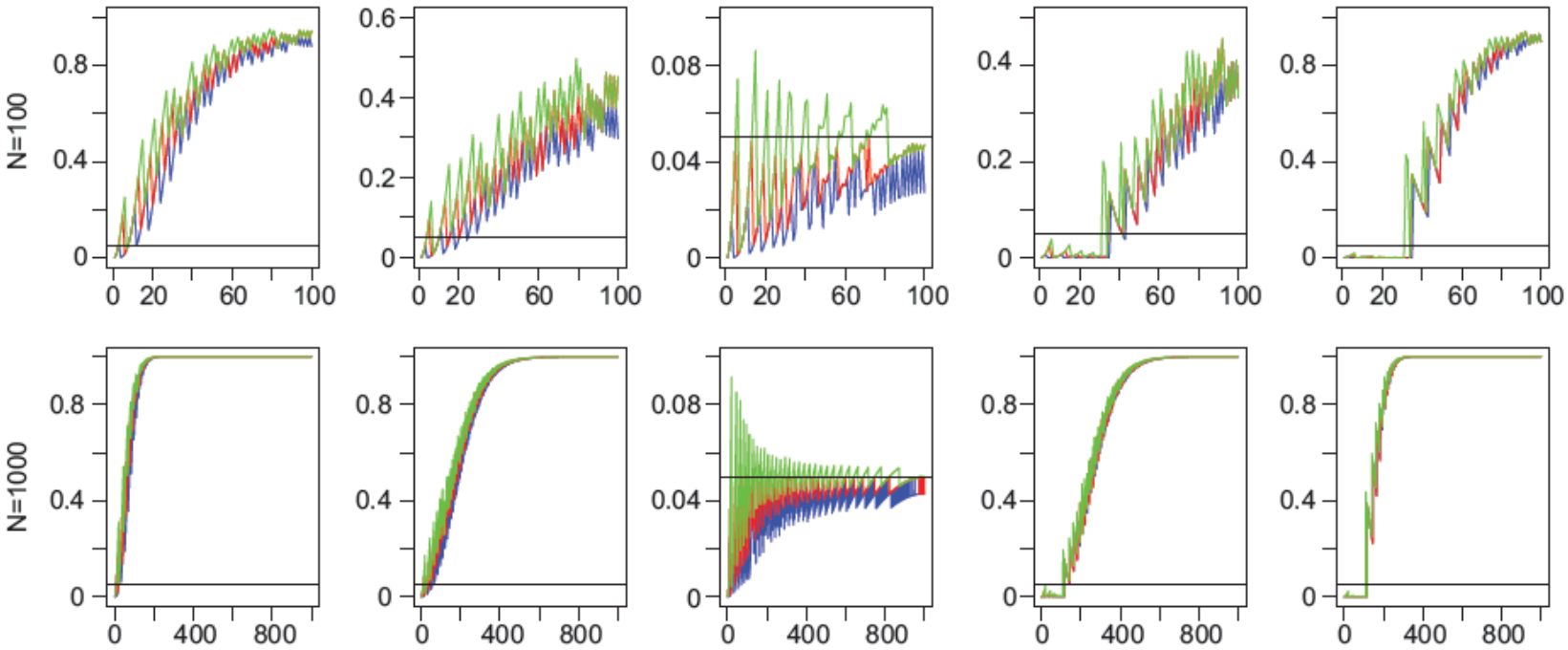


Figure 3 Power of HWP exact tests against minor allele count for different sample sizes (25, 50, 100 and 1000) and degree of disequilibrium (1, 2, 4, 8 and 16). Standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

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 about 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 $L = 10^6$, the "genome-wide significance level" is 5×10^{-8} in order for $\alpha = 0.05$.

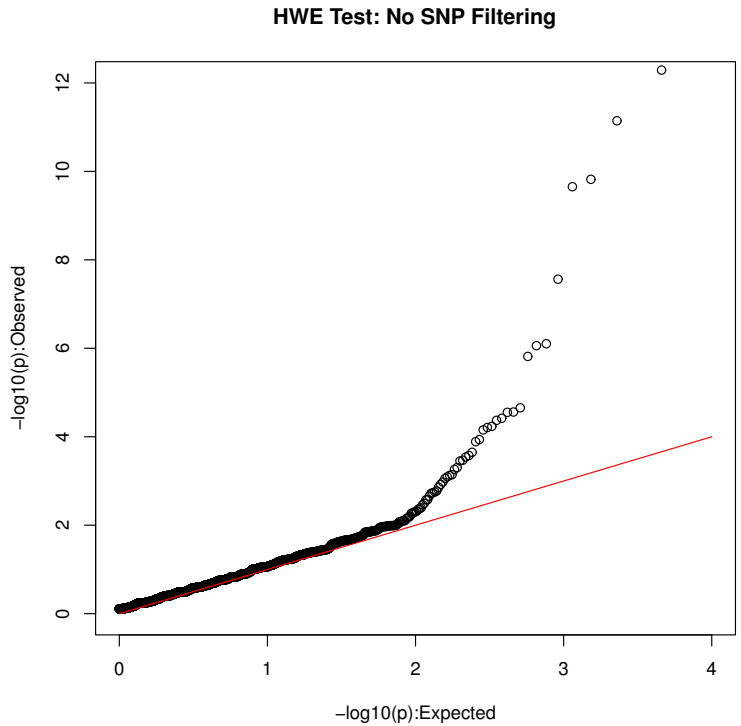
QQ-Plots

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. $1/2n, 3/n, \ldots, (2n-1)/2n$. 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.

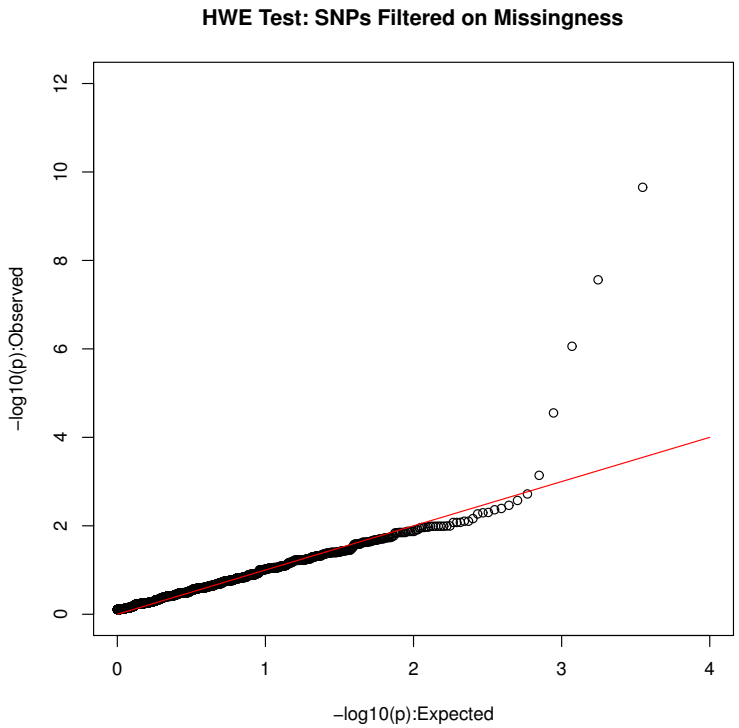
A useful diagnostic for QQ-plots is the "genomic control" quantity λ. This is the ratio of the median of the observed distribution of the test statistic to the expected median. We have calculated this from the p-values of the exact test statistics, and assumed these have a uniform distribution on [0,1], and a median of 0.5, under the null hypothesis of HWE. The ratio should be 1.

QQ-Plots



The results for 9208 SNPs on human chromosome 1 for the 50 AMD controls ($\lambda = 0.86$). 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 ($\lambda = 1.035$, closer to 1 than for all the SNPs). 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).

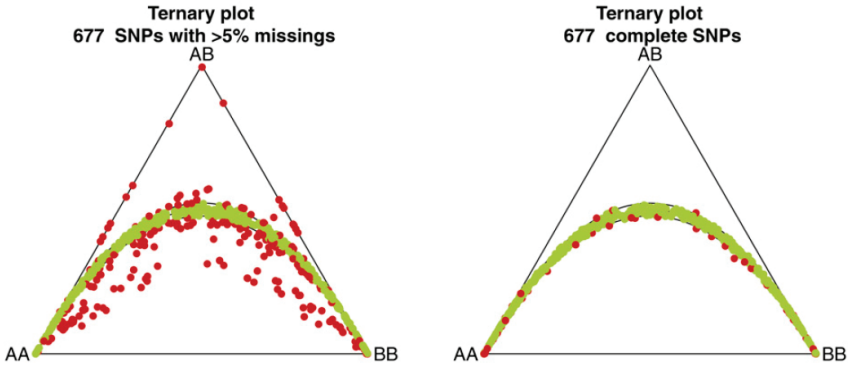
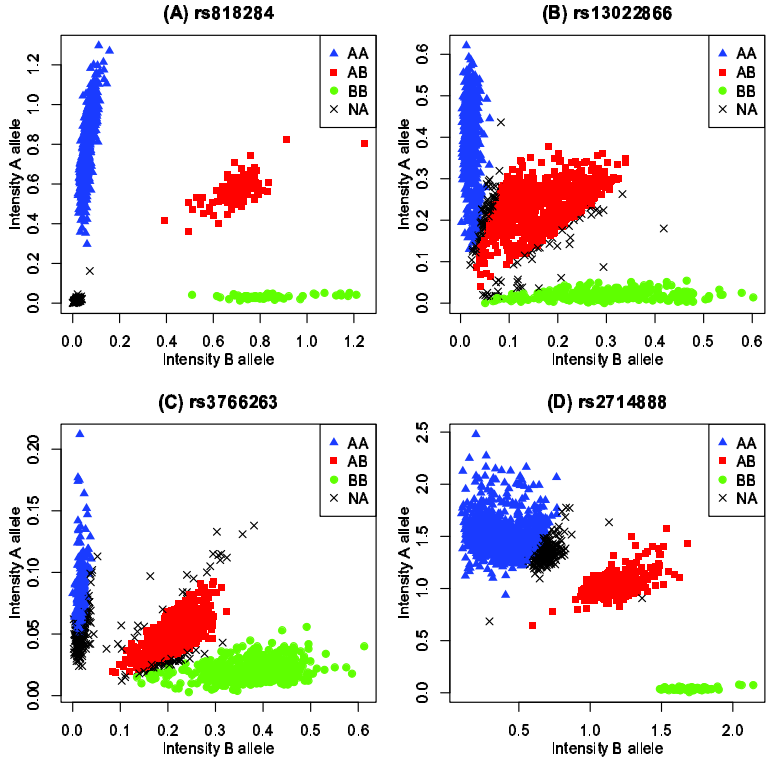


Figure 1 Left panel: ternary plot for 677 single-nucleotide polymorphisms (SNPs) with >5% missing. A total of 229 SNPs (34%) are significant in a χ^2 test. Right panel: 677 SNPs without missings taken at random. A total of 56 (8%) SNPs are significant in a χ^2 test. Significant markers are red and nonsignificant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

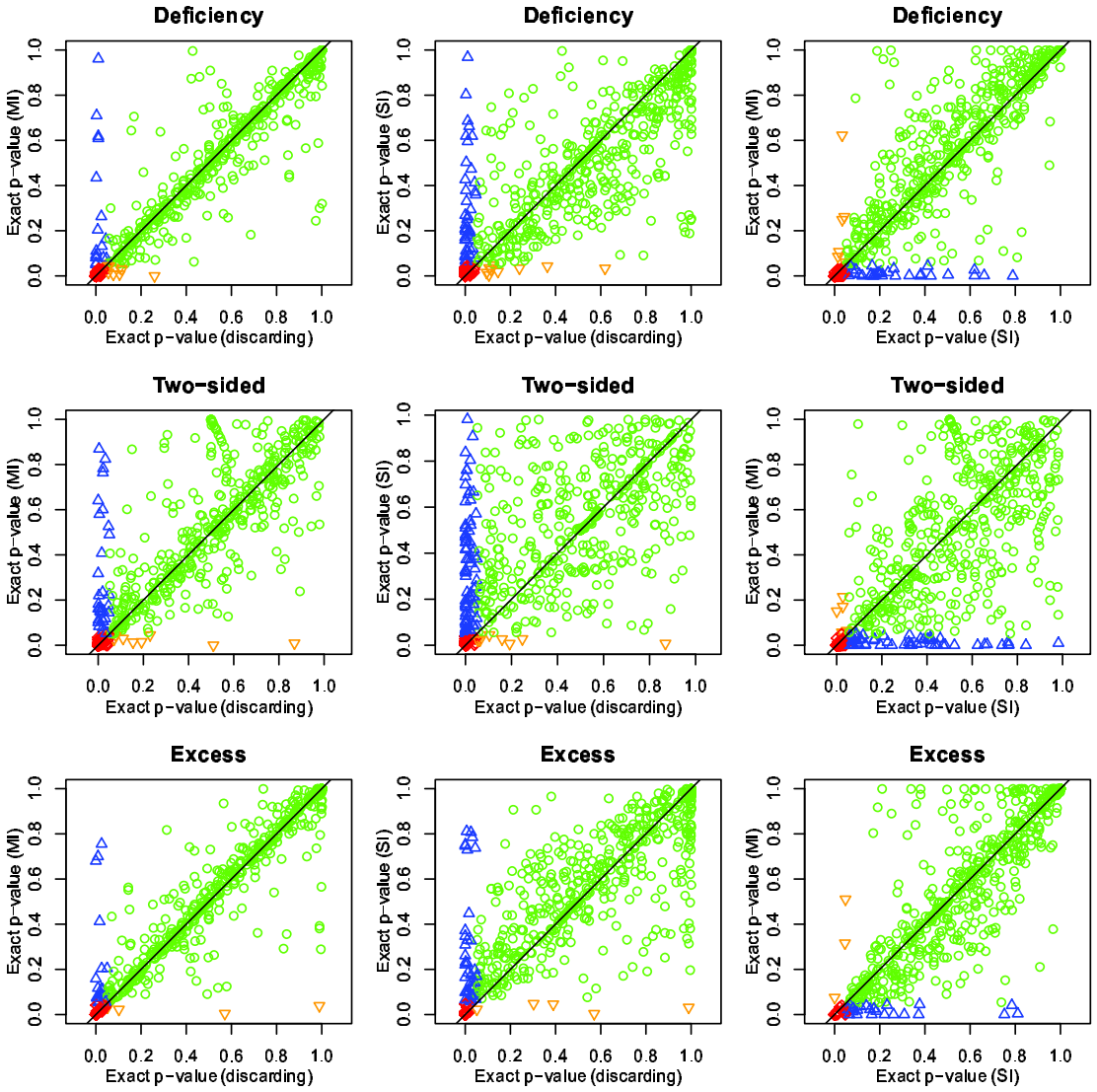
DeFinetti diagram: distance of point to side of triangle is frequency of genotype shown on opposite vertex.

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. Should 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

X-linked Markers: Real Data

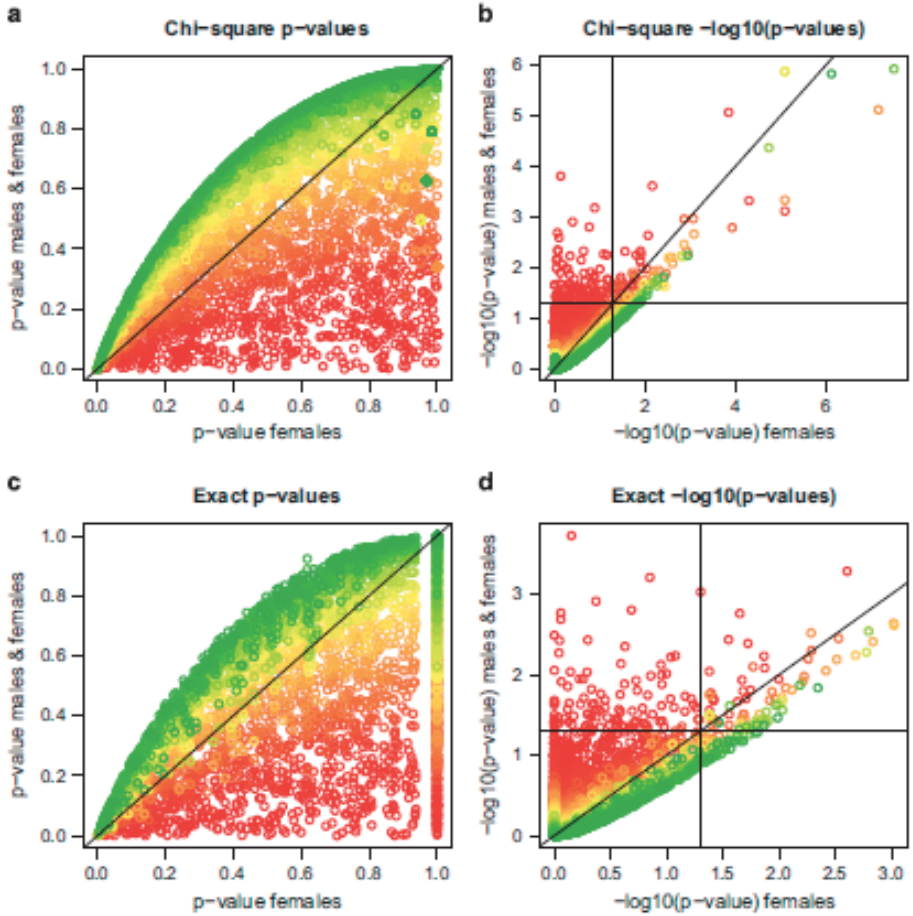


Figure 6 Scatter plots of *P*-values for χ^2 tests and exact tests for HWE using females only and using both males and females for 4158 SNPs at the X chromosome of the venous thrombosis database. The horizontal and vertical black lines in (**b**) and (**d**) corresponds to a significance level of 5%.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

X-linked Markers: Real Data

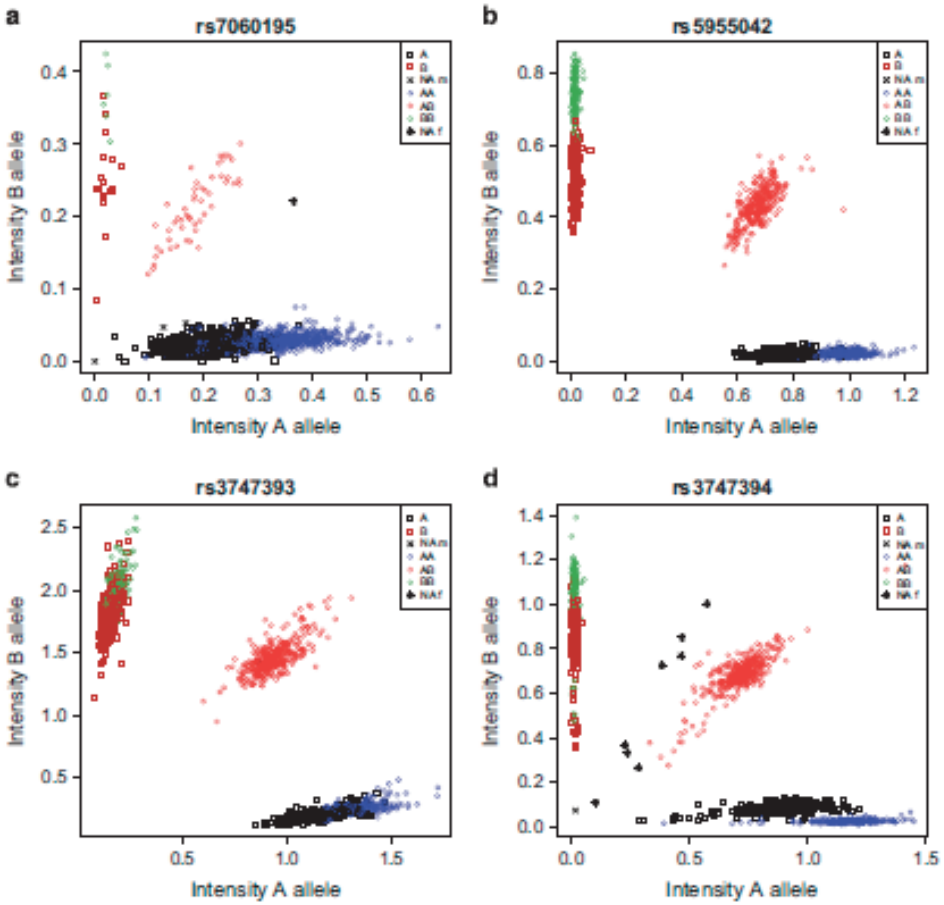


Figure 8 Cluster plots of allele intensities of four SNPs of the venous thrombosis database that are significant (**a**, **b**, **c**) in an all-individual exact test for HWE.

Separate Male and Female Autosomal Counts

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

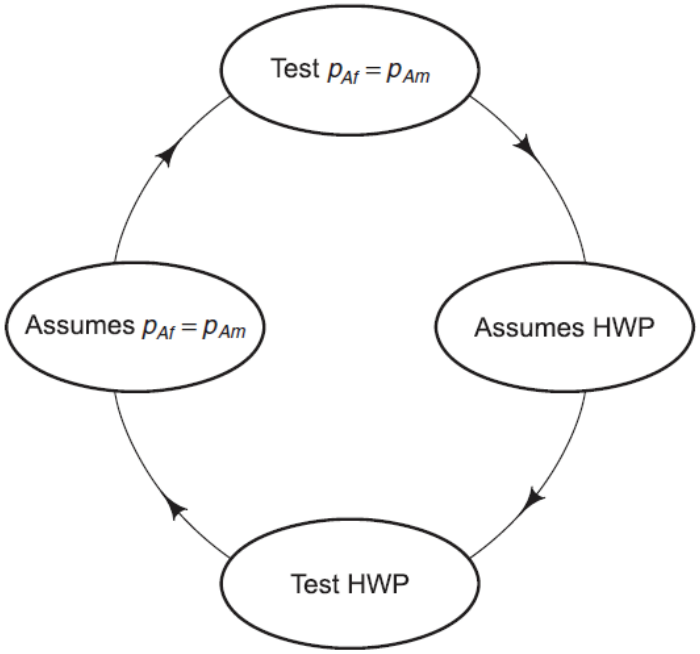


FIGURE 1 Vicious testing circle: mutual dependency of a test for EAF in males and females and a test for HWP

Notes: *A* allele frequencies in males and females are represented by p_{Am} and p_{Af}, respectively.

Graffeleman J, Weir BS. 2018. Genetic Epidemiology 42:24-48.

Separate M&F Counts: Scenarios

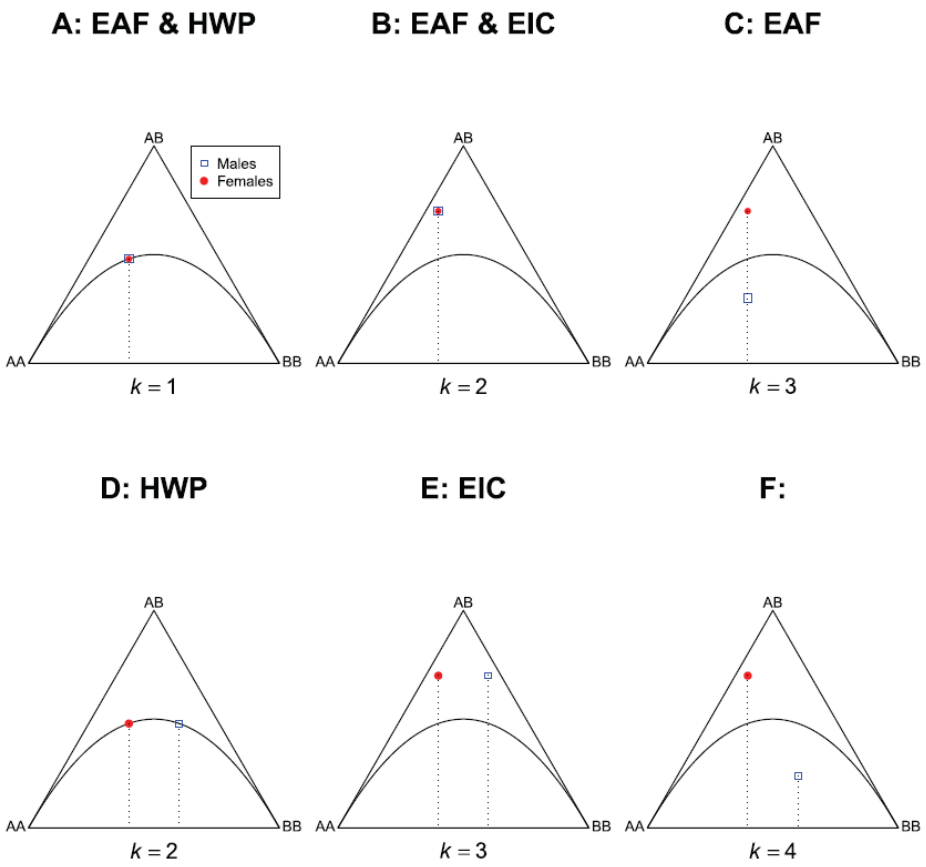


FIGURE 2 Ternary diagrams for male and female genotype frequencies

Notes: (A) HWP and EAF. (B) Equality of inbreeding coefficients, EAF, and both sexes out of HWP. (C) Unequal inbreeding coefficients, both sexes out of equilibrium but with equal allele frequencies. (D) Both sexes in HWP but with different allele frequencies. (E) Each sex out of equilibrium with identical inbreeding coefficients and different allele frequencies. (F) Both sexes out of equilibrium, with different inbreeding coefficients and different allele frequencies. The number of free parameters k is given below the basis of each scenario.

Separate M&F Counts: Joint Exact Test

To test for both Equal Allele Frequencies (EAF) and Hardy-Weinberg Proportions (HWP):

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

m_{AA}, m_{AB}, m_{BB}	genotype counts in males
f_{AA}, f_{AB}, f_{BB}	genotype counts in females
$n_m = m_{AA} + m_{AB} + m_{BB}$	number of males
$n_f = f_{AA} + f_{AB} + f_{BB}$	number of females
$n = n_m + n_f$	total sample size
$m_A = 2m_{AA} + m_{AB}, m_B = 2m_{BB} + m_{AB}$	numbers of A, B alleles in males
$f_A = 2f_{AA} + f_{AB}, f_B = 2f_{BB} + f_{AB}$	numbers of A, B alleles in females
$n_A = m_A + f_A, n_B = m_B + f_B$	total numbers of A, B alleles

Separate M&F Counts: HWP Exact Test

To test for HWP:

$$\Pr(n_{AB}|n, n_A) = \frac{n_A! n_B! n! 2^{n_{AB}}}{n_{AA}! n_{AB}! n_{BB}!}$$

n_{AA}, n_{AB}, n_{BB}	total genotype counts in males and females
$n = n_{AA} + n_{AB} + n_{BB}$	total sample size
$n_A = 2n_{AA} + n_{AB}, n_B = 2n_{BB} + n_{AB}$	total numbers of A, B alleles

Separate M&F Counts: EAF Exact Test

To test for EAF:

$$\Pr(n_A|n, m_A) = \frac{n_A!n_B!n_m!n_f!}{m_A!m_B!f_A!f_B!}$$

m_A, m_B	numbers of A, B alleles in males
f_A, f_B	numbers of A, B alleles in females
$n_m = m_A + m_B$	total number of male alleles
$n_f = f_A + f_B$	total number of female alleles
$n_A = m_A + f_A, n_B = m_B + f_B$	total numbers of A, B alleles
$n = n_m + n_f = n_A + n_B$	total number of alleles in males and females

Separate M&F Counts: 1000 Genomes Result

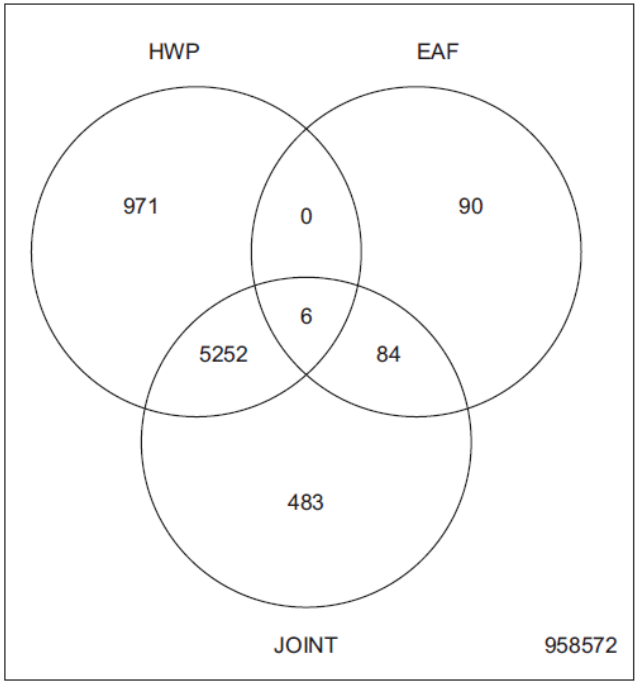
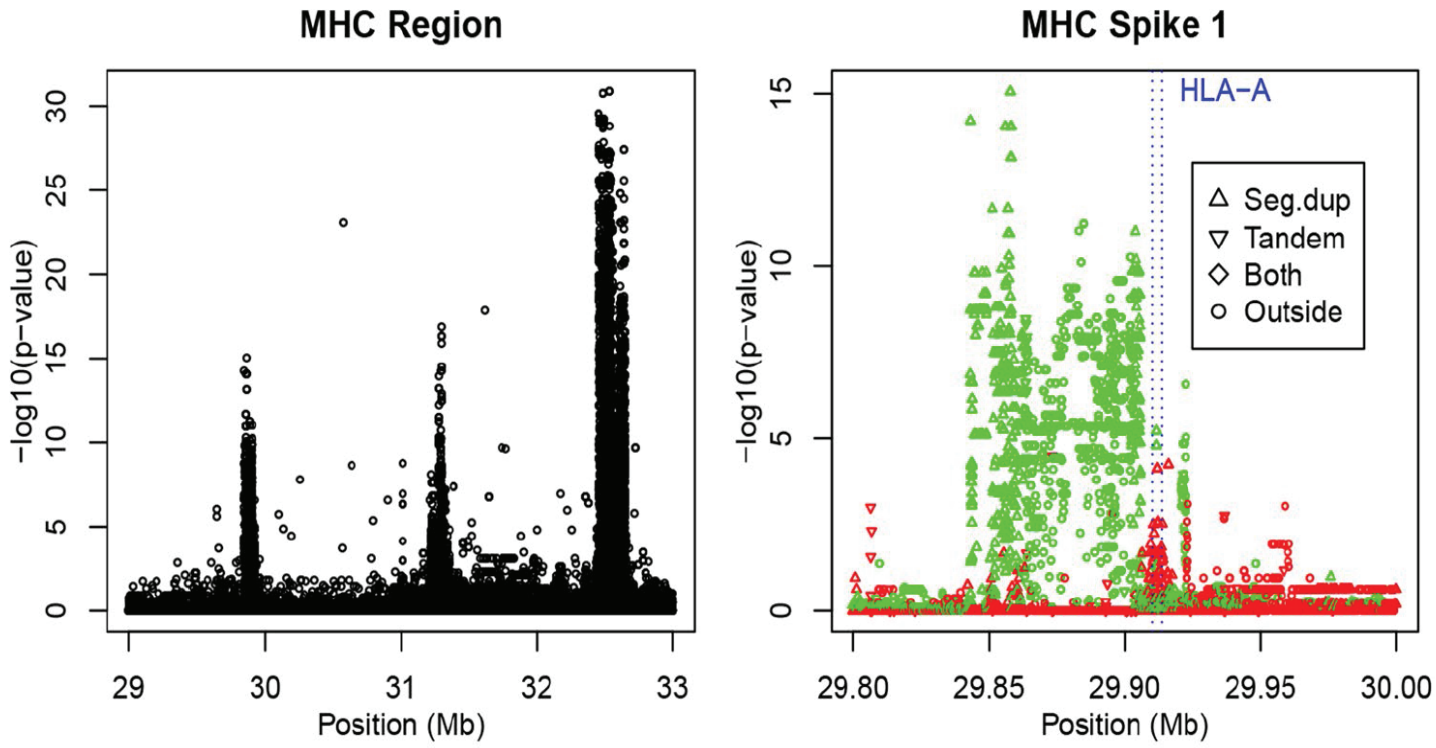


FIGURE 6 Venn diagrams of HWP, EAF, and joint exact test results for all nonmonomorphic complete SNPs on chromosome 1 of the JPT sample

Notes: Circles enclose the number of significant SNPs (at $\alpha = 0.001$) for the different tests.

MHC Region HWE Tests



Green: heterozygote deficiency. Red: heterozygote excess.

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

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

Aside: Gametic Linkage Disequilibrium

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

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

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

$$\mathrm{Cov}(x_1, y_1) = \mathrm{Cov}(x_2, y_2) \;\; = \;\; P_{AB} - p_A p_B = D_{AB}$$

$$\mathrm{Corr}(x_1, y_1) = \mathrm{Corr}(x_2, y_2) \;\; = \;\; D_{AB} / \sqrt{[p_A(1-p_A)p_B(1-p_B)]} = \rho_{AB}$$

Estimation of LD

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

$$\Pr(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = \frac{n!(P_{AB})^{n_{AB}}(P_{Ab})^{n_{Ab}}(P_{aB})^{n_{aB}}(P_{ab})^{n_{ab}}}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!}$$

The data are the counts of four gamete types, so there are three degrees of freedom. There are three parameters: p_A, p_B, D_{AB} so Bailey's method leads directly to MLE's:

$$\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B$$

$$\hat{\rho}_{AB} = r_{AB} = \frac{\hat{D}_{AB}}{\sqrt{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b}}$$

Testing LD

Writing the MLE of D_{AB} as

$$\hat{D}_{AB} = \frac{1}{n^2}(n_{AB}n_{ab} - n_{Ab}n_{aB})$$

where n is the number of gametes in the sample, allows the use of the "Delta method" to find

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

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

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

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

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

Goodness-of-fit Test

The test statistic for the 2×2 table

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

has the value

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

$$= \frac{n\hat{D}_{AB}^2}{\tilde{p}_A \tilde{p}_a \tilde{p}_B \tilde{p}_b}$$

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

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

Aside: Composite Linkage Disequilibrium

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

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

Write

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

Composite LD and Allele Dosage

Now set $X = x_1 + x_2, Y = y_1 + y_2$, the allelic dosages at each locus, 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

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

Composite Linkage Disequilibrium Test

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

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	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

$$
\begin{aligned}
n_{AB} &= 2 \times 0 + 0 + 1 + \frac{1}{2}(3) = 2.5 \\
n_A &= 12, \ \tilde{p}_A = 0.4 \\
n_B &= 6, \ \tilde{p}_B = 0.2 \\
\widehat{f}_A &= 1 - \frac{8/15}{0.48} = -0.11 \\
\widehat{f}_B &= 1 - \frac{4/15}{0.32} = 0.17
\end{aligned}
$$

Example

The estimated composite disequilibrium coefficient is

$$\hat{\Delta}_{AB} = \frac{2.5}{15} - 2(0.4)(0.2) = 0.0067$$

The test statistic is

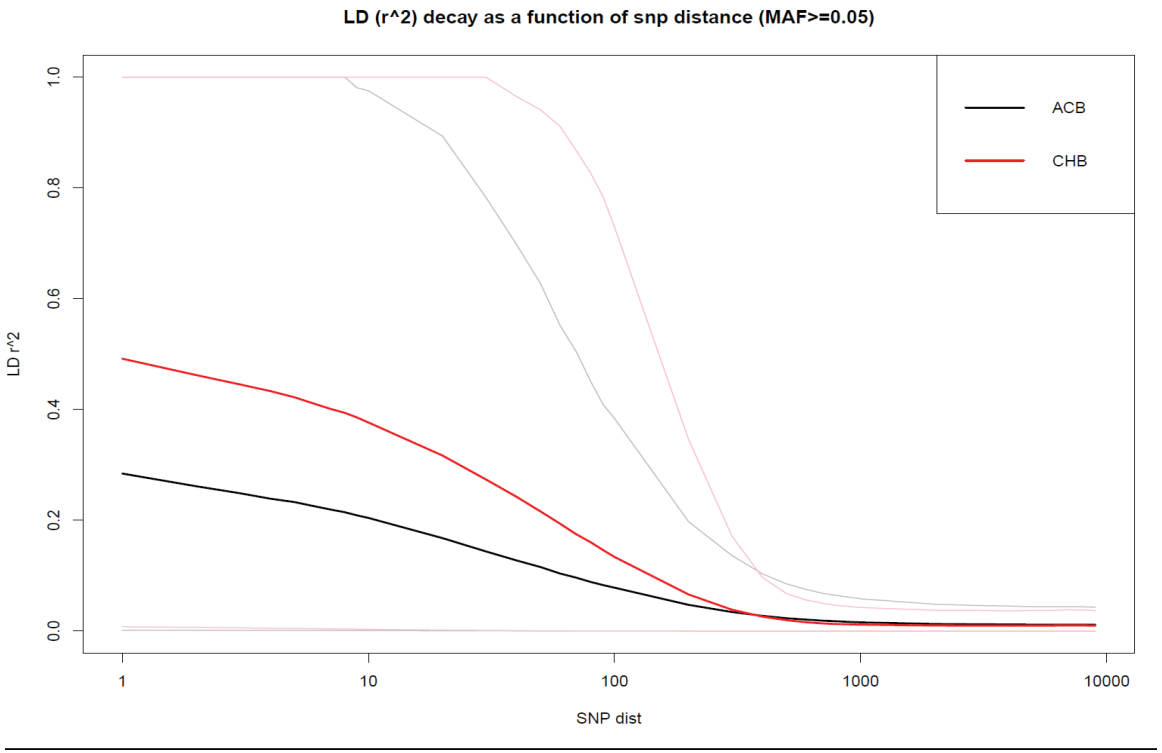
$$X^2 = \frac{15 \times (0.0067)^2}{0.24 \times 0.89 \times 0.16 \times 1.17} = 0.02$$

Previous work on EM algorithm, assuming HWE, estimated p_{AB} as 0.0893 so

$$\hat{D}_{AB} = 0.0893 - 0.4 \times 0.2 = 0.0093$$

$$X^2 = \frac{30 \times (0.0093)^2}{0.4 \times 0.6 \times 0.2 \times 0.8} = 0.07$$

1000 Genomes Example



Allele dosage squared correlations for pairs of SNPs on chromosomes 21 and 22 of the 1000 Genomes ACB and populations. Heavy lines: means. Light lines: 5th and 95th percentiles.

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 two loci with alleles A_u, B_v, the entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv})$$

In the absence of linkage disequilibrium $\tilde{P}_{uv} = \tilde{p}_u \tilde{p}_v$ so

$$\begin{aligned} H_{AB} &= -\sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] \\ &= H_A + H_B \end{aligned}$$

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 Marker	Entropy		
	Single	Multi	Cond.
DYS385ab	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

Most-discriminating loci may not contribute to the most-discriminating haplotypes. No additional discriminating power beyond 10 loci.

Population Structure and Relatedness

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 III 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)

Additional Questions of Interest

If genotypic data are available, individual inbreeding and kinship values can be estimated:

- What is the Genetic Relatedness Matrix? (association mapping)
- How do social behaviors evolve?
- How should captive breeding programs be managed? (conservation genetics)
- Are these remains from a person in this family? (disaster victim identification)

Statistical Analysis

Possible to approach these data from purely statistical viewpoint.

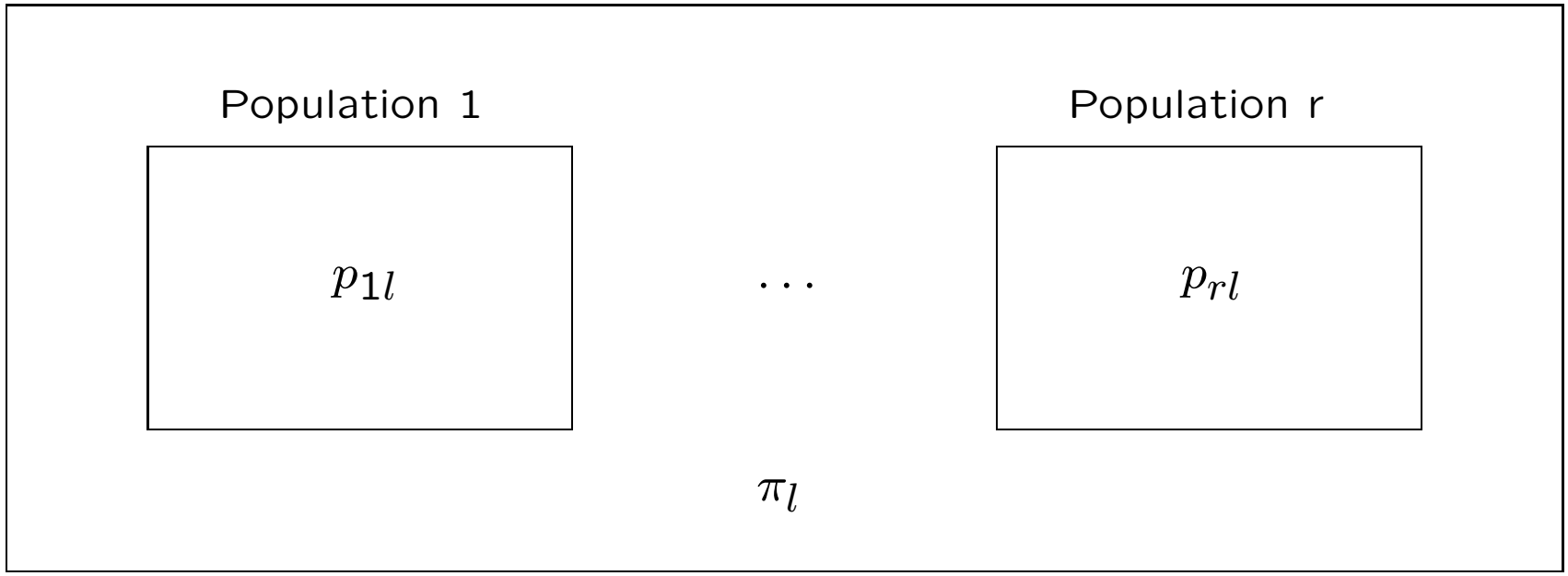
Could test for differences in allele frequencies among populations.

Could use various multivariate techniques to cluster populations.

These analyses may not answer the biological questions.

Notation

Genetic Analysis: SNP l Allele Frequencies

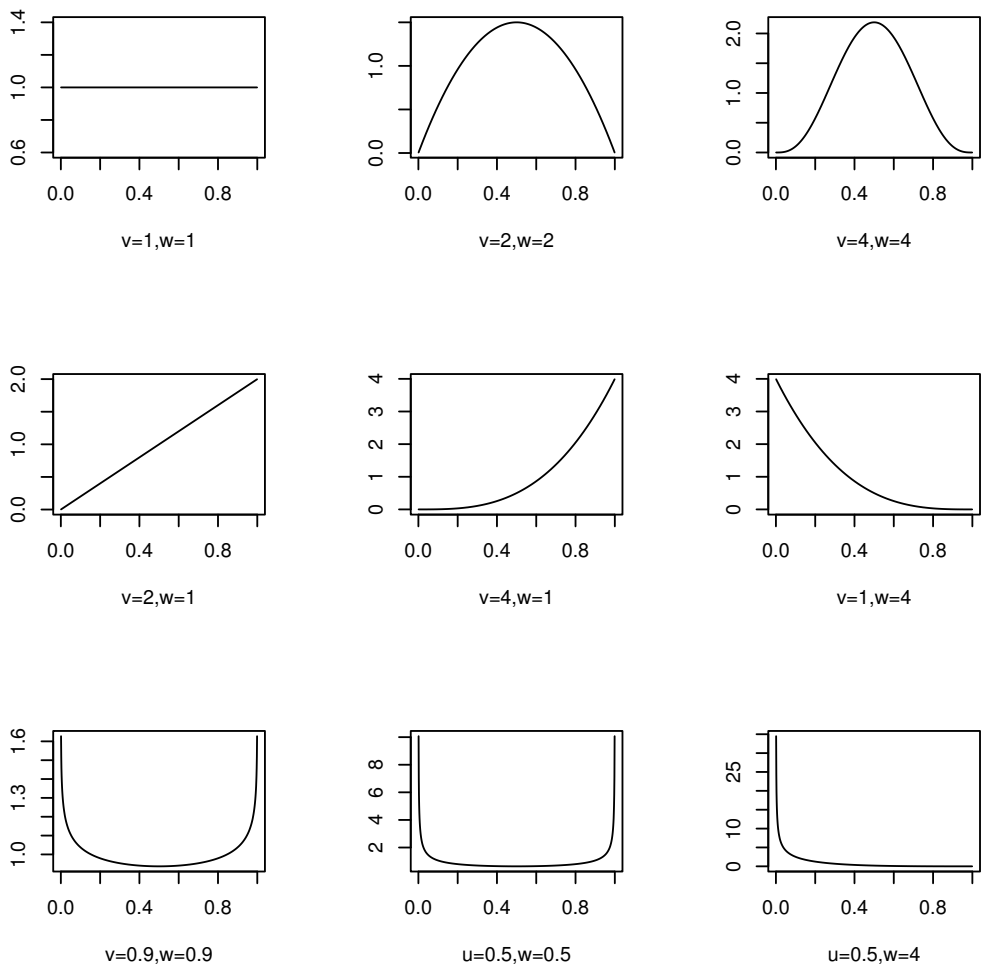


Among samples of n_i alleles from population i: counts for SNP l reference allele follow a binomial distribution with mean p_{il} and variance $n_i p_{il}(1-p_{il})$. Sample allele frequencies $\tilde{p}_{il}$ have expected values p_{il} and variances $p_{il}(1-p_{il})/n_i$.

Among replicates of population i: p_{il} values follow a distribution with mean π_l and variance $\pi_l(1-\pi_l)\theta^i$. Distribution sometimes assumed to be Beta.

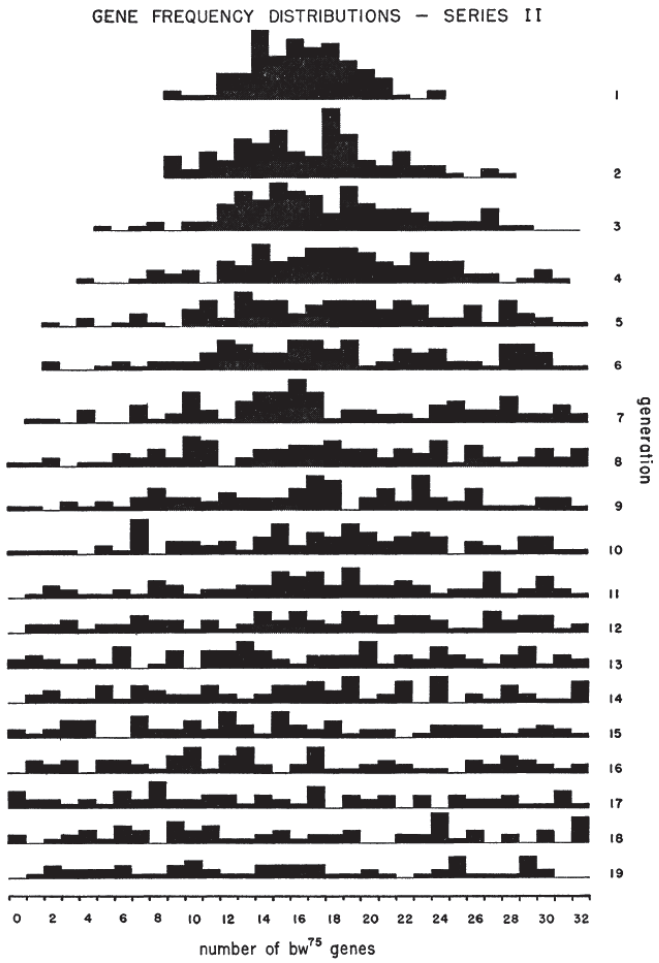
Beta distribution: Theoretical

The beta probability density is proportional to $p^{v-1}(1-p)^{w-1}$ 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).



What is θ?

Two ways of thinking about θ.

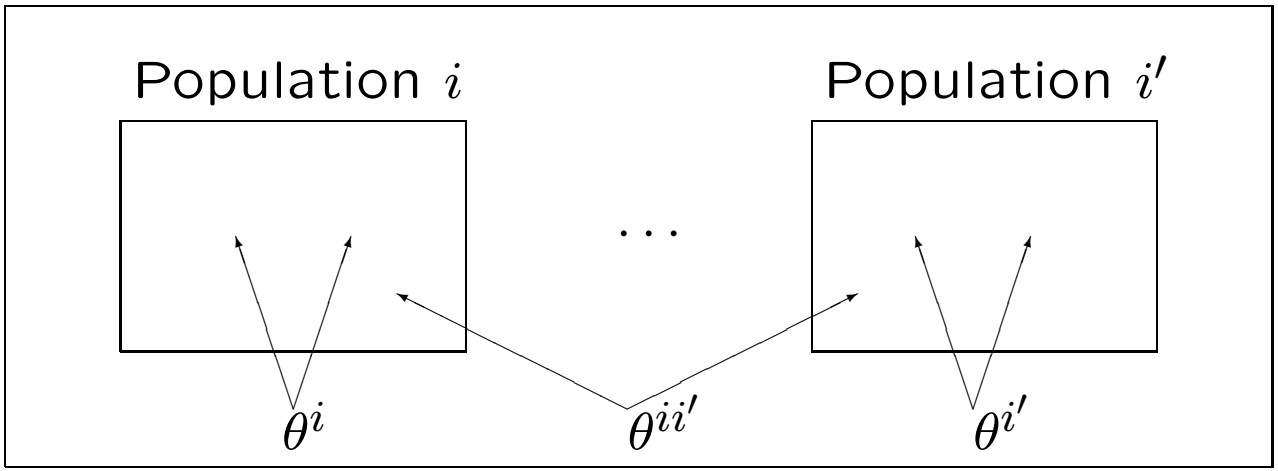
It measures the probability a pair of alleles are identical by descent: and this is with respect to some reference population.

The target alleles may be in specified populations, and this leads to characterization of population structure, of they may be in specified individuals and this leads to characterization of inbreeding and relatedness.

θ also describes the variance of allele frequencies among populations, or among evolutionary replicates of a single population.

Weir BS, Goudet J. 2017. A unified characterization of population structure and relatedness. Genetics 206:2085-2103.
Goudet J, Kay T, Weir BS. 2018. How to estimate kinship. Molecular Ecology 27:4121-4135.

Allele-level θ's



θ's are ibd probabilities for pairs of alleles from specified populations.

θ_W^i is average of the within-population probabilities θ^i. Average over populations of θ_W^i is θ_W.

θ_B is average of the between-population-pair probabilities $\theta^{ii'}$.

Allelic Measure Predicted Values

Predicted Values of the θ's: Pure Drift

The 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.

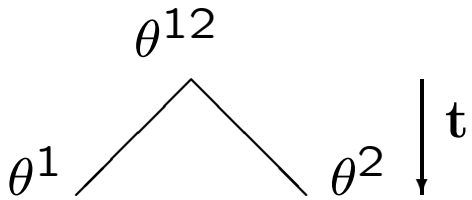
In the case of pure drift, where population i has constant size N_i and there is random mating, t generations after the population began drifting from an ancestral population in which $\theta^i = 0$

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

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):

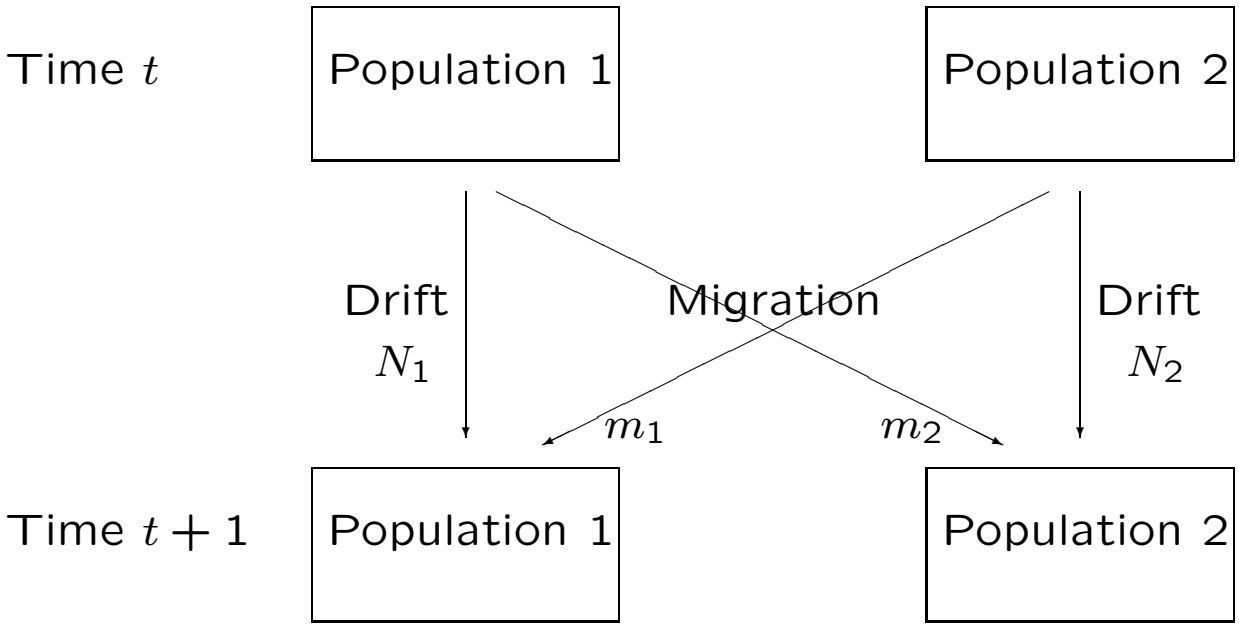


$$\theta^i \;=\; 1-(1-\theta^{12})\left(\frac{2N_i-1}{2N_i}\right)^t,\;\; i=1,2$$

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 $\beta^1 = (\theta^1 - \theta^{12})/(1-\theta^{12})$ and $\beta^2 = (\theta^2 - \theta^{12})/(1-\theta^{12})$ are positive, or that one of them is negative and the other one positive.

Drift and Mutation

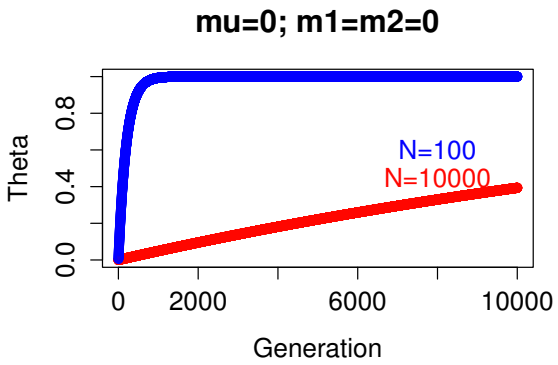
If there is no migration, the θ's tend to equilibrium values of

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

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

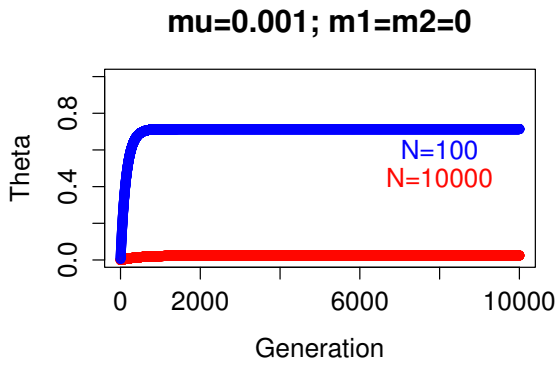
Drift, Mutation and Migration

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



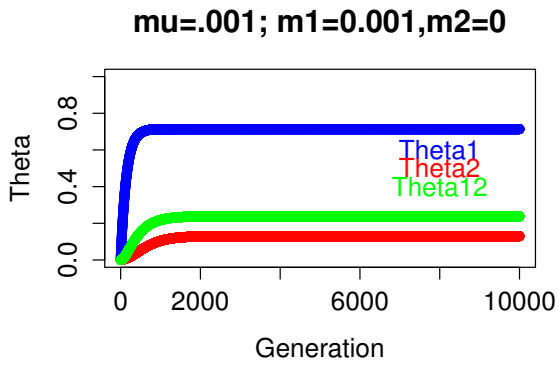
Drift Only

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



Drift and Mutation

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



Drift, Mutation and Migration

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

Multiple Populations

For random union of gametes, when pairing of alleles into individuals is not needed, the ibd probability θ_W^i for any distinct pair of alleles within population i relative to the ibd probability between populations is

$$\beta_{WT}^i = \frac{\theta_W^i - \theta_B}{1 - \theta_B}$$

This is the population-specific F_{ST}^i for alleles.

Averaging over populations:

$$\beta_{WT} = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

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

Genotypic Measure Predicted Values

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'.

Kinship vs Inbreeding

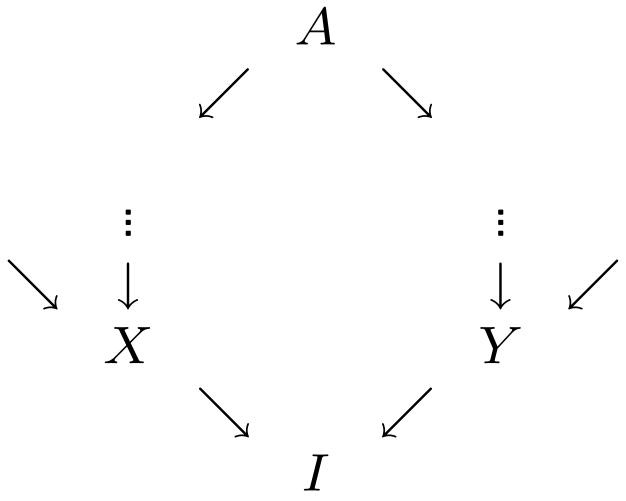
The kinship of individuals j, j' in population i is the probability an allele from j is ibd to an allele from j'. This is $\theta^i_{jj'}$.

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



If there are n individuals (including X, Y, A) in the path linking the parents through A, then the inbreeding F_I of I, or the kinship θ_{XY} of X and Y, is

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

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

Average Kinships

The average over all pairs of distinct individuals, $j \neq j'$, of the kinships $\theta^i_{jj'}$ 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.

Within-population 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-individual 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}.

Total 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

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

The average over individuals within this population is the population-specific F_{IT}^i. The average of these over all populations is the total inbreeding coefficient $F_{IT} = (F_I - \theta_B)/(1 - \theta_B)$.

Within-population Kinship

For population i, the kinship of individuals j, j' relative to the kinship 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.

Genotypic Measures

When individuals are distinguished:

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

$$F_{IS} = \frac{F_{IT} - F_{ST}}{1 - F_{ST}}$$

This classic result also holds for population-specific values

$$(1 - F^i_{IT}) = (1 - F^i_{IS})(1 - F^i_{ST})$$

$$F^i_{IS} = \frac{F^i_{IT} - F^i_{ST}}{1 - F^i_{ST}}$$

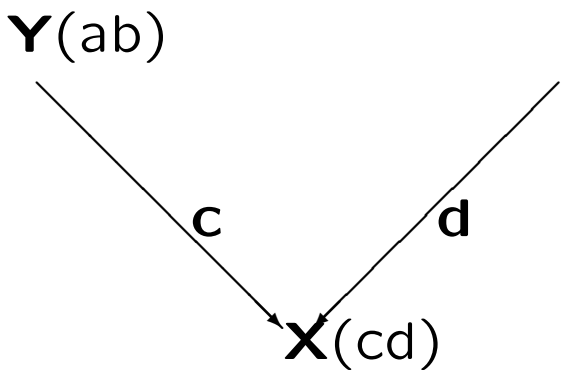
k-coefficients

The kinship 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.

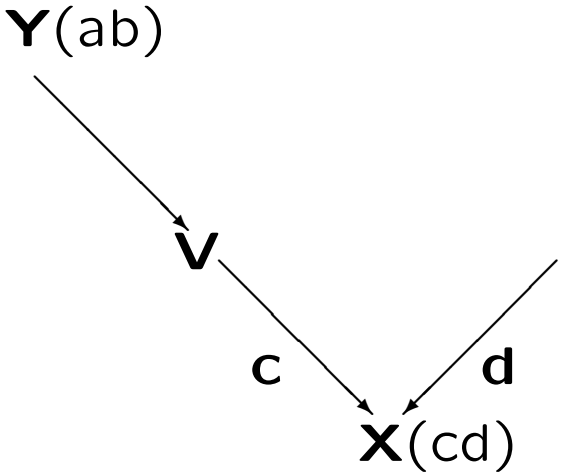
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are written as k_0, k_1, k_2 and $\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$.

Parent-Child



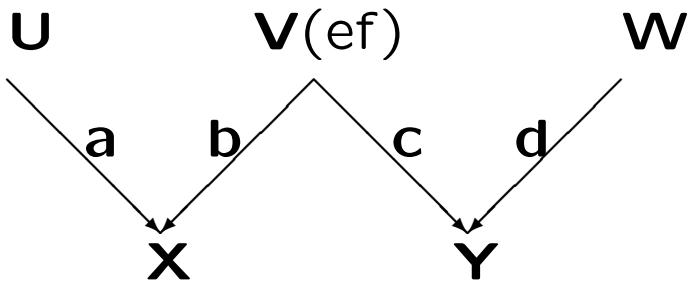
$$\Pr(c \equiv a) = 0.5, \quad \Pr(c \equiv b) = 0.5, \quad k_1 = 1$$

Grandparent-grandchild



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

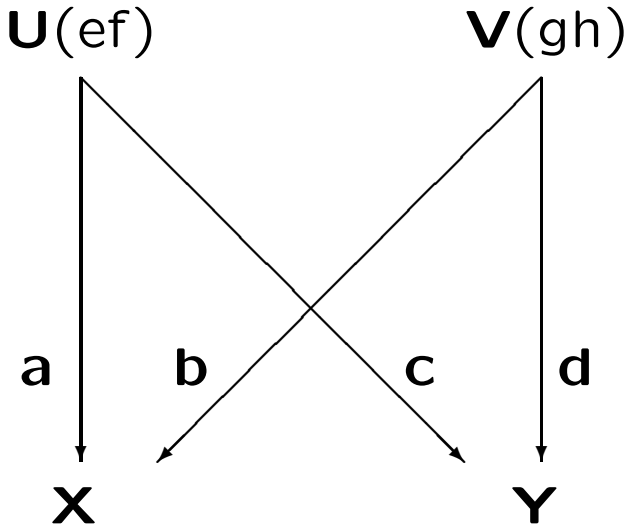
Half sibs



		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 $k_0 = 0.5$.

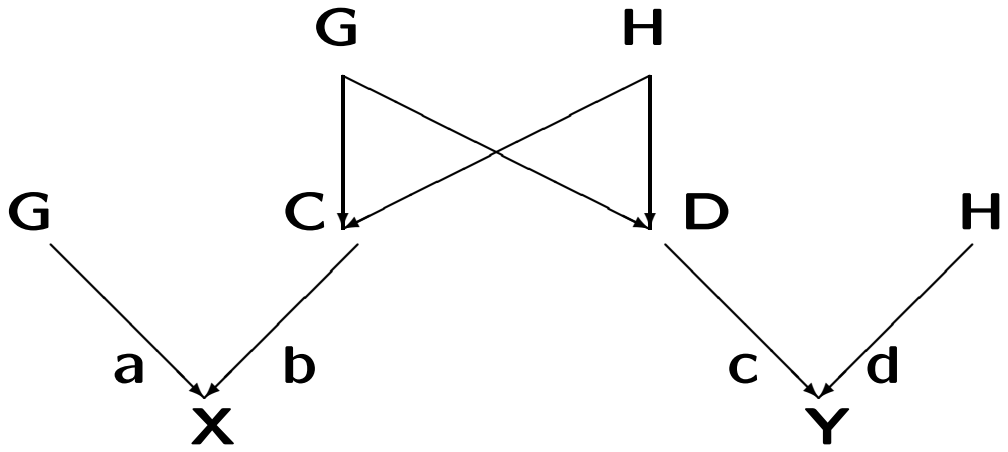
Full sibs



		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

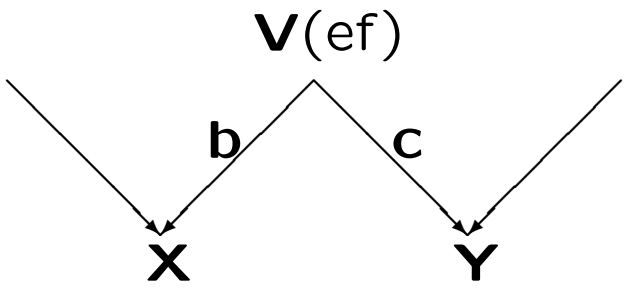


Non-inbred Relatives

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).

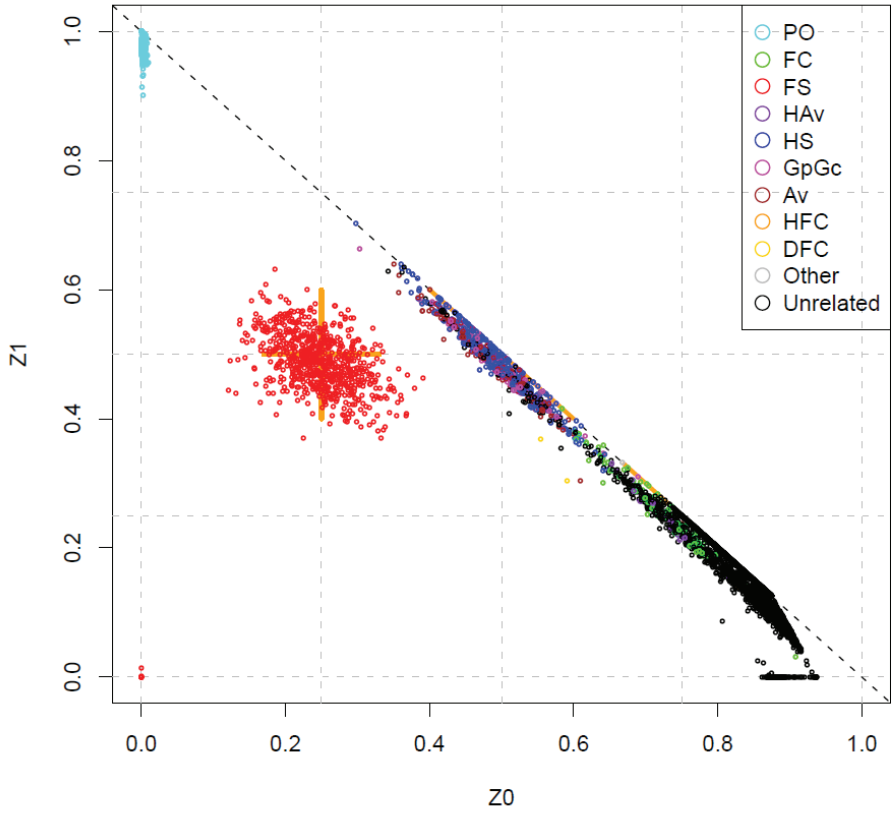
Predicted vs Actual Kinship



For half-sibs, for example, the predicted kinship, is $(1/2)^3 = 1/8$.

However, alleles b, c are equally likely to be ibd or not ibd (ibd if they are both copies of e or f) so the actual kinship is either 0.25 (with probability 1/2) or 0 (with probability 1/2). The actual kinship of X, Y has an expected value of 1/8 and a standard deviation of 1/8. Over the whole genome, the standard deviation is 0.013. The estimate from observed marker genotypes will be of the actual ("gold standard") kinship. [Hill and Weir, Genet Res 2011]

PLINK Example



Shows variation of actual k's around predicted k's.

Individual Inbreeding Estimation

Allele Matching Approach

We work with observed allelic matching $\tilde{M}_j$ within individual j, and $\tilde{M}_{jj'}$ between individuals j, j'. For SNPs, these proportions are:

		$\tilde{M}_j$
j	AA	1
	AB	0
	BB	1

	$\tilde{M}_{jj'}$	j'		
		AA	AB	BB
j	AA	1	0.5	0
	AB	0.5	0.5	0.5
	BB	0	0.5	1

These are compared to the average matching for all pairs of individuals: $\tilde{M}_S$ for all pairs in the same sample or $\tilde{M}_B$ for all pairs from different samples.

Allele Matching

Our model specifies that the expectation over evolutionary replicates for a matching proportion at SNP l, $\tilde{M}_l$ is $M_l + (1 - M_l)\theta$ where θ is the ibd probability for the pair(s) of alleles being matched and M_l is a nuisance parameter:

$$M_l \;=\; \pi_l^2 + (1-\pi_l)^2 = 1 - 2\pi_l(1-\pi_l)$$

Our estimates for inbreeding and kinship are

$$\widehat{\beta}_j = \frac{\tilde{M}_j - \tilde{M}_S}{1 - \tilde{M}_S} \quad , \quad \widehat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_S}{1 - \tilde{M}_S}$$

We combine over SNPs with weighted averages

$$\widehat{\beta}_j = \frac{\sum_l (\tilde{M}_{j_l} - \tilde{M}_{S_l})}{\sum_l (1 - \tilde{M}_{S_l})} \quad , \quad \widehat{\beta}_{jj'} = \frac{\sum_l (\tilde{M}_{jj'_l} - \tilde{M}_{S_l})}{\sum_l (1 - \tilde{M}_{S_l})}$$

Allele Matching

We find that our estimates behave well for estimating the parameters

$$\beta_j = \frac{F_j - \theta_S}{1 - \theta_S} \quad , \quad \beta_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

Individuals less inbred than the average kinship have negative β values.

The average over pairs of individuals j, j' in one population, of either the estimates $\widehat{\beta}_{jj'}$ or the parameters $\beta_{jj'}$, gives zero. Some estimates and parameters are negative and some are positive.

Alternative Estimators: Heterozygosity

The heterozygosity indicator $\tilde{H}_{jl}$ at SNP l for individual j is 1 if the individual is heterozygous and 0 if it is homozygous. Hall et al. [Genet Res 2012] and Yengo et al. [PNAS 2017] gave individual-specific estimates:

$$\widehat{f}_{\mathrm{Hom}_j} = 1 - \frac{\tilde{H}_{jl}}{2\tilde{p}_l(1-\tilde{p}_l)}$$

and used weighted averages over SNPs:

$$\begin{aligned}\widehat{f}_{\mathrm{Hom}_j} &= 1 - \frac{\sum_l \tilde{H}_{jl}}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)} \\ &= 1 - \frac{H_{\mathrm{Obs}}}{H_{\mathrm{Exp}}}\end{aligned}$$

This estimator was called f_{PLINK} by Gazal et al.[Hum Hered 2014]. Note the similarity to the MLE for the within-population inbreeding coefficient f given earlier - that quantity is the average over individuals of the $\widehat{f}_{\mathrm{Hom}_j}$ quantities.

Alternative Estimators: Heterozygosity

What do the usual inbreeding estimators actually estimate under genetic sampling?

$$\mathcal{E}(\hat{f}_{\mathsf{Hom}_j}) \;=\; 1 - \frac{1 - F_j}{(1-\theta_S) - \frac{1}{2n}(1 + F_W - 2\theta_S)}$$

For large sample sizes, this reduces to

$$\mathcal{E}(\hat{f}_{\mathsf{Hom}_j}) \;=\; \frac{F_j - \theta_S}{1 - \theta_S}$$

In other words, $\hat{f}_{\mathsf{Hom}_j}$ is an (almost) unbiased estimate of $\beta_j = (F_j - \theta_S)/(1 - \theta_S)$, the individual-specific version of Wright's F_{IS} [Wright, Am Nat 1922].

Averaging over individuals gives the usual estimate for $f = F_{IS}$ for the population, and $F_{IS} = (F_{IT} - F_{ST})/(1 - F_{ST})$.

Aside: Expectation of $2\tilde{p}_l(1-\tilde{p}_l)$

Expectations of allele frequencies in a sample of n individuals:

$$\begin{aligned}
\mathcal{E}(\tilde{p}_l) &= \pi_l \\
\mathcal{E}(\tilde{p}_l^2) &= \pi_l^2 + \pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right] \\
\mathcal{E}[2\tilde{p}_l(1-\tilde{p}_l)] &= 2\pi_l(1-\pi_l]\left[(1-\theta_S) - \frac{1}{2n}(1 + F_W - 2\theta_S)\right] \\
&\approx 2\pi_l(1-\pi_l](1-\theta_S)
\end{aligned}$$

Alternative Estimators: GCTA

If X_{jl}, the allele dosage, is the number of copies of the reference allele for SNP l carried by individual j, Yang et al. [Am J Hum Genet 2011] introduced $\hat{F}^{III}$, called $\hat{F}_{\mathsf{Uni}}$ by Yengo et al. and f_{GCTA3} by Gazal et al:

$$\hat{F}^{u}_{\mathsf{Uni}_j} \;=\; \frac{1}{L}\sum_{l=1}^{L}\left(\frac{X_{jl}^2-(1+2\tilde{p}_l)X_{jl}+2\tilde{p}_l^2}{\tilde{p}_l(1-\tilde{p}_l)}\right)$$

For large samples this has an expected value under genetic sampling of

$$\mathcal{E}(\hat{F}_{\mathsf{Uni}_j}) \;=\; \frac{F_j-2\psi_j+\theta_S}{1-\theta_S}$$

where ψ_j is the average kinship of individual j with other members of the study sample,

$$\psi_j \;=\; \frac{1}{n-1}\sum_{\substack{j'=1\\ j\neq j'}}^{n}\theta_{jj'}$$

Alternative Estimators: GCTA

The inclusion of the ψ term means that the ranking of $\widehat{F}_{\mathsf{Uni}_j}$ expected values can be different from the ranking of F_j values. The rankings of $\widehat{f}_{\mathsf{Hom}_j}$ expected values are the same as those for F_j.

Yang et al. also discussed

$$\mathsf{GCTA}_j \;=\; \frac{1}{L}\sum_{l=1}^{L}\frac{(X_{jl}-2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)}-1$$

For large samples, these estimates have expected values

$$\mathcal{E}(\mathsf{GCTA}_j) \;=\; \frac{F_j-4\psi_j+3\theta_S}{1-\theta_S}$$

This has behavior close to that of $\widehat{F}_{\mathsf{Uni}_j}$.

Alternative Estimators: MLE

Hall et al. used EM to give MLEs for f_j, assuming π_l's were known (and equal to $\tilde{p}_l$), using

$$\begin{aligned} \Pr(\tilde{H}_{jl} = 1) &= 2\tilde{p}_l(1-\tilde{p}_l)(1-f_j) \\ \Pr(\tilde{H}_{jl} = 0) &= 1 - 2\tilde{p}_l(1-\tilde{p}_l)(1-f_j) \end{aligned}$$

but it is easier to use a grid search to maximize the likelihood $\mathrm{Lik}(f_j)$, or its logarithm:

$$\mathrm{Lik}(f_j)] = \prod_l [1 - 2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)]^{1-\tilde{H}_{jl}} [2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)]^{\tilde{H}_{jl}}$$

These estimates are close in value to $\widehat{f}_{\mathrm{Hom}_j}$.

Alternative Estimators: Runs of Homozygosity

Estimators so far use single SNP statistics and average over SNPs.

Runs of homozygosity, with a large number of SNPs, are likely to represent regions of identity by descent. The inbreeding coefficient can be estimated as the proportion of windows of SNPs that are completely homozygous.

Requires judgment in deciding window length, degree of window overlap, allowance for some heterozygotes, and (possibly) minor allele frequency [McQuillan et al., Am J Hum Genet 2006; Joshi et al., Nature 2015].

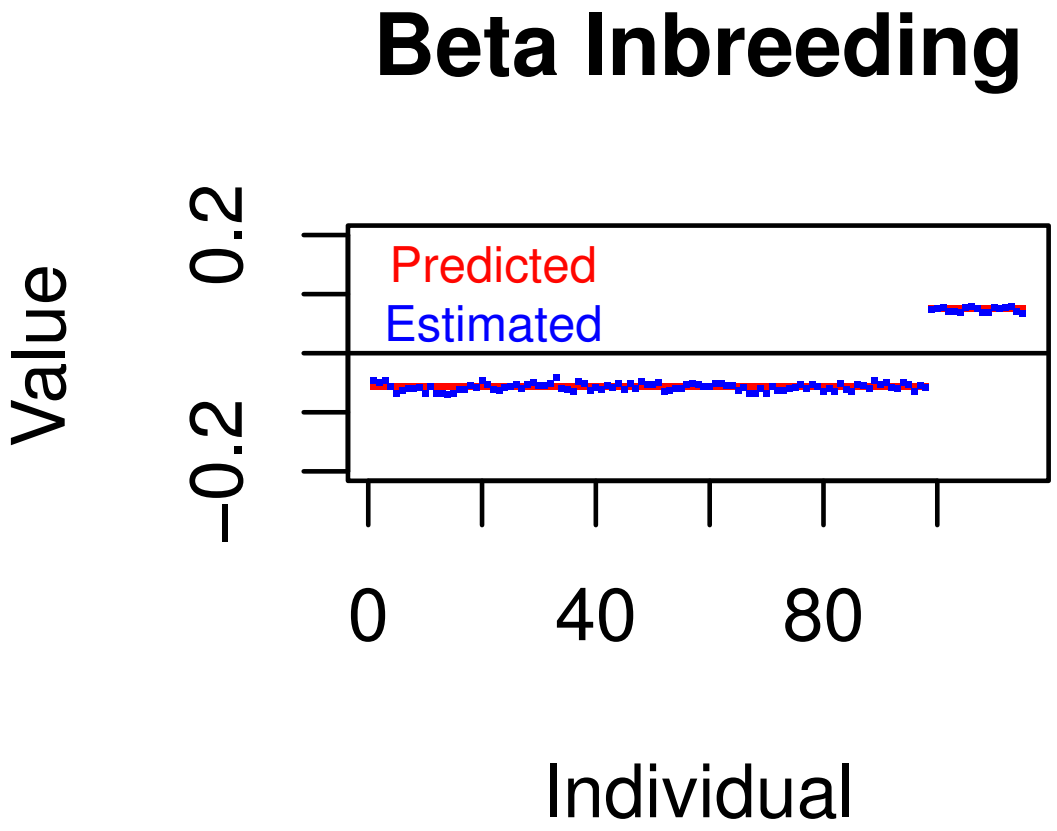
Example

The β inbreeding estimator was applied to a set of 115 individuals simulated and typed at 79,069 polymorphic SNPs [Weir and Goudet, Genetics 2017].

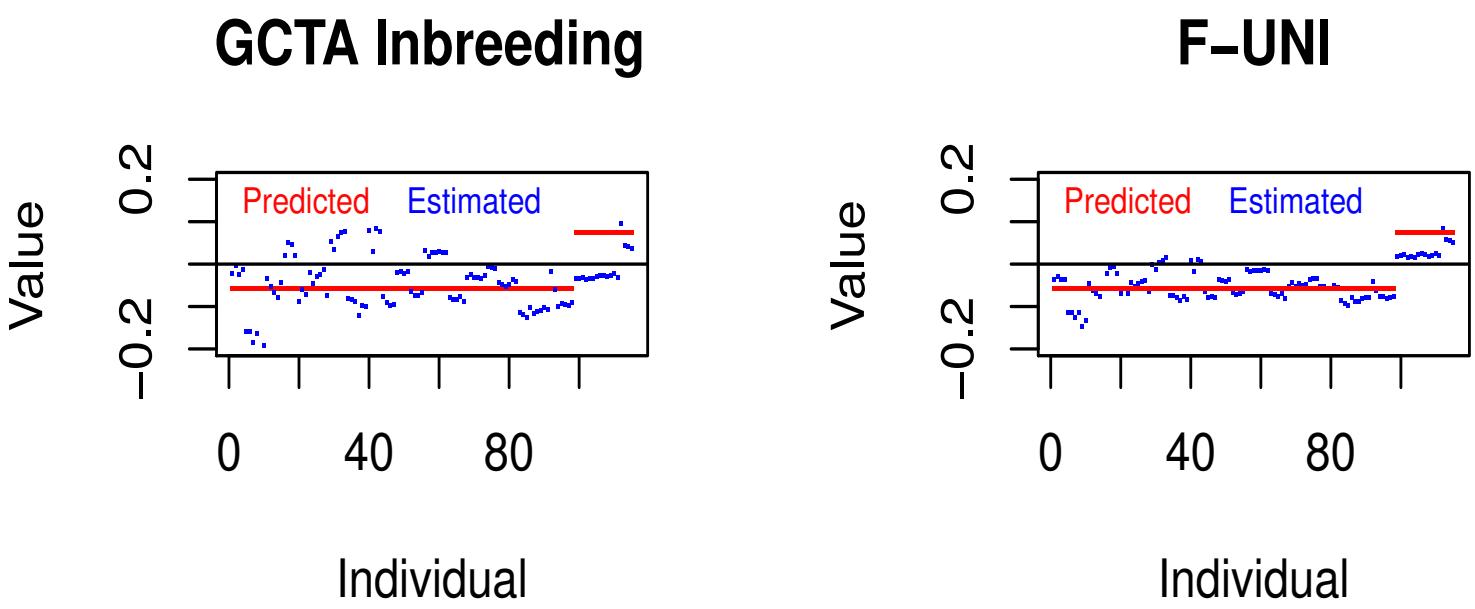
Among the 6,555 pairs of individuals the kinship values have an average value of $\theta_S = 0.0427$. There are 17 individuals with values of $F = 0.125, \beta = 0.0860$ and 98 with $F = 0, \beta = -0.0446$ predicted from the pedigree.

The $\widehat{\beta}_j$ values are very close to the $\beta_j = (F_j - \theta_S)/(1 - \theta_S)$ values, as shown on the next slide:

Example: Beta values



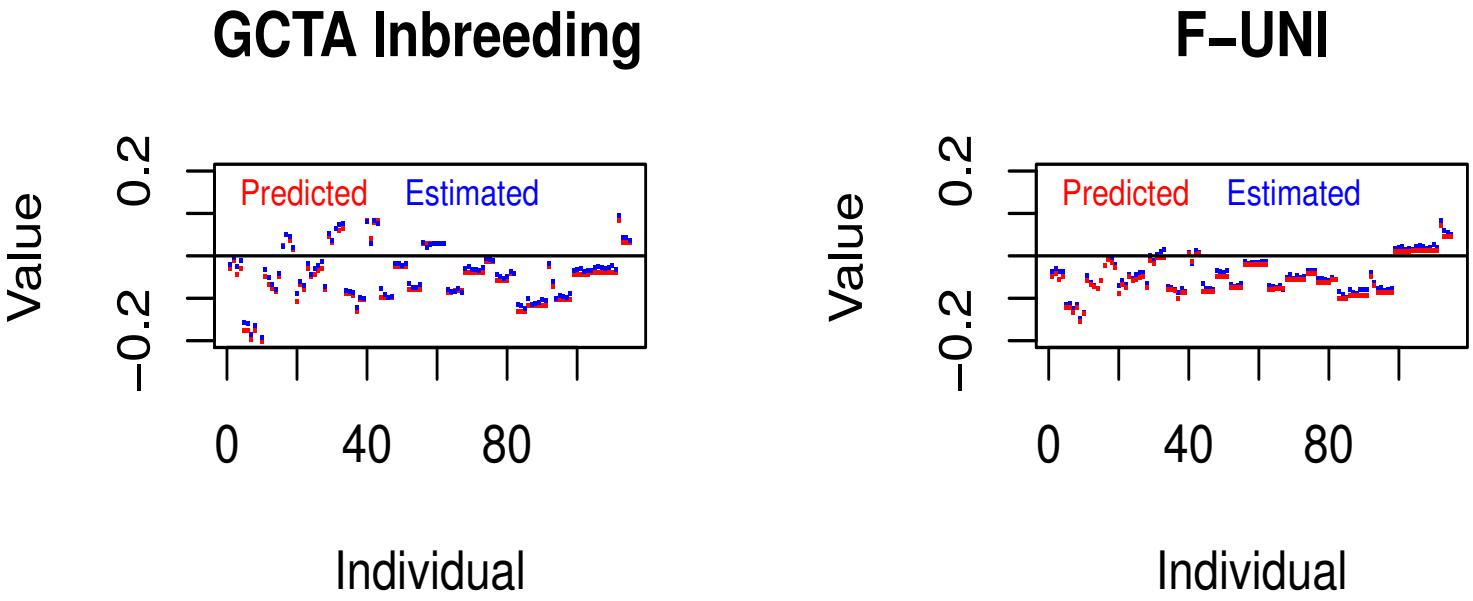
Example: GCTA values



The problem is that these estimates use $\tilde{p}$'s instead of π's.

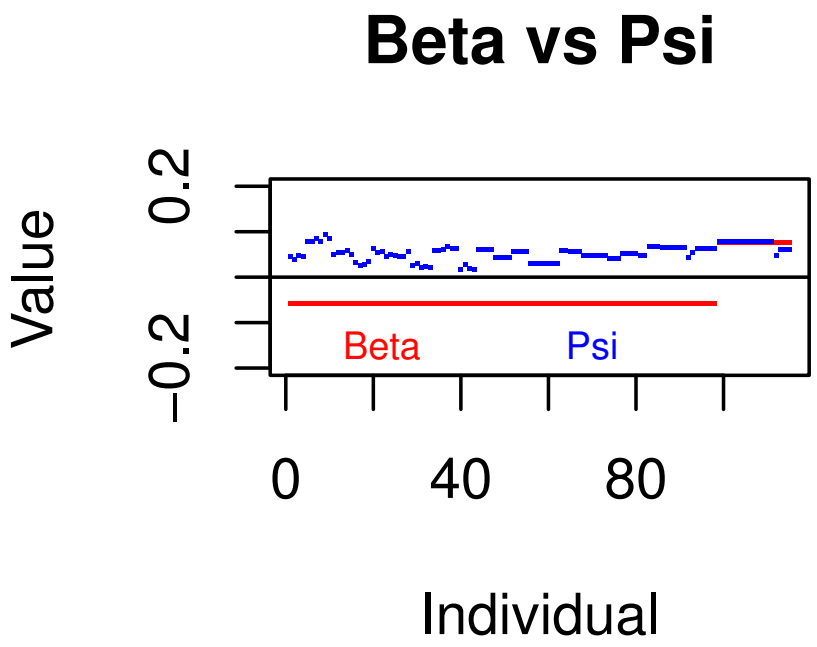
Example: GCTA Expected values

The GCTA estimators are close to their expected vales, but not to F or to β.



Example: Beta vs Psi

Individuals with the same F_j will have the same β_j but can have quite different ψ_j values:



Comparison of Estimators: Simulations

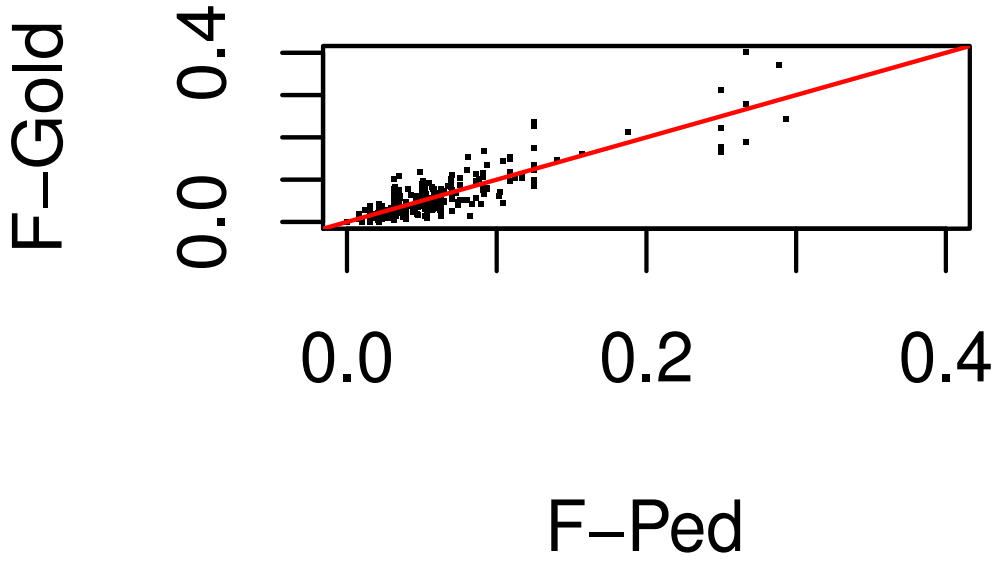
We generated 50 founder individuals, with 100,000 SNPs over a 20 Morgan map.

We then used our own quantiNemo software [Neuenschwander et al. Bioinformatics 2008] to generate eight subsequent generations of 50 individuals per generation and it is these 400 descendants we use for subsequent analysis.

The mating system was 80% monogamous and 20% random mating. Each of the 100 alleles per SNP among the founders was given a unique identifier so that subsequent identity by descent could be tracked. The average ibd proportion over loci, within individuals and between each pair of individuals, provided "gold standard" or actual inbreeding and kinship coefficients, as opposed to the pedigree-based values from path counting.

Simulated Pedigree vs Actual Inbreeding

100K SNPs



Comparison of Estimators: Notation

Fped, Bped: pedigree values of F and β.

Fgold, Bgold: actual values of F and β.

Froh: runs of homozgosity estimate.

Fmle: maximum likelihood estimate of F.

Fhom: $1 - \tilde{H}/2\tilde{p}(1-\tilde{p})$

Fbet: allele-matching estimates of β,

Ugold: actual value of F_{Uni}.

Funi: GCTA estimates of F_{Uni}.

Comparison of Estimators: Correlations

	Fped	Bped	Fgold	Bgold	Froh	Fmle	Fhom	Fbet	Ugold	Funi
Fped	1.000	1.000	0.902	0.901	0.879	0.790	0.836	0.836	0.707	0.642
Bped	1.000	1.000	0.902	0.902	0.879	0.790	0.836	0.836	0.707	0.642
Fgold	0.902	0.902	1.000	1.000	0.975	0.889	0.918	0.918	0.829	0.743
Bgold	0.901	0.902	1.000	1.000	0.975	0.889	0.918	0.918	0.829	0.743
Froh	0.879	0.879	0.975	0.975	1.000	0.929	0.952	0.952	0.819	0.779
Fmle	0.790	0.790	0.889	0.889	0.929	1.000	0.976	0.976	0.838	0.876
Fhom	0.836	0.836	0.918	0.918	0.952	0.976	1.000	1.000	0.747	0.781
Fbet	0.836	0.836	0.918	0.918	0.952	0.976	1.000	1.000	0.747	0.781
Ugold	0.707	0.707	0.829	0.829	0.819	0.838	0.747	0.747	1.000	0.917
Funi	0.642	0.642	0.743	0.743	0.779	0.876	0.781	0.781	0.917	1.000

Estimation of Kinship

Estimation of Kinship

We have a general estimator for the kinship of individuals j, j' in the same sample:

$$\widehat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_R}{1 - \tilde{M}_R}$$

Here $\tilde{M}_{jj'}$ is the allele matching for the target pair of individuals, and $\tilde{M}_R$ is for a reference set.

- if R is all pairs of individuals in the same sample, $\tilde{M}_R$ is the average matching over jj' pairs, and the estimates have an average of zero.

Estimation of Kinship

- if R is a set of populations, say in the continent to which the target pair of individuals belong, $\tilde{M}_R$ is the average matching for all pairs of alleles, one from each of two populations in this same set of populations. (Continental Reference)

- if R is all populations for which data are available, $\tilde{M}_R$ is the average matching for all pairs of alleles, one from each of any two of these populations. (World Reference)

The averages of these two sets of estimates over all pairs of indivduals in one population can be positive or negative.

Kinship is relative, not absolute

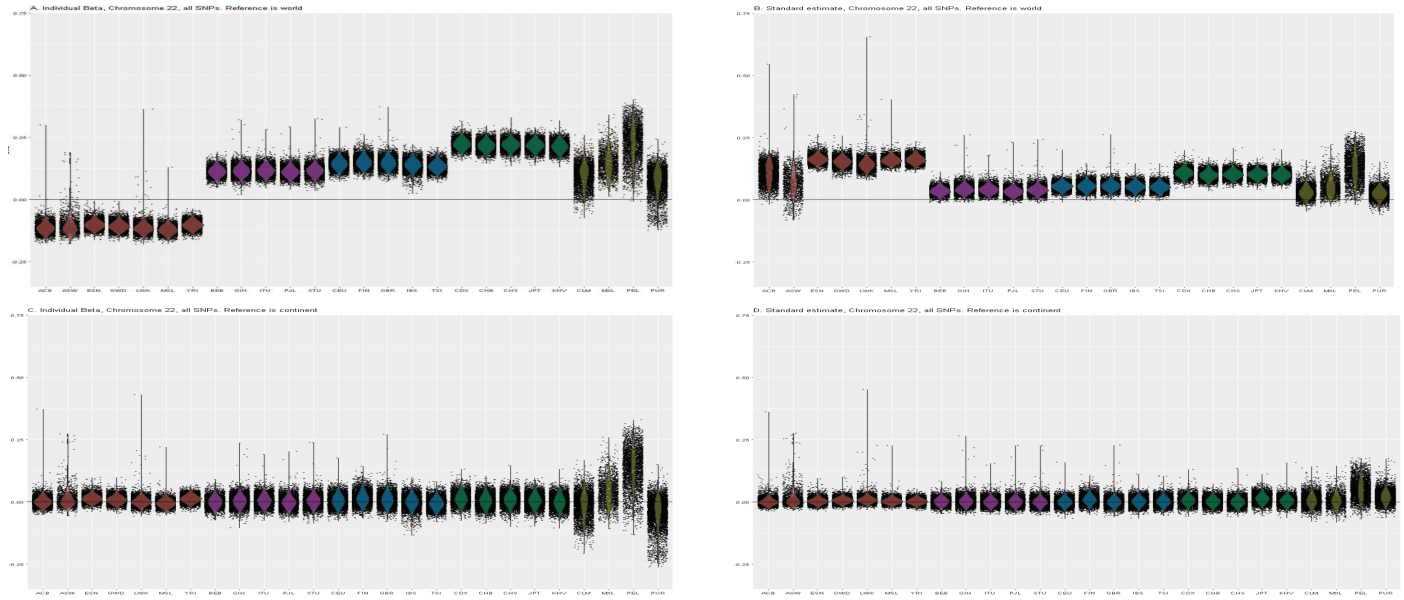
The β kinship estimates have been applied to 1000 Genomes data, and compared to standard estimates, shown on next slide.

For the whole world, all 26 populations, as reference the β 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 GCTA estimates - those estimates showed higher kinship 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 kinship, except for the admixed AMR.

Kinship is relative, not absolute

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



Beta estimates GCTA 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

Estimators for Populations

Matching Proportions for Populations

If kinships for pairs of individuals within a population are estimated by comparing their allele matching proportions to matching between populations, the average over all pairs is the population-specific F_{ST}.

Alternatively, we can work directly with sample allele frequencies, as on next slides.

Matching Proportions Within Populations

When the genotypic structure of data is ignored, or not known, allelic data can be used.

If $2n_{il}$ alleles at SNP l are observed for population i, and if r_{il} of these are the reference type, the observed matching proportion of allele pairs (reference or non-reference) within this sample, is

$$\begin{aligned}\tilde{M}^i_{Wl} &= \frac{1}{2n_{il}(2n_{il}-1)}[r_{il}(r_{il}-1)+(2n_{il}-r_{il})(2n_{il}-r_{il}-1)] \\ &\approx \tilde{p}^2_{il}+(1-\tilde{p}_{il})^2\end{aligned}$$

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

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}^i_{Wl}) = M_l + (1-M_l)\theta^i_W$$

where $M_l = \pi_l^2 + (1-\pi_l)^2$.

Matching Proportions between Populations

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

$$\begin{aligned}\tilde{M}_{Bl}^{ii'} &= \frac{1}{4n_i n_{i'}} \sum_{\substack{j=1 \\ j \neq j'}}^{2n_i} \sum_{j'=1}^{2n_{i'}} x_{ju} x_{j'u} \\ &= \tilde{p}_{il}\tilde{p}_{i'l} + (1-\tilde{p}_{il})(1-\tilde{p}_{i'l})\end{aligned}$$

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}_{Bl}^{ii'}) = M_l + (1-M_l)\theta_B^{ii'}$$

and, averaging over all pairs of populations

$$\mathcal{E}(\tilde{M}_{Bl}) = M_l + (1-M_l)\theta_B$$

Allele-based Estimate of F_{ST}

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

$$\hat{\beta}^i_{WT} = \hat{F}^i_{ST} = \frac{\sum_l (\tilde{M}^i_{Wl} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})}$$

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:

$$\hat{F}_{WT} = \hat{\beta}_{WT} = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B}$$

and the parametric global value $F_{WT} = (\theta_W - \theta_B)/(1 - \theta_B)$.

Simple Computing Equations for F_{ST}

For large sample sizes and r populations:

$$\begin{aligned}\tilde{M}_W^i &\approx \sum_l [\tilde{p}_{il}^2 + (1-\tilde{p}_{il})^2] \\ \tilde{M}_W &= \frac{1}{r}\sum_{i=1}^{r} \tilde{M}_{Wl}^i = \sum_l [\bar{p}_l^2 + (1-\bar{p}_l)^2 + 2\frac{r-1}{r}s_l^2]\end{aligned}$$

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

For all sample sizes:

$$\begin{aligned}\tilde{M}_B^{ii'} &= \sum_l [\tilde{p}_{il}\tilde{p}_{i'l} + (1-\tilde{p}_{il})(1-\tilde{p}_{i'l})] \\ \tilde{M}_B &= \frac{1}{r(r-1)} \sum_{\substack{i=1 \\ i\neq i'}}^{r} \sum_{i'=i}^{r} \sum_l \tilde{M}_{Bl}^{ii'} \\ &= \sum_l [\bar{p}_l^2 + (1-\tilde{p}_l)^2 - 2\frac{1}{r}s_l^2]\end{aligned}$$

SNP-allele-based Estimates for F_{ST}

The population-specific estimates are

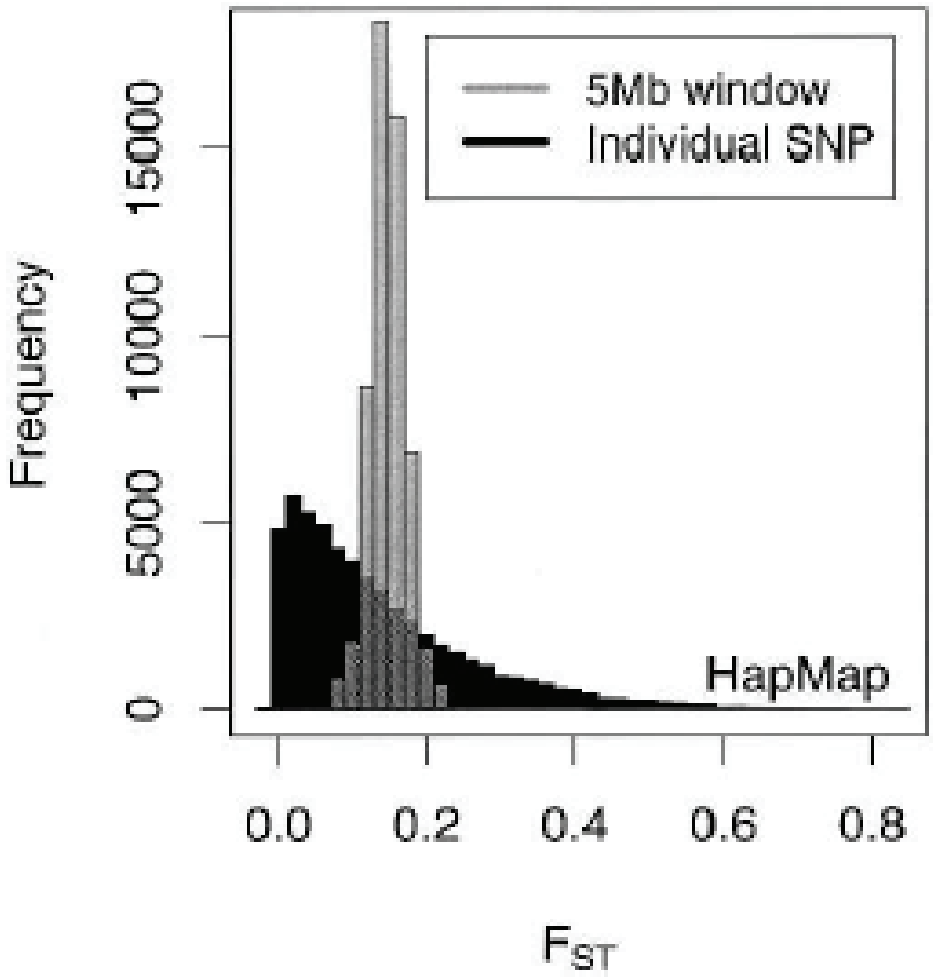
$$\widehat{F}^{i}_{WT} \;=\; 1 - \frac{\sum_l \tilde{p}_{il}(1-\tilde{p}_{il})}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

The global estimates are

$$\widehat{F}_{WT} \;=\; \frac{\sum_l (s_l^2)}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

The classical expression $s^2/\bar{p}(1-\bar{p})$ is fine if there is a large number of populations, but not for $r = 2$.

Effect of Number of Loci



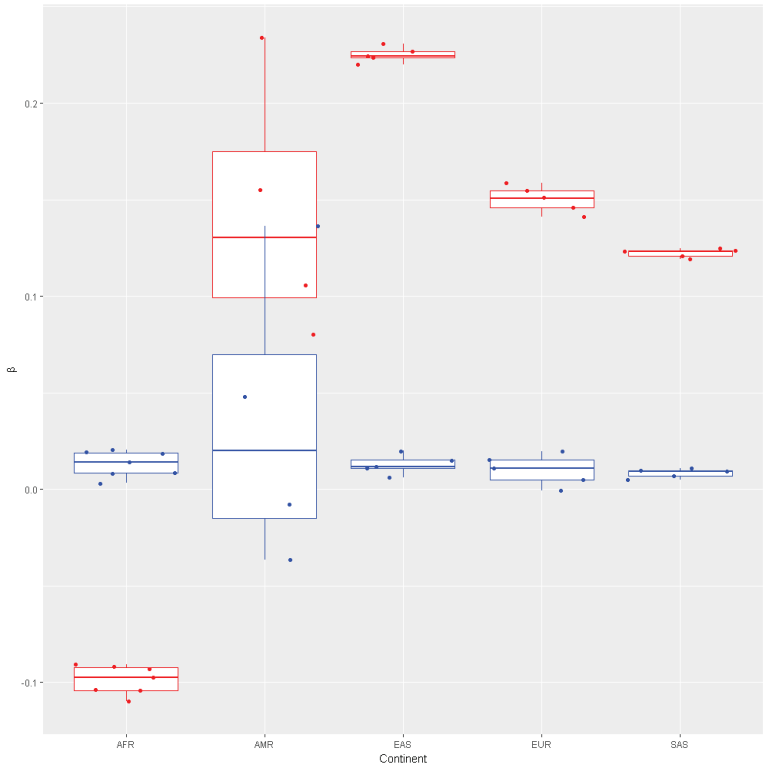
F_{ST} is relative, not absolute

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.

SNP F_{ST}'s are relative, not absolute



Blue box: Population relative to pairs of populations in same continent.

Red box: Population relative to pairs of populations in whole world.

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:

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

These have expected values

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

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

$$\hat{\theta}_{WC} = \frac{\sum_l(\mathrm{MSB}_l - \mathrm{MSW}_l)}{\mathrm{MSB}_l + (n_c-1)\mathrm{MSW}_l}$$

Weir & Cockerham 1984 Estimator

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

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

where

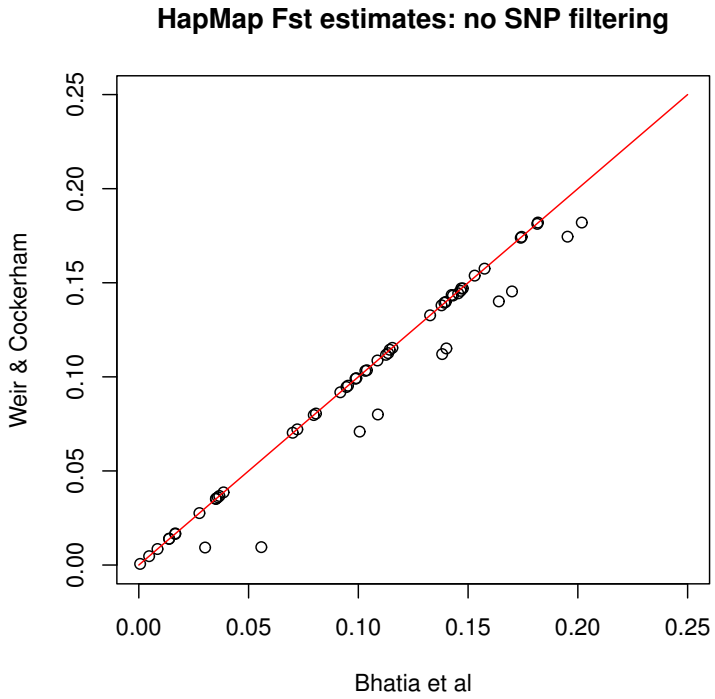
$$\theta_{Wc} = \frac{\sum_i n_i^c \theta^i}{\sum_i n_i^c} \quad , \quad \theta_{Bc} = \frac{\sum_{i \neq i'} n_i n_{i'} \theta^{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

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

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

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

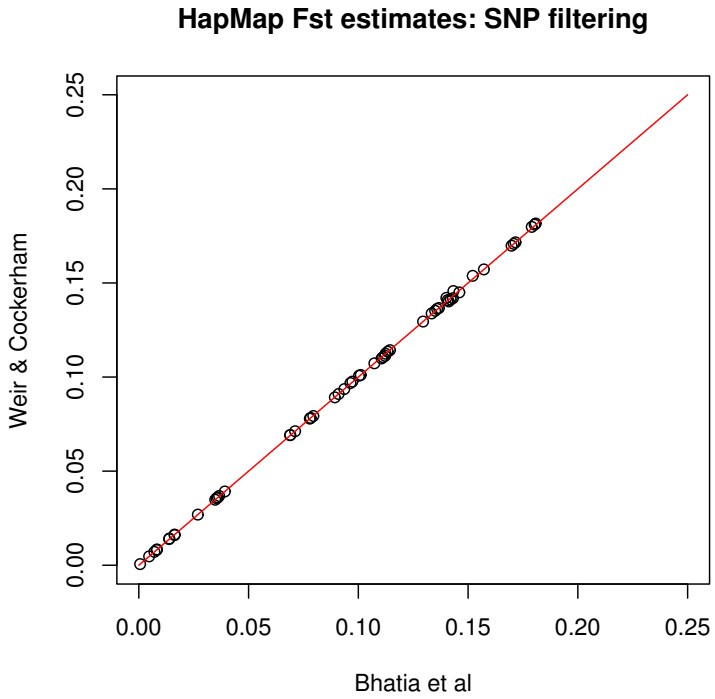
WC84 vs Beta Allele-based Estimators



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



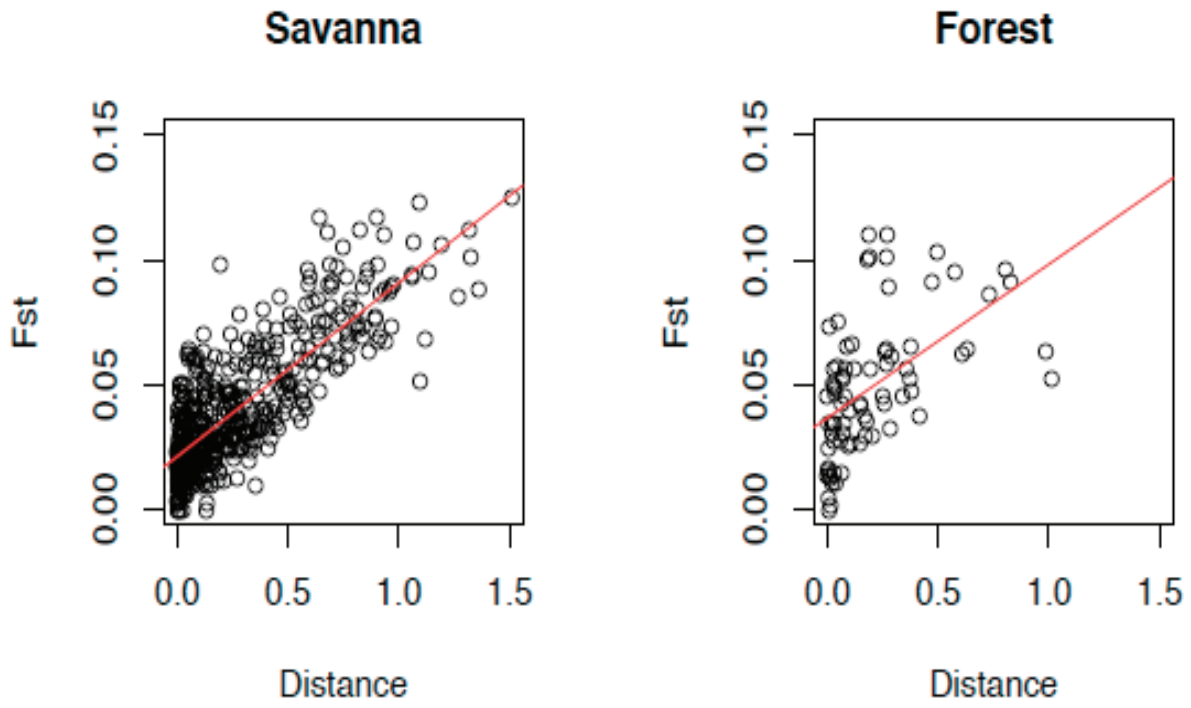
F_{WT} estimates for HapMap III, using the 42,463 SNPs on chromosome 1 that have at least five copies of the minor allele in samples from all 11 populations.

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

Evolutionary Inferences

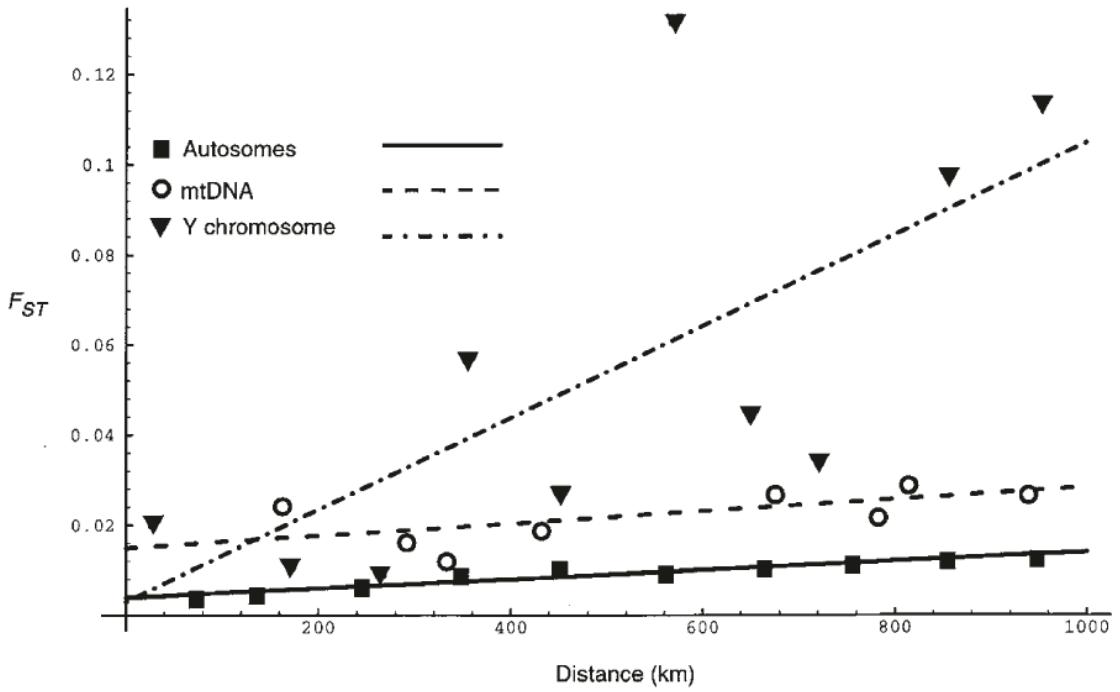
Geographic and Genetic Distances

From Slides 172 and 173, equilibrium values of F_{ST} for pairs of populations serve as measures of genetic diatance between populations, and so may reflect geographic distances also.



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

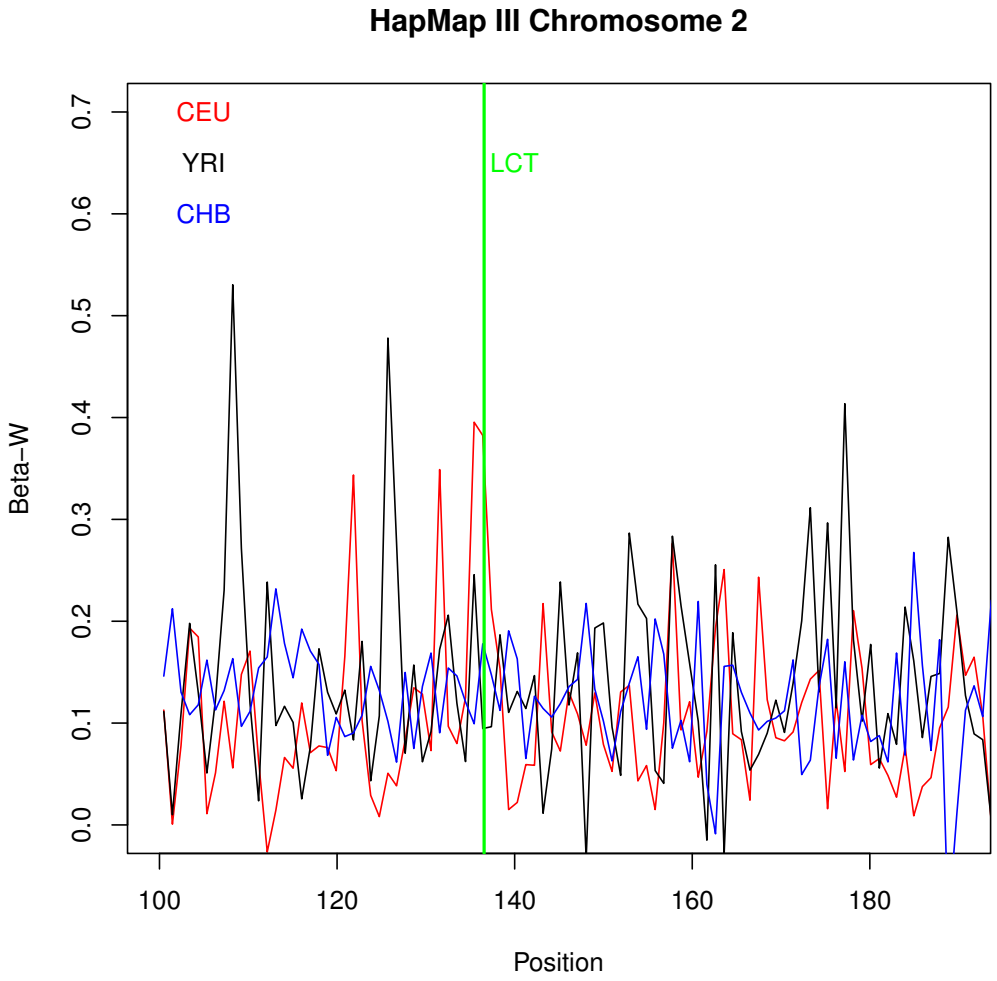
Human Migration Rates



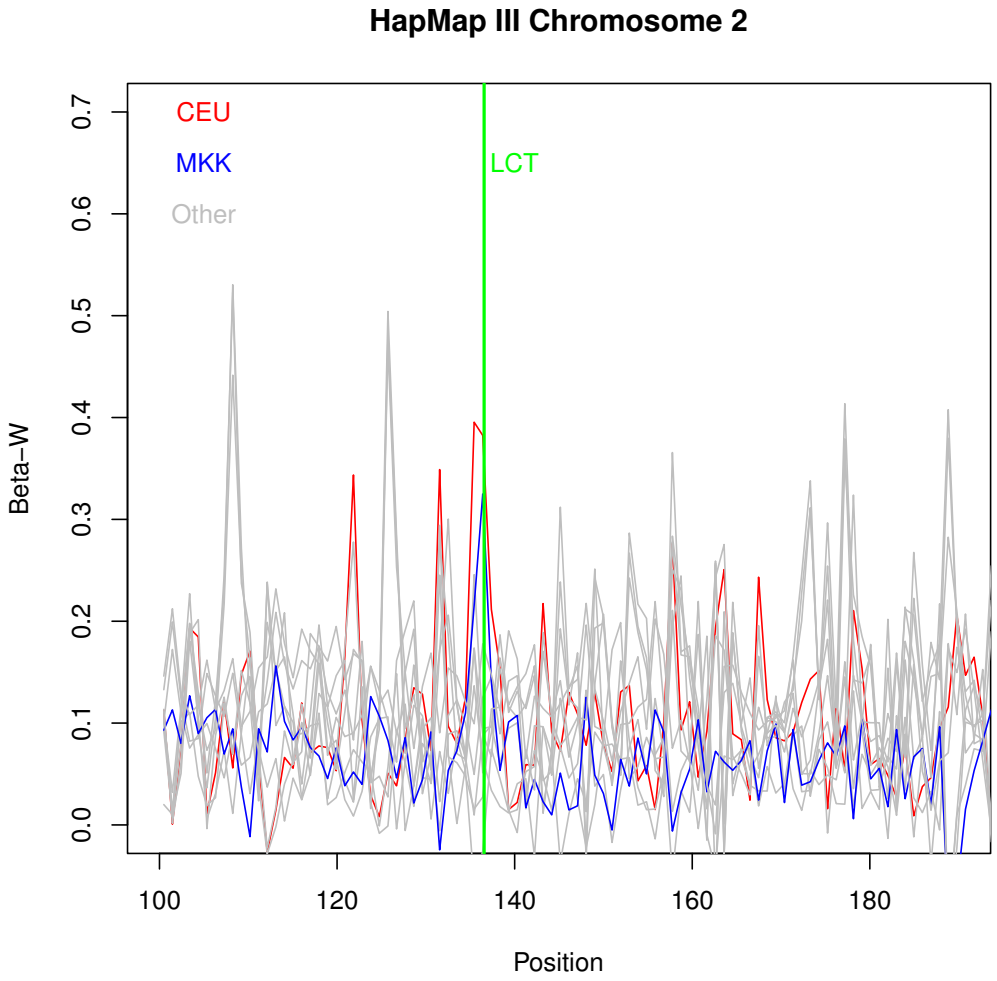
Suggests higher migration rate for human females among 14 African populations.

[Seielstad MT, Minch E, Cavalli-Sforza LL. 1998. Nature Genetics 20:278-280.]

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

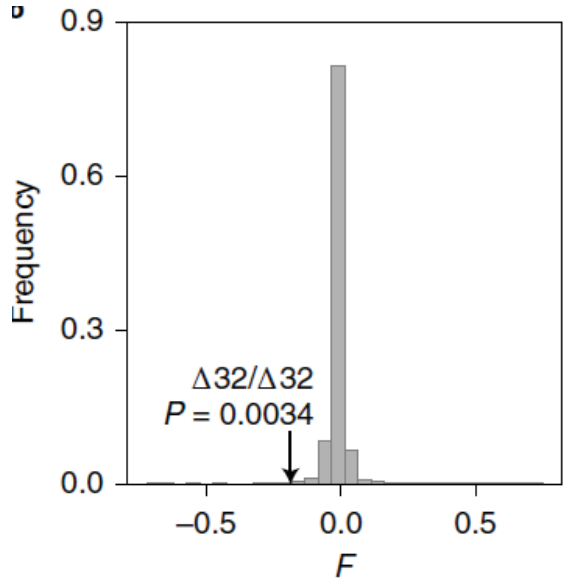
Analysis of HapMap 3 data using Fixation Index (Fst) identified genomic regions and single nucleotide polymorphisms (SNPs) as strong candidates for recent selection for lactase persistence and cholesterol regulation in 143156 founder individuals from the Maasai population in Kinyawa, Kenya (MKK). The 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]

CCR5-Δ32

Write the CCR5-Δ32 allele as A. The within-population inbreeding coefficient $\widehat{f}$ at CCR5 is lower than for other SNPs, suggesting selection against homozygosity.

If the CCR5-Δ32 allele is written as A, f is modified to $f(1-\tilde{p}_A)/\tilde{p}_A$.



[Wei X, Nielsen R. 2019. Nature Medicine 25:909-910.]