SISG
2014

SISG Module 10: Population Genetics and Association Mapping

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Module Overview
Refresher:
Genetic Data and Probability

Summer Institute in Statistical Genetics 2014

Module 10

Topic 1

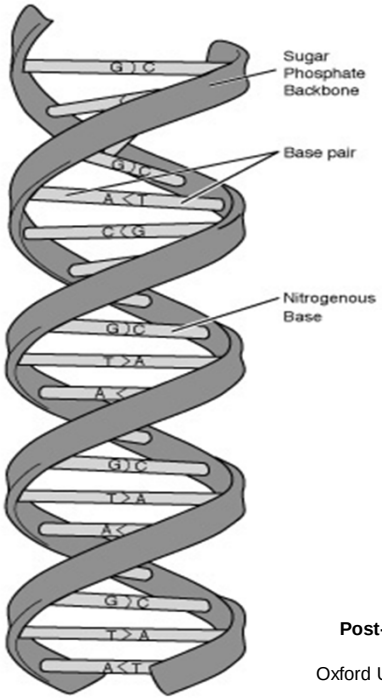
Topics

1 Introductory Material: genetic terminology, probability, Mendelain genetics (Kerr)
2 Allele frequencies and Hardy-Weinberg Equilibrium (Kerr)
3 Introduction to Linkage and Linkage disequilibrium (Kerr)
4 Case-Control Association (Thornton)
5 Quantitative traits and heritability (Thornton)
6 Identity by Descent, Kinship Coef, Coef of Fraternity (Kerr)
7 Testing with Related Samples (Thornton)
8 Estimating Relatedness (Thornton)
9 Population structure (Thornton)
10 Methods for Cryptic structure (Thornton)

Software for Genetic Studies: R, Haploview, Locuszoom, PLINK

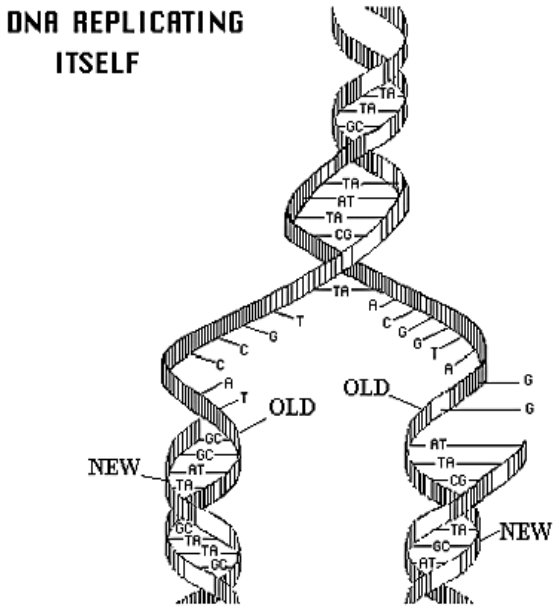
Instructors

Kathleen Kerr katiek@u.washington.edu
Tim Thornton tathornt@u.washington.edu



...ATGATGATCCAGC...
...TACTACTAGGTCG...

Post-genome Informatics
Minoru Kanehisa,
Oxford University Press, 2000



Central Dogma

"Genes are perpetuated as sequences of nucleic acid, but function by being expressed in the form of proteins."

(Lewin, GENES VII, page 31)

DNA

↓ transcription

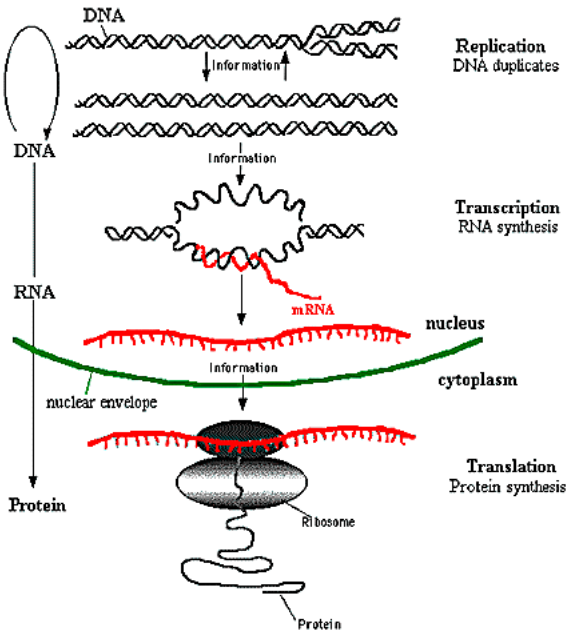
RNA

↓ translation

Protein

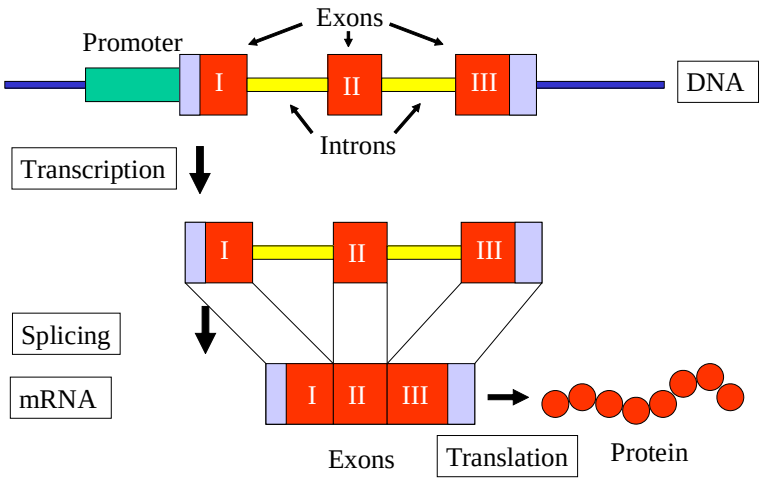
↓

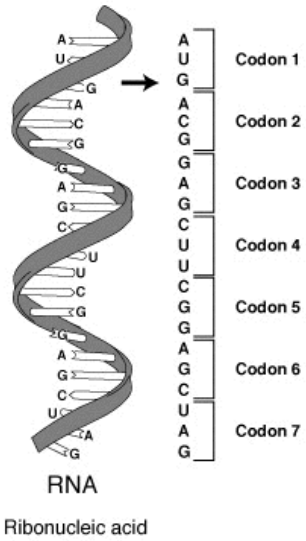
Phenotype



The Central Dogma of Molecular Biology

Post-genome Informatics
Minoru Kanehisa,
Oxford University Press, 2000



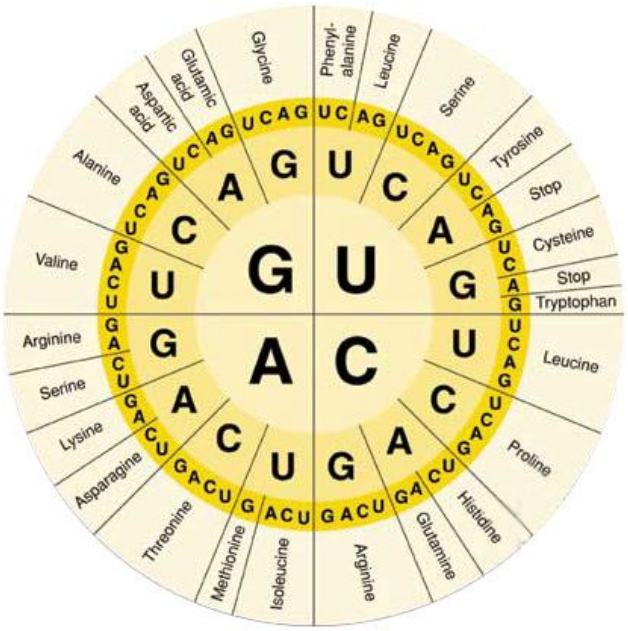


The genetic code defines a mapping between 3-tuples of nucleotides (called **codons**) and the 20 amino acids that comprise proteins.

There are also “start” and “stop” codons.

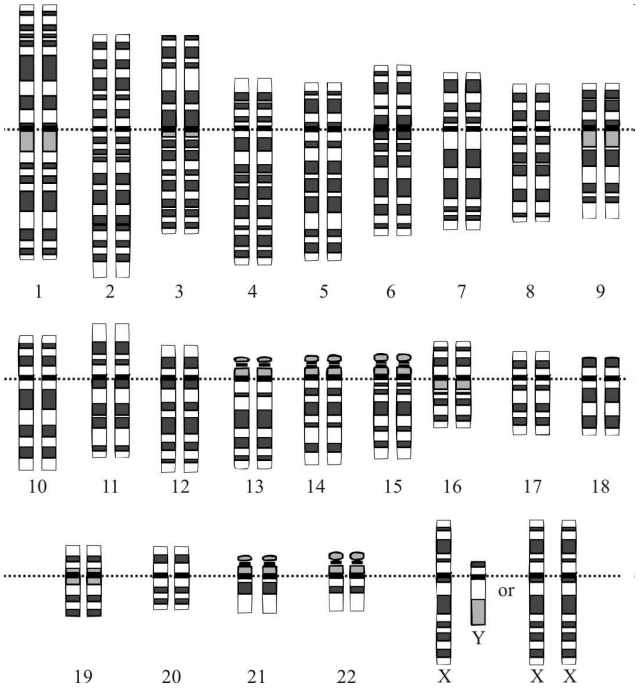
The genetic code has redundancy but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them specifies any other amino acid (no ambiguity).

The Genetic Code



Some Terminology

- DNA is organized into **chromosomes.** Humans have 22 pairs of **autosomes** plus the sex chromosomes. Mammalian cells are **diploid**, meaning that chromosomes come in pairs.
- Chromosomes have two “arms” connected by the **centromere**. The shorter of the two arms extending from the centromere is called **p**. The the longer is called **q**.



More Terminology

- A specific location on a chromosome, for instance the location of a gene, a **SNP** (single-nucleotide polymorphism), or another genetic marker, is a **locus** (plural: **loci**). There can be more than one form of a locus. These forms are called **alleles**. When there is more than one allele at a locus, the locus is said to be **polymorphic**.

Still More Terminology

- **Mitosis** is cell division that yields two identical **diploid** cells, which have two of each chromosome. **Meiosis** is a special type of cell division that happens in reproductive tissue yielding **haploid** cells, which have one of each chromosome, called **gametes**. In females, the gametes are the egg cells and in males the gametes are the sperm cells.

and Some More Terminology

- When two haploid gametes unite, the complete diploid number of chromosomes is reinstated. In sexual reproduction, each individual has one chromosome of **maternal** origin and one chromosome of **paternal** origin. Thus at any locus an individual has one allele of maternal origin and one allele of paternal origin. These define the individual's **genotype** at that locus. If an individual has two copies of the same allele, then that individual is **homozygous** at that locus. If an individual has two different alleles at a locus, then s/he is **heterozygous**.

Genetics Data Explosion

- Sequence data
  - Human Genome Project
  - 1000s genomes project
  - Mouse Genome Project
  - etc. etc.
- Genotype data
  - HapMap Project
  - SNP chips
  - Copy Number variation

Data Explosion

- Gene Expression data
  - Microarray gene expression data
  - RNA-seq data
- Proteomics data
- Metabalomics data

The Foundation of Statistics: Probability

Probability

Probability provides the language of data analysis. We should distinguish between a precise definition of probability, e.g.

- **Equiprobable outcomes definition**
  - Probability of E is number of outcomes in E divided by the total number of equally likely outcomes, e.g. probability of a head = 1/2
- **Long-run frequency definition**
  - If E occurs n times in N identical experiments, the probability of E is the limit of n/N as N goes to infinity

...and a subjective probability ("I'm 90% sure I turned off the lights")

Conditional Probabilities

All probabilities are conditional – probabilities depend on the context.

If E is an event or proposition of interest, and I is information or data, then P(E|I) is the "probability of E conditional on I" or "probability of E given I."

Conditional Probability Example

Role a fair die.

E=get a 4

I=get an even number

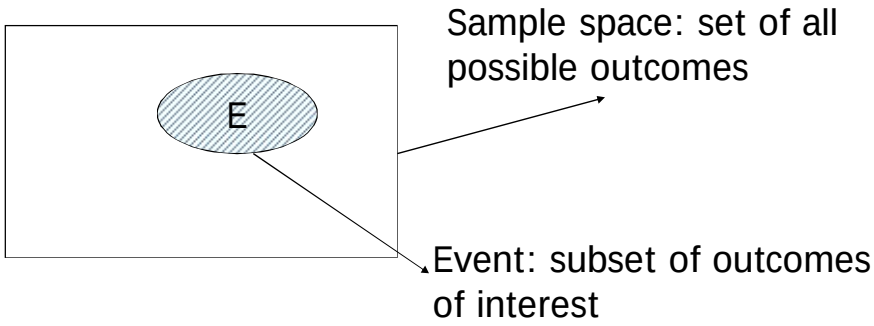
P(E | role a fair die)=1/6

P(E | role a fair die and I)=1/3

Comment: Conditional Probabilities and Hypothesis Testing

- A p-value results from statistical hypothesis testing
- Loosely, p-value ≈ P(data | null hypothesis)
- <u>Incorrectly</u>, a p-value is sometimes interpreted as P(null hypothesis | data)

Rules of Probability



Probability of event E:

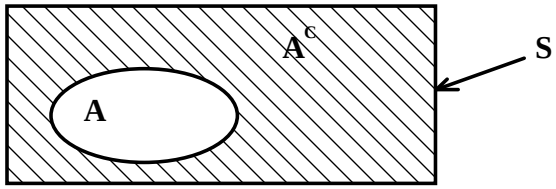
$$0 \le P(E) \le 1$$

Example: Events

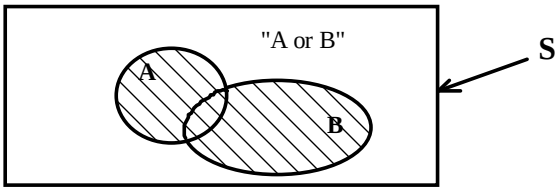
- A father's blood type is A and a mother's blood type is B.

- Event: "child has blood type A"
- Event: "child has an 'O' allele"

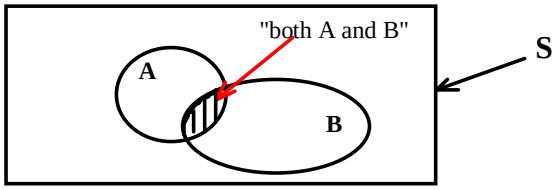
Complement event
A^c

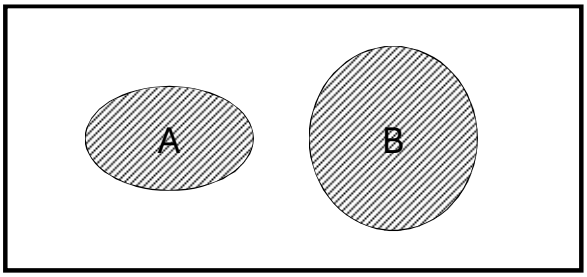


Union event
$A \cup B$ [A or B]



Intersection event
$A \cap B$ [A and B] [AB]





Disjoint or Mutually Exclusive events
[$A \cap B = \varnothing$ ("empty set")]

Example: A father's blood type is A and a mother's blood type is B.
A="first child has blood type A"
B="first child has blood type B"

Example: A and A^c are always mutually exclusive events

Rules of Probability

- Let A and B denote events and S denote the sample space

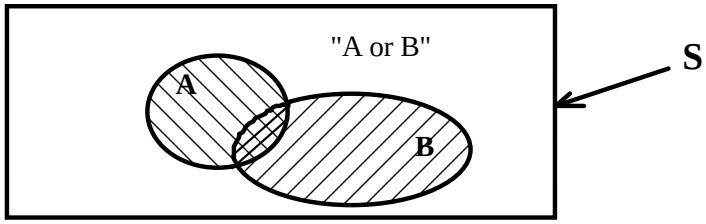
- Rule 1: $0 \leq P(A) \leq 1$
- Rule 2: $P(S) = 1$
- Rule 3: Addition rule for <u>disjoint</u> events

  $P(A \text{ or } B) = P(A) + P(B)$
- Rule 4: Complement rule

  $P(A^c) = 1 - P(A)$

Rules of Probability

- **<u>General addition rule</u>**: What is P(A or B)?



$$P(A \text{ or } B) = \frac{\#(A \text{ or } B)}{\#S} = \frac{(\#A) + (\#B) - (\#AB)}{\#S}$$
$$= P(A) + P(B) - P(AB)$$

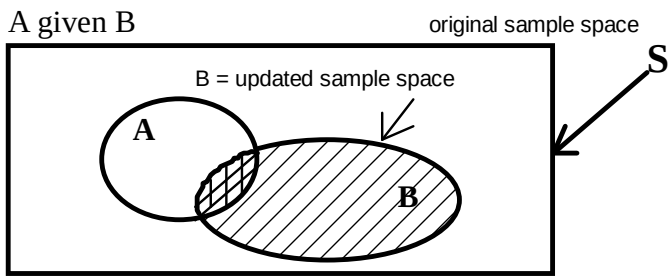
If events are disjoint, then P(A B) = 0 and

P(A or B) = P(A) + P(B)

Rules of Probability

- **<u>Conditional probability</u>**: What is P(A|B)?

  "the probability of A given B"



$$P(A|B) = \frac{P(A\ B)}{P(B)}$$

Rules of Probability

$$P(A|B) = \frac{P(A\ B)}{P(B)}$$

Equivalently: $$P(A\ B) = P(A|B)\ P(B)$$

Rules of Probability

<u>Definition:</u> Two events *A* and *B* are <u>independent</u> if

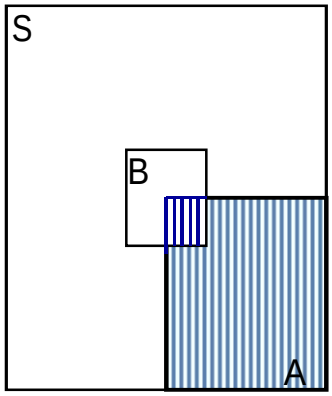
$$P(A|B)=P(A) \text{ or, equivalently, if } P(B|A)=P(B).$$

Interpretation:

If two events do not influence each other (that is, if knowing one has occurred does not change the probability of the other occurring), the events are independent.

Using the multiplicative rule, if two events *A* and *B* are independent, then

$$P(AB) = P(A)\,P(B).$$



Independence is harder to visualize

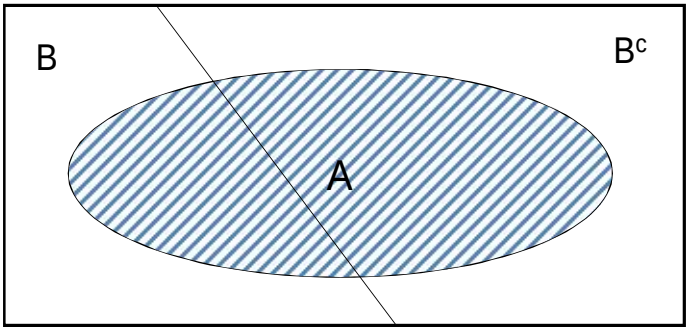
Independence means that P(A|B)=P(A)
[area of A in relation to S is the same as the area of AB in relation to B]

Caution

- "Independent" and "mutually exclusive" sound similar in everyday language. But in probability they mean VERY different things.

Rules of Probability

- Partitioning



$$A = (A \ \ and \ \ B) \ or \ (A \ \ and \ \ B^c)$$

Disjoint!

$$P(\mathrm{A}) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

Rules of Probability

- **<u>Bayes' theorem/Bayes' rule</u>**

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \text{not } B)P(\text{not } B)}$$

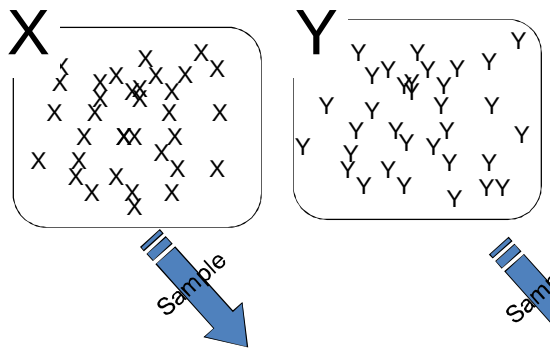
Definition of conditional probability

Partitioning

- Bayes' rule… not just for Bayesians!
- "The term 'controversial theorem' sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom." -Bradley Efron, *Science* 2013

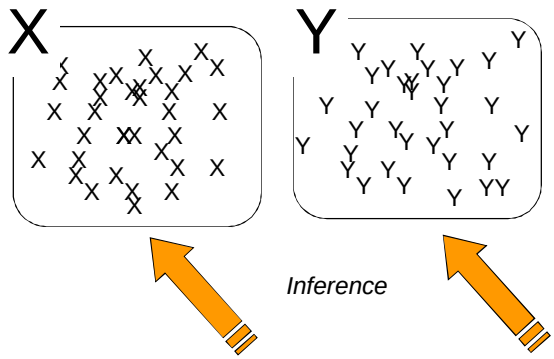
The Paradigm of Statistics: Statistical Inference

Basic Statistics: Two Populations to Compare (unobserved)
For example, we may wish to compare smokers to non-smokers; or wild-type individuals with mutant genotypes



observed:

$X_1, X_2, X_3, \ldots, X_m.$ $\quad$ $Y_1, Y_2, Y_3, \ldots, Y_n.$



statistical analysis:

$X_1, X_2, X_3, \ldots, X_m.$ $\quad Y_1, Y_2, Y_3, \ldots, Y_n.$

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir Ronald A. Fisher
1890-1962

Elements of Statistical Inference

- Estimator (point estimates)
- Confidence Intervals
- Hypothesis Testing

Estimator

- In statistics, we use data on a sample to estimate population parameters. For example, for a specific genetic variant, a population parameter is the frequency of that variant in the population.
- We like estimators that are:
  - unbiased: expected value is the parameter
  - consistent: increasing accuracy as sample size increases
  - efficient: small variance

Statistical Estimates

•Point estimates: We like point estimates that have low bias, are consistent, make efficient use of the data

population μ sample $X_1, X_2, X_3, X_4, X_5, X_6, X_7, \ldots$ data

Estimate of μ, the population mean	Unbiased?	Consistent?	Efficient?
median	Biased*	No*	NA
X_1	unbiased	No	No
X_{odd}	unbiased	Yes	No
X	unbiased	Yes	Yes

Components of Inference

•Confidence intervals: The set of parameter values for which the observed data are "typical" / the set of parameter values for which the observed data are "consistent"

•Example: I test a drug on 100 mice with a particular disease. 79% of the mice are cured. I estimate that the drug has a cure rate of 0.79 with 95% confidence interval 0.70 to 0.86.

Components of Inference

- p-values: What is the probability of observing data like our sample when there is nothing going on?
  - "Nothing going on" = null hypothesis
  - p-value ≈ P(data | null hypothesis)

Hypothesis Tests and P-values

- Null hypothesis H_0 and Alternative hypothesis H_A.
  - E.g., H_0: SNP is not associated with height.

    H_A: SNP is associated with height.

    (Hypotheses are statements about nature or about the population, not about the sample or about data)
- Test statistic T
- Statistical theory tells us the sampling distribution of T if the null hypothesis is true
- Sometimes, we do not know the distribution of T under H_0 and try to approximate it using the data (permutations, etc)
- The result of a Hypothesis Test is "reject H_0" or "fail to reject H_0"

Example, continued: an existing treatment for a disease has a well-established cure rate of 0.75. Do I have compelling evidence that my treatment is better?

H_0: cure rate=0.75

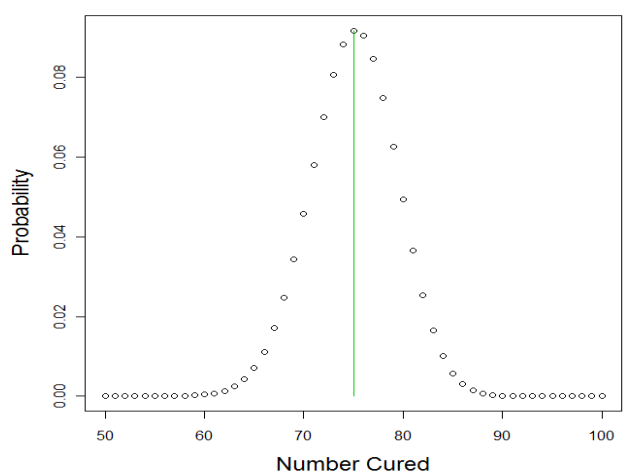
H_A: cure rate≠0.75

If the true cure rate is 0.75 and I observe 100 treated individuals, then the number cured has a Binomial distribution 100 "trials" and probability 0.75.

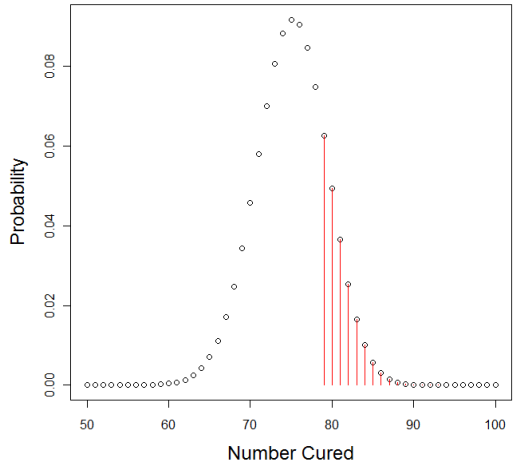
p-value=P(observe 79 or more extreme | H_0)

Binomial distribution for n=100 "trials" with probability 0.75 of "success"

For example, the probability of 75 successes is about 0.092.



The one-sided p-value is the probability of getting a result as extreme or more extreme than the observed data, 79. In this case the p–value≈0.21.



Hypothesis Testing: errors

•When we perform a hypothesis test, there are two ways we an make a mistake.

•**Type I error**: Reject the null hypothesis when it is true. We usually conduct a test in a way that ensures that this type of error is not too large. A popular threshold is 5%. In this case the "**size**" or "**α-level**" of a hypothesis test is 0.05 and we reject the null hypothesis if the p-value<0.05.

•**Type II error**: Fail to reject the null hypothesis when it is false.

- Next, we “warm up” by applying the basic rules of probability to some early experiments in genetics.

Module Warm-Up: Mendelian Genetics

Mendel's Laws

Mendel's first law: each character is controlled by a pair of genes. Two genes segregate into different gametes during meiosis.

Mendel's second law: When two or more pairs of genes segregate, they do so independently.

Classical Mendelian experiments use inbred strains of animals or self-fertilized plants so that individuals in the starting generation are homozygous at every locus and genetically identical.

Example 1: Peas experiment

round X wrinkled

↓

round

↓

round X wrinkled

3 : 1

AA X aa

↓

Aa

↓

AA : Aa : aa

1 : 2 : 1

round wrinkled

Example 2 (Fisher): Plant stem length

tall X short

↓

tall F1

↓

tall : short F2

3 : 1

F2 generation: AA : Aa : aa

Example 2

Question: How to distinguish the AA and Aa genotypes?

Fisher's answer: grow 10 offspring from each tall F2 plant. If all offspring are tall, classify at AA; otherwise classify as Aa.

P(an offspring is tall | Aa)= $\frac{3}{4}$

P(all 10 offspring are tall | Aa)= $\left(\frac{3}{4}\right)^{10} \approx 0.056$

Therefore, about 5% of Aa plants in F2 will be misclassified as AA.

Among tall plants in the F2 generation, the expected ratio of AA:Aa is 1:2. However, among those labeled "AA" and "Aa," the expected ratio is about 37:63

Example 3: Rabbits

grey X albino

↓

grey F1

↓

grey : black : albino F2

9 : 3 : 4

F2 generation: AA : Aa : aa – single gene model with two alleles cannot explain these data

Example 3: Rabbits

Color is controlled by two genes
1.Presence of color C, no color c
2.Grey G or black b

CC or Cc	→	GG or Gb	grey rabbit
		bb	black rabbit
cc	→	albino rabbit regardless of other locus	

Example 3: Rabbits

CCGG X ccbb

↓

CcGb

↓

F_2	CG	Cb	cG	cb
CG	grey	grey	grey	grey
Cb	grey	black	grey	black
cG	grey	grey	albino	albino
cb	grey	black	albino	albino

Example 4: Mice

grey X chocolate

↓

grey F1

↓

grey : black : chocolate F2

12 : 3 : 1

Example 4: Mice

Genetic Model
1. Grey allele G dominates allele g
2. Black allele B dominates chocolate allele b

GG or Gg	→	grey mouse
gg and (BB or Bb)	→	black mouse
gg and bb	→	chocolate mouse

Example 4: Exercise

Show how this genetic model explains the observed phenotype relative frequencies.

F_2	GB	Gb	gB	gb
GB				
Gb				
gB				
gb				

Example 5: Sweet Pea Flower Color

white X white (two different strains)

↓

purple F1

↓

purple : white F2

9 : 7

Genetic Model: A dominates a
B dominates b

Need at least one A and one B to have purple; otherwise, white

Example 5: Sweet Pea Flower Color

AAbb X aaBB

↓

AaBb

↓

F_2	AB	Ab	aB	ab
AB				
Ab	purple		purple	
aB	purple	purple		
ab	purple			

Example 6: Bean Flower Color

Flowers come in shades from white to purple. Quantify color: white (0) to purple (10)

10 X 0
purple X white
↓
5
↓

color	10	9	8	7	6	5	4	3	2	1	0
relative count	1	0	2	2	1	4	1	2	2	0	1

Genetic Model: Two genes with additive effects:

Gene 1: A=3, a=0

Gene 2: B=2, b=0

Example 6: Bean Flower Color

AABB X aabb
↓
AaBb
↓

F_2	AB	Ab	aB	ab
AB	10	8	7	5
Ab	8	6	5	3
aB	7	5	4	2
ab	5	3	2	0

Example 7: Human Blood Groups

Four phenotypes in the ABO blood group: A, B, AB and O

Two loci theory (Von Dungen & Hirszfeld 1910, Otlenberg 1923).
Locus 1 has A dominant to a
Locus 2 has B dominate to b

phenotype	genotype
O	aabb
A	AAbb or Aabb
B	aaBB or aaBb
AB	AABB or AaBB or AABb or AaBb

Example 7: Human Blood Groups

Four phenotypes in the ABO blood group: A, B, AB and O

One locus theory
One locus with three alleles A, B, and O

phenotype	genotype
O	OO
A	A A or A O
B	B B or B O
AB	A B

Example 7: Exercise

What are the possible blood types among offspring from each type of mating?

Parental Blood Types	Two loci model	Single locus model
O x O		
O x A		
O x B		
A x A		
A x B		
B x B		
O x AB		
A x AB		
B x AB		
AB x AB		

Allele Frequencies and Hardy-Weinberg Equilibrium

Summer Institute in Statistical Genetics 2014

Module 10

Topic 2

Allele Frequencies and Genotype Frequencies

How do allele frequencies relate to genotype frequencies in a population?

If we have genotype frequencies, we can easily get allele frequencies.

Example

Cystic Fibrosis is caused by a recessive allele. The CF locus is on the long arm of chromosome 7 (7q31). Of 10,000 Caucasian births, 5 were found to have Cystic Fibrosis and 442 were found to be heterozygous carriers of the mutation that causes the disease. Denote the Cystic Fibrosis allele with cf and the normal allele with N. Based on this sample, how can we estimate the allele frequencies in the population?

$$\text{In the sample, } \frac{5}{10000} \text{ are } cf, cf$$
$$\frac{442}{10000} \text{ are } cf, N$$
$$\frac{9553}{10000} \text{ are N, N}$$

Example, con't

So we use 0.0005, 0.0442, and 0.9553 as our estimates of the genotype frequencies in the population. The only assumption we have used is that the sample is a random sample. Starting with these genotype frequencies, we can estimate the allele frequencies without making any further assumptions:

$$\text{Out of } 20{,}000 \text{ alleles in the sample,}$$
$$\frac{442+10}{20{,}000} = .0226 \text{ are } cf$$
$$1 - .0226 = 0.9774 \text{ are } N$$

Hardy-Weinberg Assumptions

In contrast, going from allele frequencies to genotype frequencies requires more assumptions.

Hardy-Weinberg model

- infinite population
- discrete generations
- random mating
- no selection
- no migration in or out of population
- no mutation
- equal initial genotype frequencies in the two sexes

Consider a locus with two alleles A and a

1st generation

genotype	frequency
AA	u
Aa	v
aa	w

$u+v+w=1$

From these genotype frequencies, we can quickly calculate allele frequencies:

$P(A)=u+\frac{1}{2}v$

$P(a)=w+\frac{1}{2}v$

2nd generation

mating type	mating frequency*	expected progeny
AA x AA	u^2	AA
AA x Aa	2uv	½ AA + ½ Aa
AA x aa	2uw	Aa
Aa x Aa	v^2	¼ AA + ½ Aa + ¼ aa
Aa x aa	2vw	½ Aa + ½ aa
aa x aa	w^2	aa

*check that $u^2 + 2uv + 2uw + v^2 + 2vw + w^2 = (u+v+w)^2 = 1^2 = 1$

For generation 2:

$p \equiv P(AA) = u^2 + \frac{1}{2}(2uv) + \frac{1}{4}v^2 = (u + \frac{1}{2}v)^2$

$q \equiv P(Aa) = uv + 2uw + \frac{1}{2}v^2 + vw = 2(u + \frac{1}{2}v)(\frac{1}{2}v + w)$

$r \equiv P(aa) = \frac{1}{4}v^2 + \frac{1}{2}(2vw) + w^2 = (w + \frac{1}{2}v)^2$

For generation 3:

$P(AA) = (p + \frac{1}{2}q)^2 = [(u + \frac{1}{2}v)^2 + \frac{1}{2}2(u + \frac{1}{2}v)(\frac{1}{2}v + w)]^2$

$= [(u + \frac{1}{2}v)[(u + \frac{1}{2}v) + (\frac{1}{2}v + w)]]^2$

$= [(u + \frac{1}{2}v)(u + v + w)]^2$

$= [(u + \frac{1}{2}v)(1)]^2$

$= [u + \frac{1}{2}v]^2$

$= p$ ... the same as generation 2

Similarly, in generation 3 P(Aa)=q and P(aa)=r.

<u>Equilibrium</u> is reached after one generation of mating under the Hardy-Weinberg assumptions. Genotype frequencies remain the same from generation to generation.

Hardy-Weinberg Genotype Frequencies

When a population is in Hardy-Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population. We have the following relationship between genotype frequencies and allele frequencies for a population in Hardy-Weinberg equilibrium:

$P(AA) = P(A)P(A)$

$P(Aa) = 2P(A)P(a)$

$P(aa) = P(a)P(a)$

For example, consider a diallelic locus with alleles A and B with frequencies 0.85 and 0.15, respectively. If the locus is in HWE, then the genotype frequencies are:

$P(AA) = 0.85 * 0.85 = 0.7225$

$P(AB) = 0.85*0.15 + 0.15*0.85 = 0.2550$

$P(BB) = 0.15*0.15 = 0.0225$

Example

Establishing the genetics of the ABO blood group system was one of the first breakthroughs in Mendelian genetics. The locus corresponding to the ABO blood group has three alleles, A, B and O and is located on chromosome 9q34. The alleles A and B are dominant to O. This leads to the following genotypes and phenotypes:

Genotype	Blood type
AA, AO	A
BB, BO	B
AB	AB
OO	0

Mendel's first law allows us to quantify the types of gametes an individual can produce. For example, an individual with type AB produces gametes A and B with equal probability (1/2).

Example, con't

From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be P(A)=0.2877, P(B)=0.1065 and P(O)=0.6057. If an individual has blood type B, what gametes can be produced and with what frequency? (Note where HWE is invoked in the following)

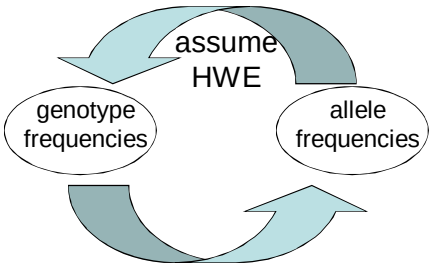
If a person has blood type B, then the genotype is BO or BB.

$$\gamma_1 = P(\text{genotype } BO \,|\, \text{blood type } B) = \frac{2p_O p_B}{2p_O p_B + p_B^2} = 0.92$$

$$\gamma_2 = P(\text{genotype } BB \,|\, \text{blood type } B) = \frac{p_B^2}{2p_O p_B + p_B^2} = 0.08$$

$$P(B \text{ gamete} \,|\, \text{blood type } B) = 1 \times \gamma_2 + \frac{1}{2}\gamma_1 = 0.54$$

$$P(O \text{ gamete} \,|\, \text{blood type } B) = 1 - 0.54 = 0.46$$



Why should we be skeptical of the HW assumptions?

- Small population sizes. Chance events can make a big difference.
- Deviations from random mating.
  - Assortive mating. Mating between genotypically simlar individuals increases homozygosity for the loci involved in mate choice without altering allele frequencies.
    - E.g., Assortive mating has been hypothesized to be a cause of rising rates of autism (very controversial hypothesis)
  - Inbreeding. Mating between close relatives increases homozygosity for the whole genome without affecting allele frequencies.
  - Population sub-structure
  - Disassortive mating. Mating between dissimilar individuals increases heterozygosity without altering allele frequencies.
- Mutation
- Migration
- Selection

Testing Hardy-Weinberg Equilibrium

When a locus is not in HWE, then this suggests one or more of the Hardy-Weinberg assumptions is false. Departure from HWE has been used to infer the existence of natural selection, argue for the existence of assortive (non-random) mating, and infer genotyping errors.

It is therefore of interest to test whether a population is in HWE at a locus. We will discuss the two most popular ways of testing HWE

1. Chi-Square test
2. Exact test

Chi-Square Goodness-Of-Fit Test

Compares observed genotype counts with the values expected under Hardy-Weinberg. For a locus with two alleles, we might construct a table as follows:

Genotype	Observed	Expected
AA	n_{AA}	np^2
Aa	n_{Aa}	$2np(1-p)$
aa	n_{aa}	$n(1-p)^2$

where $p \equiv p(A) = (n_{Aa} + 2\, n_{AA})/2n$

The test statistic is:

$$X^2 = \sum_{genotypes} \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

Observed	Expected if HWE true
N_{AA}	$N(p_A)^2$
N_{Aa}	$N2p_A(1-p_A)$
N_{aa}	$N(1-p_A)^2$

In the application to HWE, a convenient form for the test statistic is:

$$X^2 = n\left(\frac{4n_{AA}n_{aa} - n_{Aa}^2}{(2n_{AA} + n_{Aa})(2n_{aa} + n_{Aa})}\right)^2.$$

The sampling distribution of the test statistic under the null hypothesis is approximately a χ^2 distribution with 1 degree of freedom.

There is a rule of thumb for such χ^2 tests: the expected count should be at least 5 in every cell. If allele frequencies are low, and/or sample size is small, and/or there are many alleles at a locus, this may be a problem.

Exact Test

The Hardy-Weinberg exact test is based on calculating probabilities

P(genotype counts | allele counts)

under HWE.

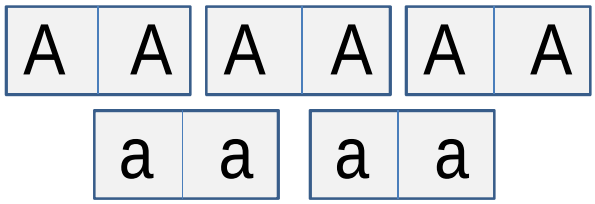
Example: Exact Test

Suppose we have a sample of 5 people and we observe genotypes AA, AA, AA, aa, and aa.

If five individuals have among them 6 ‘A’ alleles and 4 ‘a’ alleles, what genotype configurations are possible?

Permutation Test

- Make a set of five index cards to represent the 5 observed genotypes:



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

aa	Aa	AA	empirical Probability (random permutations)	theoretical probability
2	0	3		0.048
1				0.571
0				0.381

Example

Suppose we have a sample of 100 individuals and 21 ‘a’ alleles are observed (so 200-21=179 ‘A’ alleles).

Note that specifying the number of heterozygotes determines the number of AA and aa genotypes.

aa	Aa	AA	probability
10	1	89	<<.000001
9	3	88	<<.000001
			<.000001
			.000001
			.000047
			.000870
			.009375
			.059283
			.214465
1	19	80	.406355
0	21	79	.309604

Wiggington, Cutler, Abecasis, AJHG 2005

The formula is:

$$P(n_{Aa} \mid n_A, n_a, \text{HWE}) = \frac{n!}{n_{AA}!\,n_{Aa}!\,n_{aa}!}\,\frac{2^{n_{Aa}}\,n_A!\,n_a!}{(2n)!}$$

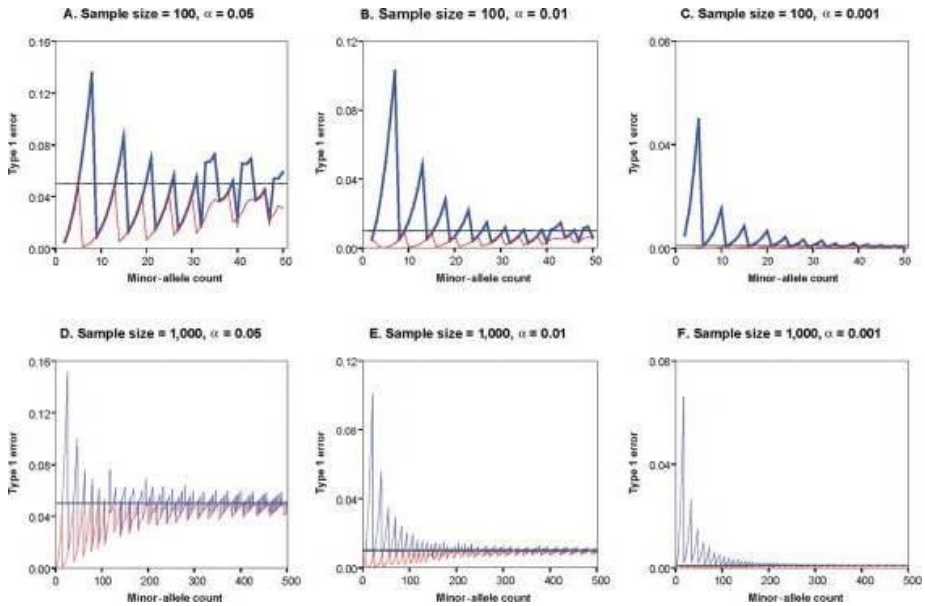
If we had actually observed 13 heterozygotes in our sample, then the exact test p-value would be

≈.009375+.000870+.000047+.000001≈0.010293

(To get the p-value, we sum the probabilities of all configurations with probability equal to or less that the observed configuration.)

How do the exact test and the χ^2 test compare?

Next slide is Figure 1 from Wigginton et al (AJHG 2005). The upper curves give the type I error rate of the chi-square test; the bottom curves give the type I error rate from the exact test. The exact test is always conservative; the chi-square test can be either conservative or anti-conservative.



Tests of HWE: Which one is best?

- The Exact Test should be preferred for smaller sample sizes and/or multiallelic loci, since the χ^2 test is *prima facie* not valid in these cases (rule of thumb: must expect at least 5 in each cell)
- The coarseness of Exact Test means it is conservative. In Example 4, we reject the null hypothesis that HWE holds if 13 or fewer heterozygotes are observed. But the observed p-value is actually 0.010293. Thus to reject at the 0.05 level, we actual have to see a p-value as small as 0.010293.

Tests of HWE: Which one is best?

- The χ^2 test can have inflated type I error rates. Suppose we have 100 genes for which HWE holds. We conduct 100 χ^2 tests at level 0.05. We expect to reject the null hypothesis that HWE holds in 5 of the tests. However, the results of Wiggington et al (AJHG, 2005) say, on average, it can be more than 5 depending on the minor allele count. Although it is not desirable for a test to be conservative (Exact Test), an anti-conservative test is considered unacceptable.
  - Wiggington et al (AJHG, 2005) give an extreme example with a sample of 1000 individuals. At a nominal α=0.001, the true type I error rate for the χ^2 test exceeds 0.06.

Tests of HWE: Which one is best?

- The χ^2 test is a two-sided test. In contrast, the Exact Test can be made one-sided, if appropriate. Specifically, one can test for a deficit of heterozygotes (if one suspects inbreeding or population stratification); test for an excess of heterozygotes (which indicate genotyping errors for some genotyping technologies).
- For both tests, p-values do not have a uniform distribution under the null hypothesis. This is problematic for making inference when conducting lots of tests (e.g. qq plots).

Tests of HWE: Which one is best?

- Summary/Conclusion: The Exact Test is better, but it is not great. It tends to be conservative; has limited power with typical sample sizes; and p-values are not uniformly distributed. However, it is valid with small sample sizes.

Power for Testing HWE

The Chi-Square test, though not the preferred test for small samples, provides a convenient way to investigate power.

In the two allele case, it can be shown that the test statistic

$$X^2 = \sum_{genotypes} \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

is algebraically equal to

Power for Testing HWE

$$X^2 = \sum_{genotypes} \frac{(O - \mathrm{E})^2}{E} = n\hat{f}^2$$

where

$$\hat{f} = 1 - \frac{2n \times n_{Aa}}{n_A n_a}$$

f is also the "inbreeding coefficient" of the population (more later).

Power for Testing HWE

When HWE holds, X^2 has a chi-square distribution with 1 df.

When HWE does not hold, X^2 has a non-central chi-square distribution with non-centrality parameter nf^2.

The cut-off for significance at the 5% level of a chi-square with 1 df is 3.84. That is, our p-value will be less than 0.05 if we observe a test statistic greater than 3.84.

In order to be at least 90% sure of rejecting HWE when HWE is false, the non-centrality parameter should be at least 10.51.

Power for Testing HWE

$$nf^2 \geq 10.51$$

$$n \geq \frac{10.51}{f^2}$$

If f=0.01, then n has to be over 100,000.

Observing Phenotypes

- What if we cannot see genotypes? The observed data are phenotypes, some of which correspond to multiple genotyeps.

Example: HWE and Human Blood Types

Suppose we have a sample of size N from a population and the data are the counts of the phenotypes n_O, n_A, n_B, and n_{AB} ($n_O + n_A + n_B + n_{AB}=N$). If we had genotypes, it would be easy to estimate allele frequencies. But we have observed phenotypes, which are aggregates of genotypes.

Given the phenotype data, how do we estimate the allele frequencies r, p, and q?

allele	O	A	B
frequency	r	p	q

$r+p+q=1$

phenotype	genotype	genotype frequency under HWE
O	OO	r^2
A	AO or AA	$2pr + p^2$
B	BO or BB	$2qr + q^2$
AB	AB	$2pq$

- If we could observe all genotype counts (i.e., n_{AO} and n_{AA} not just n_A; n_{BO} and n_{BB} not just n_B), then our estimates of allele frequencies would be:

$$\hat{p} = \frac{2n_{AA} + n_{AO} + n_{AB}}{2N}$$

$$\hat{q} = \frac{2n_{BB} + n_{BO} + n_{AB}}{2N}$$

$$\hat{r} = \frac{2n_{O} + n_{AO} + n_{BO}}{2N}$$

- On the other hand, if we knew p,q,r, we could estimate n_{AO} , n_{AA} , n_{BO} and n_{BB} .

Gene-counting algorithm (an EM algorithm)

Gene-Counting Algorithm:

1. Select starting estimates p_0, q_0, r_0
2. Estimate n_{AO} , n_{AA} , n_{BO} and n_{BB}
3. Use estimates of n_{AO} , n_{AA} , n_{BO} and n_{BB} to get new estimates of p,q,r: p_1, q_1, r_1.
4. Repeat step 2 (estimation step) and step 3 (which is really a maximization step). Stop when the estimates of p,q,r do not change more than a tiny amount.

Note that completing step 2 requires assuming HWE.

Linkage and Linkage Disequilibrium

Summer Institute in Statistical Genetics 2014

Module 10

Topic 3

Linkage

Linkage in a simple genetic cross

In the early 1900's Bateson and Punnet conducted genetic studies using sweet peas. They studied two characters: petal color (P purple dominates p red) and pollen grain shape (L elongated is dominant to l disc-shaped).

PPLL X ppll

↓

PpLl

Plants in the F_1 generation were intercrossed: PpLl X PpLl. According to Mendel's second law, "during gamete formation, the segregation of one gene pair is independent of other gene pairs."

The table helps us calculate the expected relative frequencies of the four types of plants in the F_2 generation according to Mendel's second law.

	PL	Pl	pL	pl
PL	Purple Long	Purple Long	Purple Long	Purple Long
Pl	Purple Long	Purple disc-shaped	Purple Long	Purple disc-shaped
pL	Purple Long	Purple Long	red Long	red Long
pl	Purple Long	Purple disc-shaped	red Long	red disc-shaped

Here are the expected relative frequencies of the four phenotypes of plants:

	Elongated	disc-shaped
Purple	9	3
red	3	1

Here are the observed data:

	Elongated	disc-shaped
Purple	284	21
red	21	55

The observed data clearly do not fit what is expected under the model. The reason is that the loci are linked.

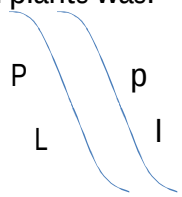
Linkage

Genes are physically arranged in linear strands of DNA and grouped into chromosomes. When a gamete is formed, chunks of a chromosome are passed on. Suppose we have two genes, one with alleles A1 and A2 and another with alleles B1 and B2, that are physically close on a chromosome. Suppose an individual is heterozygous at both loci and, furthermore, the **phase** is as follows:

A1	A2
B1	B2

If the genes are closely linked, a gamete is much more likely to contain (A1,B1) or (A2,B2) - "non- recombinants." If there is recombination, a gamete contains (A1, B2) or (A2,B1), but these two possibilities are less likely. (In contrast Mendel's second law says that all four possibilities are equally likely.)

In sweet peas the loci controlling petal color and pollen grain shape are physically close together on the same chromosome. In the F_1 generation, the **phase** of the PpLl plants was:



Both dominant alleles were located on the same chromosome. A sweet pea was equally likely to pass on a gamete containing P as a gamete containing p, and equally likely to pass on a gamete containing L as a gamete containing l. However, two events such as "pass on P" and "pass on L" were NOT independent because of linkage. A gamete is much more likely to contain PL or pl than to contain Pl or pL. A gamete contains Pl or pL only if there is a recombination event between the two loci when the gamete is formed.

Genotypes and Haplotypes

Haplotype: A combination of alleles or DNA bases on the same (physical) chromosome.

Terminology	
Genotype AaBb "phase is unknown"	**Haplotype** AB/ab or Ab/aB "phase is known"

Recombination Fraction

When two loci follow Mendel's Second law, recombinants and non-recombinants are produced with equal frequency. When loci are physically close to one another on a chromosome, there is a deviation from this relationship. This deviation is summarized by the **recombination fraction c** (sometimes denoted by θ).

$$c = P(\text{recombinant gamete})$$

When loci are unlinked, c=1/2.

When loci are completely linked, c=0.

For c in between 0 and 1/2, the loci are said to be "linked" or "in genetic linkage."

How might we estimate the recombination fraction from data from Bateson and Punnet's sweet pea intercross?

P(recombinant gamete)=c

		½ (1-c)	½ c	½ c	½ (1-c)
		PL	Pl	pL	pl
½ (1-c)	PL	Purple Long	Purple Long	Purple Long	Purple Long
½ c	Pl	Purple Long	Purple disc-shaped	Purple Long	Purple disc-shaped
½ c	pL	Purple Long	Purple Long	red Long	red Long
½ (1-c)	pl	Purple Long	Purple disc-shaped	red Long	red disc-shaped

We can form a likelihood for the data that is a function of the recombination fraction c. We can find the value of c that maximizes this likelihood.

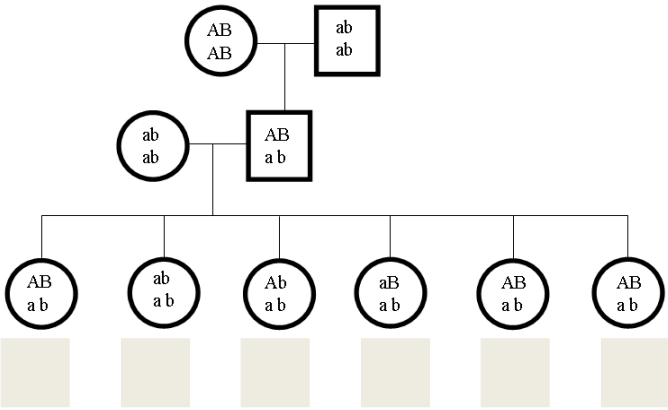
P(red, disc-shaped)=1/4 $(1-c)^2$

P(red, long)=(½ c)(½ c) + (½ c)(½ (1−c)) + (½ c)(½ (1−c))

et cetera

Exercise 1

two loci and a human pedigree



Which children are **recombinants**? Which are **non-recombinants**?
How could you estimate the recombination fraction for loci A and B?

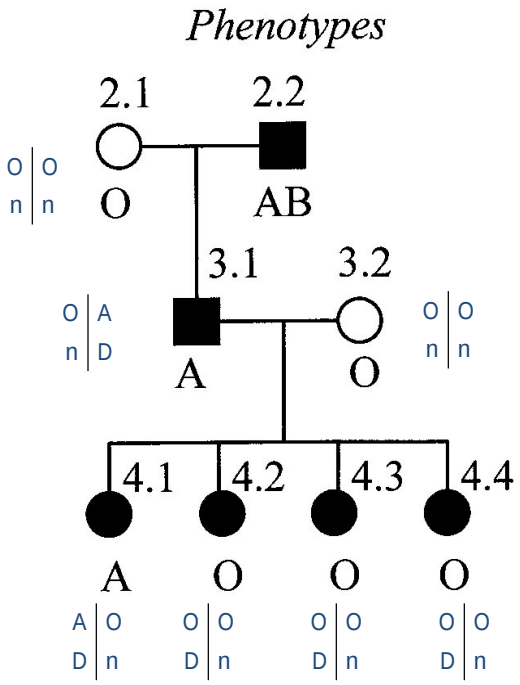
In the previous example, we have all the genotypes for both markers and know the phase for the genotypes. That is, we know how the alleles are coupled on each chromosome for the third generation. We could therefore tell without ambiguity which offspring have a paternal chromosome that is a recombinant.

Example

simple linkage analysis: phase unknown

Charcot-Marie-Tooth neuropathy (CMT) is a disease with autosomal dominant inheritance. Consider the following pedigree (Ott J (1999) Analysis of human genetic linkage, 3rd edition. Johns Hopkins University Press). Shading represents an individual with CMT and the ABO blood types are given below each individual. As we know, ABO blood types are determined by a single gene with three alleles. We might be interested in estimating the recombination fraction between the ABO blood type locus and the gene for CMT.

The first thing that we need to do is determine all genotypes with phase.



In the example we were able to directly count the number of recombinants in the offspring generation. This is an idealized situation. Note that the grandparent generation was necessary for establishing phase, i.e. a nuclear family would not have been sufficient to determine recombinants directly.

Note that we assumed a very specific mode of inheritance: dominant single allele, fully penetrant. We also used a single recombination fraction for males and females.

Suppose we had a set of CMT/ABO families for which we counted N non-recombinants and R recombinants. Then we would have a likelihood of:

$$L(c) = \Pr(data \mid c) = \binom{N+R}{R}(1-c)^N c^R$$

Under the null hypothesis of no linkage, i.e. H_0: c=0.5:

$$L(c=\frac{1}{2}) = \mathrm{P}(data \mid c=\frac{1}{2}) = \binom{N+R}{R}\frac{1}{2}^{(N+R)}$$

LOD Scores

In linkage analysis it is conventional to use LOD scores.
We get a LOD score by taking $\log_{10}$ of the likelihood ratio

$$LOD = \log_{10}\left(\frac{\Pr(data \mid alternative)}{\Pr(data \mid null)}\right)$$

The LOD scores for this example is:

$$LOD(c) = N\log(1-c) + R\log c - (N+R)\log\left(\frac{1}{2}\right)$$

Linkage Analysis

In linkage analysis on human pedigrees, we find evidence for the location of a gene for a trait of interest if we find a gene (or marker) of known location with which it is tightly linked. Linkage analysis on human pedigrees is based on the kind of analysis we just did. However, real-life linkage analysis is hugely more complicated because much more is unknown.

- missing genotypes through the pedigree
- phase may not be able to be determined without ambiguity
- decreased penentrance: P(disease | disease genotype) < 1
- Misspecified relationships, e.g. false paternity
- Inbreeding, creating “loops” in the pedigree

Allelic Association

"The excessive co-occurrence of certain combinations of alleles in the same gamete because of tight linkage, or for other reasons, is known as allelic association"

-Pak Sham, Statistics in Human Genetics 1998, p. 145

Consider two loci A and B, with alleles $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ occurring at frequencies $p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population. We can consider an individual's two **haplotypes** with respect to these loci, one of maternal origin and one of paternal origin.

How many possible haplotypes are there?

The haplotypes can be denoted

$A_1B_1, A_1B_2, \ldots, A_mB_n$ with frequencies $h_{11}, h_{12}, \ldots, h_{mn}$.

In a population, if the occurrence of allele A_i and the occurrence of allele B_j in a haplotype are independent events, then, by definition $h_{ij} = p_i q_j$.

Example of LD: Lactase gene

- Samples from three populations
  - CEU – Eurpean ancestry, living in Utah
  - YRI – Yoruban origin from Nigeria
  - CHB/JPT – Chinese and Japanese, living in Beijing and Tokyo
- LCT (Lactase gene): a particular mutation appears to allow the ability to digest milk into adulthood. The high frequency of the mutation in thought to be the result of positive selection in Europeans; the mutation is much less common in non-pastorialist populations.

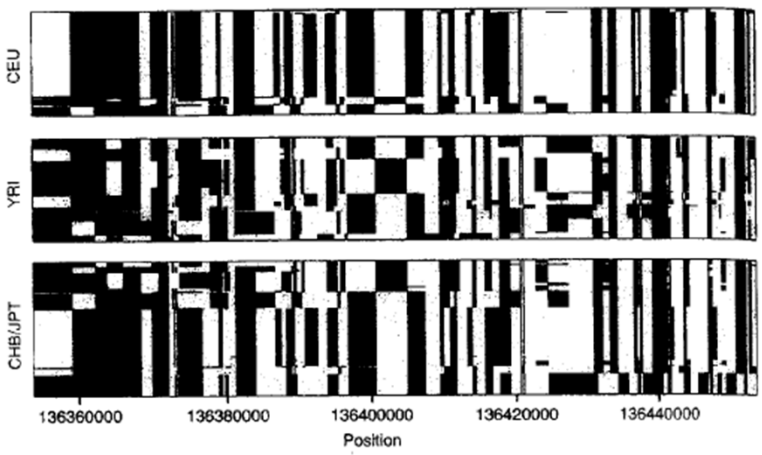


Figure 27.2 Haplotype structure in a 100-kb region surrounding the *Lactase* gene on human chromosome 2 for the four HapMap populations (The International HapMap Consortium, 2005). CEU = Individuals of European origin from Utah; YRI = Yoruba from Nigeria; CHB/JPT = Han Chinese from Beijing and Japanese from the Tokyo region (60 unrelated individuals in each of CEU and YRI, 90 in CHB/JPT). The CEU panel is dominated by a single haplotype that extends over the entire region. Much higher haplotype diversity is found in the other populations (see also Table 27.1).

Terminology

"independent alleles"	$h_{ij} = p_i q_j$
"positively associated alleles"	$h_{ij} > p_i q_j$
"negatively associated alleles"	$h_{ij} < p_i q_j$

Exercise 2

allelic associations

Consider two biallelic loci A and B. There are four possible haplotypes: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Suppose that the frequencies of these four haplotypes in a large population are 0.4, 0.1, 0.2, and 0.3, respectively. Are there any allelic associations between these loci? What are they?

Exercise 2

Linkage Disequilibrium (LD)

"The term linkage disequilibrium (LD) is broadly used to refer to the non-random sharing (or lack thereof) of combinations of [genetic] variants. "

-G. McVean, "Linkage Disequilibrium, Recombination, and Selection" , in Handbook of Statistical Genetics, Eds. Balding, Bishop, Cannings)

Linkage Disequilibrium (LD)

Consider two loci A and B, with alleles A_1, A_2, ... , A_m and B_1, B_2, ... , B_n. If there is allelic association between any pair of alleles in the different loci, then the loci are said to be in **linkage disequilibrium**. Notice that linkage disequilibrium is a population-level characteristic.

We will discuss three measures of Linkage Disequilibrium, D, D', and R^2.

Linkage Disequilibrium Coefficient D

For ease of notation, we define D for two biallelic loci with alleles A and a at locus 1; B and b at locus 2:

$$D_{AB}=P(AB)-P(A)P(B)$$

Note that $D_{aB} = P(aB)-P(a)P(B)$

$= P(aB)-(1-P(A))P(B)$

$= P(aB)-P(B)+P(A)P(B)$

$= P(aB)-P(aB) -P(AB)+P(A)P(B)$

$= P(A)P(B) -P(AB)$

$= -D_{AB}$

Thus, the magnitude of the coefficient is important, not the sign.

Linkage Disequilibrium Coefficient D

LD is a property of two loci, not their alleles. Therefore it is good that the magnitude of D does not depend on the choice of alleles.

The range of values the linkage disequilibrium coefficient can take on varies with allele frequencies. Allele frequencies cannot be negative, and we know, for example, that P(AB) must be less than both P(A) and P(B). This gives us relations such as:

(1) $0 \leq p_{AB} = p_A p_B + D_{AB} \leq p_A, p_B$
(2) $0 \leq p_{aB} = p_a p_B - D_{AB} \leq p_a, p_B$
(3) $0 \leq p_{Ab} = p_A p_b - D_{AB} \leq p_A, p_b$
(4) $0 \leq p_{ab} = p_a p_b + D_{AB} \leq p_a, p_b$

For example, to derive expression (1):

From the definition: $p_{AB} = p_A p_B + D_{AB}$

Since p_{AB} is a probability, we have $0 \leq p_{AB}$

From the rules of probability: $p_{AB} \leq p_A, p_B$

So putting these together we have (1):
$0 \leq p_{AB} = p_A p_B + D_{AB} \leq p_A, p_B$

For (2)-(4), use relationships $D_{AB} = -D_{aB}$, etc.

Take the first inequality from (1) through (4):

(1) $0 \leq p_A p_B + D_{AB}$
(2) $0 \leq p_a p_B - D_{AB}$
(3) $0 \leq p_A p_b - D_{AB}$
(4) $0 \leq p_a p_b + D_{AB}$

(1) $-p_A p_B \leq D_{AB}$
(2) $D_{AB} \leq p_a p_B$
(3) $D_{AB} \leq p_A p_b$
(4) $-p_a p_b \leq D_{AB}$

Linkage Disequilibrium Coefficient D

The inequalities on the previous slide lead to bounds for D_{AB} :

$$-p_A p_B, -p_a p_b \leq D_{AB} \leq p_a p_B, p_A p_b$$

Normalized LD Coefficient D'

We have just seen that the possible values of D depend on allele frequencies. This makes D difficult to interpret. For reporting purposes, the **normalized linkage disequilibrium coefficient D'** is often used.

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\max(-p_A p_B, -p_a p_b)} & \text{if } D_{AB} < 0 \\ \\ \dfrac{D_{AB}}{\min(p_a p_B, p_A p_b)} & \text{if } D_{AB} > 0 \end{cases}$$

Exercise 3

Consider two biallelic loci, A,a and B,b. What is the theoretical range of the linkage disequilibrium coefficient D_{AB} and its absolute value $|D_{AB}|$ under the follow scenarios?

(a)P(A)=1/2, P(B)=1/2

Part 3

(b) P(A)=.95, P(B)=.95

(c) P(A)=.95, P(B)=.05

(d) P(A)=1/2, P(B)=.95

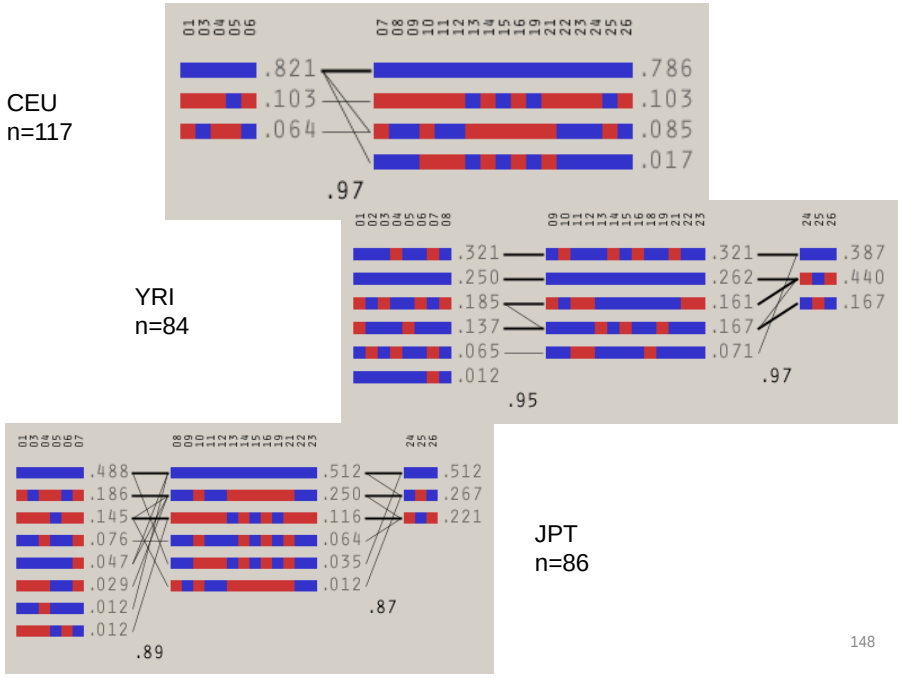
Exercise 4

Under what circumstances might D_{AB} reach its theoretical maximum value? Suppose D_{AB} =P(a)P(B). What does this imply? Why does this make sense?

Example: LCT haplotype data and multiallelic D'

- The haplotype display on the next slide shows observed haplotypes for SNPs in the LCT gene for samples from three populations. The haplotype data are summarized as "haplotype blocks" and the display gives sample frequencies and connections from one block to the next. In the crossing areas, a value of multiallelic D' is shown. This represents the level of recombination between the two blocks. The value of multiallelic D' is computed for only the haplotypes ("alleles") currently displayed.

SNP haplotypes in the LCT gene – HapMap data – Chr 2 136260→136310



Estimating D

Suppose we have a situation where we can consider single chromosomes to have been sampled from a population of interest. The data might be arranged in a table such as:

		First Locus		
		B	b	
Second Locus	A	n_{AB}	n_{Ab}	n_A
	a	n_{aB}	n_{ab}	n_a
		n_B	n_b	N

We would like to estimate D_{AB} from the data. The maximum likelihood estimate of D_{AB} is

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

Estimating D

...where the population frequencies are just estimated by the sample frequencies. The MLE turns out to be slightly biased. If N gametes have been sampled, then

$$E(\hat{D}_{AB}) = \frac{N-1}{N} D_{AB}$$

The variance of this estimate depends on both the true allele frequencies and the true level of linkage disequilibrium:

$$Var(\hat{D}_{AB}) = \frac{1}{N}[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]$$

Testing for LD with D

Since D_{AB}=0 corresponds to the status of no linkage disequilibrium, it is sometimes of interest to test the null hypothesis H_0: D_{AB}=0. One way to do this is to use a chi-square statistic. It is constructed by squaring the z-statistic:

$$z = \frac{\hat{D}_{AB} - E(\hat{D}_{AB})}{\sqrt{Var(\hat{D}_{AB})}}$$

after setting D_{AB}=0 in both the expectation and the variance. The test statistic is compared against a Chi-Square distribution with one degree of freedom.

R^2

The correlation between any two quantitative variables Y And Z is $r_{YZ} = \frac{cov(Y,Z)}{\sqrt{var(Y)var(Z)}}$.

Define a random variable X_A to be 1 if the allele at the first locus is A and 0 if the allele is a.

Define a random variable X_B to be 1 if the allele at the second locus is B and 0 if the allele is b.

Then $r_{AB} = \frac{cov(X_A,X_B)}{\sqrt{var(X_A)var(X_B)}}$.

cov: $E[X_A X_B]-E[X_A]E[X_B]=P(X_A=X_B=1)-P(X_A=1)\ P(X_B=1)=D_{AB}$.

var: $var(X_A)=E[X_A^2]-E^2[X_A]= P(X_A=1) - P(X_A=1)^2=f_A f_a$.

R^2

.

So $r_{AB} = \frac{cov(X_A,X_B)}{\sqrt{var(X_A)var(X_B)}} = \frac{D_{AB}}{\sqrt{f_A f_a f_B f_b}}$.

It is usually more common to consider the square:

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}$$

R^2

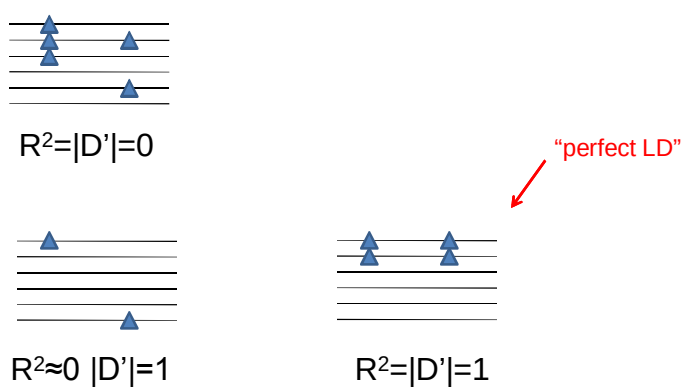
- Has the same value however the alleles are labeled
- Relationship between R^2 and the power of association studies
- Tests for LD: A natural test statistic to consider is the contingency table test. Compute a test statistic using the Observed haplotype fequencies and the Expected frequency if there were no LD:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Under the null hypothesis of No LD, the statistic is asymptotically χ^2 distributed with 1 df. It turns out that

$$X^2 = n\hat{r}^2.$$

Is LD high or low?



When there are rare alleles it becomes more important to choose our metric carefully.

“Perfect LD”

haplotype	frequency
AB	98%
Ab	0%
aB	0%
ab	2%

D=P(AB)-P(A)P(B)=.98-(.98)(.98)= 0.0196
Maximum value for these allele frequencies:
min(.98 · .02, .02 · .98)= 0.0196
D’=1

$R^2 = (0.0196)^2/[(.98)(.98)(.02)(.02)]=1$

(This is sometimes called “perfect LD”: two alleles at two loci have the same allele frequency and don’t ever “separate”.)

when D’ and R^2 “disagree”

haplotype	frequency
AB	90%
Ab	5%
aB	5%
ab	0%

D=P(AB)-P(A)P(B)=.9-(.95)(.95)= -0.0025
Maximum value for these allele frequencies:
max(-.95 · .95, -0.05 · .05)=max(-.9025, -.0025)=-.0025
D’=1

(When D’=1 then at least one haplotype doesn’t exist in the data.)

$R^2 = (-0.0025)^2/[(.95)(.95)(.05)(.05)]=0.0028$

How are allelic associations generated?

- Genetic drift: In a finite population, the gene pool of one generation can be regarded as a random sample of the gene pool of the previous generation. As such, allele and haplotypes frequencies are subject to sampling variation – random chance. The smaller the population is, the larger the effects of genetic drift are.

- Mutation: If a new mutation appears in a population, alleles at loci linked with the mutant allele will maintain an association for many generations. The association lasts longer when linkage is greater (that is, the recombination fraction is much smaller than ½ – very close to 0).

How are allelic associations generated?

•Founder effects: Applies to a population that has grown rapidly from a small group of ancestors. For example, the 5,000,000 Finns mostly descended from about 1000 people who lived about 2000 years ago. Such a population is prone to allelic disequilibrium.

•Selection: When an individual's genotype influences his/her reproductive fitness. For example, if two alleles interact to decrease reproductive fitness, the alleles will tend to be negatively associated.

•Stratification: Some populations consist of two or more subgroups that, for cultural or other reasons, have evolved more or less separately. Two loci that are in linkage equilibrium for each subpopulation may be in linkage disequilibrium for the larger population.

Exercise 5

population stratification

Consider a population with three subpopulations.
Consider two biallelic loci, the first with alleles A and a; the second with alleles B and b.

N	A allele frequency	B allele frequency	AB haplotype frequency
1000	0.3	0.5	0.15
2000	0.2	0.4	0.08
10000	0.05	0.1	0.005

Exercise 5

Do any of three subpopulations show allelic association?

Exercise 5

Does the larger population show allelic association?

How are allelic associations maintained?

- selection
- non-random mating (e.g., pop stratification)
- Linkage:

Consider again two loci A and B, with alleles

$A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ with frequencies

$p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population. The haplotypes are $A_1B_1, A_1B_2, \ldots, A_mB_n$, with frequencies

$$h_{11}^0, h_{12}^0, \ldots, h_{mn}^0,$$

in generation 0. What is the frequency of haplotype A_iB_j, in the next generation? In the following calculation we assume random mating in the population.

$$
\begin{aligned}
h_{ij}^1 &= P(haplotype^1 = A_i B_j) \\
&= P(haplotype^1 = A_i B_j \mid \text{no recombination}) P(\text{no recombination}) \\
&+ P(haplotype^1 = A_i B_j \mid \text{recombination}) P(\text{recombination}) \\
&= P(haplotype^1 = A_i B_j \mid \text{no recombination})(1-c) \\
&+ P(haplotype^1 = A_i B_j \mid \text{recombination}) c \\
&= h_{ij}^0 (1-c) + p_i q_j c
\end{aligned}
$$

From this we can deduce that the difference in haplotype frequency between the generations is:

$$h_{ij}^1 - h_{ij}^0 = c(p_i q_j - h_{ij}^0)$$

When will this difference be 0? That is, when are haplotype frequencies stable?

When c=0 or no LD.

We can also use this expression to characterize the difference between the true haplotype frequency and what the haplotype frequency would be under equilibrium:

$$h_{ij}^1 - p_i q_j = (1-c)(h_{ij}^0 - p_i q_j)$$

Extending this to the kth generation, we get:

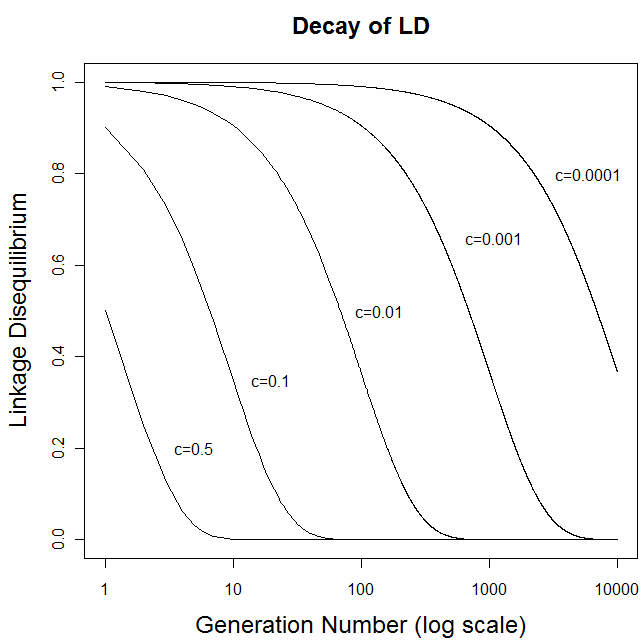
$$h_{ij}^k - p_i q_j = (1-c)^k (h_{ij}^0 - p_i q_j)$$

Alternatively we can write:

$$D_{ij}^{1} = (1-c)D_{ij}^{0}$$

$$D_{ij}^{k} = (1-c)^{k} D_{ij}^{0}$$

The figure on the next slide shows the decline of linkage disequilibrium in a large, randomly-mating population for several different values of c.



Case-Control Association Testing

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 4

Introduction

- Association mapping is now routinely being used to identify loci that are involved with complex traits.
- Technological advances have made it feasible to perform case-control association studies on a genome-wide basis with hundreds of thousands of markers in a single study.
- We consider testing a genetic marker for association with a disease in a sample of unrelated subjects.
- Case-control association methods essentially test for independence between trait and allele/genotype.

Case-Control Association Testing

- Allelic Association Tests
  - Allele is treated as the sampling unit
  - Typically make an assumption of Hardy-Weinberg equilibrium (HWE). Alleles within an individual are conditionally independent, given the trait value.
- Genotypic Association Tests
  - Individual is the sampling unit
  - Does not assume HWE

Pearson's χ^2 Test for Allelic Association

- The classical Pearson's χ^2 test is often used for allelic association testing.
- This test looks for deviations from independence between the trait and allele.
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled "1" and "2"
- Let N_{ca} be the number of cases and N_{co} be the number of controls with genotype data at the marker.

Pearson's χ^2 Test for Allelic Association

- Below is a 2×2 contingency table for trait and allelic type

	Cases	Controls	Total
Allele 1	n_1^{ca}	n_1^{co}	n_1
Allele 2	n_2^{ca}	n_2^{co}	n_2
Total	$2N_{ca}$	$2N_{co}$	T

- n_1^{ca} is the number of type 1 alleles in the cases and $n_1^{ca} = 2 \times$ the number of homozygous $(1,1)$ cases $+$ the number of heterozygous (1,2) cases
- n_2^{co} is the number of type 2 alleles in the controls and $n_2^{co} = 2 \times$ the number of homozygous $(2,2)$ controls $+$ the number of heterozygous (1,2) controls
- Hypotheses
  - H_0: there is *no association* between the row variable and column variable
  - H_a: there *is* an association between the two variables

Pearson's χ^2 Test for Allelic Association

- Can use Pearson's χ^2 test for independence. The statistic is:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

- What is the the expected cell number under H_0? For each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1) = (2-1)(2-1) = 1$ degree of freedom

LHON Example: Pearson's χ^2 Test

- From Phasukijwattana et al. (2010), Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- Corresponding 2×2 contingency table for allelic type and case-control status

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Intuition for the test: Suppose H_0 is true, allelic type and case-control status are independent, then what counts would we expect to observe?

LHON Example: Pearson's χ^2 Test

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Let n be the total number of alleles in the study. Assuming independence, the expected number of case alleles that are of type T is:

$$n \times P(\text{Allele is from a Case and Allelic type is T})$$

$$= nP(\text{Allele is from a Case})P(\text{Allelic type is T})$$

$$= 656\left(\frac{178}{656}\right)\left(\frac{550}{656}\right) = \frac{(178)(550)}{656} = 149.2378$$

LHON Example: Pearson's χ^2 Test

- Fill in the remaining cells for the expected counts

	Cases	Controls	Total
Allele T	149.2378		
Allele C			
Total			

- Calculate the X^2 statistic

$$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + \cdots + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- What is the p-value?

$$P(\chi_1^2 \geq 4.369) = .037$$

LHON Example: Pearson's χ^2 Test

- Fill in the remaining cells for the expected counts

	Cases	Controls	Total
Allele T	149.2378	400.7622	550
Allele C	28.7622	77.2378	106
Total	178	478	656

- Calculate the X^2 statistic

$$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + \cdots + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- What is the p-value?

$$P(\chi_1^2 \geq 4.369) = .037$$

Fisher's Exact Test for Allelic Association

- For contingency tables that have cells with less than 5 observations
- Consider the table below

	Cases	Controls	Total
Allele T	21	14	35
Allele C	3	10	13
Total	24	24	48

- The marginal counts of the table are fixed: There are 24 case alleles, 24 control alleles, 35 T alleles, and 13 C alleles
- Let X be the number of case alleles that are of type T. A test based on X can be constructed.
- Under the null hypothesis, X will have a hypergeometric distribution where the probability that $X = x$ is

$$\binom{35}{x}\binom{13}{24-x}\Big/\binom{48}{24}$$

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	11	12	13	14	15	16	17	18	19	20	21	22	23	24
P(X=x)														

- $P(X = 11)$ is

$$\binom{35}{11}\binom{13}{13} \Big/ \binom{48}{24} = .00001$$

- $P(X = 12)$ is

$$\binom{35}{12}\binom{13}{12} \Big/ \binom{48}{24} = .0003$$

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	P(X=x)
11	.00001
12	.0003
13	.004
14	.021
15	.072
16	.162
17	.241
18	.241
19	.162
20	.072
21	.021
22	.004
23	.0003
24	.00001

- Construct a rejection region for a two-sided test with $\alpha = .05$.
- Can the null hypothesis be rejected at the .05 level for the observed value $X = 21$?

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	P(X=x)
11	.00001
12	.0003
13	.004
14	.021
15	.072
16	.162
17	.241
18	.241
19	.162
20	.072
21	.021
22	.004
23	.0003
24	.00001

- So, a rejection region for a two-sided test with $\alpha = .05$ would consist of the following values for X: 11, 12, 13, 14, 21, 22, 23, and 24.
- The observed X value of 21 for the data falls in this region, so the test would reject at the level .05.

The Armitage Trend Test for Genotypic Association

- The most common genotypic test for unrelated individuals is the Armitage trend test (Sasieni 1997)
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled "1" and "2"
- Let $Y_i = 2$ if individual i is homozygous (1,1), 1 if the i is heterozygous, and 0 if i is homozygous (2,2)
- Let $X_i = 1$ if i is a case and 0 if i is a control.
- A simple linear regression model of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

The Armitage Trend for Genotypic Association

- To test this hypothesis, the Armitage trend test statistic is

$$A_r = \frac{\hat{\beta}_1^2}{VAR(\hat{\beta}_1)} = Nr_{xy}^2$$

  where r_{xy}^2 is the squared correlation between genotype variable Y and phenotype variable X.
- Note that the variance estimate for Y that is used in the calculation of the Armitage trend test is the sum of the squared deviations of Y from the fitted values of Y for regression with only an intercept term.
- Under the null hypothesis, A_r will follow an approximate χ^2 distribution with 1 degree of freedom.
- The Armitage trend test can be shown to be valid when HWE does not hold.

LHON Example: Armitage Trend Test

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- The Armitage test statistic for this data is

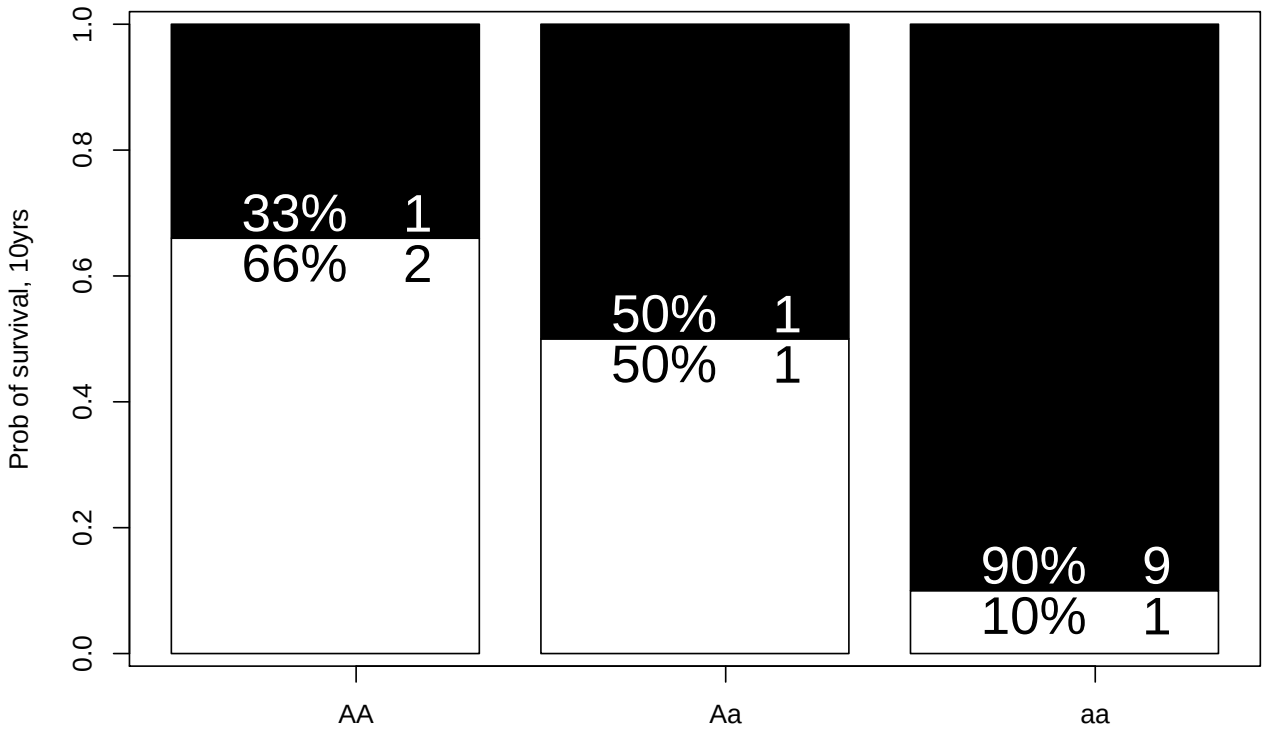
$$A_r = Nr_{xy}^2 = 328(.0114) = 3.74$$

- The p-value is

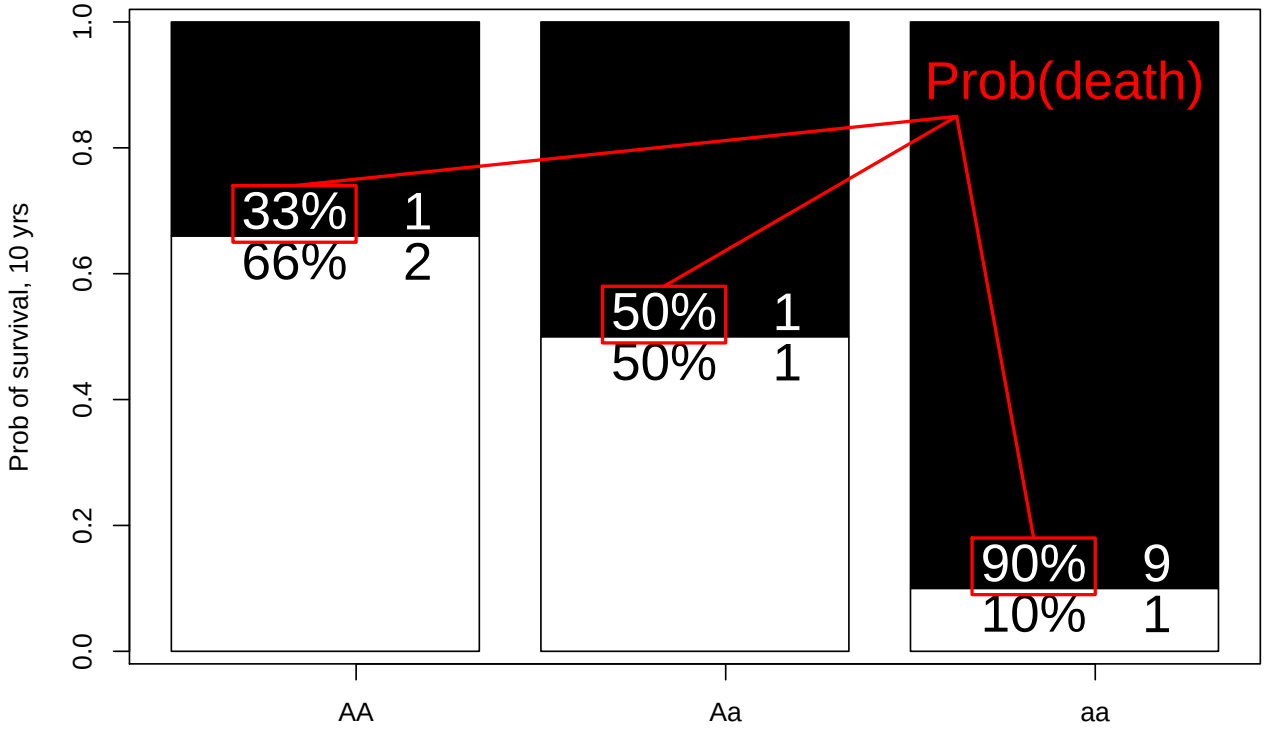
$$P(\chi_1^2 \geq 3.743) = .053$$

Odds Ratios (ORs) for Genotypes

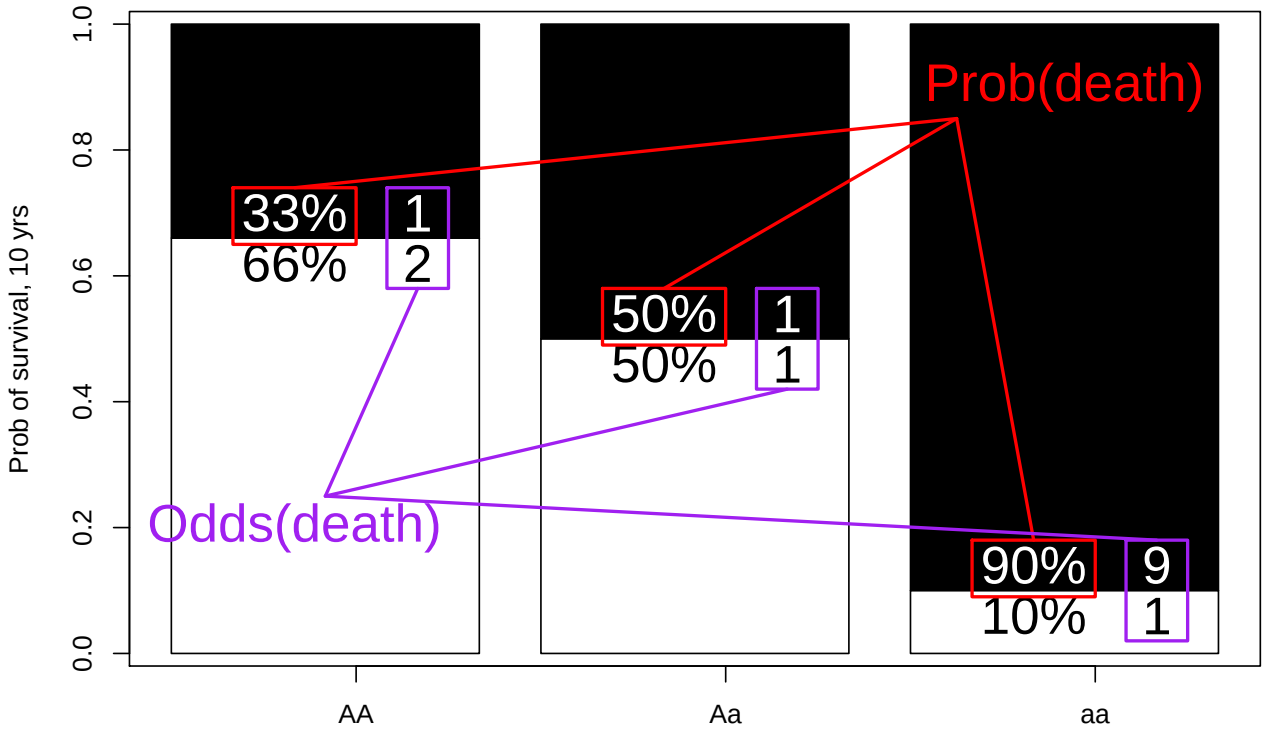
- What are **odds**? Really just **probability**...
- Odds are a [gambling-friendly] measure of chance;



Odds Ratios (ORs) for Genotypes



Odds Ratios (ORs) for Genotypes



– so what are **odds ratios**?

Odds Ratios (ORs) for Genotypes

	Cases	Controls
TT	A	B
CT	A'	B'
CC	C	D

- Typically choose a reference genotype.

$$OR_{TT} = \frac{\text{odds of disease in an individual with the TT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

$$OR_{CT} = \frac{\text{odds of disease in an individual with the CT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

- $OR_{TT} = 1$ implies no association with and disease. Similarly for OR_{CT}.
- $OR_{TT} > 1$ or $OR_{TT} < 1$ implies association with the disease.

Odds Ratios (ORs) for Genotypes

- Logistic regression is generally used to get odds ratios and confidence intervals for genotypes.
- Let π_i be the probability that individual i is affected with the disease and let G_i be the genotype for individual i at the SNP:

$$\log(\text{odds of disease for individual } i | G_i)$$

$$= log\left(\frac{\pi_i}{1-\pi_i}\middle| G_i\right)$$

$$= \beta_0 + \beta_{CT} I\{G_i = CT\} + \beta_{TT} I\{G_i = TT\}$$

where $I\{G_i = CT\}$ is 1 if $G_i = CT$ and 0 otherwise, and similarly for $I\{G_i = TT\}$.

Odds Ratios (ORs) for Genotypes

- The coefficient estimates for $\hat{\beta}_{CT}$ and $\hat{\beta}_{TT}$ can be used to calculate odds ratios:

$$OR_{CT} = exp(\hat{\beta}_{CT})$$

$$OR_{TT} = exp(\hat{\beta}_{TT})$$

- 95% CI for OR_{CT} is

$$exp(\hat{\beta}_{CT} \pm 1.96 \times s.e.(\hat{\beta}_{CT}))$$

Odds Ratios for LHON Example

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- We will use the R software package to obtain odds ratios and confidence intervals for this data set (as well as Pearson's χ^2 test and Armitage Trend tests).
- Exercises and some commands for analyzing the LHON data with R can be found on the following webpage:

http://faculty.washington.edu/tathornt/SISG_MODULE8.html

Odds Ratios (ORs) based on Allele Counting

- We can also obtain allelic odds ratios
- Odds ratios based on an allele counting model essentially assumes an additive model
- Genotype TT has twice the risk (or protection) of heterozygous genotype CT.
- Same risk (or protection) for the comparison of heterozygous CT genotype and homozygous CC genotype.

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

$$OR_T = \frac{\text{odds of disease with T allele}}{\text{odds of disease with C allele}}$$
$$= \frac{(n_A/n_B)}{(n_C/n_D)} = \frac{n_A \times n_D}{n_B \times n_C}$$

Odds Ratios (ORs) Allele Counting

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

- $OR_T = 1$ implies no association with and disease
- $OR_T > 1$ or $OR_T < 1$ implies association with the disease

Confidence Intervals for Odds Ratios (ORs)

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

$$OR = \frac{n_A \times n_D}{n_B \times n_C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_C} + \frac{1}{n_D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{n_A \times n_D}{n_B \times n_C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_C} + \frac{1}{n_D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

LHON Example: Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{158 \times 86}{392 \times 20} = 1.7332$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{158} + \frac{1}{392} + \frac{1}{20} + \frac{1}{86}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

$$= exp(log(1.7332) - 1.96 \times 0.2665) = 1.03$$

- Upper limit of 95% CI $= 2.92$

References

- Phasukijwattana N, Kunhapan B, Stankovich J, Chuenkongkaew WL, Thomson R, Thornton T, Bahlo M, Mushiroda T, Nakamura Y, Mahasirimongkol S, et al. (2010). Genome-wide linkage scan and association study of PARL to the expression of LHON families in Thailand. *Hum. Genet.* **128**, 39-49.
- Sasieni P (1997). From genotypes to genes: doubling the sample size. *Biometrics* **5**, 1254-1261.

Association Testing with Quantitative Traits: Common and Rare Variants

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 5

Introduction to Quantitative Trait Mapping

- In the previous session, we gave an overview of association testing methods when the trait of interest is binary (e.g. 1/0, affected/unaffected, dead/alive),
- Phenotypes of interest are often quantitative, and in this session we focus on the topic of genetic association testing with quantitative traits.
- The field of **quantitative genetics** is the study of the inheritance of continuously measured traits and their mechanisms.
- Vast amounts of literature on this topic!

Introduction to Quantitative Trait Mapping

- Quantitative trait loci (QTL) mapping involves identifying genetic loci that influence the phenotypic variation of a quantitative trait.
- QTL mapping is commonly conducted with GWAS using common variants, such as variants with minor allele frequencies $\geq 1\% - 5\%$
- There generally is no simple Mendelian basis for variation of quantitative traits
- Some quantitative traits can be largely influenced by a single gene as well as by environmental factors

Introduction to Quantitative Trait Mapping

- Influences on a quantitative trait can also be due to a number of genes with similar (or differing) effects
- Many quantitative traits of interest are complex where phenotypic variation is due to a combination of both multiple genes and environmental factors
- Examples: Blood pressure, cholesterol levels, IQ, height, weight, etc.

Quantitative Genetic Model

- The classical quantitative genetics model introduced by Ronald Fisher (1918) is $Y = G + E$, where Y is the phenotypic value, G is the genetic value, and E is the environmental deviation.
- G is the combination of all genetic loci that influence the phenotypic value and E consists of all non-genetic factors that influence the phenotype
- The mean environmental deviation E is generally taken to be 0 so that the mean genotypic value is equal to the mean phenotypic value, i.e., $E(Y) = E(G)$

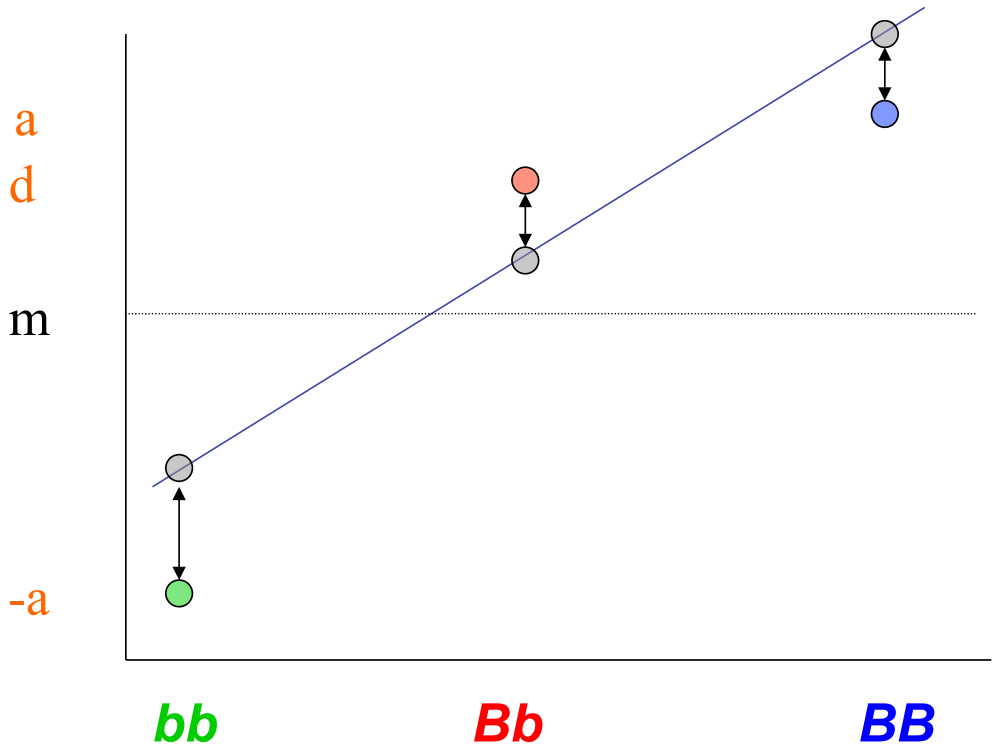
Quantitative Genetic Model

- Consider a single locus. Fisher modeled the genotypic value G with a linear regression model (least squares) where the genotypic value can be partitioned into an additive component (A) and deviations from additivity as a result of dominance (D), where

$$G = A + D$$

Linear Regression Model for Genetic Values

Falconer model for single biallelic QTL



Var (X) = Regression Variance + Residual Variance
= Additive Variance + Dominance Variance

Components of Genetic Variance

- From the properties of least squares, the residuals are orthogonal to the fitted values, and thus $Cov(A, D) = 0$. So we have that

$$Var(G) = Var(A) + Var(D)$$

or

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

- σ_A^2 is the **additive genetic variance**. It is the genetic variance associated with the average additive effects of alleles
- σ_D^2 is the **dominance genetic variance**. It is the genetic variance associated with the dominance effects.

Heritability

- The heritability of a trait is written in terms of the components of variances of the trait.
- Remember that $Y = G + E = A + D + E$
- The following ratio of variance components

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

  is defined to be the **narrow-sense heritability** (or simply heritability)
- h^2 is the proportion of the total phenotypic variance that is due to additive effects.
- Heritability can also be viewed as the extent to which phenotypes are determined by the alleles transmitted from the parents.

Heritability

- The **broad-sense heritability** is defined to be

$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2}$$

- H^2 is the proportion of the total phenotypic variance that is due to all genetic effects (additive and dominance)
- There are a number of methods for heritability estimation of a trait.
- Module 12 (Mixed Models in Quantitative Genetics) and Module 17 (Human Complex Traits) cover the topic of heritability in more detail.

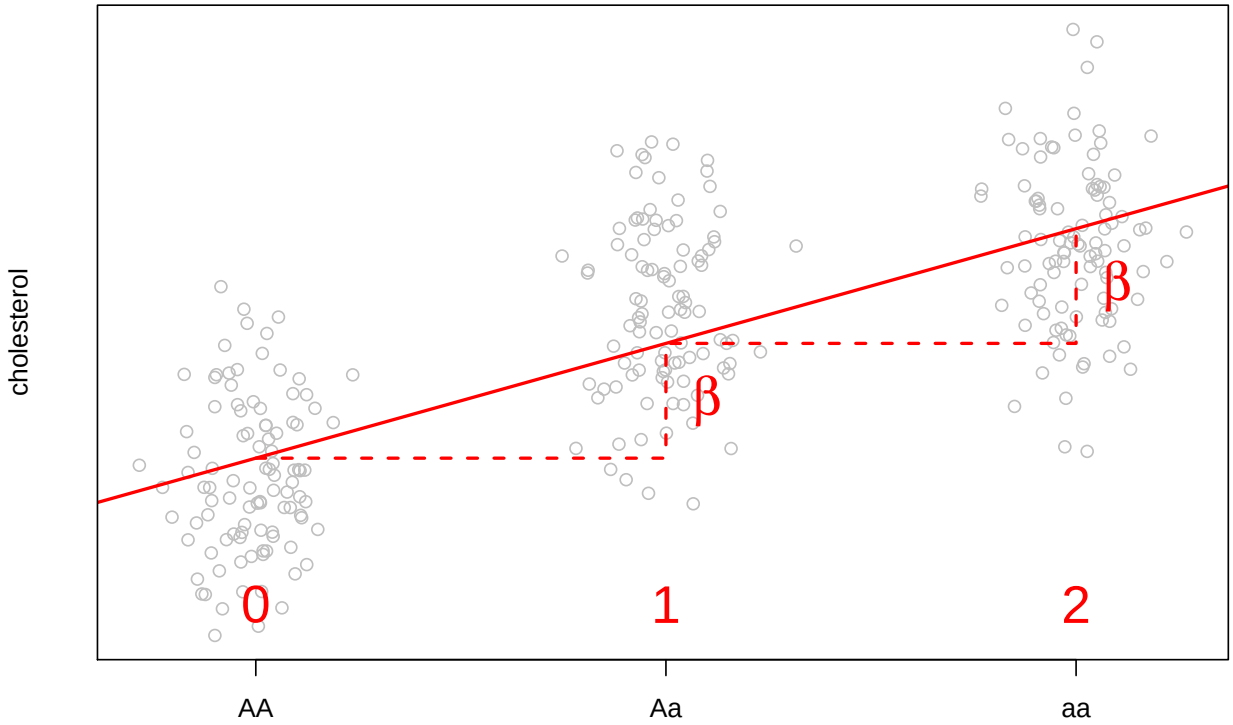
QTL Mapping

- For traits that are heritable, i.e., traits with a non-negligible genetic component that contributes to phenotypic variability, identifying (or mapping) QLT that influence the trait is often of interest.
- Linear regression models are commonly used for QTL mapping
- Linear regression models will often include a single genetic marker (e.g., a SNP) as predictor in the model, in addition to other relevant covariates (such as age, sex, etc.), with the quantitative phenotype as the response

Linear regression with SNPs

Many analyses fit the 'additive model'

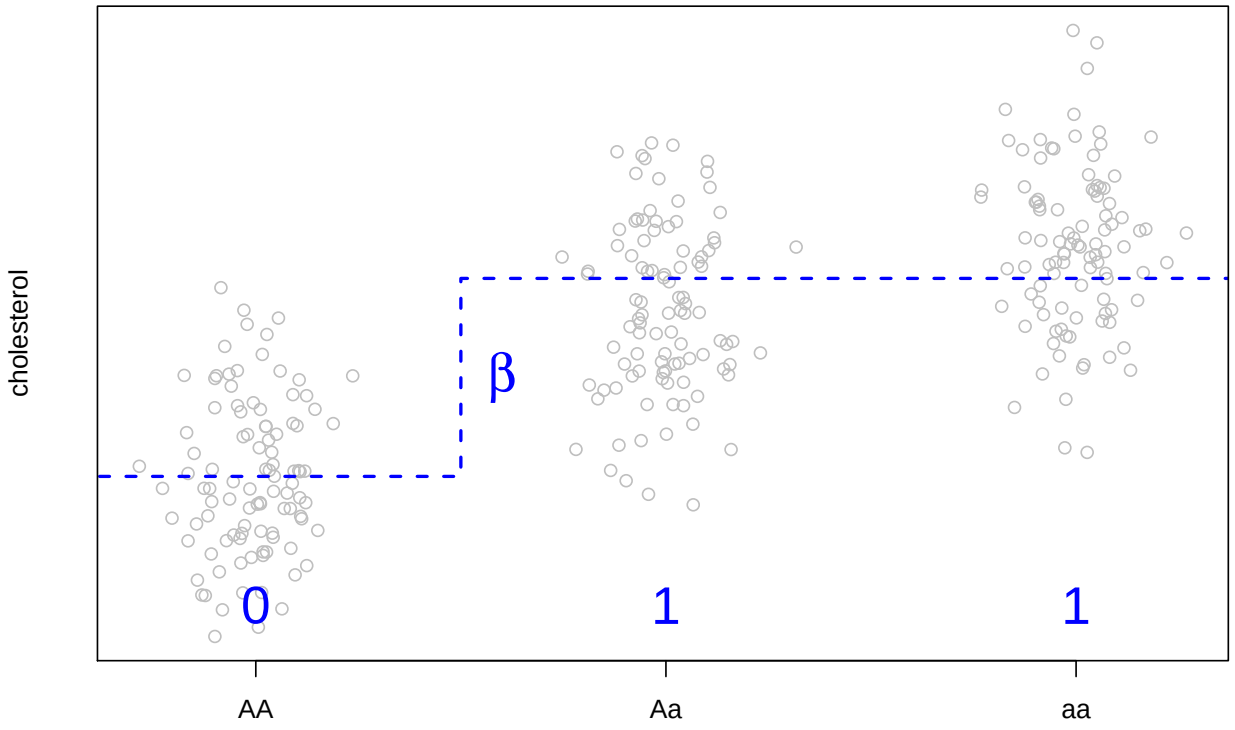
$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

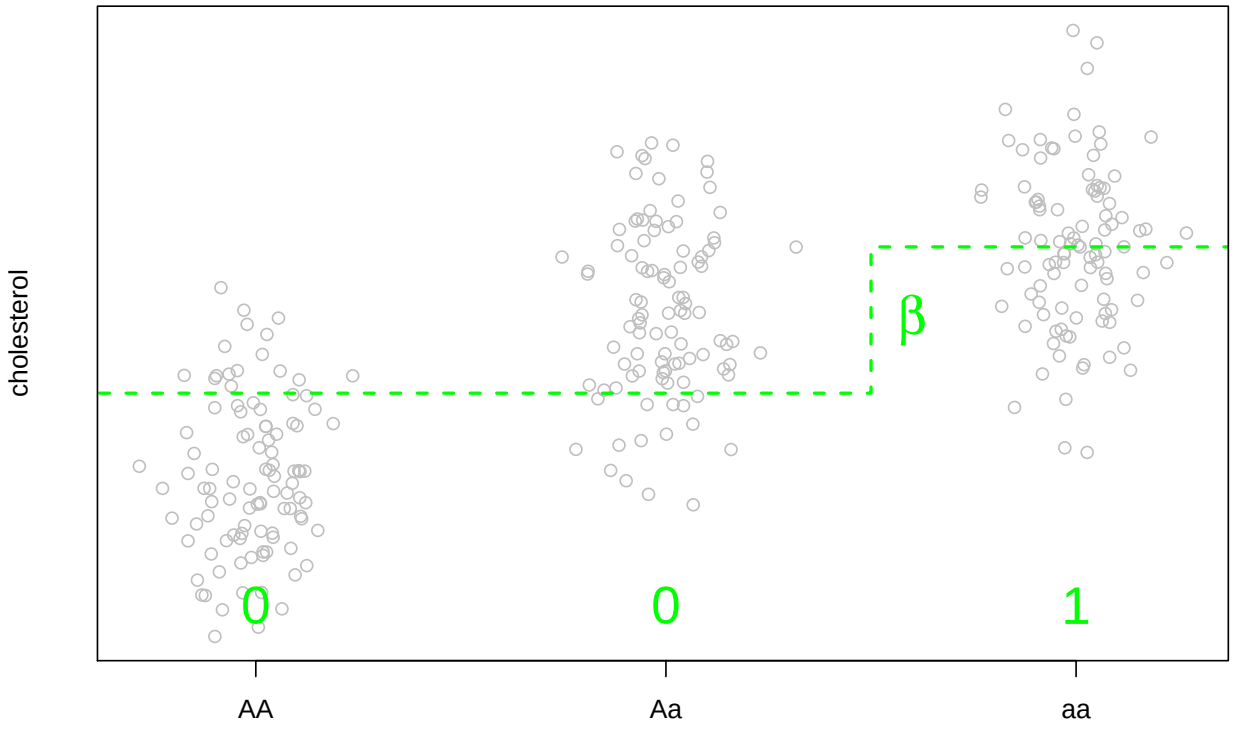
$$y = \beta_0 + \beta \times (G \neq AA)$$



Linear regression, with SNPs

or the 'recessive model';

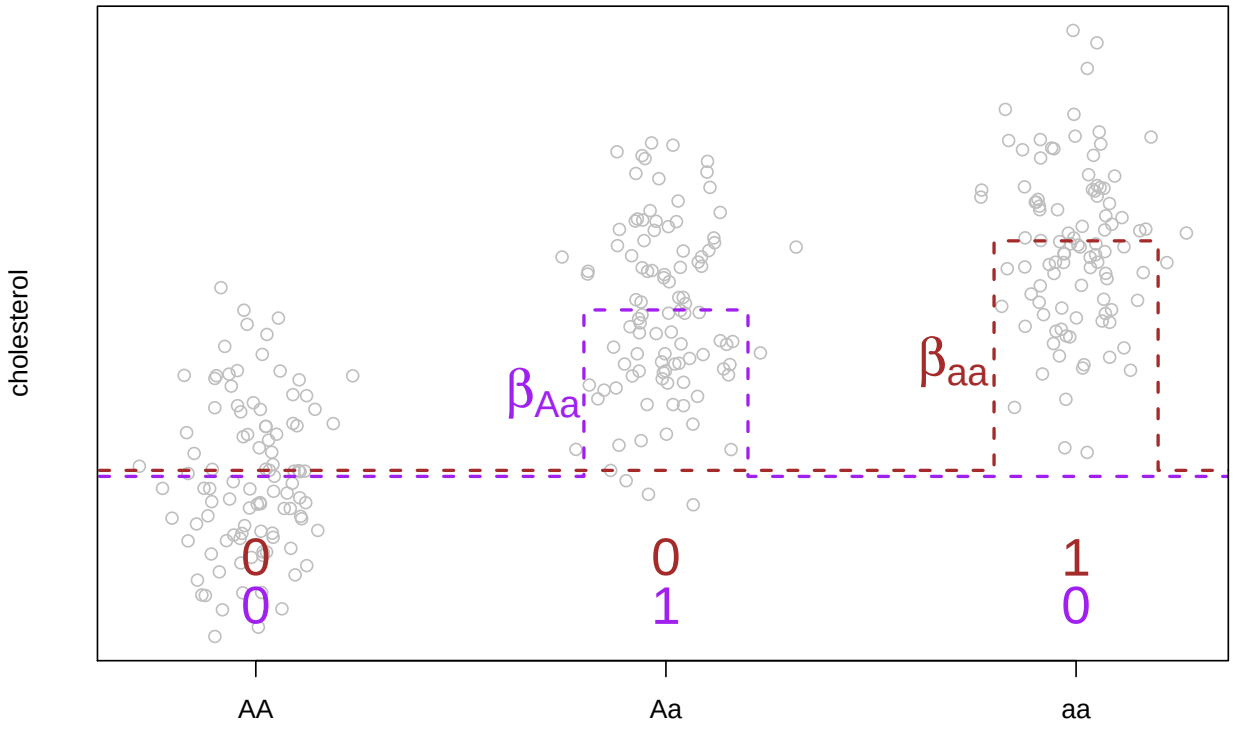
$$y = \beta_0 + \beta \times (G == AA)$$



Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$

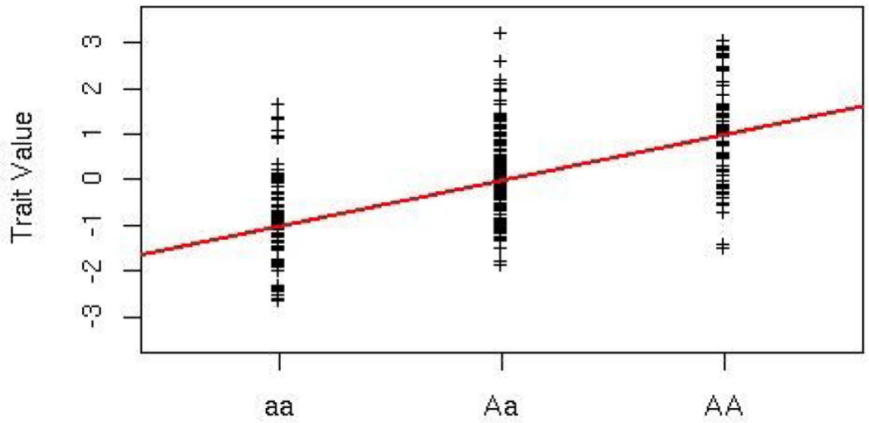


Additive Genetic Model

- Most GWAS perform single SNP association testing with linear regression assuming an additive model.

Unrelated Samples

$$\hat{y}_i = \mu + \hat{\beta} x_i$$



Additive Genetic Model

- The additive linear regression model also has a nice interpretation, as we saw from Fisher's classical quantitative trait model!
- The coefficient of determination (r^2) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the "SNP heritability"

Additive Genetic Model

- Consider the following additive model for association testing with a quantitative trait and a SNP with alleles A and a:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

  where X is the number of copies of the reference allele A.
- What would your interpretation of ϵ be for this particular model?

Association Testing with Additive Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Two test statistics for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim \mathbf{t}_{N-2} \approx N(0,1) \text{ for large } N$$

$$T^2 = \frac{\hat{\beta}_1^2}{var(\hat{\beta}_1)} \sim \mathbf{F}_{1,N-2} \approx \chi_1^2 \text{ for large } N$$

where

$$var(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{XX}}$$

and S_{XX} is the corrected sum of squares for the X_i's

Statistical Power for Detecting QTL

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- We can also calculate the power for detecting a QTL for a given effect size β_1 for a SNP.
- For simplicity, assume that Y has been a standardized so that with $\sigma_Y^2 = 1$.
- Let p be the frequency of the A allele in the population

$$\sigma_Y^2 = \beta_1^2 \sigma_X^2 + \sigma_\epsilon^2 = 2p(1-p)\beta_1^2 + \sigma_\epsilon^2$$

- Let $h_s^2 = 2p(1-p)\beta_1^2$, so we have $\sigma_Y^2 = h_s^2 + \sigma_\epsilon^2$
- Interpret h_s^2 (note that we assume that trait is standardized such that $\sigma_Y^2 = 1$)

Statistical Power for Detecting QTL

- Also note that $\sigma_\epsilon^2 = 1 - h_s^2$, so we can write $Var(\hat{\beta}_1)$ as the following:

$$var(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{XX}} \approx \frac{\sigma_\epsilon^2}{N\left(2p(1-p)\right)} = \frac{1-h_s^2}{2Np(1-p)}$$

- To calculate power of the test statistic T^2 for a given sample size N, we need to first obtain the expected value of the non-centrality parameter λ of the chi-squared (χ^2) distribution which is the expected value of the test statistic T squared:

$$\lambda = [E(T)]^2 \approx \frac{\beta_1^2}{var(\hat{\beta}_1)} = \frac{Nh_s^2}{1-h_s^2}$$

since $h_s^2 = 2p(1-p)\beta_1^2$

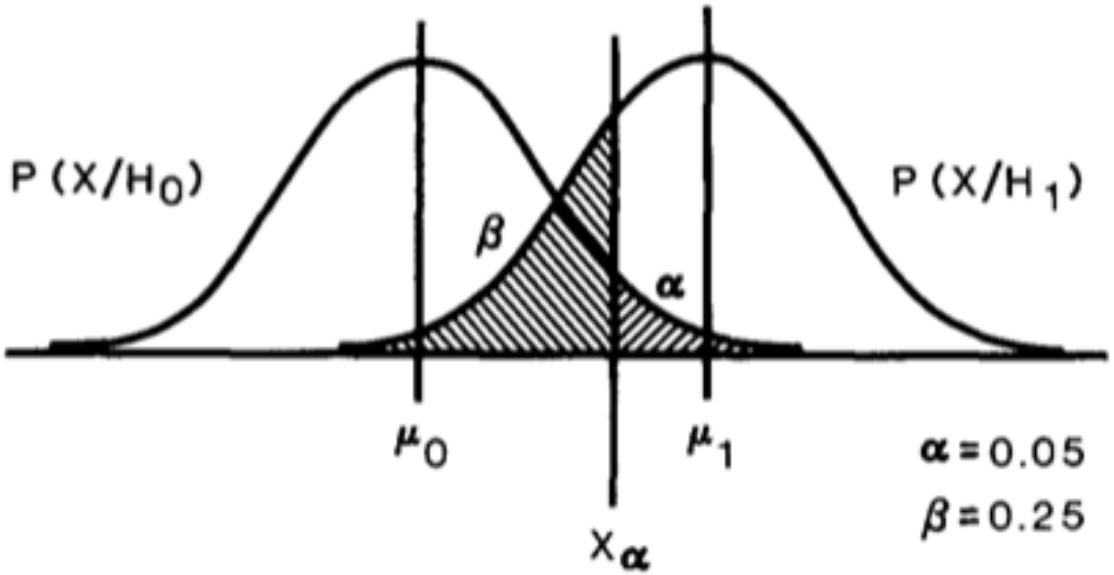
Required Sample Size for Power

- Can also obtain the required sample size given type-I error α and power $1-\beta$, where the type–II error is β :

$$N = \frac{1-h_s^2}{h_s^2}\left(z_{(1-\alpha/2)} + z_{(1-\beta)}\right)^2$$

where $z_{(1-\alpha/2)}$ and $z_{(1-\beta)}$ are the $(1-\alpha/2)$th and $(1-\beta)$th quantiles, respectively, for the standard normal distribution.

Statistical Power for Detecting QTL



Genetic Power Calculator (PGC)
http://pngu.mgh.harvard.edu/~purcell/gpc/

Genetic Power Calculator

Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to *Shaun Purcell*.

If you use this site, please reference the following Bioinformatics article:

```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Genetic Power Calculator

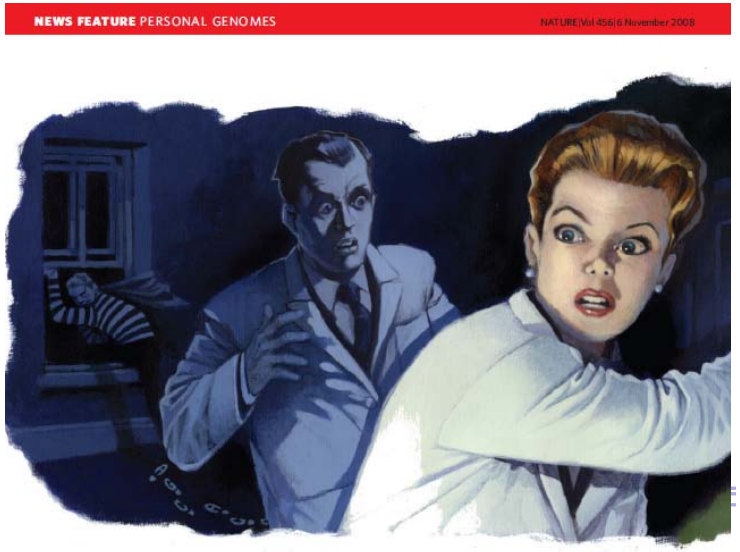
QTL Association for Sibships

```
Total QTL variance                : [    ] (0 - 1)
Dominance : additive QTL effects  : [    ] (0 - 1) [✓] No dominance (* see below)
QTL increaser allele frequency    : [    ] (0 - 1)
Marker M1 allele frequency        : [    ] (0 - 1)
Linkage disequilibrium (D-prime)  : [    ] (0 - 1)
Sibling correlation               : [    ] (0 - 1) (* see below)

Sample Size                       : [        ] (0 - 10000000) (N=families, not individuals)
Sibship Size                      : [Pairs   ] [✓] Both parents genotyped

User-defined type I error rate    : [0.05    ] (0.00000001 - 0.5)
User-defined power: determine N   : [0.80    ] (0 - 1)
(1 - type II error rate)
```

[Process] [Reset]

Missing Heritability

Disease	Number of loci	Percent of Heritability Measure Explained	Heritability Measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenotypic variance
Fasting glucose	4	1.5%	Phenotypic variance



- GWAS works
- Effect sizes are typically small
  - Disease: OR ~1.1 to ~1.3
  - Quantitative traits: % var explained <<1%

Genetic Power Calculator (Shaun Purcell)

http://pngu.mgh.harvard.edu/~purcell/gpc/

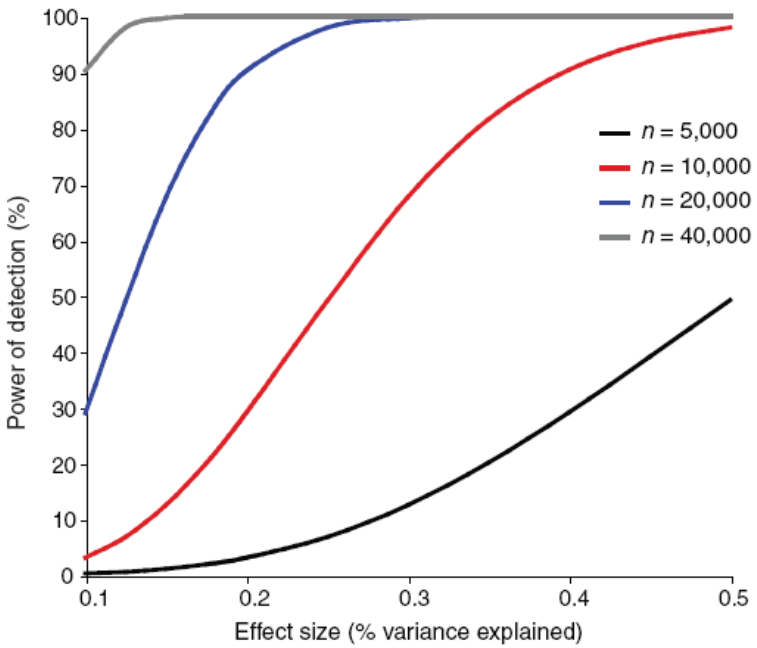
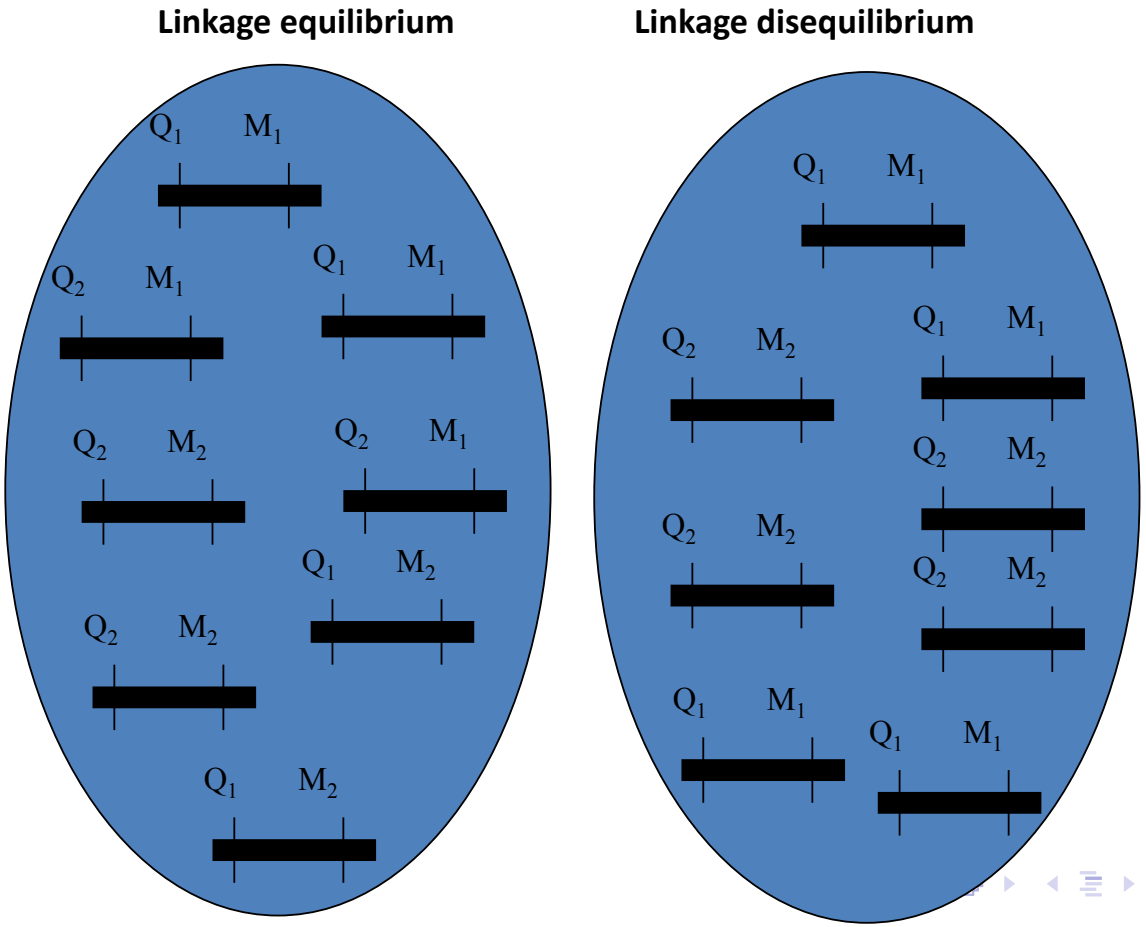


Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical for a GWAS with a sample size of n = 5,000–40,000.

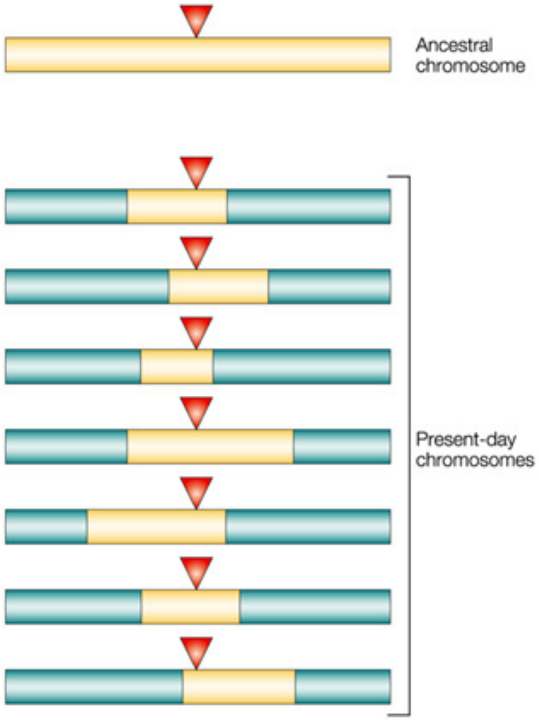
LD Mapping of QTL

- For GWAS, the QTL generally will not be genotyped in a study



LD Mapping of QTL

Linkage disequilibrium around an ancestral mutation



LD Mapping of QTL

- r^2 = LD correlation between QTL and genotyped SNP
- Proportion of variance of the trait explained at a SNP $\approx r^2 h_s^2$
- Required sample size for detection is

$$N \approx \frac{1 - r^2 h_s^2}{r^2 h_s^2} \left(z_{(1-\alpha/2)} + z_{(1-\beta)} \right)^2$$

- Power of LD mapping depends on the experimental sample size, variance explained by the causal variant and LD with a genotyped SNP

Rare Variants and Sequencing Studies

- The uncovered variants from GWAS have largely been of small effect and explain only a small fraction of trait heritability.
- Rare variants, defined here as variants with minor allele frequencies less than $1\% - 5\%$, likely play a significant role in many complex traits
- Some of the missing heritability not explained by the common variants identified through GWAS potentially can be explained by causal rare variants

Uncovering Rare Variants through Sequencing

- Detecting rare variant associations from GWAS data is difficult due to rare variants having low LD with common variants SNP genotyping arrays
- Advancements in high-throughput sequencing technologies allow for rare variants to be identified from sequence data
- Whole-genome and whole-exome sequencing studies are now routinely conducted for the identification of rare-variants that are associated with complex traits.

Association Testing with Rare Variants

- The single-variant association tests previously discussed have essentially no power for detecting associations with rare variants due to their low frequencies.
- A popular strategy for detecting rare variant associations from sequencing data is to jointly consider all rare variants in a genetic region or gene in the association analysis.

Association Testing with Rare Variants

- Consider a sample with n individuals with quantitative phenotype data available and that have been sequenced in a genetic region with m variant sites.
- For individual i in the sample, let y_i denote the quantitative trait value for individual i in the sample.
- Let $\mathbf{G_i} = (g_{i1}, \ldots, g_{im})$ denote the genotypes for individual i at the m variant sites where g_{ij} takes values of 0, 1, or 2 according to the minor allele count at the j-th variant site.
- Most association methods for rare variants from sequencing data can be classified into two groups: burden association tests and kernel association tests

Burden Association Tests for Rare Variants

- Burden tests aggregate all rare variants across a genetic region into a single value for each individual.
- Weights for the m variant sites, $w_1, \ldots, w_m$, can also be used:
  - variant sites can all be given the same weight
  - weighting can be done according to minor allele frequency
  - weights can be assigned based on other features of the variants, such as functionality.

Burden Association Tests for Rare Variants

- A general linear regression model for burden tests is:

$$y_i = \beta_0 + \beta_1 \sum_{j=1}^{m} w_j g_{ij} + \epsilon_i$$

- A sample individual's aggregate value across a genetic region into a single value, $\sum_{j=1}^{m} w_j g_{ij}$, which can be viewed as the individuals genetic burden score.
- Association of phenotype and burden scores (or the variant-sum across the region) can be obtained by testing the null hypothesis $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ using a score statistic.

Burden Association Tests for Rare Variants

- Burden tests have been demonstrated to have reasonable power when a large proportion of the rare variants in a region are causal and with effects on phenotype that are in the same direction
- Can also perform burden tests for a case-control phenotype:

$$logit(\pi_i) = \beta_0 + \beta_1 \sum_{j=1}^{m} w_j g_{ij}$$

  where π_i is the disease probability for individual i.
- A number of burden tests have been proposed including the combined multivariate and collapsing method (Li and Leal, 2008) and the weighted sum test (Madsen and Browning, 2009) and .

Kernal Association Tests for Rare Variants

- Kernel association methods do not aggregate variants into a single value as the burden tests do.
- Kernel-based approaches aggregate individual variant statistics measuring strength of association with each variant site.
- The sequence kernel association test (SKAT) proposed by Wu et al. (2011) is a widely used kernel rare variant association test.

SKAT for rare-variant association testing

- SKAT is based on the following linear mixed model:

$$y_i = \alpha_0 + \mathbf{G_i^T}\boldsymbol{\beta} + \epsilon_i$$

where

  - $\mathbf{G_i} = (g_{i1}, \ldots, g_{im})^T$ is the genotype vector for individual i at the m variant sites,
  - $\boldsymbol{\beta}$ is an $m \times 1$ vector of random effects for the m variant sites with $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau\mathbf{W})$
  - τ is the variance component for the variants
  - $\mathbf{W}$ is an m-dimensional diagonal matrix of pre-specified weights for the variant sites.

- The SKAT association statistic is a score statistic for testing the null hypothesis $H_0 : \tau = 0$ versus $H_A : \tau > 0$

SKAT for rare-variant association testing

- This is equivalent to testing the null hypothesis of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$ but without requiring an m-degree of freedom test for detecting genetic associations for all variant sites, which would have little to no power when testing multiple variants with small effect sizes.
- Note that the variance-covariance matrix of the phenotype data $\mathbf{Y} = (y_1, \ldots, y_n)$ is a function of the linear "kernel" matrix $\mathbf{K} = \mathbf{GWG^T}$, where $\mathbf{G} = (\mathbf{G_1}, \ldots, \mathbf{G_n})$ is an $n \times m$ matrix of the genetic vectors for the n sample individuals at the m variant sites.
- Other non-linear kernel matrices have also been proposed.

SKAT-O: A unified rare-variant association approach

- Kernal-based methods have higher power when a large proportion of the variants are non-causal or if there is a combination of both risk and protective variants in a region that are influencing the phenotype (Wu et al., 2011).
- Lee et al. (2012) proposed the SKAT-O method (where the "O" is for "optimal")
- SKAT-O is a unified rare-variant association approach that essentially a weighted average of burden and SKAT score tests. The optimal weighting, based on the data, is chosen for increased power.
- SKAT-O has been demonstrate to work well in a variety of rare-variant association settings.

References

- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. **Am J Hum Genet** 91:224-237.
- Li B, Leal SM (2008) Methods for Detecting Associations with Rare Variants for Common Diseases: Application to the Analysis of Sequence Data. **Am J Hum Genet** 83: 311-321.
- Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. **PLoS Genet** 5(2):e1000384.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. **Am J Hum Genet** 89: 82-93.

Identity by Descent, the Kinship Coefficient, and the Coefficient of Fraternity

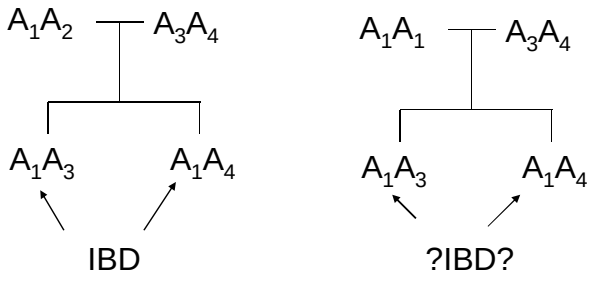
Summer Institute in Statistical Genetics 2014

Module 10

Topic 6

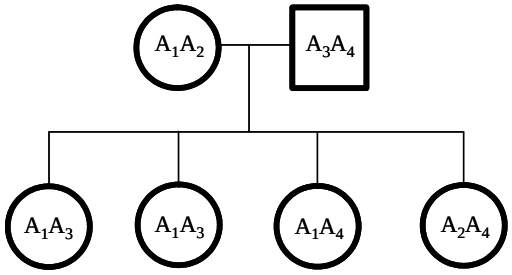
Identical by Descent (IBD)

Two alleles are IBD if they are copies of the same ancestral gene. (The definition of "ancestor" is somewhat dependent on your pedigree.)



Exercise 1

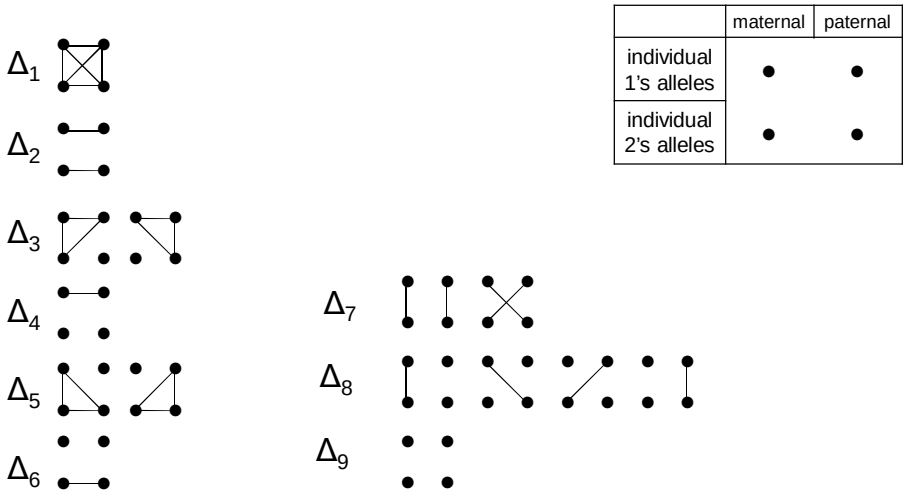
For the following nuclear family, there are six pairs of siblings. For each pair, how many alleles are shared IBD?



Consider a single locus in two individuals. There are four alleles. We can use a schematic to describe IBD status. Start with a grid of four nodes:

	maternal	paternal
individual 1's alleles	•	•
individual 2's alleles	•	•

Connect a pair of nodes with a line if the alleles are IBD. There are 15 possible configurations of IBD (next slide). We do not usually care about the distinction between maternal and paternal alleles, so the 15 possible configurations of IBD reduce to nine "identity states."



For a given pair of individuals there are probabilities, Δ_1 through Δ_9 associated with each of the nine states. Δ_i is the probability the two individuals have IBD status in state i.

The first six states only occur if there is inbreeding. In large, random-mating populations the probability of the first six states is essentially 0. That is, Δ_1 through Δ_6 are 0. Δ_7 through Δ_9 are the probabilities of interest.

Note that Δ_7 can be characterized as the probability that two individuals share <u>both</u> of their alleles IBD. Δ_7 is also called the **coefficient of fraternity**.

θ_{ij} is the **kinship coefficient**, also called the **coefficient of coancestry**, and is the probability that for a given genetic locus a randomly selected allele from individual i and a randomly selected allele from individual j are identical by descent (IBD).

The coefficients of kinship and fraternity characterize the IBD sharing for a pair of relatives. The more distant two relatives are, the less likely it is that they will have alleles IBD. The table on the next slide summarizes these coefficients for a number of relationships.

Coefficients of kinship and fraternity for common relationships in a non-inbred population:

Relationship	θ	Δ_7
Parent-offspring		
Grandparent-grandchild	1/8	0
Great grandparent-great grandchild	1/16	0
Half sibs	1/8	0
Full sibs, dizygotic twins		
Uncle(aunt) – nephew(niece)	1/8	0
First cousins	1/16	0
Double first cousins	1/8	1/16
Second cousins	1/64	0
Monozygotic twins		

Exercise: Fill in the missing values in the table

- Monozygotic (“identical”) twins

- Parent-offspring

- Full siblings

Inbreeding

- Two individuals that have IBD alleles are **related**.
- We defined θ = probability that an allele taken at random from one individual is IBD to an allele from the same locus taken at random from another individual (coancestry coefficient).
- The **inbreeding coefficient** of an individual is the coancestry of its parents.

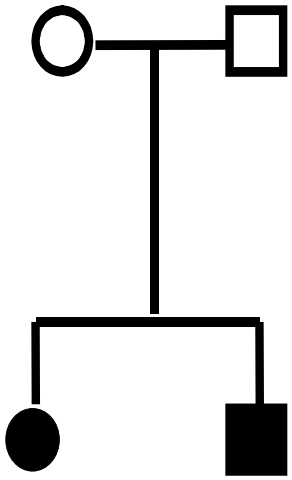
Estimating Relatedness in Homogenous Populations

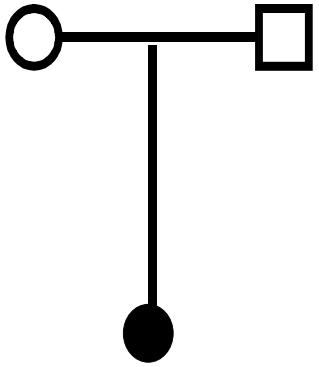
Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 7: Part I

Incomplete Genealogy

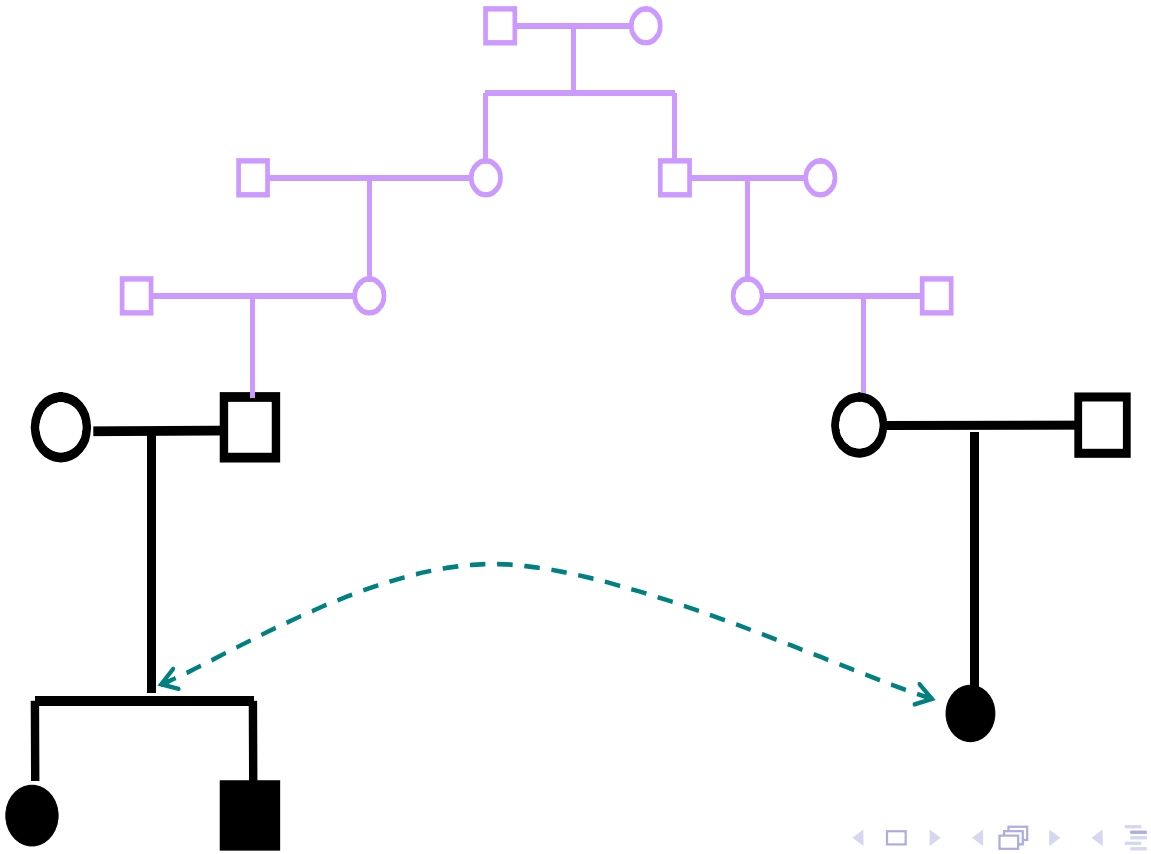
- Many statistical methods for genetic data, e.g. linkage and association methods, are based on assumptions of independent samples or samples with known relationships.





Incomplete Genealogy

- Misspecified and cryptic relationships can invalidate many of these methods.



Identifying Relative Pairs

- In principle, could determine the relationship between two individuals by simply looking at the percentage of IBD sharing in the genome for the two
  - parent-offspring sharing: 50% of genome
  - sibs: 50% of genome (on average)
  - avuncular: 25% of genome (on average)
- However, we do not directly observe IBD sharing. We only observe DNA sequences.

Genome Screen Data to Identify Relative Pairs

- It is now common to have genome screen data on hundreds of thousands of genetic markers.
- Genome screen data can be used to infer genealogical relationships.
- Example: Suppose we are interested in identifying the relationship between two individuals and assume for now that haplotype phase is known.
- Observed sequence on a chromosome from individual 1:

  ...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...
- Observed sequence on a chromosome from from individual 2:

  ...GGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...
- If haplotype phase is known, blocks of identical DNA sequences can be used to infer relationships.

Genome Screen Data to Identify Relative Pairs

- Stanley F Nelson (UCLA Department of Human Genetics): IBD sharing between relatives: rapid drop in number of blocks yet size drops asymptotically:
  - 1st cousins: n=20-30, average size$\sim$20-30mb
  - 2nd cousins: n=5-8, average size$\sim$20mb
  - 3rd cousins: n=1-3, average size $\sim$18mb
  - 4th cousin: n=0-1, average size $\sim$16mb
  - 5th cousins: n=0-1, average size $\sim$14mb
  - 6th cousins: n=0-1, average size$\sim$12mb

Hidden Markov Model for Identifying Relative Pairs

- McPeek and Sun (2000) developed approximate likelihood method to identify relative pairs for close relationships
- Stankovich et al. (2005) extended method for more distantly related pairs (degree 13: 6th cousin). Software is GBIRP
- Uses a 2-state Hidden Markov model for IBD status (yes/no) to approximate the likelihood
- Likelihood is a function of the distance between genetic markers, frequency of alleles between the markers, and relationship of individuals

Hidden Markov Model for Identifying Relative Pairs

- Find pairwise relationship that maximizes the log likelihood ratio for the observed genome screen data (g_1, g_2) over various types of relationships (up to 6th cousins)

$$log \frac{P(g_1, g_2 | related)}{P(g_1, g_2 | unrelated)}$$

- High power to identify relationships up to degree eight (third cousins once removed)
- Typical error in degree for relationship $\leq$ eight is 1

GBIRP Results for Known Relationships

Table: GBIRP MS Pairs

ID1	ID2	Truth	Estimate
20001	30001	2	2
23908	24501	3	3
5809	3701	3	3
45101	45201	4	4
6807	9603	5	6
4801	3701	5	5
8201	42204	5	6
7202	7804	5	7
31001	7603	6	6
4801	5809	6	6
6802	21006	6	6
30602	20503	7	7
30603	9803	7	7
133505	30103	7	9
32204	1303	8	7
33404	4204	8	8
23804	1303	8	8
30501	7037	9	9
2901	602	9	∅
6202	602	9	∅
8003	1704	10	∅
4902	42204	10	∅
20503	1203	11	9
24001	32801	11	12
30501	7902	13	∅

IBD Sharing Probabilities

- IBD sharing probabilities are another measure of relatedness for pairs of individuals
- For any pair of outbred individuals i and j, let δ_k be the probability that i and j share k alleles IBD at a locus where k is 0, 1, or 2.

IBD Sharing Probabilites for Outbreds

Relationship	δ_2	δ_1	δ_0
Parent-Offspring	0	1	0
Full Siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Half Siblings	0	$\frac{1}{2}$	$\frac{1}{2}$
Uncle-Nephew	0	$\frac{1}{2}$	$\frac{1}{2}$
First Cousins	0	$\frac{1}{4}$	$\frac{3}{4}$
Double First Cousins	$\frac{1}{16}$	$\frac{6}{16}$	$\frac{9}{16}$
Second Cousins	0	$\frac{1}{16}$	$\frac{15}{16}$
Unrelated	0	0	1

Estimating IBD Sharing Probabilities: EM Algorithm

- It is often not be possible to determine exactly how many alleles a pair share IBD.
- Can estimate IBD sharing probabiliting using genetic marker data across the genome.
- Choi, Wijsman, and Weir (2009) proposed using an EM algorithm to estimate the IBD probabilities for this problem.

Estimating IBD Sharing Probabilities: EM Algorithm

- Suppose the data consists of N genetic markers accross the genome
- Assume for now that at we observe IBD sharing at each marker for individuals i and j in the sample
- Let X_k be the number of markers for which i and j share k alleles IBD, and let let δ_k be the probability that i and j share k alleles IBD at a merek where k is 0, 1, or 2..
- If the IBD sharing process at the markers is observed, what would the likelihood function be?

Estimating IBD Sharing Probabilities: EM Algorithm

- The likelihood function for the IBD sharing process would have the following multinomial distribution

$$L(X_0, X_1, X_2) = \frac{N!}{X_0! X_1! X_2!} \delta_0^{X_0} \delta_1^{X_1} \delta_2^{X_2}$$

  where $X_k = \sum_{r=1}^{N} I\{\ i \text{ and } j \text{ share k alleles IBD at marker } r\}$
- Could estimate the δ_k's using the X_k's, which are the sufficient statistics: The MLE is $\hat{\delta}_k = \frac{X_k}{N}$ for $k = 0, 1, 2$.
- The IBD process, however is not observed.
- What is the complete data and what is the observed data?

Expectation Step of EM Algorithm

- The X_k values are the unobserved complete data.
- The observed data is the genotype data for individuals i and j at the N markers, and the X_k values are the missing data
- The E step of the EM algorithm calculates the expected value of X_k conditioned on the observed genotype data.
- Remember that initial values for the δ_k's need to be given for the EM algorithm.
- Let $\delta^0 = (\delta_0^0, \delta_1^0, \delta_2^0)$ be the initial values.
- Let $\mathbf{G} = (G_1, \ldots G_r, \ldots G_N)$, where $G_r = (G_{i_r}, G_{j_r})$ is the genotype data at marker r for i and j.

Expectation Step of EM Algorithm

- $X_2 = \sum_{r=1}^{N} I\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\}$
- $E\left[X_2 | \mathbf{G}, \delta^0\right] =$

$$\sum_{r=1}^{N} E\left[I\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\} | \mathbf{G}, \delta^0\right]$$

$$= \sum_{r=1}^{N} E\left[I\{\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\} | G_r, \delta^0\right]$$

$$= \sum_{r=1}^{N} P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r | G_r, \delta^0\right)$$

$$= \sum_{r=1}^{N} \frac{P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r | \delta^0\right)}{P\left(G_r | \delta^0\right)}$$

Expectation Step of EM Algorithm

- The numerator of the summand is
  $P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r|\delta^0\right)$

$$= P\left(G_r|\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, \delta^0\right) \times$$

$$P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r|\delta^0\right)$$

$$= P\left(G_r|\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, \delta^0\right) \delta_2^0$$

- $P\left(G_r|\ i \text{ and } j \text{ share 2 alleles IBD at marker } r\right)$ will be based on the population allele frequency distribution at marker r.

Expectation Step of EM Algorithm

- For simplicity, assume that marker r is a SNP with the 2 allelic types labeled "0" and "1'"
- Let p_r be the frequency of allelic type 1 in the population at marker k, where $0 < p_r < 1$.
- If the genotype of i is (1,1) and the genotype of j is (1,1) at marker r, then
  $P\left(G_r \middle| \ i \text{ and } j \text{ share 2 alleles IBD at marker } r\right) = p_r^2$ (if HWE is assumed).
- What is the probability if the genotype of i is (1,2) and the genotype of j is (2,2) at marker r?
- What is the probability if the genotype of i is (1,2) and the genotype of j is (1,2) at marker r?

Expectation Step of EM Algorithm

- From these probabilities, we can obtain $E\left[X_2|\mathbf{G},\delta^0\right] =$

$$\sum_{r=1}^{N} \frac{P\left(\ i \text{ and } j \text{ share 2 alleles IBD at marker } r, G_r|\delta^0\right)}{P\left(G_r|\delta^0\right)}$$

- Can similarly obtain $E\left[X_1|\mathbf{G},\delta^0\right]$ and $E\left[X_0|\mathbf{G},\delta^0\right]$, where

$$X_1 = \sum_{r=1}^{N} I\left\{\ i \text{ and } j \text{ share 1 alleles IBD at marker } r\right\}$$

and

$$X_0 = \sum_{r=1}^{N} I\left\{\ i \text{ and } j \text{ share 0 alleles IBD at marker } r\right\}$$

Maximization Step of EM Algorithm

- The M step involves maximizing the expected value of the log-likelihood (obtained in the E step) with respect to the δ_k parameters.
- The MLE is:
  - $\hat{\delta}_0 = \frac{E\left[X_0|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
  - $\hat{\delta}_1 = \frac{E\left[X_1|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
  - $\hat{\delta}_2 = \frac{E\left[X_2|\mathbf{G},\delta^0\right]}{E[X_0|\mathbf{G},\delta^0]+E[X_1|\mathbf{G},\delta^0]+E[X_2|\mathbf{G},\delta^0]}$
- The next step is to set $\delta^1 = \hat{\delta}$ and then return to the E step of the algorithm.
- Continue iterating between the E and M step until the $\hat{\delta}^i$ values converge.

Estimating IBD Sharing Probabilities: Method of Moments

- Purcell et al. (2007) proposed a method of moments estimator for IBD sharing probabilities
- Estimate IBD sharing probabilities based on IBS sharing for pairs of individuals
- Implements the IBD sharing method of moments estimator in their software package PLINK

Estimating Kinship Coefficients

- Kinship coefficients can also be used to quantify relationships between two individuals.

Table: Kinship Coefficients

Relationship	ϕ
Parent-Offspring	1/4
Full Siblings	1/4
Half Siblings	1/8
Uncle-nephew	1/8
First Cousins	1/16
Double First Cousins	1/8
Second Cousins	1/64
unrelated	0

- Note that $\phi = \frac{1}{2}\delta_2 + \frac{1}{4}\delta_1$

Estimating Kinship Coefficients

- Thornton and McPeek (2010) propose a method to estimate kinship coefficients using genetic marker data
- Consider once again a marker r with 2 allelic types labeled "0" and "1"
- Let p_r be the frequency of allelic type 1, where $0 < p_r < 1$.
- Consider two individuals i and j. For individual i, let $Y_{i_r} = \frac{1}{2} \times$ (the number of alleles of type 1 in individual i at marker r). So the value of Y_{i_r} is 0, $\frac{1}{2}$, or 1. Similarly define Y_{j_r} for individual j.
- It can be shown that $Cov(Y_{i_r}, Y_{j_r}) = p_r(1 - p_r)\phi_{ij}$, where ϕ_{ij} is the kinship coefficient for i and j.
- Rearrange terms to see that $\phi_{ij} = \frac{Cov(Y_{i_r}, Y_{j_r})}{p_r(1-p_r)}$

Estimating Kinship Coefficients

- This relationship will hold for markers across the genome (with the allele frequency distribution changing for each marker).
- Can use data across the genome to estimate kinship coefficients for pairs of individuals
- Let N be the total number of markers in the data.
- For any pair of individuals i and j, can estimate ϕ_{ij} with

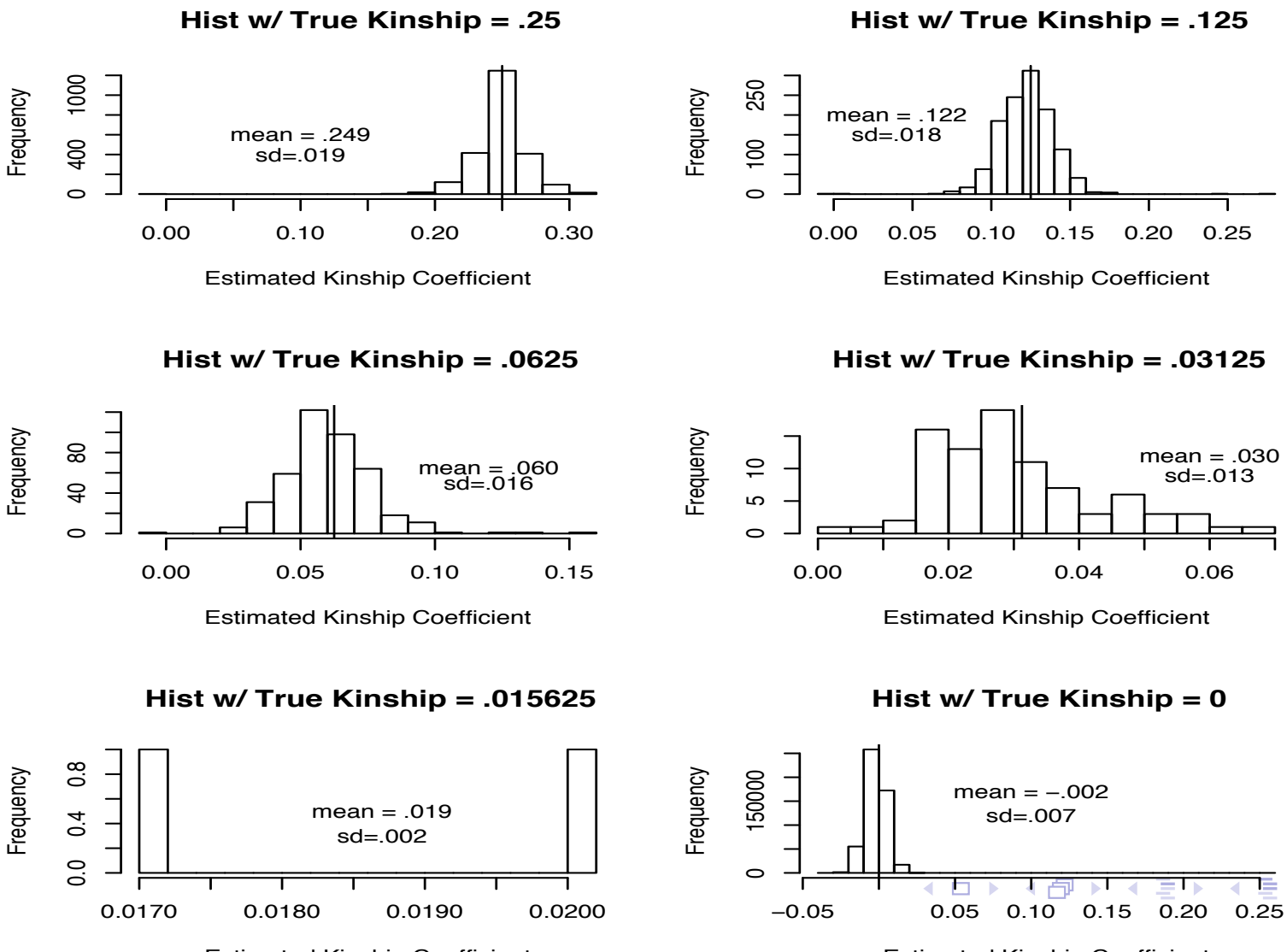
$$\hat{\phi}_{ij} = \frac{1}{N} \sum_{r=1}^{N} \frac{(Y_{i_r} - \hat{p}_r)(Y_{j_r} - \hat{p}_r)}{\hat{p}_r(1 - \hat{p}_r)}$$

where $\hat{p}_r$ is an allele frequency estimate for the type 1 allele at marker r

Estimating Kinships Using GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals. Individuals labeled as "white, non-Hispanic" were considered.
- 10K SNP array (10,081 SNPs) on 22 autosomal chromosomes
- Estimated kinship coefficients using genome-screen data

Estimating Kinships Using COGA Data



Estimating Kinships Using COGA Data

- From the given pedigrees, two pairs of individuals that should have a kinship coefficient of .25 appear to be unrelated (estimated kinship coefficients of -0.006 and -0.003, respectively)
- Two pairs of individuals that should have a kinship coefficient of .125 appear to be unrelated (estimated kinship coefficients of -0.003 and 0.002, respectively)
- 9 pairs of "unrelated" individuals have a kinship coefficient around .125
- 2 pairs of "unrelated" individual have a kinship coefficient around .25

References

- Choi Y, Wijsman EM, Weir BS (2009). Case-control association testing in the presence of unknown relationships. *Genet. Epi.* **33**, 668-678.
- McPeek MS and Sun L (2000). Statistical Tests for Detection of Misspecified Relationships by Use of Genome-Screen Data, *Am. J. Hum. Genet.* **66**, 1076-1094.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575.

References

- Stankovich J, Bahlo M, Rubio JP, Wilkinson CR, Thomson R, Banks A, Ring M, Foote SJ, Speed TP (2005). Identifying nineteenth century genealogical links from genotypes. *Hum. Genet.* **117**, 188-199
- Thornton T, McPeek MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.

Estimating Relatedness in Populations with Admixed Ancestry

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 7: Part II

Relatedness Inference in Structured Populations

- Popular algorithms for relationship inference are based on a strong assumption of population homogeneity
- This assumption is often untenable. GWAS often have cryptic population structure (or ancestry differences among the sample individuals)
- In samples with population structure, relationship estimation methods that assume homogeneity can give extremely biased results
- The degree of relatedness among related and unrelated sample individuals with similar ancestry are systematically inflated

Structured Populations with Distinct Ancestral Subpopulations

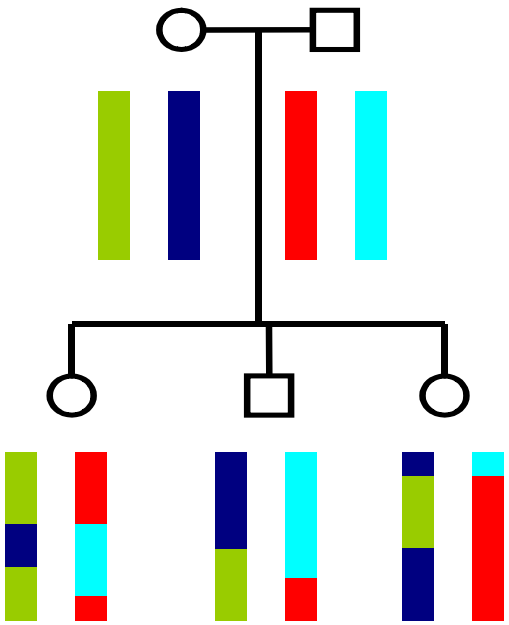
- Manichaikul A et al. (2010) propose an estimator, KING-robust, which stands for Kinship-based INference for Genome-wide association studies
- Estimates kinship coefficients in for individuals from ancestrally distinct subpopulations
- KING-robust estimates kinship coefficients for a pair of individuals by using the shared genotype counts as a measure of the genetic distance between the pair.
- Method does not require allele frequency estimates at the marker: is based on allele sharing counts for individuals
- Gives biased kinship estimates for individuals with different ancestry

Admixed Populations

- Genetic models used to identify related individuals from large scale genetic data often make simplifying assumptions about population structure – either random mating or simple structures.
- In reality, human populations do not mate at random nor are there simple endogamous subgroups.
- While GWAS have primarily examined populations of European ancestry, more recent studies involve admixed populations.
- A number of populations, including the two largest minority populations in the United States, Hispanics and African Americans, are known to have ancestral admixture of chromosomes from different continents.

Ancestry Admixture

- Consider two admixed parents, where each are admixed from different ancestral populations.
- In the picture below, positions on the chromosomes that are the same color are from the same ancestral population.



Relatedness Inference in Admixed Samples

- Thornton et al. (2012) proposed REAP (Relatedness Estimation in Admixed Populations) for relatedness inference in samples from populations with admixed ancestry
- Consider the problem of estimating relatedness in a set N of outbred individuals who are sampled from a population with admixture from K subpopulations
- Let $\mathbf{q}^s = (q_1^s, \ldots, q_K^s)^T$ denote the vector of subpopulation-specific allele frequencies at SNP s, where q_k^s is the allele frequency of SNP s in subpopulation k, $1 \leq k \leq K$.
- Define $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})^T$ to be the genome-wide ancestry vector for $i \in N$, where a_{ik} is the proportion of ancestry from subpopulation k for i, $a_{ik} \geq 0$ for all k, and $\sum_{k=1}^{K} a_{ik} = 1$.

Estimating Relatedness in an Admixed Population

- Let Y_i^s be the genotype variable for individual i, where $Y_i^s = \frac{1}{2} \times$ (the number of alleles of type 1 at SNP s in individual i). Similarly define Y_j^s for individual j.
- Conditional on $\mathbf{q}^s$, we assume alleles of an outbred individual i are independent, identically-distributed (i.i.d.) Bernoulli random variables , a modeling assumption made by other commonly-used models of population structure (Balding-Nichols model with admixture).
- We denote $\mu_i^s = E[Y_i^s | \mathbf{a}_i, \mathbf{q}^s]$ to be the expected value of Y_i^s conditional on $\mathbf{q}^s$ and $\mathbf{a}_i$ where

$$\mu_i^s = \mathbf{a}_i^T \mathbf{q}^s = \sum_{k=1}^{K} a_{ik} q_k^s,$$

- The variance of Y_i^s conditional on $\mathbf{q}^s$ and $\mathbf{a}_i$ is $.5\mu_i^s(1 - \mu_i^s)$.

Estimating Kinship Coefficients: Admixed Population

- For i and j from a homogenous populations, it can be shown that $\phi_{ij} = \frac{1}{2}\rho_{Y_iY_j}$ for i and j , where $\rho_{Y_iY_j}$ is the correlation of Y_i^s and Y_j^s.
- For estimating ϕ_{ij} in structured populations with admixture, we propose to similarly calculate the correlation of Y_i^s and Y_j^s
- Propose using a correlation that is calculated conditional on the admixture ancestry proportions of i and j as well as the subpopulation allele frequencies.

Estimating Kinship Coefficients: Admixed Population

- The conditional correlation that we estimate for inference on ϕ_{ij} is $\rho_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s}$, which is the correlation of Y_i^s and Y_j^s conditional on $\mathbf{a}_i$, $\mathbf{a}_j$, and $\mathbf{q}^s$.
- When genome-screen data is available for i and j we estimate ϕ_{ij} in the presence of population structure with admixture with the REAP estimator

$$\hat{\phi}_{ij}^A = \frac{1}{2}\hat{\rho}_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s}$$

where

$$\hat{\rho}_{Y_i Y_j|\mathbf{a}_i,\mathbf{a}_j,\mathbf{q}^s} = \frac{1}{|\mathcal{S}_{ij}|}\sum_{s\in\mathcal{S}_{ij}} \frac{(Y_i^s - \hat{\mu}_i^s)(Y_j^s - \hat{\mu}_j^s)}{\sqrt{.5\hat{\mu}_i^s(1-\hat{\mu}_i^s)}\sqrt{.5\hat{\mu}_j^s(1-\hat{\mu}_j^s)}},$$

Estimating IBD Sharing Probabilities: Admixed Populations

- Can also extend estimating IBD sharing probabilities in admixed populations.
- Define Z_{ij}^s as before to be an indicator for i and j sharing 0 alleles IBD at SNP s
- Can use the conditional expectation of Z_{ij}^s given $\mathbf{a}_i, \mathbf{a}_j, \mathbf{q}^s$ to obtain a method of moments estimator for δ_{ij}^0 in the the presence of admixture.
- For any pair of individuals i and j from an admixed population,we have that

$$E(Z_{ij}^s|\mathbf{a}_i, \mathbf{a}_j, \mathbf{q}^s) = \left[(\mu_i^s)^2(1-\mu_j^s)^2 + (1-\mu_i^s)^2(\mu_j^s)^2\right]\delta_{ij}^0$$

Estimating IBD Sharing Probabilities: Admixed Populations

- Let $\mathcal{S}_{ij}$ be the set of markers in the genome screen for which both i and j have nonmissing genotype data.
- Our REAP method of moments for δ_{ij}^0 in the presence of admixture is

$$\hat{\delta}_{ij}^{0A} = \frac{\sum_{s \in \mathcal{S}_{ij}} Z_{ij}^s}{\sum_{s \in \mathcal{S}_{ij}} \left[(\hat{\mu}_i^s)^2 (1 - \hat{\mu}_j^s)^2 + (1 - \hat{\mu}_i^s)^2 (\hat{\mu}_i^s)^2 \right]}$$

Estimating IBD Sharing Probabilities: Admixed Populations

- The remaining two IBD sharing probabilities, δ_{ij}^1 and δ_{ij}^2, can be written as a function of δ_{ij}^0 and ϕ_{ij}
- Estimate $\delta_{ij}^{1^A}$ with $\hat{\delta}_{ij}^{1^A} = 2 - 2\hat{\delta}_{ij}^{0^A} - 4\hat{\phi}_{ij}^A$
- Estimate $\delta_{ij}^{2^A}$ with $\hat{\delta}_{ij}^{2^A} = \hat{\delta}_{ij}^{0^A} + 4\hat{\phi}_{ij}^A - 1$.

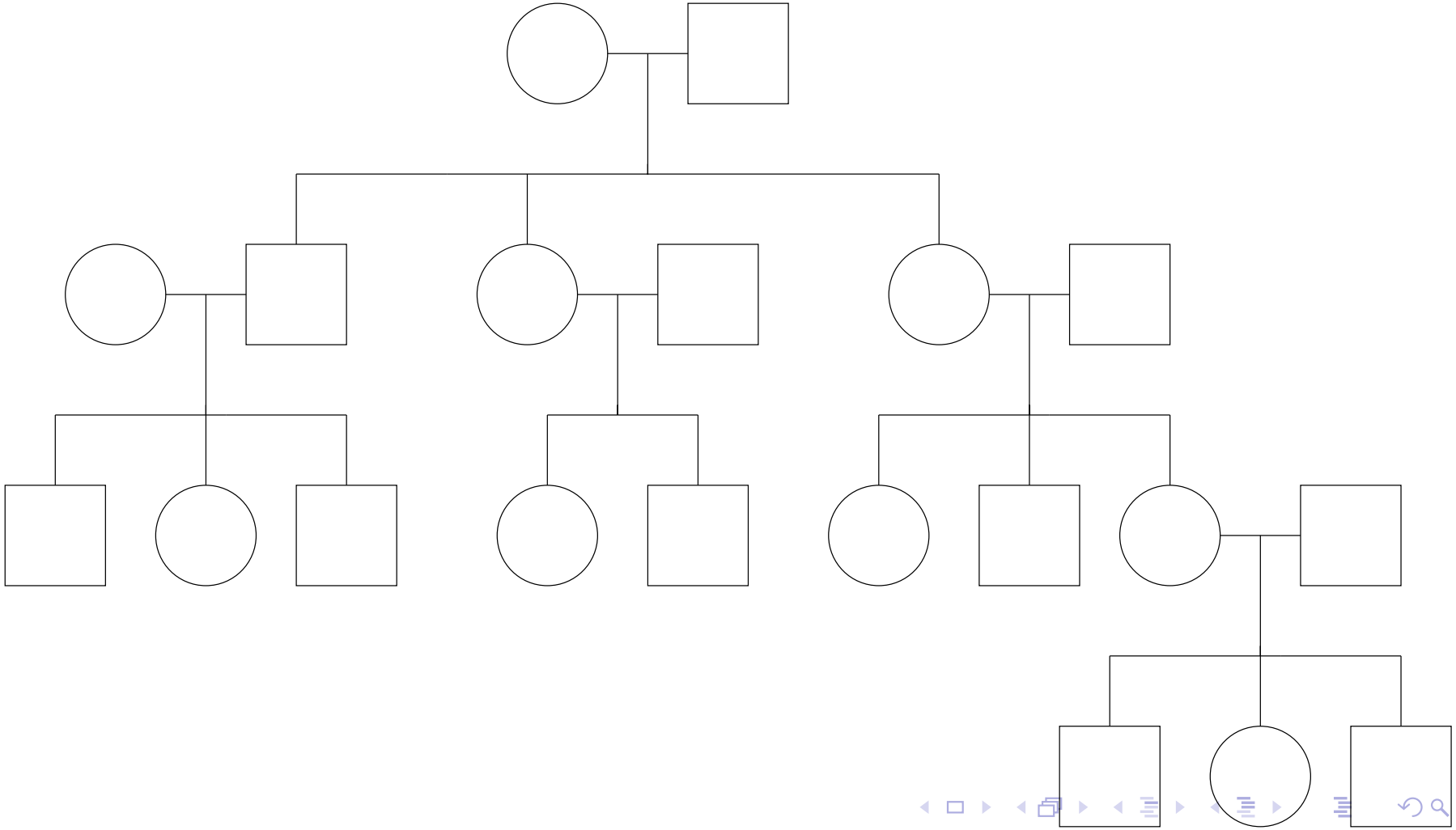
Simulation Studies: Relatedness and Population Structure

- Perform simulation studies, in which population structure and related individuals are simultaneously present
- The population structure settings used in the simulation studies are based on the Balding-Nichols model.
- For each SNP, an ancestral population allele frequency p was drawn from the uniform distribution on [0.1,0.9].
- We set $F_{ST} = .2$ in the Balding-Nichols model to simulate two highly divergent subpopulations.

Simulation Studies: Relatedness and Population Structure

- We consider population structure settings where individuals from an admixed population formed from two divergent subpopulations.
- Population structure setting 1 has individuals sampled from an admixed population formed from ancestral populations and where there is assortative mating.
- Population structure setting 2 has individuals sampled from an admixed population formed from ancestral populations where there is random mating
- We sample 400 individuals from 20 outbred pedigrees containing 1st, 2nd, 3rd, and 4th-degree relationships.

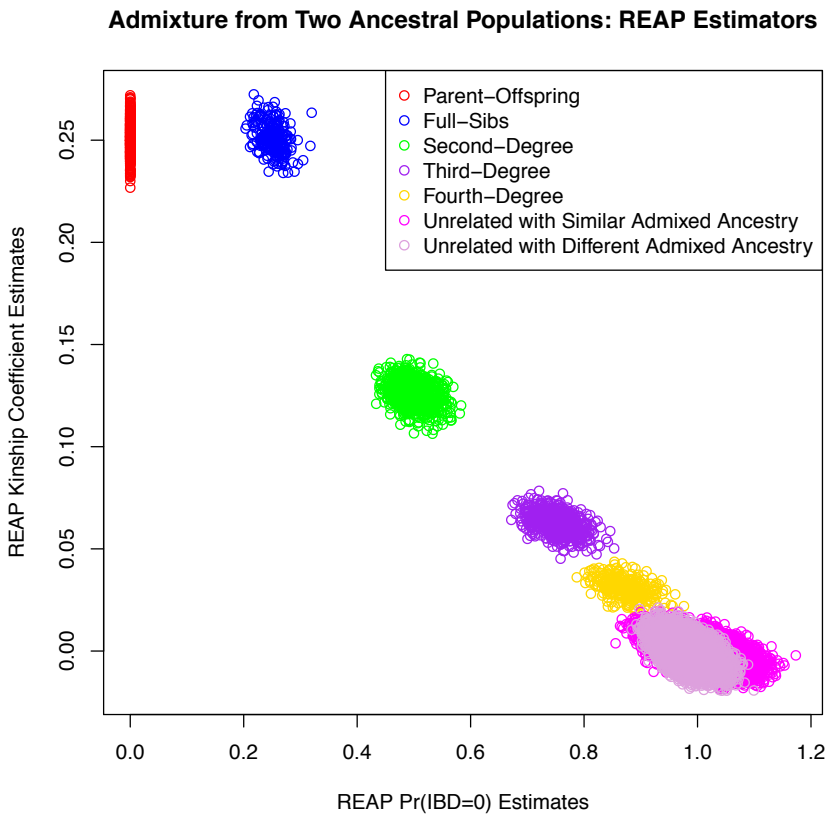
Pedigree Configuration

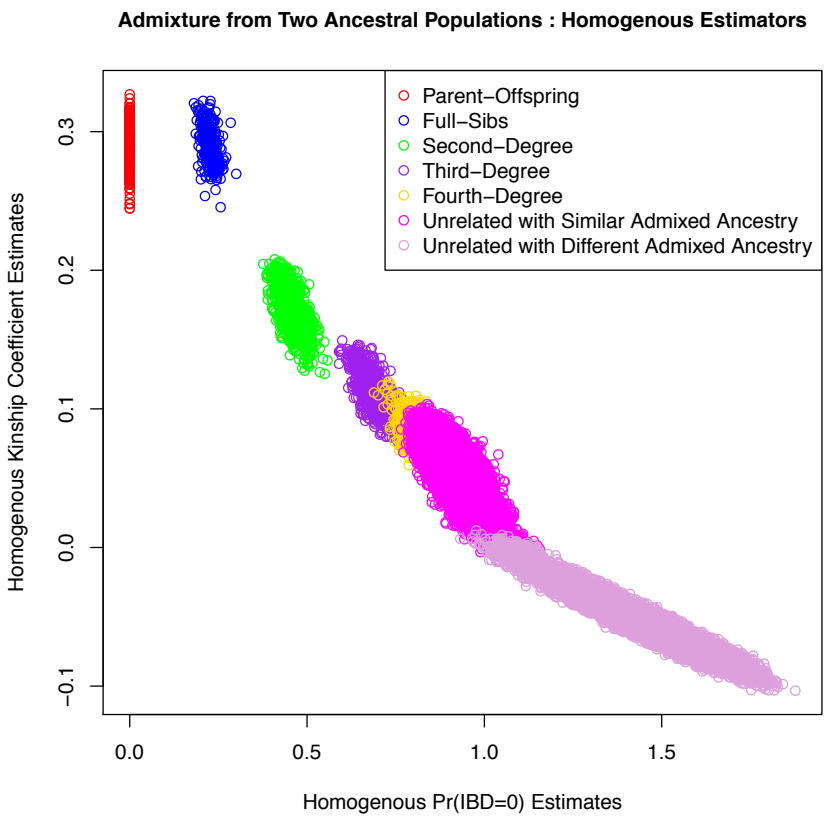


Simulation Studies: Relatedness and Population Structure

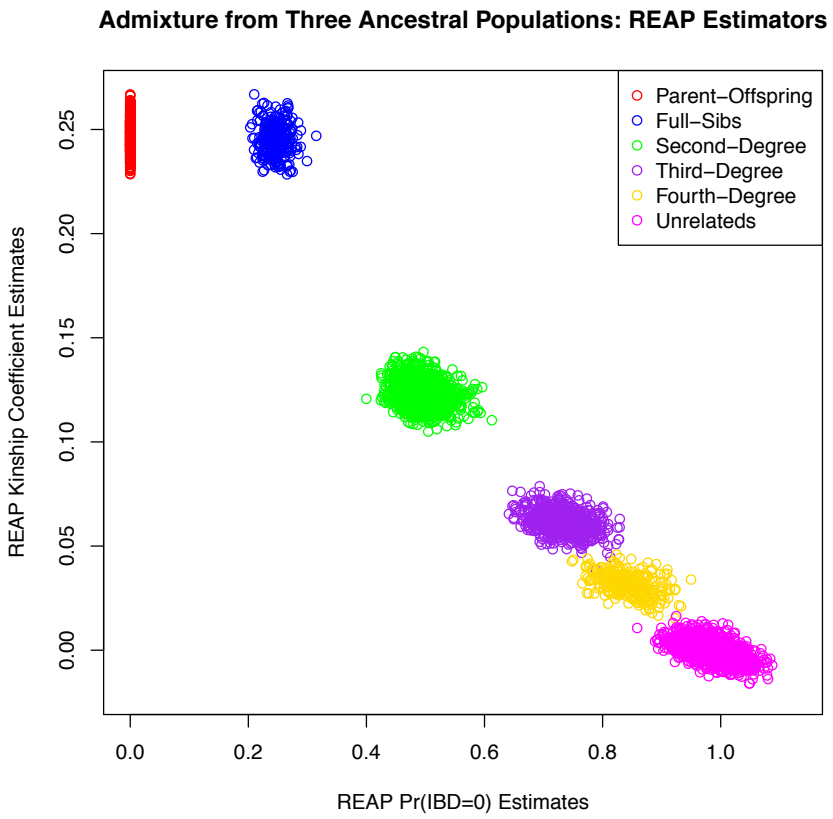
- For each of the two population structure settings we generate genotype data for 10,000 random SNPs.
- Genome-wide ancestry estimates used by REAP for the sample individuals were obtained by the *frappe* software program
- *frappe* implements an EM algorithm for simultaneously inferring each individuals ancestry proportion and allele frequencies in the ancestral populations.

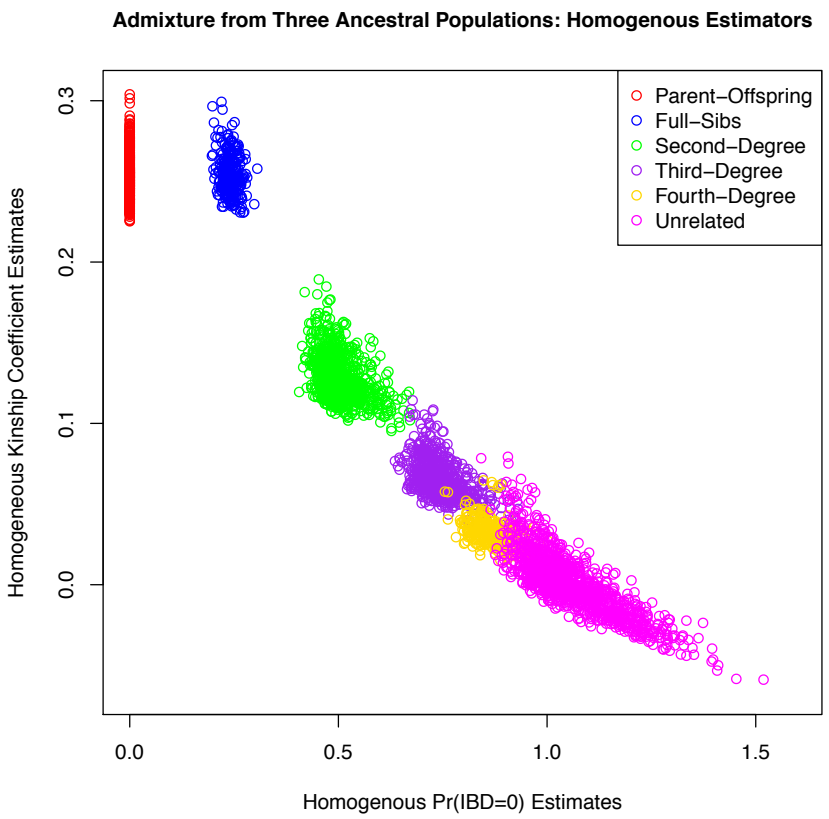
Setting 1: Admixture from Two Ancestral Populations and Assortative Mating





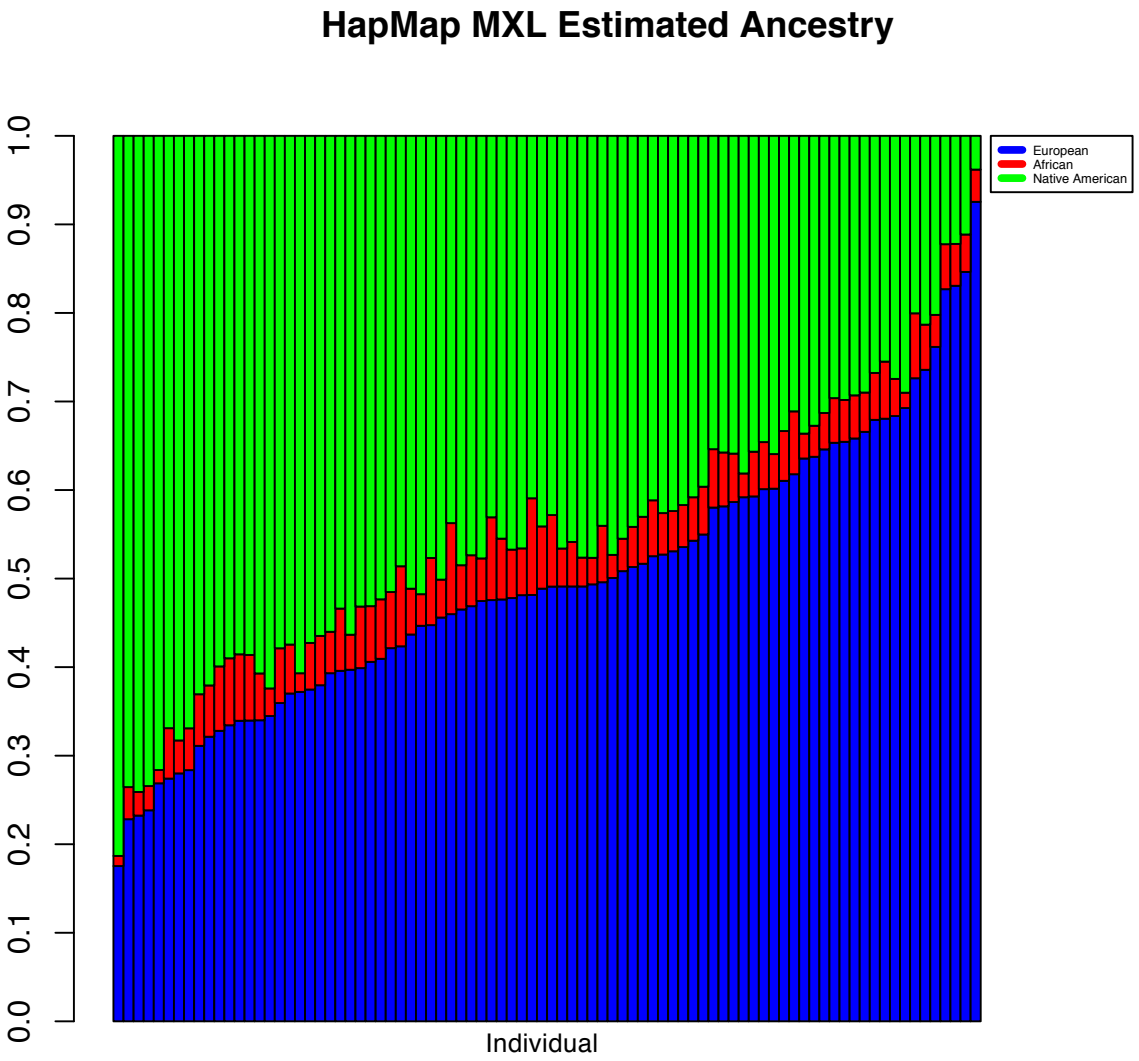
Setting 2: Admixture from Three Ancestral Populations and Random Mating

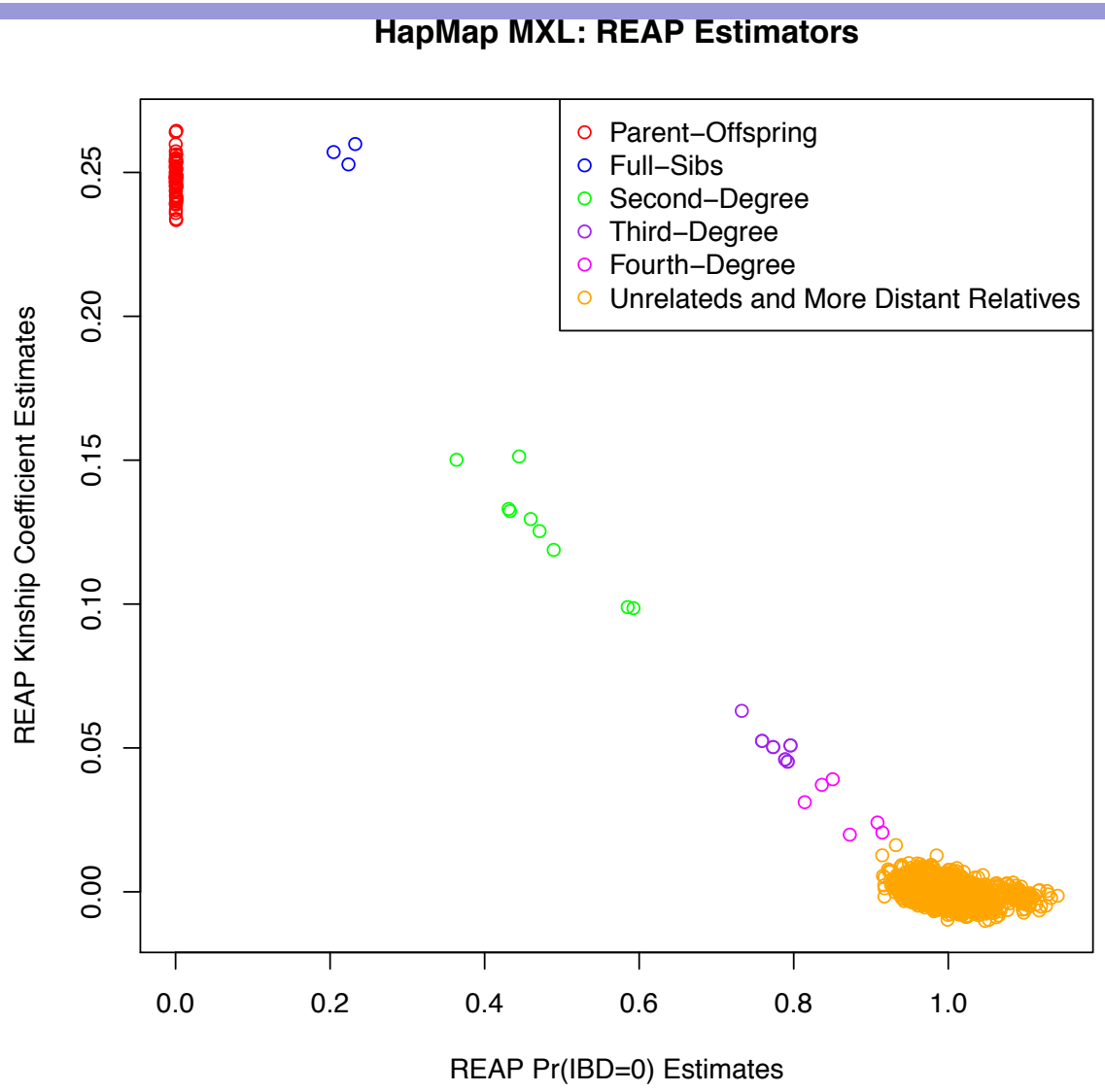




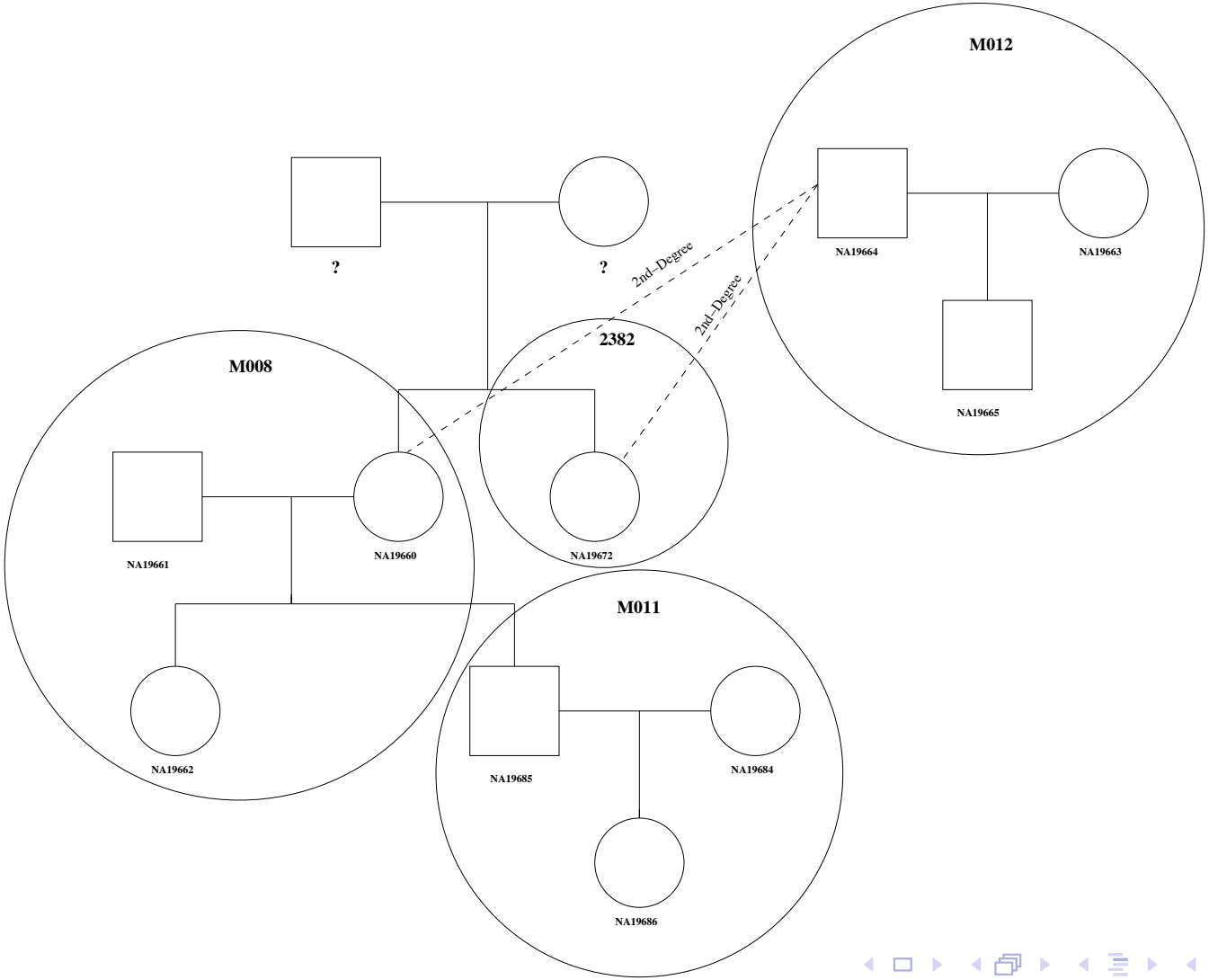
Estimating Kinship: HapMap Mex Sample

- Estimate estimating kinship coefficients and IBD sharing probabilities in the HapMap Mexicans in Los Angeles (MXL) sample of release 3 of phase III..
- Used *frappe* to estimate genome-wide ancestry for the 86 individuals in the sample
- We set the number of ancestral populations $K = 3$
  - HapMaP YRI for African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.


HapMap MXL Estimated Ancestry
European
African
Native American
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Individual


HapMap MXL: REAP Estimators
Parent−Offspring
Full−Sibs
Second−Degree
Third−Degree
Fourth−Degree
Unrelateds and More Distant Relatives
REAP Kinship Coefficient Estimates
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
REAP Pr(IBD=0) Estimates

Reconstructed HapMap MXL Extended Pedigree



Women's Health Initiative

- The Womens Health Initiative (WHI) is a national health study focusing on strategies for preventing chronic diseases in postmenopausal women.
- A total of 161,808 women aged 50-79 yrs. old were recruited from 40 clinical centers in the US between 1993 and 1998.
- The WHI cohort included
  - Two clinical trials of postmenopausal hormone therapy (estrogen alone and estrogen plus progestin)
  - A clinical trial of calcium and vitamin D supplements, and a dietary modification trial.

Genetic analysis of WHI-SHARe Minority Cohort

- Minority populations have largely been underrepresented in genetic studies despite bearing a disproportionately high burden for disease.
- WHI study opens up tremendous new possibilities for the identification of genetic risk factors associated with a number of clinical outcomes in the two largest minority populations in the U.S.
- The WHI SNP Health Association Resource (SHARe) minority cohort includes 8421 self-identified African American women from and 3587 self-identified Hispanic women
- 909,622 single nucleotide polymorphisms (SNPs) across the genome

Ancestry Estimation: WHI-SHARe data

- Used *frappe* to estimate genome-wide ancestry of every individual in the sample
- We set the number of ancestral populations $K = 4$
  - HapMaP YRI for African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.
  - HGDP East Asian samples for East Asian Ancestry

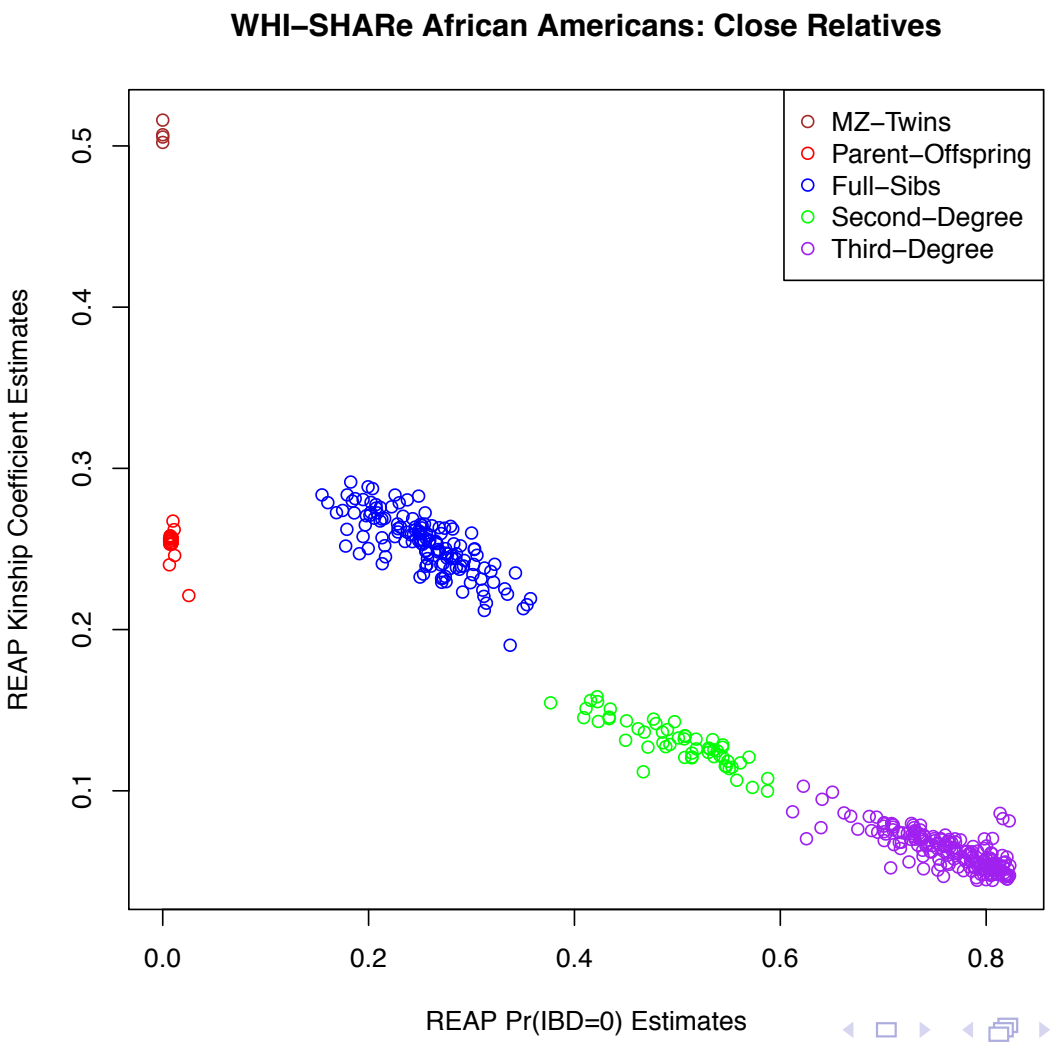
Relatedness Inference in WHI-SHARe

- No available genealogical information for the WHI-SHARe sample
- Used REAP to estimated relationships for all possible pairs:

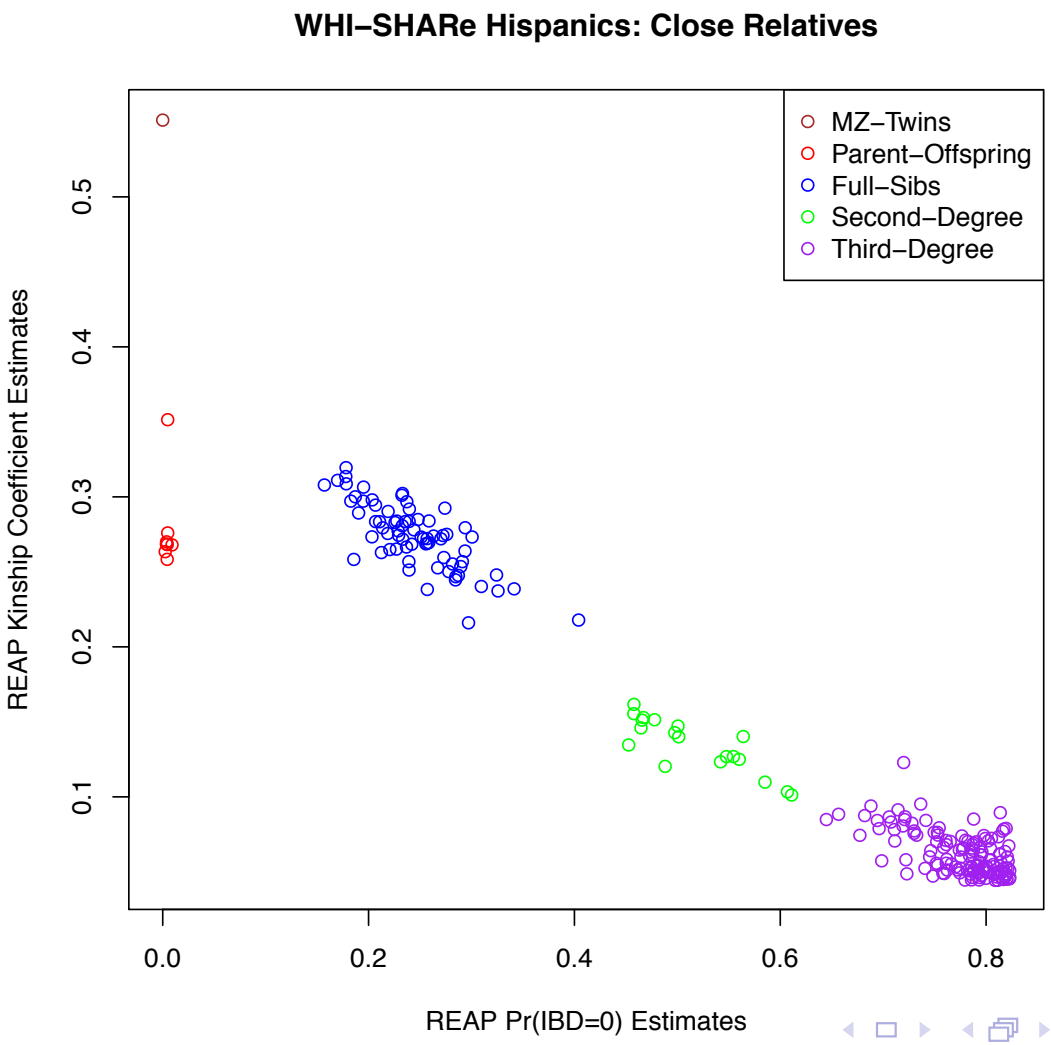
$$\binom{12008}{2} = 7,209,028$$

- Obtained estimates for kinship coefficients and IBD sharing probabilities

WHI-SHARe African Americans



WHI-SHARe Hispanics



Relatedness Inference in WHI-SHARe

- Also used the PLINK software (Purcell et al., 2007) method of moments kinship coefficient estimator: 8,932 pairs are identified to be either first or second degree relatives
- Our REAP kinship estimator that adjusts for individual specific ancestry identifies 344 individuals with kinship coefficients that are consistent with either first or second degree relatives

Relatedness Inference in WHI-SHARe

- Interestingly, there are individuals who are identified as second- and third-degree relative pairs by REAP but who have a different self-reported race/ethnicity, e.g. one individual is a self-report African American and the other is a self-report Hispanic.
- An advantage of the REAP approach is that robust relatedness estimates can be obtained for all individuals, even for individuals who have different admixed ancestry distributions and self-identify in different ethnic or nationality groups.

References

- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies.*Bioinformatics* **26**, 2867-2873.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575.
- Thornton T, Tang H, Hoffman TJ, Ochs-Balcom HM, Baan BJ, and Risch N (2012) Estimating Kinship in Admixed Populations *Am. J. Hum. Genet.* **91**

Association Testing with Related Samples

Timothy Thornton and Katie Kerr

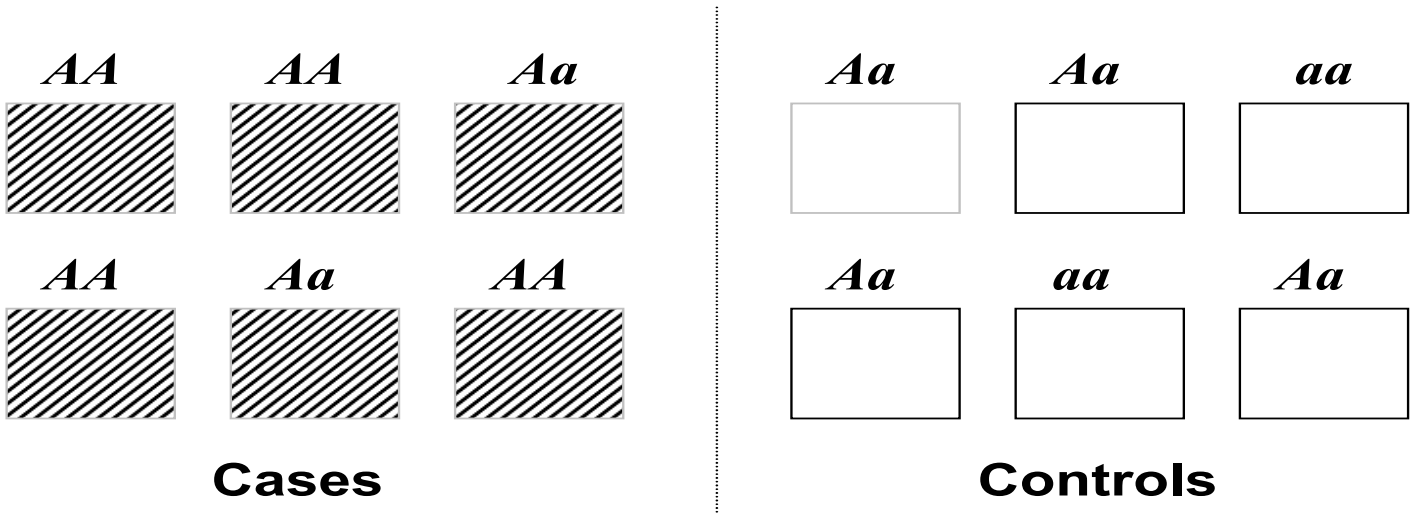
Summer Institute in Statistical Genetics 2014
Module 10
Lecture 8: Part I

Introduction

- Consider testing for association between a disease and a genetic marker
- Another way of looking at case-control association testings is as a comparison of allele/genotype frequencies.
- Idea is to look for an association by identifying genetic markers with large frequencies differences between the cases and controls.

Association Testing

- Below is a simple example to illustrate association testing at a genetic marker with two allelic types, **A** and **a**



Association Testing with Related Samples

- We consider the problem of testing for association between a binary trait and a genetic marker in a case-control design in which some individuals are related with known relationships.
- Advantages in using related individuals in association studies
  - Enrichment of susceptibility alleles in affected individuals with affected relatives (power may be improved)
  - Unaffected relatives of cases can be used as controls to protect against potential problems of unknown population substructure.

Case-Control Association

- For a given disease or trait, suppose case and control individuals have been sampled and genotyped.
- Consider a single marker with 2 allelic types labeled "0" and "1"
- Let p be the frequency of allelic type 1, where $0 < p < 1$.
- Let N be the number of individuals in the study.
- $Y = (Y_1, \ldots Y_i, \ldots Y_N)^T$ where $Y_i = \frac{1}{2} \times$ (the number of alleles of type 1 in individual i). So the value of Y_i is 0, $\frac{1}{2}$, or 1.

Common Form for Association Tests

- Measures for allele frequency differences between cases and controls generally have the following common form:

$$T = \hat{p}_{case} - \hat{p}_{null}$$

- where $\hat{p}_{case}$ is estimator of case allele frequency, $\hat{p}_{null}$ is estimator of allele frequency calculated under assumption of no association,
- A simple measure of T is

$$T = \frac{1}{N_{ca}} \sum_{i \in \text{cases}} Y_i - \frac{1}{N_{co}} \sum_{i \in \text{controls}} Y_i$$

- $E(T) = 0$ and T approximately follows a normal distribution under the null hypothesis of no association between trait and genetic marker.

Common Form for Association Tests

- $W = \frac{T^2}{Var_0(T)}$ can be used to test for association between the trait and the marker, where $Var_0(T)$ is the variance of T under the null hypothesis
- Many association statistics have this form including the classical Pearson's χ^2 test
- Under the null hypothesis, W approximately follows a χ_1^2 distribution
- $Var_0(T)$ is a function of $\sum V(Y_i)$ and $\sum Cov(Y_i, Y_j)$, for all i and j.
- $V(Y_i) = \frac{1}{2}p(1-p)$ for all outbred i.

Covariance Structure with Known Genealogy

- If i and j are unrelated, then $Cov(Y_i, Y_j) = 0$
- If i and j are related, then they have correlated genotypes
- For any pair of related individuals i and j, $Cov(Y_i, Y_j) = p(1-p)\phi_{ij}$
- ϕ_{ij} is the kinship coefficient between individuals i and j
- The kinship coefficient for i and j is the probability that a random allele selected from i and a random allele selected from j are identical by descent (IBD).
- ϕ_{ij} is also the correlation between Y_i and Y_j

Covariance Structure with Known Genealogy

- $Var_0(T)$ is essentially a correction factor that is a function of the Kinship Matrix $\mathbf{\Phi}$. For individuals $1, 2, \ldots N$

$$\mathbf{\Phi} = \begin{pmatrix} 1+h_1 & 2\phi_{12} & \ldots & 2\phi_{1N} \\ 2\phi_{12} & 1+h_2 & \ldots & 2\phi_{2N} \\ \vdots & \ldots & \ldots & \vdots \\ 2\phi_{N1} & 2\phi_{N2} & \ldots & 1+h_N \end{pmatrix},$$

where h_i is the inbreeding coefficient of individual i, and ϕ_{ij} is the kinship coefficient between individuals i and j.

Covariance Structure with Known Genealogy

Table: Kinship Coefficients

Relationship	ϕ
parent-offspring	$1/4$
full siblings	$1/4$
aunt-niece	$1/8$
first cousins	$1/16$
unrelated	0

Association Methods for Related Samples

Allelic Tests

- Risch and Teng (1998)
  - Compare allele frequencies of DNA pools of unaffecteds to frequencies in affected related individuals
  - Derived for particular cases of close relatives
- Corrected χ^2 ($W_{\chi^2_{corr}}$) test: Bourgain et al. (2003)
  - Pearson's χ^2 test with a variance correction.
  - Correction factor takes into account the correlations between related individuals to obtain correct type I error
  - Arbitrarily complex pedigrees.
- A quasi-likelihood score (W_{QLS}) test: Bourgain et al. (2003)
  - Instead of giving individuals equal weights and then correcting variance, weights depend on the correlations between related individuals (to maximize non-centrality parameter within a class of linear statistics).
  - Arbitrarily complex pedigrees

Association Methods for Related Samples

- Modified quasi-likelihood score (M_{QLS}) test: Thornton and McPeek (2007)
  - Improves power by capitalizing on the enrichment of predisposing alleles in affected related individuals
  - Optimal in a general class of linear statistics for arbitrary 2-allele disease models
  - Arbitrarily complex pedigrees

Genotypic Tests

- Slager and Schaid (2001) W_{SS} test
  - Treats individuals as unrelated and applies a correction factor
  - Correction factor calculated conditional on genotype data in region.
  - Derived for particular cases of close relatives
  - Similar to $W_{\chi^2_{corr}}$ allelic test

Association Methods for Related Samples

- All of the methods have the following form:

$$W = \frac{\left(\hat{p}_{case} - \hat{p}_{null}\right)^2}{Var_0\left(\hat{p}_{case} - \hat{p}_{null}\right)}$$

- $\hat{p}_{case}$ is estimator of case allele frequency, $\hat{p}_{null}$ is estimator of allele frequency calculated under assumption of no association, and $Var_0\left(\cdot\right)$ denotes variance calculated assuming that there is no association.
- Risch and Teng (1998), $W_{\chi^2_{corr}}$, and Slager and Schaid (2001): $\hat{p}_{case}$ is sample mean of cases and $\hat{p}_{null}$ is the sample mean based on everyone
- W_{QLS}: $\hat{p}_{case}$ is BLUE of allele frequency for cases and $\hat{p}_{null}$ is the BLUE of allele frequency for everyone
- M_{QLS}: $\hat{p}_{case}$ is estimator based on phenotypes and pedigree for all individuals and $\hat{p}_{null}$ is the BLUE for everyone

Association Methods for Cryptic Relatedness

- Choi, Wijsman, and Weir (2009) estimate relatedness using an EM algorithm (discussed in lecture 9). Use IBD sharing probability estimates to correct the variance of the Pearson χ^2 test.
- Thornton and McPeek (2010) use known pedigree information to improve power. Correct the variance of various association statistics using an empirical covariance matrix to adjust for known, unknown, and misspecified relationships.

References

- Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C , McPeek MS (2004). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am. J. Hum. Genet.* **73**, 612-626.
- Choi Y, Wijsman EM, Weir BS (2009). Case-control association testing in the presence of unknown relationships. *Genet. Epi.* **33**, 668-678.
- Risch N, Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research* **8**, 1273-1288.

References

- Slager SL, Schaid D (2001). Evaluation of candidate genes in case-control studies, a statistical method to account for related subjects. *Am. J. Hum. Genet.* **68**, 1457-1462.
- Thornton T, McPeek MS (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81**, 321-337.
- Thornton T, McPeek MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.

X-Chromosome Association Testing with Related Samples

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 8: Part II

Association Testing on the X-chromosome

- Association methods have primarily been developed for the analysis of markers on the autosomal chromosomes; generally not directly applicable to markers on the X-chromosome
- Females and males have different copy numbers of the X-chromosome, where females have two copies and males have only one

Recent Work for X-chromosome Association Testing

- Clayton D (*Biostatistics* 2008) proposed 1 and 2 degrees-of-freedom χ^2 tests for the association of bi-allelic markers on the X-chromosome in samples with unrelated individuals
- Uh et al. (*BMC Genetics* 2009) proposed an X-chromosome method for samples with related individuals.
  - Proposed analyzing males and females separately, and then combining results for a 2 degrees-of-freedom χ^2 tests.
  - Use autosomal kinship coefficients to account for the correlated alleles/genotypes in the sample.

Association Testing on the X-chromosome

- We consider the problem of testing for association between a binary trait and a X-linked genetic marker in a case-control design in which some individuals are related with known relationships
- We focus on the analysis of markers from the non-pseudoautosomal region of the X-chromosome, where there is no recombination between the X and Y chromosomes.
  - For markers on the pseudoautosomal region of the X and Y chromosomes, autosomal association methods for related individuals can be used to analyze data of this type.
- As in our previous work for autosomal chromosomes, we do not put constraints on how the individuals might be related, allowing for individual relationships to be quite complex, including inbred populations

Association Testing on the X-chromosome

- Challenges for methodology development
  - Accounting for allele copy number differences in males and females for X-linked markers
  - Appropriately adjusting for correlations among same and different gender relatives
  - Computationally feasible for thousands of markers and/or extremely complex pedigrees.
- For a given disease or trait, suppose case and control individuals have been sampled and genotyped.
- Consider a single X-linked marker with 2 allelic types labeled "0" and "1"
- Let p be the frequency of allelic type 1, where $0 < p < 1$.
- Let N be the number of individuals in the study.

Association Testing on the X-chromosome

- We define $\mathbf{Y^X} = (Y_1^X, \ldots Y_i^X, \ldots Y_N^X)^T$ where

$$Y_i^X = \begin{cases} \frac{1}{2} \times (\text{the number of alleles of type 1 in individual } i) & \text{for } i \text{ female} \\ I\{\text{the allele in individual } i \text{ is of type 1}\} & \text{for } i \text{ male} \end{cases}$$

- So the value for Y_i^X is 0, $\frac{1}{2}$, or 1 for females, and 0 or 1 for males.
- We construct Y_i^X in such a way such the under the null, $E[Y_i^X] = p$ for i male or female

Dosage Compensation

- Y_i^X is 0, $\frac{1}{2}$, or 1 for females, and 0 or 1 for males.
- Y_i^X also corresponds nicely with dosage compensation model
- Dosage compensation is a special mechanism in human cells where there is an inactivation of one of the X chromosomes in females.
- If the trait were caused by the X-linked marker, then under dosage compensation trait model:
  - A male with 1 allele of type 1 will have an equivalent phenotypic expression to that of a homozygous (1,1) female
  - A male with 1 allele of type 0 will have an equivalent phenotypic expression to that of a homozygous (0,0) female

Null Variance

- Let $\mathbf{\Sigma_0}$ denote the covariance matrix of $\mathbf{Y^X}$ under the null hypothesis of no association.
- Under the assumption that Hardy-Weinberg equilibrium (HWE) holds at the marker, it can be shown that $\mathbf{\Sigma}_0 = p(1-p)\mathbf{\Phi^X}$ where

$$\mathbf{\Phi^X} = \begin{pmatrix} \phi^X_{11} & \phi^X_{12} & \cdots & \phi^X_{1n} \\ \phi^X_{12} & \phi^X_{22} & \cdots & \phi^X_{2n} \\ \vdots & \cdots & \cdots & \vdots \\ \phi^X_{n1} & \phi^X_{n2} & \cdots & \phi^X_{nn} \end{pmatrix}$$

where ϕ^X_{ij} is the X-kinship coefficient for i and j.

Outbred Kinship Coefficients

Table: X and Autosomal Kinship Coefficients for a Nuclear Family

Relationship of j to i	ϕ_{ij}	ϕ_{ij}^X
Mother of i	$\frac{1}{4}$	$\frac{1}{4}$ if i is female, $\frac{1}{2}$ if i is male
Father of i	$\frac{1}{4}$	$\frac{1}{2}$ if i is female, 0 if i is male
Full sister of i	$\frac{1}{4}$	$\frac{3}{8}$ if i is female, $\frac{1}{4}$ if i is male
Full brother of i	$\frac{1}{4}$	$\frac{1}{4}$ if i is female, $\frac{1}{2}$ if i is male
i (Self-kinship)	$\frac{1}{2}$	$\frac{1}{2}$ if i is female, 1 if i is male

ϕ_{ij}^X is the X-chromosome kinship coefficient for i and j
ϕ_{ij} is the autosomal kinship coefficient for i and j

Relaxing the Hardy Weinberg Assumption

- We provide another approach to the estimation of the covariance matrix of $\mathbf{Y}^{\mathbf{X}}$.
- In the absence of HWE, we estimate the variance for an outbred female with $\hat{\sigma}_2^2$, where $\hat{\sigma}_2^2$ is the sample variance for outbred females
- For males, the variance is $p(1-p)$ as before, since males only have 1 allele at an X-chromosome marker.

The X_M Association test

- Thornton et al. (*Genet Epidemiol*, 2012) developed the X_M test for X-chromosome association analysis
- X_M, that appropriately adjusts for correlated alleles and genotypes among same and different gender relatives
- The method also accounts for allele copy number differences in females and males.
- Some of the features of the method include:
  - applicable to completely general combinations of family and case-control designs
  - can incorporate different trait prevalence values for females and males
  - allows for both unaffected controls and controls of unknown phenotype to be included in the same analysis.

The X_M Association test

- X_M is analogous to the M_{QLS} test proposed by Thornton and McPeek (*AJHG* 2007)
- Under a dosage compensation trait model, X_M is optimal for 2-allele additive models as the effect size goes to 0.

The X_χ test

- Also developed the X_χ association test for X-linked markers chro
- X_χ is an extension of the test proposed by Clayton D (*Biostatistics* 2008) for samples with related individuals.
- It is analogous to the corrected Pearson's χ^2 statistic of Bourgain et al. (*AJHG* 2004) for association testing for markers on the autosomal chromosomes in related samples.
- Both X_χ and X_M approximately follow a χ^2 distribution with 1 degree of freedom under the null hypothesis.

Simulation Studies

- Compare the power and the type I error of the X_M and X_χ tests.
- Simulated 60 extended pedigrees that consist of 16 individuals (8 males and 8 females) with 3 generations.
  - 20 pedigrees with 4 affecteds, 20 pedigrees with 5 affecteds, and 20 pedigrees with 6 affecteds
  - Modeled a situation in which families are ascertained when there are multiple affecteds
  - An individual (affected or unaffected) from an extended pedigree is included in our simulation study if at least half of the individuals' siblings, parents, and/or offspring in the extended pedigree are affected.
- 200 unaffected unrelated individuals (100 females and 100 males) from the population were used as controls in addition to the related controls

Simulation Studies

- We consider four different classes of multigene X-chromosome dosage compensation trait models.
- Model I has two unlinked causal SNPs on with epistasis between them and both of them acting dominantly.
- Model II also consists of two unlinked causal SNPs on the X-chromosome with epistasis between them, with SNP 1 following a general two-allele model and SNP 2 acting recessively.
- Model III has two unlinked causal SNPs with epistasis between them, with SNP 1 acting recessively and SNP 2 acting dominantly.
- Model IV has three unlinked causal SNPs with epistasis between them and with each SNP acting dominantly.

Simulation Studies: Type I Error

- Simulation studies were conducted under the null hypothesis to verify the accuracy of the χ_1^2 approximation for X_M and X_χ.
- Model I was used to simulated disease status, and then SNPs that were not linked or associated with the causal SNPs were tested for association.

Simulation Studies: Type I Error

Table: Empirical Type I Error of X_M and X_χ, Based on 5,000 Simulated Replicates

	Empirical Type I Error (SE) with Nominal Type I Error of			
	.05 Level		.01 Level	
p	X_M	X_χ	X_M	X_χ
.4	.048 (.003)	.054 (.003)	.010 (.0014)	.011 (.0015)
.2	.045 (.003)	.044 (.003)	.008 (.0013)	.010 (.0014)
.05	.051 (.003)	.050 (.003)	.010 (.0014)	.009 (.0013)

- Association is tested with a bi-allelic X-linked marker having minor allele frequency p.

Simulation Studies: Type I Error

Table: Empirical Type I Error, at Level .0001. Based on 100,000 Simulated Replicates

	Empirical Type I Error: .0001 Level	
p	X_M	X_χ
.4	.00010	.00012
.2	.00015	.00013
.05	.00009	.00015

- Association is tested with a bi-allelic X-linked marker having minor allele frequency p.
- Using an exact binomial calculation, empirical type 1 error rates falling in the range of .00004 to .00016 are not significantly different from the nominal .0001 level.

Simulation Studies: Power

Table: X-Chromosome Multigene Disease Models: Empirical Power and SE for X_χ and X_M. Based on 5,000 Simulated Replicates

Model	SNP1 Allele Frequency	Population Prevalence	Estimated Power (SE)	
			X_M	X_χ
I	.2	.09	.73 (.006)	.63 (.007)
II	.5	.14	.86 (.005)	.87 (.005)
III	.4	.08	.96 (.003)	.93 (.004)
IV	.3	.05	.84 (.005)	.85 (.005)

Application to GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 143 pedigrees with each pedigree containing at least 3 affected individuals.
- Individuals that were labeled as "white, non-Hispanic" were considered.
- 830 cases, 187 unaffected controls, and 13 controls with unknown phenotype.
- Affymetrix 10K Array: 310 SNPs on X-chromosome

Significant Association with the MAOA Gene

- X_M found a significant association with rs979606: $\left(P = 7.65 \cdot 10^{-6}\right.$ uncorrected, .002 corrected$\left.\right)$
- rs979606 is in an intron of the MONOAMINE OXIDASE A (MAOA) gene
- The Monoamine oxidase-A gene is involved with the production of the enzyme monoamine oxidase. The enzyme breaks down chemicals (neurotransmitters) that control mood, aggression, and pleasure.
- From Online Mendelian Inheritance in Man (OMIM): A number of studies have showed that low levels of MAOA expression are associated with
  - antisocial behavior: violent, criminal, or impulsive behavior
  - inefficiency of handling conflict

The MAOA Gene and Antisocial Behavior

- Caspi et al. (*Science* 2002) studied a population of 1,037 children from the New Zealand's Dunedin Multidisciplinary Health and Development Study, which began in 1972, who were followed and assessed from birth through age 26 years.
  - 52% of the children were male
  - For the 8% of the male children that were severely maltreated between the ages of 3 and 11 years, found that those with low MAOA activity were much more likely to develop antisocial behavior, conduct disorder, a disposition toward violent behavior, or conviction for violent offense than were those with high MAOA activity.
  - **Significant gene-environment interaction**: With the absence of abuse, having low MAOA activity did not make boys any more likely to be antisocial

The MAOA Gene association with Alcoholism and Antisocial Behavior

- Ducci et al. (*Molecular Psychiatry* 2008) conducted a study of alcoholism and antisocial personality disorder (ASPD) in a sample of 291 women, 50% of whom experienced child sexual abuse (CSA).
  - Participants included 168 alcoholics (39 with ASPD) and 123 controls (no alcoholics and no ASPD).
  - Low MAOA activity was associated with alcoholism ($P = 0.005$), particularly antisocial alcoholism ($P = 0.00009$), but only among sexually abused subjects.
  - Sexually abused women who were homozygous for the low activity allele had higher rates of alcoholism and ASPD than abused women homozygous for the high activity allele. Heterozygous women displayed an intermediate risk pattern.
  - **Significant gene-environment interaction**: Among non-sexually abused women, there was no relationship between alcoholism/antisocial behavior and MAOA .

The MAOA Gene association with Alcoholism and Antisocial Behavior

- Tikkanen et al. (*Alcohol Clin Exp Res* May 2010)
  - Studied a male sample of 174 impulsive Finnish alcoholic violent offenders
  - Found that MAOA gene expression as well as heavy drinking and childhood abuse were significant predictors of recidivism in violent behavior for these individuals
- Wang et al. (*Progress in Neuro-Psychopharmacology & Biological Psychiatry* 2007)
  - Studied a sample of 231 Han Chinese males in Taiwan, 73 of which were diagnosed with antisocial personality disorder with alcoholism (Antisocial ALC) and 158 subjects diagnosed with antisocial personality disorder without alcoholism (Antisocial Non-ALC).
  - Found a significant antisocial alcoholism association with the interaction of the MAOA and DRD2 genes

The MAOA Gene association with Alcoholism and Antisocial Behavior

- Saito et al. (*Psychiatry Research* 2002)
  - Analyzed a cohort of Finnish males with either type 1 (144 individuals) or type 2 alcoholism (58 individuals)
  - Compared the two alcoholic groups to controls (152).
  - Did not find a significant association with the MAOA gene and alcoholism when comparing each alcoholic group separately with the controls
  - Found a "trend towards significance" when comparing the combined alcoholism groups to the controls

References

- Clayton D (2008) Testing for association on the X chromosome *Biostatistics* **9**, 593-600.
- Thornton T, Zhang Q, Cai X, Ober C, and McPeek MS (2012) XM: Association Testing on the X-Chromosome in Case-Control Samples with Related Individual. *Genet Epidemiol* **36**, 438-450.

Population Structure

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 9

Nonrandom Mating

- HWE assumes that mating is random in the population
- Most natural populations deviate in some way from random mating
- There are various ways in which a species might deviate from random mating
- We will focus on the two most common departures from random mating:
  - inbreeding
  - population subdivision or substructure

Nonrandom Mating: Inbreeding

- Inbreeding occurs when individuals are more likely to mate with relatives than with randomly chosen individuals in the population
- Increases the probability that offspring are homozygous, and as a result the number of homozygous individuals at genetic markers in a population is increased
- Increase in homozygosity can lead to lower fitness in some species
- Increase in homozygosity can have a detrimental effect: For some species the decrease in fitness is dramatic with complete infertility or inviability after only a few generations of brother-sister mating

Nonrandom Mating: Population Subdivision

- For subdivided populations, individuals will appear to be inbred due to more homozygotes than expected under the assumption of random mating.
- Wahlund Effect: Reduction in observed heterozygosity (increased homozygosity) because of pooling discrete subpopulations with different allele frequencies that do not interbreed as a single randomly mating unit.

Wright's F Statistics

- Sewall Wright invented a set of measures called F statistics for departures from HWE for subdivided populations.
- F stands for fixation index, where fixation being increased homozygosity
- F_{IS} is also known as the inbreeding coefficient.
  - The correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (**I**ndividual within the **S**ubpopulation)
- F_{ST} is a measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations
  - Is defined as the correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (**S**ubpopulation within the **T**otal population).

Wright's F Statistics

- F_{IT} is not often used. It is the overall inbreeding coefficient of an individual relative to the total population (**I**ndividual within the **T**otal population).

Genotype Frequencies for Inbred Individuals

- Consider a bi-allelic genetic marker with alleles A and a. Let p be the frequency of allele A and $q = 1 - p$ the frequency of allele a in the population.
- Consider an individual with inbreeding coefficient F. What are the genotype frequencies for this individual at the marker?

Genotype	AA	Aa	aa
Frequency			

Generalized Hardy-Weinberg Deviations

- The table below gives genotype frequencies at a marker for when the HWE assumption does not hold:

Genotype	AA	Aa	aa
Frequency	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

  where $q = 1 - p$
- The F parameter describes the deviation of the genotype frequencies from the HWE frequencies.
- When $F = 0$, the genotype frequencies are in HWE.
- The parameters p and F are sufficient to describe genotype frequencies at a single locus with two alleles.

F_{st} for Subpopulations

- Example in Gillespie (2004)
- Consider a population with two equal sized subpopulations. Assume that there is random mating within each subpoulation.
- Let $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$
- Below is a table with genotype frequencies

Genotype	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$

- Are the subpopulations in HWE?
- What are the genotype frequencies for the entire population?
- What should the genotypic frequencies be if the population is in HWE at the marker?

F_{st} for Subpopulations

- Fill in the table below. Are there too many homozygotes in this population?

	Allele	Genotype		
	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
Freq. Population				
Hardy-Weinberg Frequencies				

- To obtain a measure of the excess in homozygosity from what we would expect under HWE, solve

$$2pq(1 - F_{ST}) = \frac{3}{8}$$

- What is F_{st}?

F_{st} for Subpopulations

- The excess homozygosity requires that $F_{ST} =$ _____
- For the previous example the allele frequency distribution for the two subpopulations is given.
- At the population level, it is often difficult to determine whether excess homozygosity in a population is due to inbreeding, to subpopulations, or other causes.
- European populations with relatively subtle population structure typically have an F_{st} value around .01 (e.g., ancestry from northwest and southeast Europe),
- F_{st} values that range from 0.1 to 0.3 have been observed for the most divergent populations (Cavalli-Sforza et al. 1994).

F_{st} for Subpopulations

- Nelis et al. (PLOS One, 2009) looked at the genetic structure for various populations
- Obtained pairwise F_{st} values for the four HapMap sample populations
  - Europeans (CEU) - Africans (YRI): **0.153**
  - Europeans (CEU) - Japanese (JPT): **0.111**
  - Europeans (CEU) - Chinese (CHB): **0.110**
  - Africans (YRI) - Chinese (CHB): **0.190**
  - Africans (YRI) - Japanese (JPT): **0.192**
  - Chinese (CHB) - Japanese (JPT): **0.007**

F_{st} for Subpopulations

- F_{st} can be generalized to populations with an arbitrary number of subpopulations.
- The idea is to find an expression for F_{st} in terms of the allele frequencies in the subpopulations and the relative sizes of the subpopulations.
- Consider a single population and let r be the number of subpopulations.
- Let p be the frequency of the A allele in the population, and let p_i be the frequency of A in subpopulation i, where $i = 1, \ldots, r$
- F_{st} is often defined as $F_{st} = \frac{\sigma_p^2}{p(1-p)}$, where σ_p^2 is the variance of the p_i's with $E(p_i) = p$.

F_{st} for Subpopulations

- Let the relative contribution of subpopulation i be c_i, where $\sum_{i=1}^{r} c_i = 1$.

Genotype	AA	Aa	aa
Freq. Subpop$_i$	p_i^2	$2p_iq_i$	q_i^2
Freq. Population	$\sum_{i=1}^{r} c_i p_i^2$	$\sum_{i=1}^{r} c_i 2p_iq_i$	$\sum_{i=1}^{r} c_i q_i^2$

where $q_i = 1 - p_i$

- In the population, we want to find the value F_{st} such that $2pq(1 - F_{st}) = \sum_{i=1}^{r} c_i 2p_iq_i$
- Rearranging terms:

$$F_{st} = \frac{2pq - \sum_{i=1}^{r} c_i 2p_iq_i}{2pq}$$

- Now $2pq = 1 - p^2 - q^2$ and $\sum_{i=1}^{r} c_i 2p_iq_i = 1 - \sum_{i=1}^{r} c_i(p_i^2 + q_i^2)$

F_{st} for Subpopulations

- So can show that

$$F_{st} = \frac{\sum_{i=1}^{r} c_i(p_i^2 + q_i^2) - p^2 - q^2}{2pq}$$

$$= \frac{\left[\sum_{i=1}^{r} c_i p_i^2 - p^2\right] + \left[\sum_{i=1}^{r} c_i q_i^2 - q^2\right]}{2pq}$$

$$= \frac{Var(p_i) + Var(q_i)}{2pq}$$

$$= \frac{2Var(p_i)}{2p(1-p)}$$

$$= \frac{Var(p_i)}{p(1-p)}$$

$$= \frac{\sigma_p^2}{p(1-p)}$$

Estimating F_{st}

- Let n be the total number of sampled individuals from the population and let n_i be the number of sampled individuals from subpopulation i
- Let $\hat{p}_i$ be the allele frequency estimate of the A allele for the sample from subpopulation i
- Let $\hat{p} = \sum_i \frac{n_i}{n} \hat{p}_i$
- A simple F_{st} estimate is $\hat{F}_{ST_1} = \frac{s^2}{\hat{p}(1-\hat{p})}$, where s^2 is the sample variance of the $\hat{p}_i$'s.

Estimating F_{st}

- Weir and Cockerman (1984) developed an estimate based on the method of moments.

$$MSA = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\hat{p}_i - \hat{p})^2$$

$$MSW = \frac{1}{\sum_i (n_i - 1)}\sum_{i=1}^{r} n_i\hat{p}_i(1-\hat{p}_i)$$

- Their estimate is

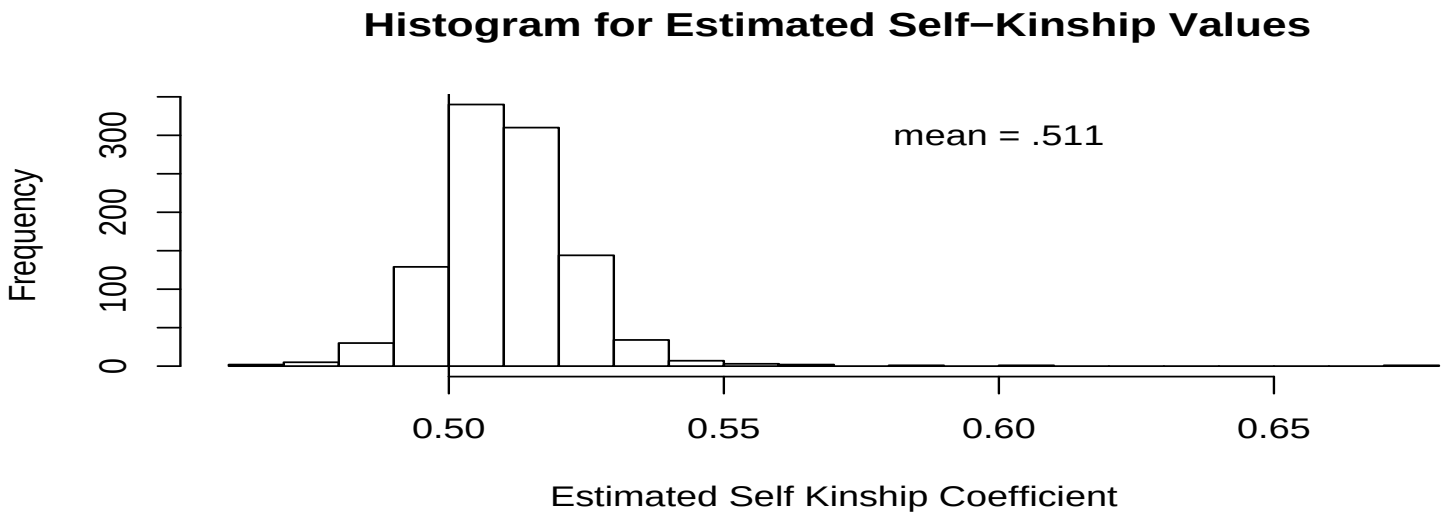
$$\hat{F}_{ST_2} = \frac{MSA - MSW}{MSA + (n_c - 1)MSW}$$

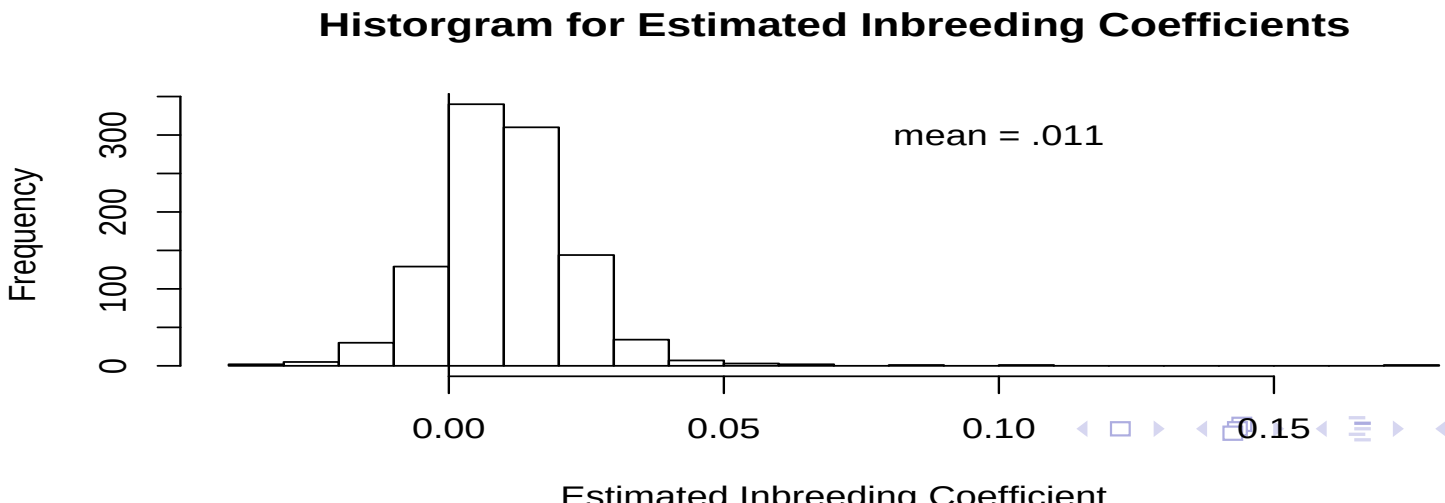
where $n_c = \sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$

GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals.
- Individuals labeled as white, non-Hispanic were considered.
- Estimated self-kinship and inbreeding coefficients using genome-screen data

COGA Data


Histogram for Estimated Self−Kinship Values
mean = .511
Frequency
0
100
200
300
0.50
0.55
0.60
0.65
Estimated Self Kinship Coefficient


Historgram for Estimated Inbreeding Coefficients
mean = .011
Frequency
0
100
200
300
0.00
0.05
0.10
0.15
Estimated Inbreeding Coefficient

References

- Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, et al. (2009) Genetic Structure of Europeans: A View from the NorthEast. *PLoS ONE* **4**, e5472. doi:10.1371/journal.pone.0005472.
- Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Methods for Cryptic Structure

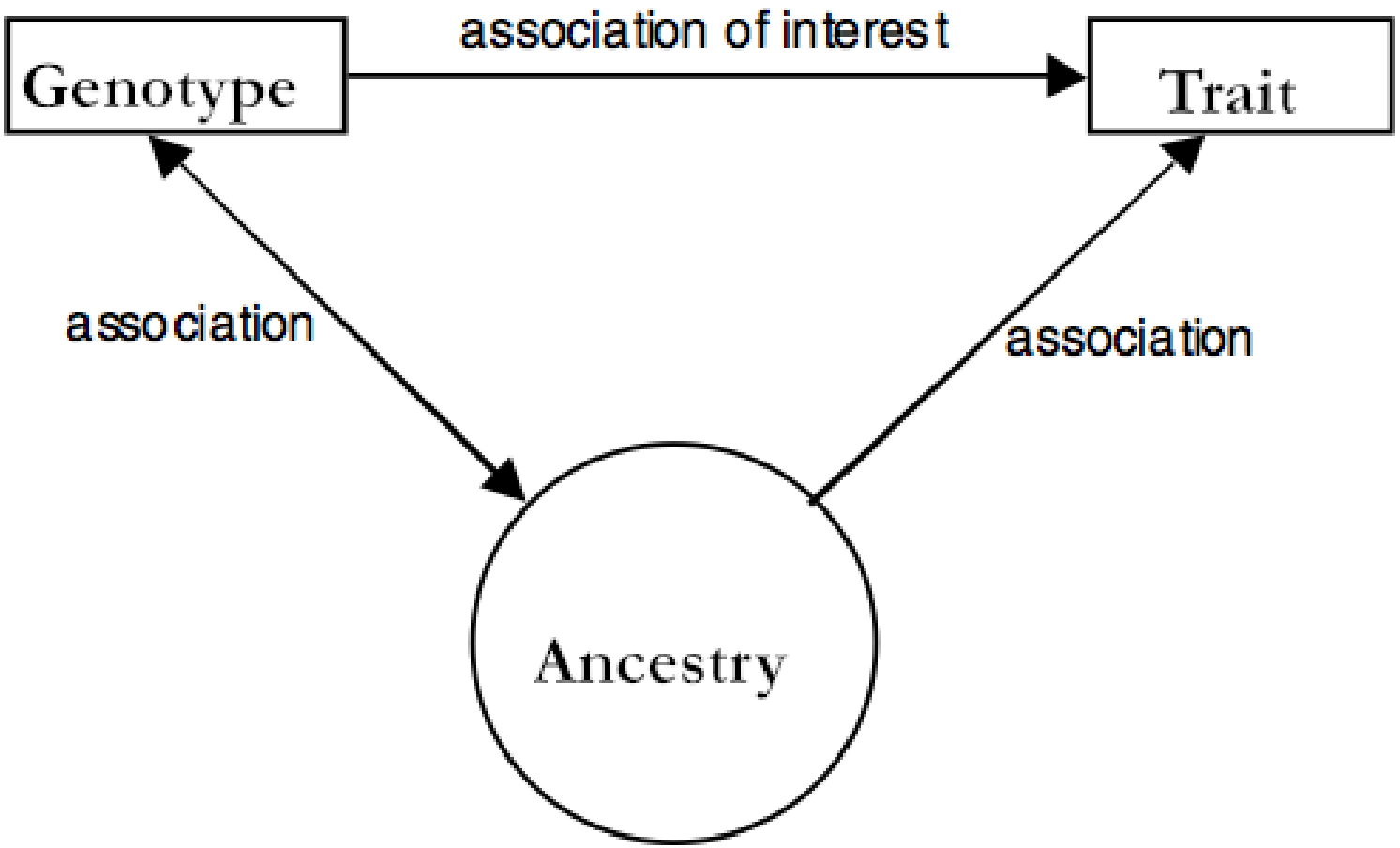
Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 10

Population Structure and Association Testing

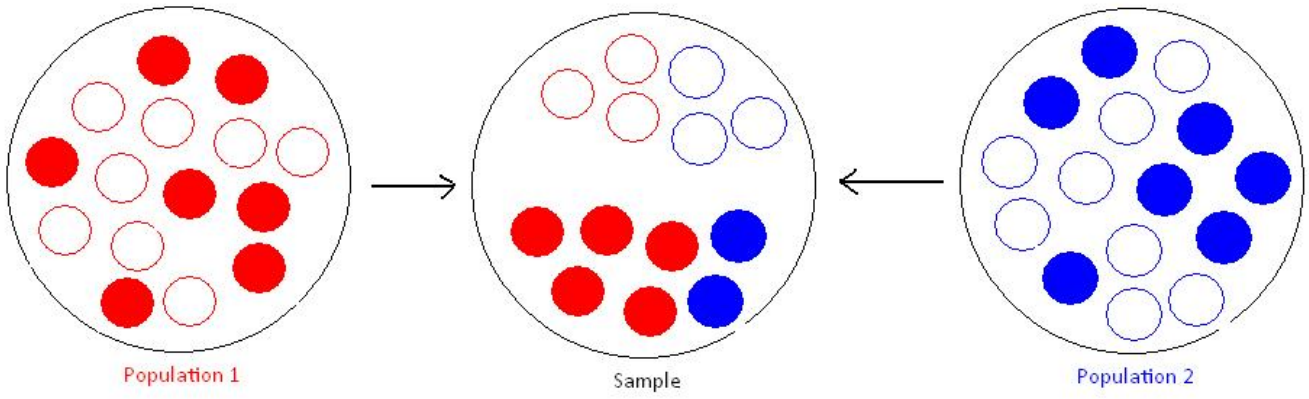
- The observations in genome-wide association studies can have several sources of dependence.
- Population structure, the presence of subgroups in the population with ancestry differences or admixed ancestry, is a major concern for association studies
- Population structure has long been recognized as a confounding factor in genetic association studies.
- Heterogeneous genomes of sample individuals can lead to both spurious association and reduced power if not properly accounted for

Confounding due to Ancestry



Spurious Association

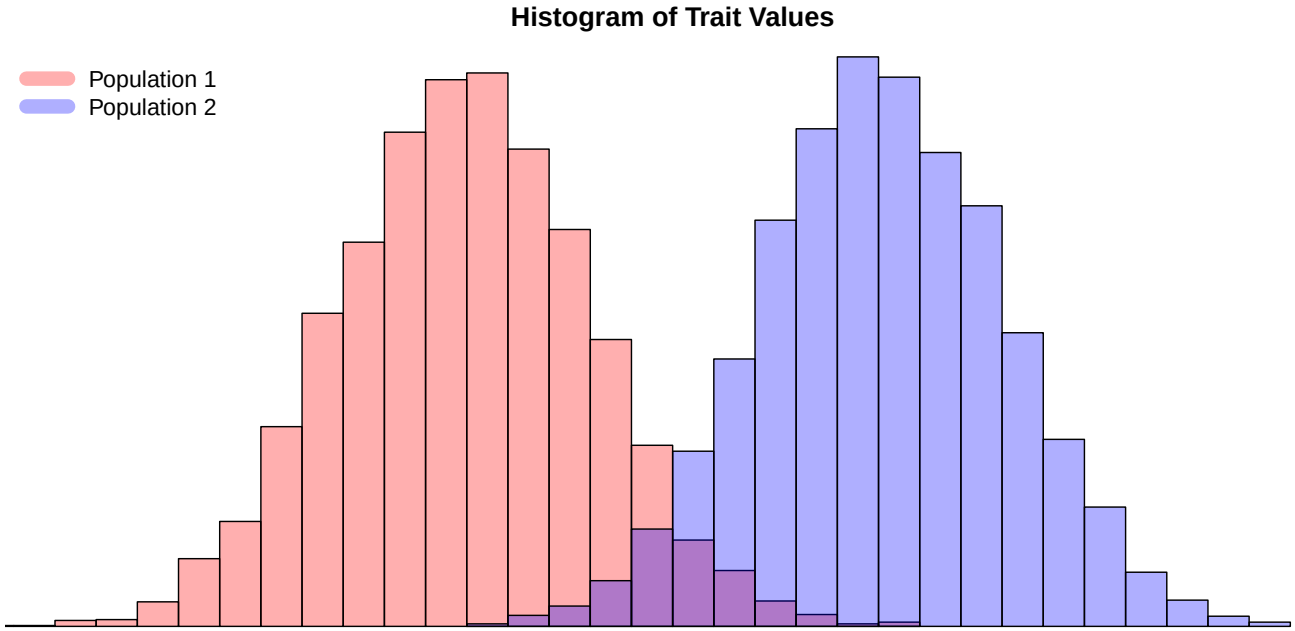
- Case/Control association test
  - Comparison of allele frequency between cases and controls.
- Consider a sample from 2 populations:



  - Red population overrepresented among cases in the sample.
  - Genetic markers that are not influencing the disease but with significant differences in allele frequencies between the populations
    $\Longrightarrow$ spurious association between disease and genetic marker

Spurious Association

- Quantitative trait association test
  - Test for association between genotype and trait value
- Consider sampling from 2 populations:



- Blue population has higher trait values.
- Different allele frequency in each population
  $\Longrightarrow$ spurious association between trait and genetic marker for samples containing individuals from both populations

Balding-Nichols Model

- A model that is often used for population structure is the Balding-Nichols model (Balding and Nichols, 1995).
- Consider unrelated outbred individuals that are sampled from a population with K subpopulations, i.e., the subpopulations are $1, 2, \ldots, K$.
- Assume that an individual can be a member of only one subpopulation, i.e., there is no admixture.
- Under the Balding-Nichols model, the allele frequency for subpopulation k, where $1 \leq k \leq K$, is a random draw from a beta distribution with parameters $p(1 - F_{st_k})/F_{st_k}$ and $(1 - p)(1 - F_{st_k})/F_{st_k}$, where $0 < p < 1$
- The parameter p can be viewed as the ancestral allele frequency and F_{st_k} can be viewed as Wright's standardized measure of variation for subpopulation k.

Balding-Nichols Model: Covariance Structure

- Consider a single bi-allelic marker (e.g. a SNP) with allele labels "0" and "1"
- Let N be the number of sampled individuals with genotype data at the marker.
- Let $X = (X_1, \ldots X_N)$ where X_i =the number of alleles of type 1 in individual i, so the value of X_i is 0, 1, or 2.
- Under the Balding-Nichols model:
  - Individual i has inbreeding coefficient equal to F_{st}
  - If individuals i and j are are both from the same subpopulation k, then $Corr(X_i, X_j) = F_{st_k}$
  - If i and j are from different subpopulations then $Corr(X_i, X_j) = 0$
- The F_{st_k} values, the number of subpopulations K, and the subpopulation memberships for the sample individuals will generally be unknown when there is cryptic population structure.

Balding-Nichols Model: Covariance Structure

If there is no structure then the covariance matrix of X will be a function of the identity matrix:

$$\mathbf{I_0} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

If there is structure then the covariance matrix of X will be a function of :

$$\mathbf{\Sigma_0} = \begin{pmatrix} 1 + F_{st_1} & F_{st_1} & \dots & 0 \\ F_{st_1} & 1 + F_{st_1} & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 + F_{st_K} \end{pmatrix},$$

Case-Control Design

- Methods have been proposed to correct for cryptic population structure in case-control studies by using data across the genome.
- Let N be the number of individuals in the study.
- Let $\mathbf{Y} = (Y_1, \ldots Y_N)$ be a phenotype indicator vector for case control status where $Y_i = 1$ if i is a case and $Y_i = 0$ if i is a control
- Let M be the number of bi-allelic markers (e.g. SNPs) in the data. Consider a marker s, where $1 \leq s \leq M$, and let $\mathbf{X}_s = (X_{1_s}, \ldots X_{N_s})$ where X_{i_s} =the number of alleles of type 1 in individual i at marker s.

Genomic Control

- Devlin and Roeder (1999) proposed correcting for substructure via a method called "genomic control."
- For each marker s, the Armitage trend statistic is calculated

$$A_{r_s} = N r^2_{X_s Y}$$

  where $r^2_{X_s Y}$ is the squared correlation between the genotype variable $\mathbf{X}_s$ for marker s and the binary phenotype variable $\mathbf{Y}$.
- If there is no population structure, the distribution of A_{r_s} will approximately follow a χ^2 distribution with 1 degree of freedom.
- If there is population structure, the statistic will deviate from a χ^2_1 distribution due to an inflated variance.

Genomic Control

- Use $\lambda = \frac{median(A_{r_1},\ldots,A_{r_s},\ldots A_{r_M})}{.456}$ as a correction factor for cryptic structure, where .456 is the median of a χ_1^2 distribution.
- The uniform inflation factor λ is then applied to the Armitage trend statistic values
$$\tilde{A}_{r_s} = \frac{A_{r_s}}{\lambda}$$
- $\tilde{A}_{r_s}$ will approximately follow a χ^2 distribution with 1 degree of freedom.

Genomic Control

- Another way to view genomic control is as follows:

$$A_{r_s} = \frac{T^2}{Var_0(T)}$$

  where T is a measure of allele frequency differences between cases and controls and $Var_0(T)$ is the variance of T under the null hypothesis
- For the Armitage statistic, $Var_0(T)$ is calculated assuming individuals are unrelated (calculation based on the identity matrix).
- Genomic control inflates this variance to account for the structure (unknown F_{st} values)

$$\tilde{A}_{r_s} = \frac{T^2}{\lambda Var_0(T)}$$

Identifying Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA is method for calculating the **principal components** that explain differences among the sample individuals in the genetic data
- To perform a PCA for population structure inference, first calculate an empirical covariance matrix $\hat{\boldsymbol{\Psi}}$ with components $\hat{\psi}_{ij}$:

$$\hat{\psi}_{ij} = \frac{1}{M} \sum_{s=1}^{M} \frac{(X_{is} - 2\hat{p}_s)(X_{js} - 2\hat{p}_s)}{\hat{p}_s(1 - \hat{p}_s)}$$

where $\hat{p}_s$ is an allele frequency estimate for the type 1 allele at marker s

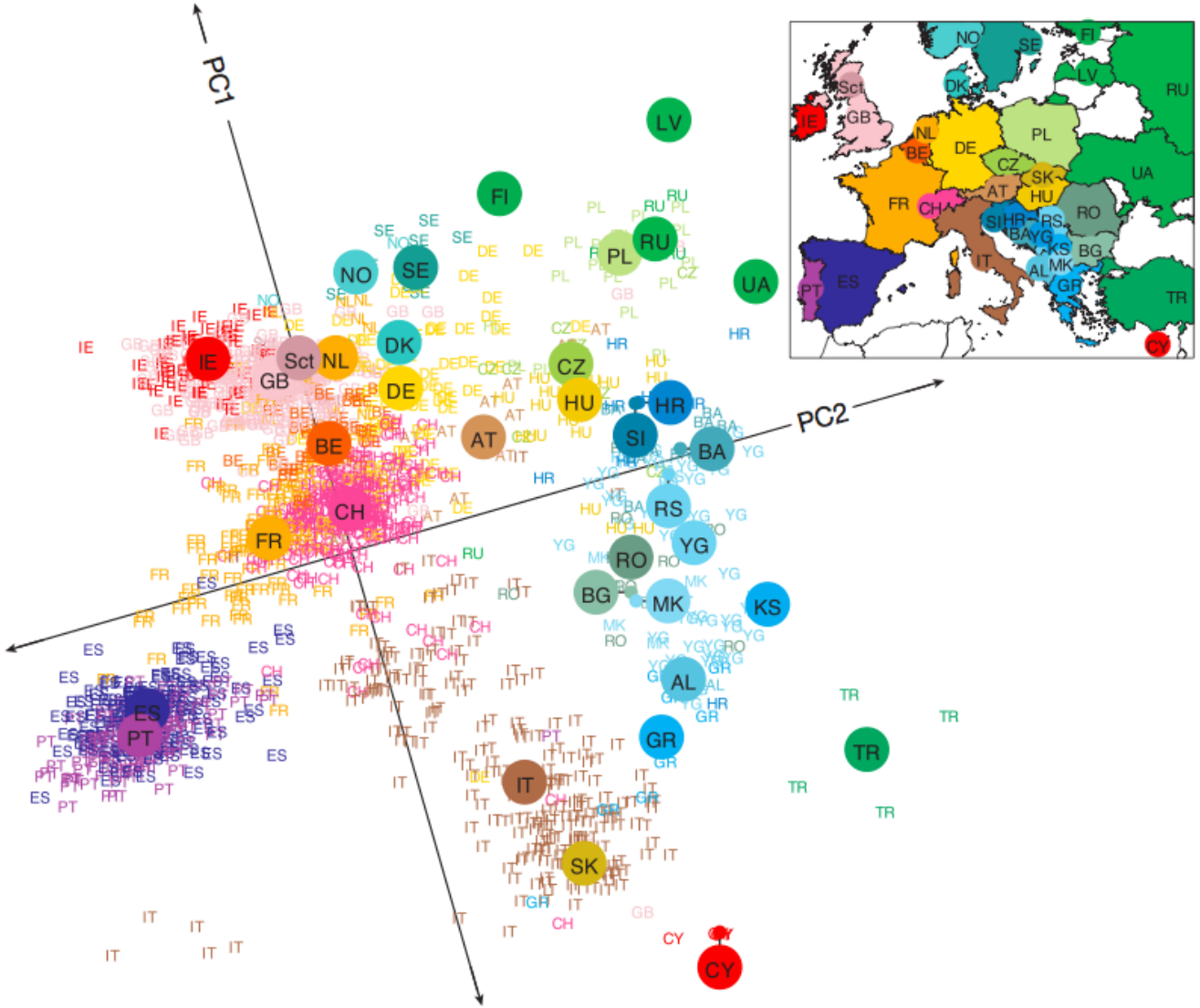
Identifying Population Structure with PCA

- Principal components (eigenvectors) for $\hat{\mathbf{\Psi}}$ can then be obtained via a singular value decomposition (SVD).
- The top principal components are viewed as continuous axes of variation that reflect subpopulation genetic variation in the sample.
- Individuals with "similar" values for a particular top principal component will have "similar" ancestry for that axes.
- Does PCA actually work in practice?

PCA of Europeans

- An application of principal components to genetic data (Novembre et al. 2008) showed that among Europeans for whom all four grandparents originated in the same country, the first two principal components computed using 200,000 SNPs could map their country of origin quite accurately in the plane

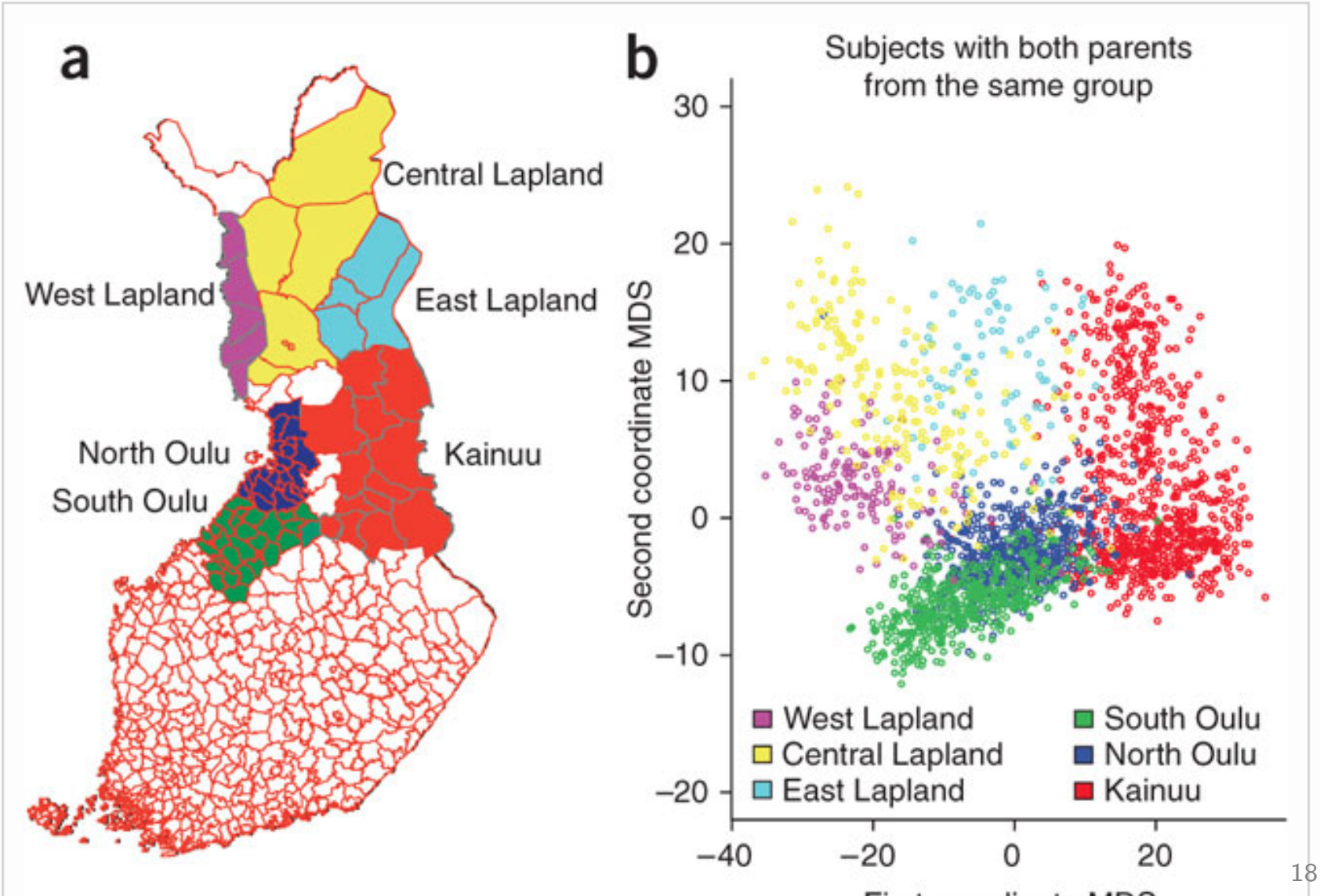
PCA of Europeans



PCA in Finland

- There can be population structure in all populations, even those that appear to be relatively "homogenous"
- An application of principal components to genetic data from Finland samples (Sabatti et al., 2008) identified population structure that corresponded very well to geographic regions in this country.

PCA in Finland



Principal Components Analysis: Association Testing

- Price et al. (2006) proposed corrected for structure in genetic association studies by using PCA
- They developed a method called EIGENSTRAT for association testing in structured populations where the top principal components (highest eigenvalues) from a PCA can be used as covariates in a multi-linear regression.

$$Y = \beta_0 + \beta_1 X + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \cdots + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

Computation of Eigenstrat

- The kth axis of variation is defined to be the kth eigenvector of $\hat{\mathbf{\Psi}}$ (that is, the eigenvector with kth largest eigenvalue).
- Thus, the ancestry a_{jk} of individual j along the kth axis of variation equals coordinate j of the kth eigenvector.
- Adjustments of genotypes for ancestry
- Let X_{js} be the genotype of individual j ($X_{js} = 0$, 1 or 2) at marker s, and let a_j be the ancestry of individual j along a given axis of variation.
- $X_{js}^{adjusted} = X_{js} - \gamma_s a_j$, where $\gamma_s = \sum_i a_i X_{is} / \sum_i a_i^2$
- γ_s is a regression coefficient for ancestry predicting genotype across individuals j with valid genotypes at marker s.
- A similar adjustment is performed for each axis of variation.

Computation of Eigenstrat

- Adjustments of phenotypes for ancestry
- $Y_j = 1$ if j is a case and $Y_j = 0$ if j is a control
- $Y_j^{adjusted} = Y_j - \delta a_i$, where $\delta = \sum_i a_i Y_i / \sum_i a_i^2$
- The EIGENSTRAT statistic adjusted for ancestry at marker s is

$$EIG_s = (N - K - 1) \times \left[corr \left(Y_s^{adjusted}, X^{adjusted} \right) \right]^2$$

  where N is the number of subjects in the sample and K is the number of axes of variation used to adjust for ancestry.
- EIG_s will approximately follow a χ^2 distribution with 1 degree of freedom.

Structured Samples with Related Individuals

- The methods discussed so far have been developed for samples with unrelated individuals
- Methods may not be valid in samples with related individuals (known and/or unknown)
- Many genetic studies have samples with related individuals
- Consider testing for association between a trait and a single genetic marker in a sample that includes related individuals.
- Assume that there are some related individuals in the sample from a structured population.
- If individuals i and j are from population k and they are related, the correlation, however, is no longer F_{st_k}!
- The correlation for the pair is now ψ_{ij}, where ψ_{ij} is a function of both the kinship coefficient for individuals i and j and F_{st_k}.

ROADTRIPS

- The ROADTRIPS approach of Thornton and McPeek (2010) incorporates an empirical covariance matrix $\hat{\boldsymbol{\Psi}}$.
- The top principal components in EIGENSTRAT are not able to capture the complicated covariance structure due to the related individuals in the samples.
- Instead of taking the top principal components of the matrix $\hat{\boldsymbol{\Psi}}$, ROADTRIPS uses the entire matrix to correct the variance of a general class of statistics for unknown structure
- ROADTRIPS extensions have been developed for a number of association tests including Pearson χ^2 test, the Armitage trend test, the corrected χ^2 test, the W_{QLS} test, and the M_{QLS} test.
- ROADTRIPS is a valid association method for partially or completely unknown population and pedigree structure.

EMMAX: Linear Mixed Effects Models For Structure

- Kang et al. [Nature Genet, 2010] proposed the EMMAX variance components method:

$$Y = \beta_0 + \beta_1 X + \mathbf{g} + \boldsymbol{\epsilon},$$

$$\text{with } \mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \hat{\boldsymbol{\Psi}}) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$$

- The variance components, σ_g^2 and σ_e^2, are then estimated based on this model using either a maximum likelihood or restricted maximum likelihood (REML) approach.
- The null hypothesis of no association is $H_0 : \beta_1 = 0$, and the alternative hypothesis is : $H_A : \beta_1 \neq 0$
- The EMMAX statistic is the score statistic for testing the null hypothesis of $\beta_1 = 0$.
- Similar mixed effects methods proposed by Lippert et al. (2011, Nat Methods) and Stephens et al. (2012, Nat Genet).

Structured Association Method

- Pritchard et al. (2000) developed the program STRUCTURE
- This method uses MCMC to cluster genetically similar individuals into "subpopulations."
- Association testing is performed within each of the subpopulations
- Assignments of individuals are highly sensitive to the number of subpopulations K in the sample, which is often unknown when there is cryptic structure
- Not computationally feasible for genome-wide association studies

Structured Association Method

- Let K be the number of populations, Let N be the number of individuals in the sample. We first consider the case when K is known.
- We assume that each individual originates in one of the K populations, i.e., there is no admixture.
- Let L denote the number of loci and let the vector X denote the genotypes of the loci for the individuals.
- Let the vector Z be the populations of origin for the individuals. Let P be the allele frequencies for the populations. The vectors Z and P are both unknown

Structured Association Method

- The elements of the vectors are as follows:

$$
\begin{aligned}
(x_l^{(i,1)}, x_l^{(i,2)}) &= \text{ genotype of ith individual at the lth locus,} \\
&\quad \text{where } i = 1, 2, ..., N \text{ and } l = 1, 2, ..., L \\
z_i &= \text{ population from which ith individual originated} \\
p_{klj} &= \text{ frequency of allele j at lth locus in populaion k} \\
&\quad \text{where } k = 1, 2, ..., K \text{ and } j = 1, 2, ..., J_l
\end{aligned}
$$

- Now, given the allele frequencies and the population of origins, we have that $x_l^{(i,a)}$ are independent and $Pr(x_l^{(i,a)} = j | Z, P) = p_{z_i lj}$.

Prior Distributions

- For the prior distribution of Z, we have that the z_i's are independent and each one follows a discrete uniform distribution where $Pr(z_i = k) = \frac{1}{K}$ for $k = 1, 2, ..., K$.
- This prior seems reasonable since we have no information on the population of origin before observing any genotype information
- For population k at locus l we have have that the prior distribution of $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l}) \sim D(\lambda_1, \lambda_1, ..., \lambda_{J_l})$ where D is a Dirichlet distribution.
- Before observing any genotype data we have no information about the allele frequency distributions. To account for our ignorance of this, we make $\lambda_1 = \lambda_2 = ... = \lambda_{J_l} = 1$.

The Gibbs Sampling MCMC Algorithm

- Let x^i denote the observed genotype data for individual i for all loci. Now

$$
\begin{aligned}
Pr(z_i = k|X, P) &= \frac{Pr(x^i|z_i = k, P)Pr(z_i = k|P)Pr(P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)Pr(z_i = k'|P)Pr(P)} \\
&= \frac{Pr(x^i|z_i = k, P)\frac{1}{K}Pr(P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)\frac{1}{K}Pr(P)} \\
&= \frac{Pr(x^i|z_i = k, P)}{\sum_{k'=1}^{K} Pr(x^i|z_i = k', P)}
\end{aligned}
$$

Note that $Pr(x_l^{(i,a)} = j|Z, P) = p_{z_i lj}$, for individual i, population z_i, locus l, and allele j. So we can easily obtain likelihood of the genotype data of individual i since $Pr(x^i|z_i = k, P) = \prod_{l=1}^{L} p_{klx_l^{(i,1)}} p_{klx_l^{(i,2)}}$

The Gibbs Sampling MCMC Algorithm

- We will now give the full condition for P. We have that $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l})$ and

$$Pr(p_{kl}|X, Z) \quad \propto \quad p_{kl1}^{\lambda_1 + n_{kl1}} \cdot p_{kl2}^{\lambda_2 + n_{kl2}} \cdots p_{klJ_l}^{\lambda_{J_l} + n_{klJ_l}}$$

  where n_{klj} is the number of copies of allele j found at locus l in individuals that are assigned to population k. So, $p_{kl} = (p_{kl1}, p_{kl2}, ..., p_{klJ_l})$ is a $D(\lambda_1 + n_{kl1}, \lambda_1 + n_{kl2}, ..., \lambda_{J_l} + n_{klJ_l})$

The Gibbs Sampling MCMC Algorithm

- Now that we have the full conditionals, we can do implement the Gibbs sampler.
- Step 0: Start with an initial value of Z by randomly drawing from the discrete uniform distribution $\frac{1}{k}$ for each component. Denote this initial value as Z^0.
- Step 1: Draw P^n from $Pr(P|X, Z^{n-1})$
- Step 2: Draw Z^n from $Pr(Z|X, P^n)$
- For n=1,2,..., repeat Step 1 and Step 2.

The Gibbs Sampling MCMC Algorithm: Example

- Two random mating populations were simulated. Each individual in the sample comes are from one of the two populations.
- Prichard et al. (2000) performed simulations under this model for 5 loci and for 15 loci.
  - The number of individuals in their sample was is over 100.
  - Algorithm performs very well in clustering individuals for the 5 loci case and is almost perfect for the 15 loci case
- Assigning individuals to originating populations depend on a number of factors which include the number of individuals in the sample which in turn affects the accuracy of P, the number of loci, and the allele frequency within the populations.
- Interest in obtaining clustering results if the number of loci is decreased to 4 and the sample size is reduced to 20 (9 from one population 1 and 11 from population 2).

The Gibbs Sampling MCMC Algorithm: Example

Simulated from the following two populations:

Allele Frequencies at Each Locus for The Two Populations

	Locus 1	Locus 2	Locus 3	Locus 4
Population 1	Allele 1: .90 Allele 2: .10	Allele 1: .20 Allele 2: .80	Allele 1: .10 Allele 2: .20 Allele 3: .70	Allele 1: .50 Allele 2: .30 Allele 3: .20
Population 2	Allele 1: .10 Allele 2: .90	Allele 1: .80 Allele 2: .20	Allele 1: .70 Allele 2: .20 Allele 3: .10	Allele 1: .20 Allele 2: .10 Allele 3: .70

The Gibbs Sampling MCMC Algorithm: Example

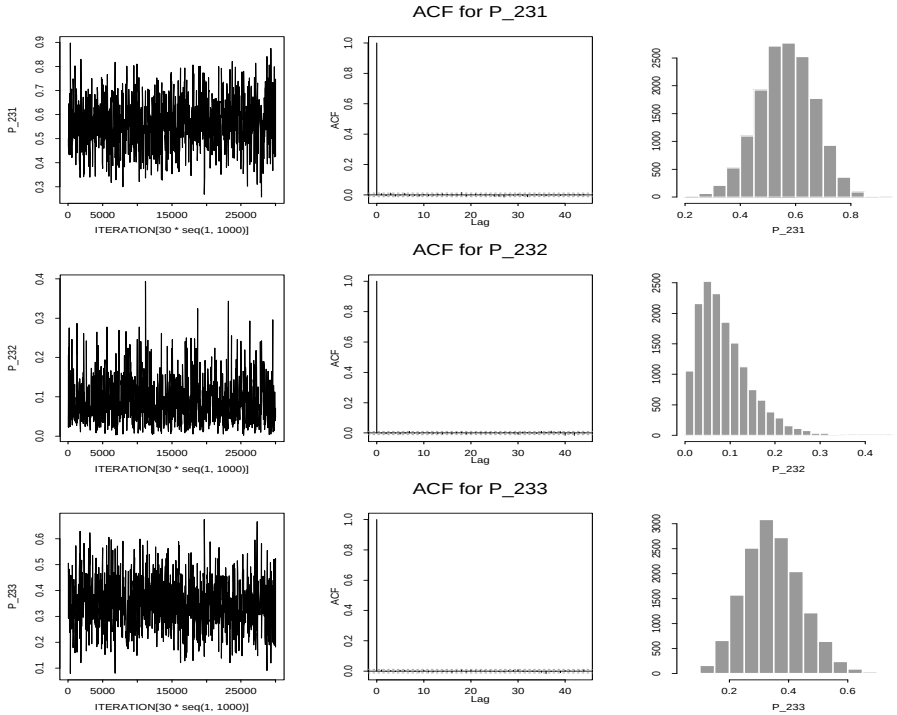
Genotypes of Sample Individuals at Each Locus

	Locus 1	Locus 2	Locus 3	Locus 4
Person 1	2, 2	1, 2	1, 1	3, 3
Person 2	2, 2	1, 1	1, 1	3, 3
Person 3	2, 2	2, 2	2, 3	1, 1
Person 4	2, 1	1, 1	1, 1	2, 3
Person 5	2, 2	1, 2	1, 1	1, 1
Person 6	2, 2	2, 1	1, 1	3, 1
Person 7	1, 1	1, 2	2, 3	3, 1
Person 8	1, 1	2, 2	1, 3	3, 1
Person 9	1, 1	2, 2	3, 1	3, 1
Person 10	1, 1	2, 1	2, 2	2, 3
Person 11	2, 2	1, 2	1, 1	3, 2
Person 12	2, 2	1, 1	2, 1	1, 3
Person 13	1, 1	2, 1	3, 3	3, 1
Person 14	2, 1	1, 2	1, 1	3, 1
Person 15	1, 1	2, 2	3, 3	3, 2
Person 16	1, 1	2, 2	3, 3	1, 1
Person 17	1, 2	1, 1	1, 1	1, 1
Person 18	2, 2	1, 1	1, 3	3, 3
Person 19	1, 1	2, 2	3, 2	2, 1
Person 20	2, 2	1, 1	3, 2	2, 3

Individuals from population 1: 3, 7, 8, 9, 10, 13, 15, 16, and 19.
Individuals from population 2: 1, 2, 4, 5, 6, 11, 12, 14, 17, 18, and 20.

The Gibbs Sampling MCMC Algorithm: Example

Below are plots for the allele frequency parameters for population 2 at locus 3. The histograms are from samples following the 15,000th iteration (30,000 total iterations).



The Gibbs Sampling MCMC Algorithm: Example

Marginal Distribution of Z

	TRUE POPULATION	Pr(Z=1)	Pr(Z=2)	Correct Infer.
Person 1	2	0.1202667	0.8797333	*YES*
Person 2	2	0.1287333	0.8712667	*YES*
Person 3	1	0.3936667	0.6063333	*NO*
Person 4	2	0.4176667	0.5823333	*YES*
Person 5	2	0.06886667	0.9311333	*YES*
Person 6	2	0.0774	0.9226	*YES*
Person 7	1	0.7792667	0.2207333	*YES*
Person 8	1	0.5047333	0.4952667	*UNSURE*
Person 9	1	0.4913333	0.5086667	*UNSURE*
Person 10	1	0.9305333	0.06946667	*YES*
Person 11	2	0.2615333	0.7384667	*YES*
Person 12	2	0.2529333	0.7470667	*YES*
Person 13	1	0.7080667	0.2919333	*YES*
Person 14	2	0.07993333	0.9200667	*YES*
Person 15	1	0.8038667	0.1961333	*YES*
Person 16	1	0.63453337	0.3654667	*YES*
Person 17	2	0.1385333	0.8614667	*YES*
Person 18	2	0.1385333	0.746	*YES*
Person 19	1	0.8957333	0.1042667	*YES*
Person 20	2	0.7332667	0.2667333	*NO*

The posterior marginal distribution of Z was obtained and used to determine whether the individuals were properly classified. An individual is considered to be properly classified if the mode of the

References

- Balding DJ, Nichols RA (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. *Genetica* **96**, 3-12.
- Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909.

References

- Pritchard J, Stephens M, Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.Y., Freimer, N. B., Sabatti, C. Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-354.
- Thornton T, McPeek MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.

Haploview

www.broad.mit.edu/mpg/haploview/

Software Demo

Summer Institute in Statistical Genetics 2014
Module 10
Topic 11

Useful for…

- Visualizing LD and “haplotype blocks”
- Haplotype population frequency estimation

- Other possible uses:
  - Simple association tests
  - Tag SNP selection
  - Visualizing GWAS results from PLINK

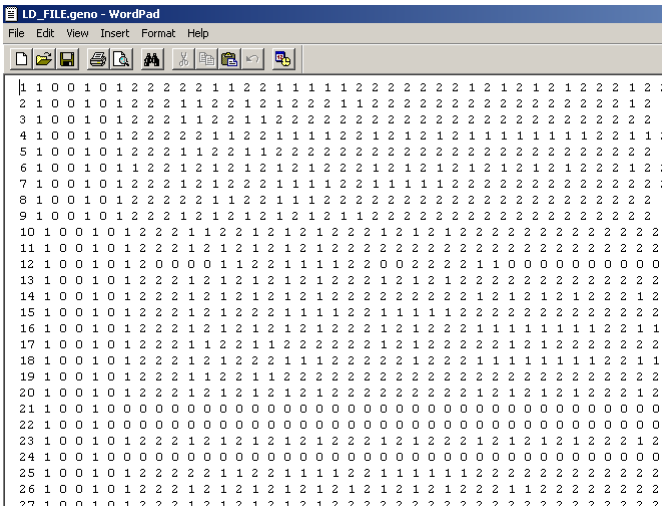
“Linkage” format data

- Pedigree Name: A unique alphanumeric identifier for this individual's family. Unrelated individuals should not share a pedigree name.
- Individual ID: An alphanumeric identifier for this individual. Should be unique within his family (see above).
- Father's ID: Identifier corresponding to father's individual ID or "0" if unknown father. Note that if a father ID is specified, the father must also appear in the file.
- Mother's ID: Identifier corresponding to mother's individual ID or "0" if unknown mother Note that if a mother ID is specified, the mother must also appear in the file.
- Sex: Individual's gender (1=MALE, 2=FEMALE).
- Affection status: Affection status to be used for association tests (0=UNKNOWN, 1=UNAFFECTED,2=AFFECTED).
- Marker genotypes


```
3   12   8   9   1   2   1 2   3 3   0 0   4 2
a    b   c   d   e   f   ----------g----------
```

- Individual 12 from pedigree 3
- Father is individual 8 and mother is individual 9
- Male
- Affected
- Genotype A C for first marker
- Genotype G G for second marker
- Missing genotype for third marker

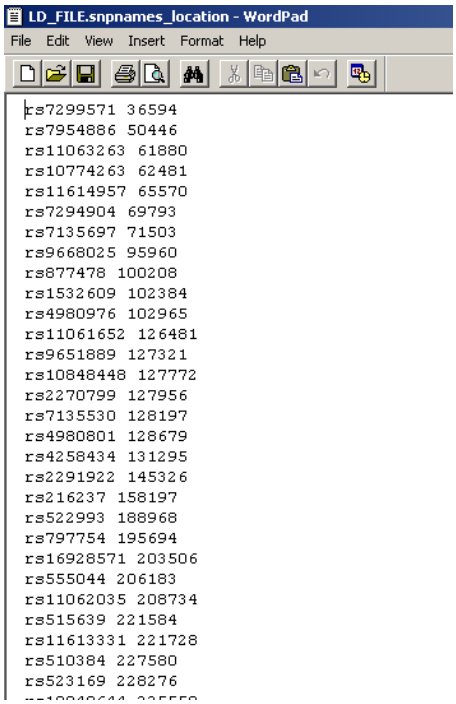
Rows: individuals of Ashkenazi Jewish ancestry
Columns: after “intro” columns, alleles for some SNPS in a stretch of chromosome12 (in this file, 1=reference allele and 2 = alternative allele)



Each marker is represented by two columns (one for each allele, separated by a space) . Standard coding is either ACGT or 1-4 where: 1=A, 2=C, 3=G, T=4.

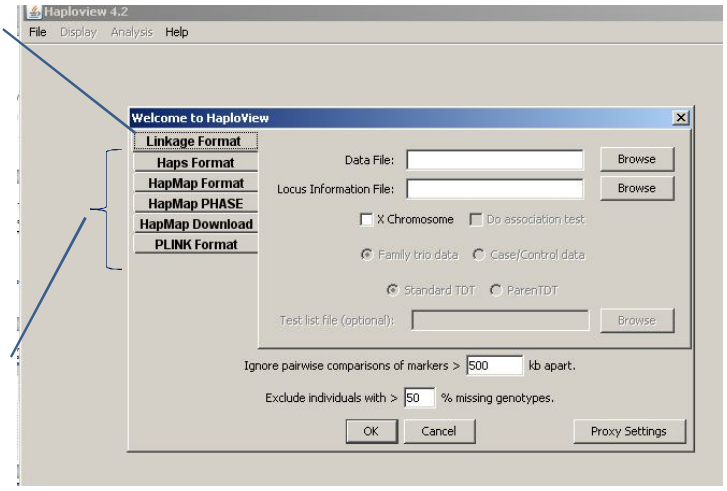
A “0” in any of the marker genotype position indicates missing data.

Names and locations of the SNPs represented in the data file

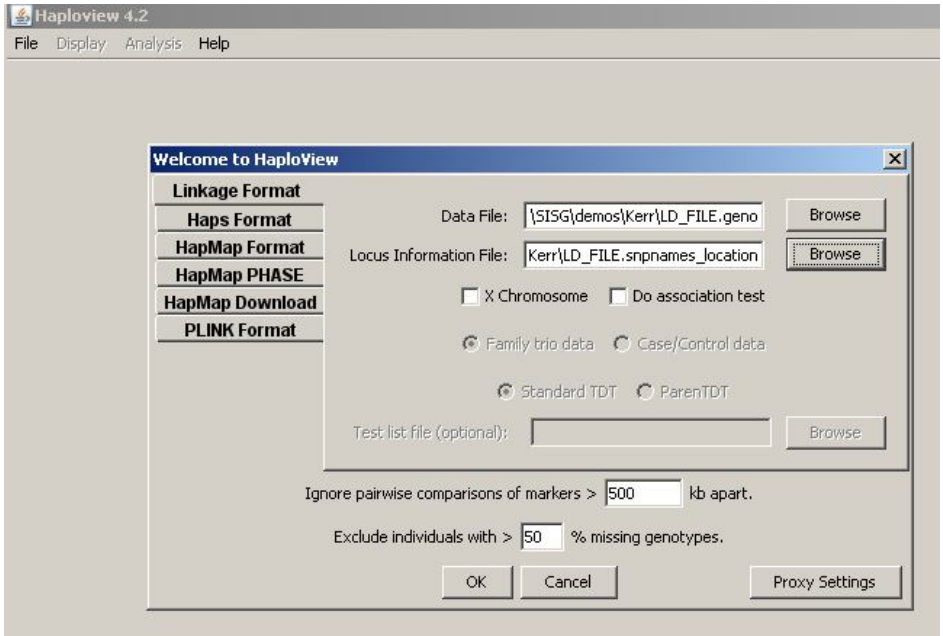


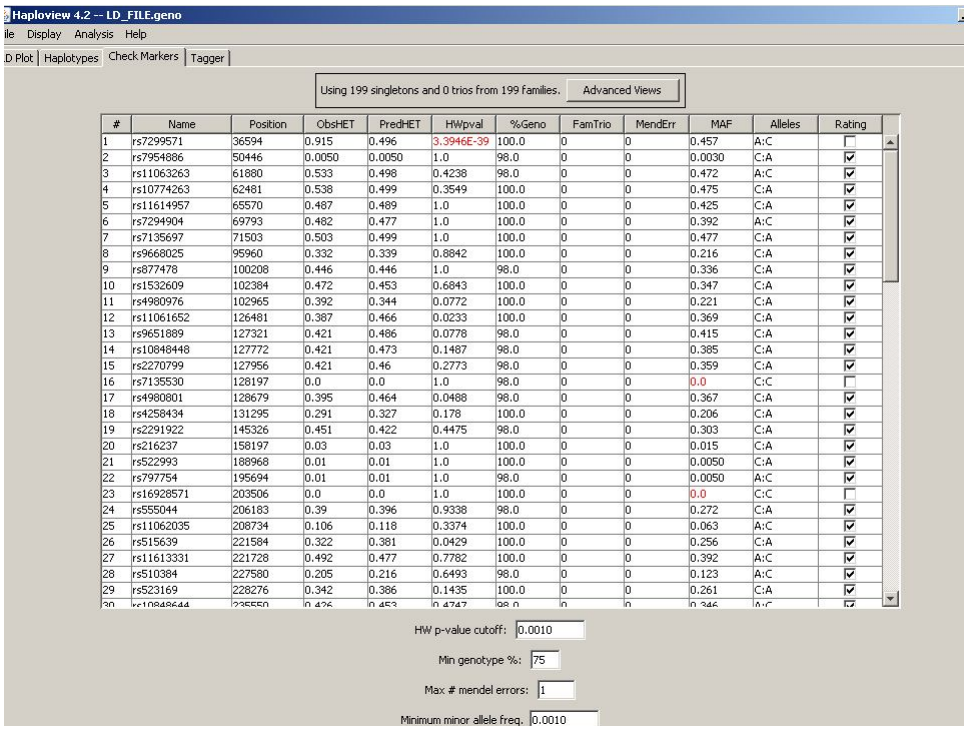
Our data are in "Linkage Format"

Other formats can be used

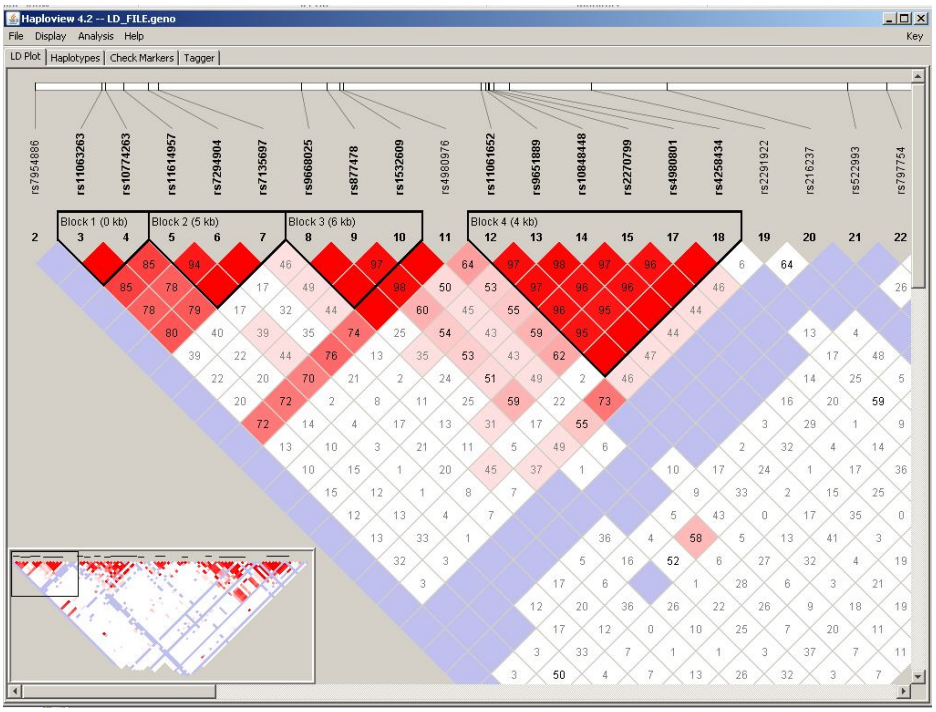


HaploView and LocusZoom





#	Name	Position	ObsHET	PredHET	HWpval	%Geno	FamTrio	MendErr	MAF	Alleles	Rating
1	rs7299571	36594	0.915	0.496	3.3946E-39	100.0	0	0	0.457	A:C	☐
2	rs7954886	50446	0.0050	0.0050	1.0	98.0	0	0	0.0030	C:A	☑
3	rs11063263	61880	0.533	0.498	0.4238	98.0	0	0	0.472	A:C	☑
4	rs10774263	62481	0.538	0.499	0.3549	100.0	0	0	0.475	C:A	☑
5	rs11614957	65570	0.487	0.489	1.0	100.0	0	0	0.425	C:A	☑
6	rs7294904	69793	0.482	0.477	1.0	100.0	0	0	0.392	A:C	☑
7	rs7135697	71503	0.503	0.499	1.0	100.0	0	0	0.477	C:A	☑
8	rs9668025	95960	0.332	0.339	0.8842	100.0	0	0	0.216	C:A	☑
9	rs877478	100208	0.446	0.446	1.0	98.0	0	0	0.336	C:A	☑
10	rs1532609	102384	0.472	0.453	0.6843	100.0	0	0	0.347	C:A	☑
11	rs4980976	102965	0.392	0.344	0.0772	100.0	0	0	0.221	C:A	☑
12	rs11061652	126481	0.387	0.466	0.0233	100.0	0	0	0.369	C:A	☑
13	rs9651889	127321	0.421	0.486	0.0778	98.0	0	0	0.415	C:A	☑
14	rs10848448	127772	0.421	0.473	0.1487	98.0	0	0	0.385	C:A	☑
15	rs2270799	127956	0.421	0.46	0.2773	98.0	0	0	0.359	C:A	☑
16	rs7135530	128197	0.0	0.0	1.0	98.0	0	0	0.0	C:C	☐
17	rs4980801	128679	0.395	0.464	0.0488	98.0	0	0	0.367	C:A	☑
18	rs4258434	131295	0.291	0.327	0.178	100.0	0	0	0.206	C:A	☑
19	rs2291922	145326	0.451	0.422	0.4475	98.0	0	0	0.303	C:A	☑
20	rs216237	158197	0.03	0.03	1.0	100.0	0	0	0.015	C:A	☑
21	rs522993	188968	0.01	0.01	1.0	100.0	0	0	0.0050	C:A	☑
22	rs797754	195694	0.01	0.01	1.0	98.0	0	0	0.0050	A:C	☑
23	rs16928571	203506	0.0	0.0	1.0	100.0	0	0	0.0	C:C	☐
24	rs555044	206183	0.39	0.396	0.9338	98.0	0	0	0.272	C:A	☑
25	rs11062035	208734	0.106	0.118	0.3374	100.0	0	0	0.063	A:C	☑
26	rs515639	221584	0.322	0.381	0.0429	100.0	0	0	0.256	C:A	☑
27	rs11613331	221728	0.492	0.477	0.7782	100.0	0	0	0.392	A:C	☑
28	rs510384	227580	0.205	0.216	0.6493	98.0	0	0	0.123	A:C	☑
29	rs523169	228276	0.342	0.386	0.1435	100.0	0	0	0.261	C:A	☑
30	rs10848644	235550	0.426	0.453	0.4747	98.0	0	0	0.346	A:C	☑

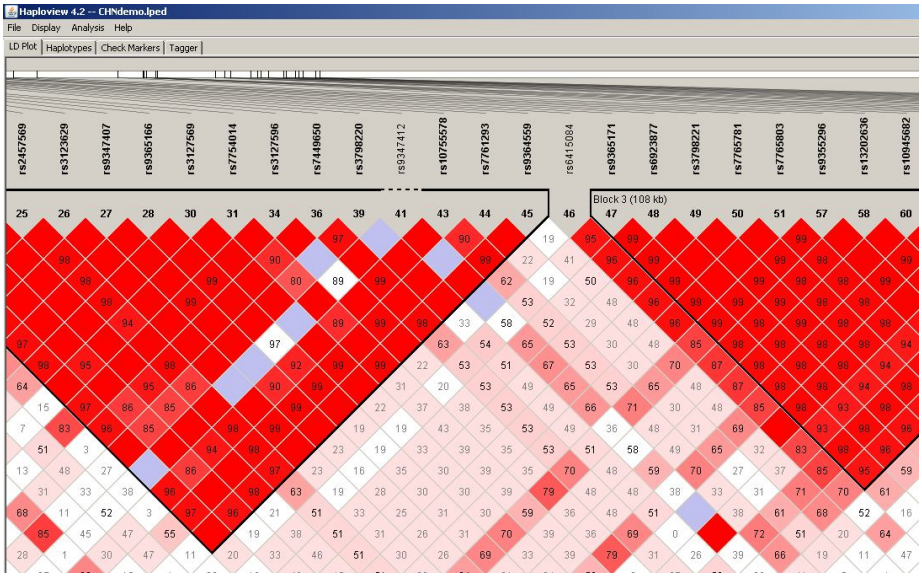


Another Example

- Data on Chinese-Americans
- CHNdemo.lped
- CHNmarkerinfo.txt

- Lots of LD

HaploView and LocusZoom

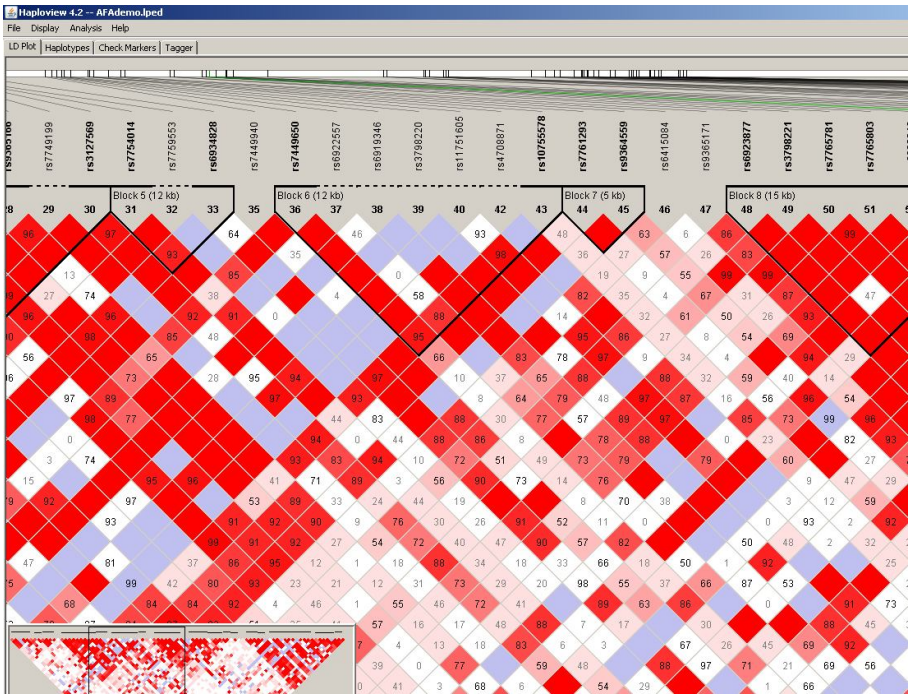


- Contrast with the same region in a sample of African-American

- AFAdemo.lped
- AFAmarkerinfo.txt

- Lots less LD

HaploView and LocusZoom



Regional Association Plots
Web based tools available at
www.broadinstitute.org/mpg/snap/ldplot.php

Software Demo

Input Data

File Edit View Insert Format Help

SNP	PValue	Position
rs791183	0.490239538691658	160530114
rs13194446	0.636718779532867	160530753
rs791184	0.0095632094857105	160531149
rs791190	0.0111350579316875	160533697
rs677985	0.00298464067989899	160539896
rs4543396	0.010271605534538	160540242
rs16891179	0.053057759496473	160542191
rs476235	0.654546090292965	160542627
rs476064	0.173640685066214	160542698
rs474244	0.119435561546612	160542875
rs3903279	0.853909493714035	160545023
rs4709408	0.288955984287674	160545221
rs2619268	0.755584720213045	160545968
rs653753	0.789866777080132	160546991
rs3861973	0.731085141958688	160548895
rs2448298	0.674813711198785	160554025

Required Columns

- SNP – must follow a standard nomenclature unless POSITION also specified
- PValue

Optional Columns

- POSITION – the coordinate of the SNP on the chromosome
  - If not supplied, then Hapmap position will be used. If not in Hapmap, then SNP won't be plotted
  - ALL SNPs must be on the same chromosome within 1Mb

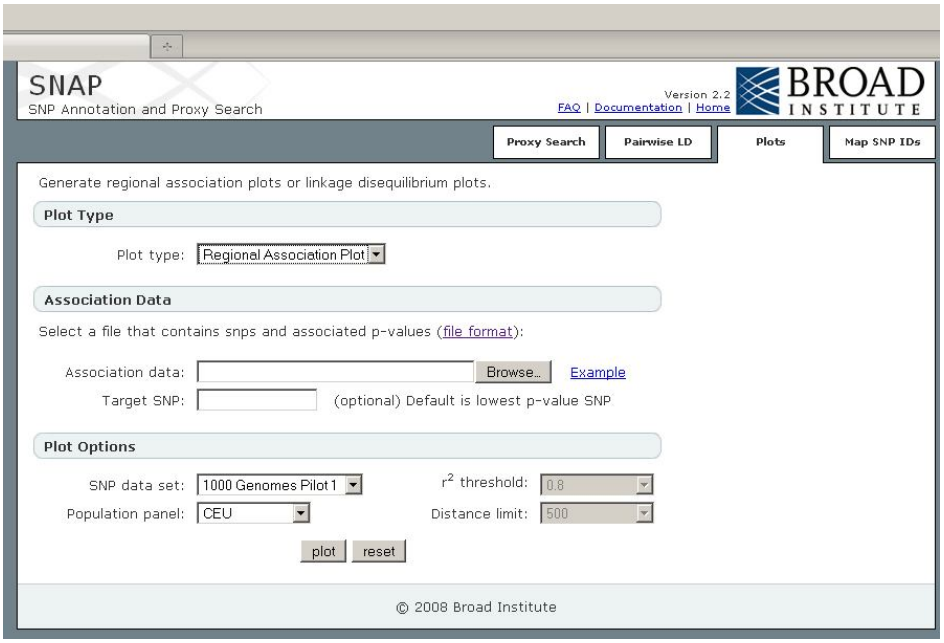
Optional Columns

- RSquared – measure of LD between this SNP and target SNP
  - If not supplied, it is estimated from Hapmap data. If one or both SNPs in not in Hapmap, then LD assumed to be 0
- SnpType – "typed" or "imputed"

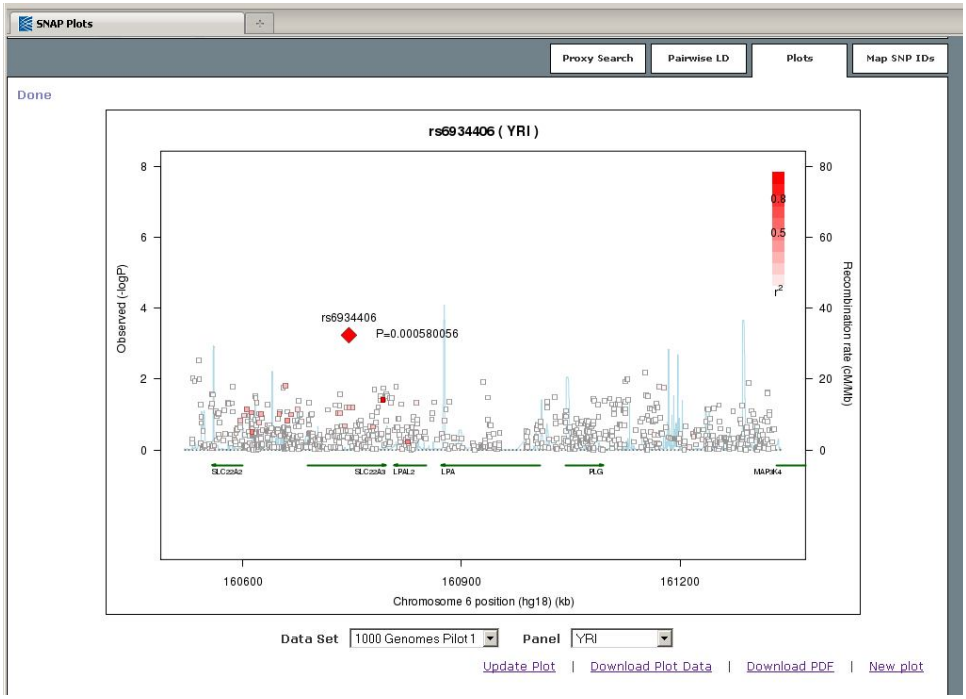
Other Inputs

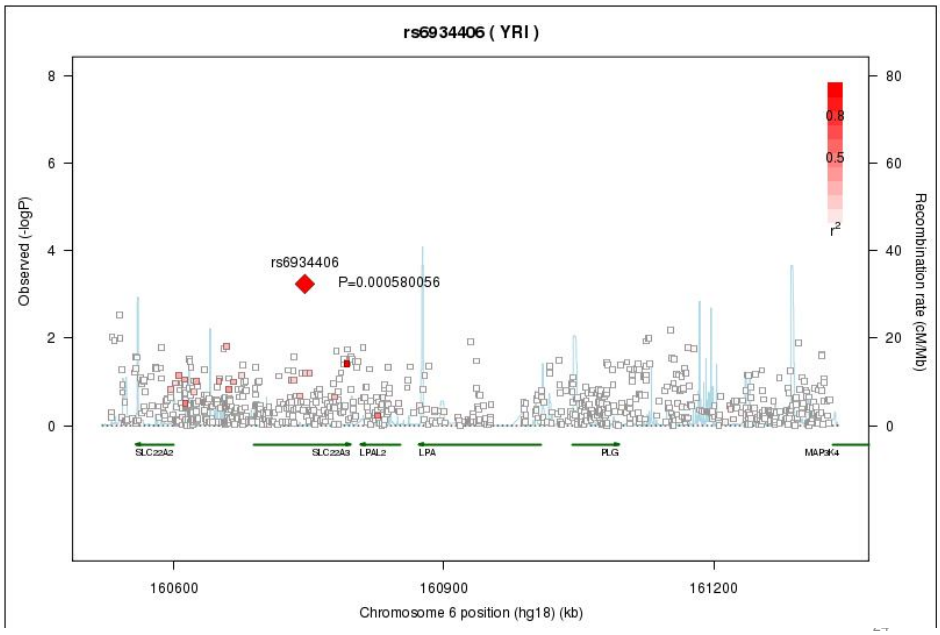
- The "target" SNP is the focus of the plot. The r^2 values are computed from this SNP to all other SNPs. If not specified, then software defaults to use the SNP with the smallest pvalue.
- Population panel. Used, for example, to estimate r^2 values if not input by user.

Input file example: AFAdata4RegionalAssocPlot1.txt



HaploView and LocusZoom

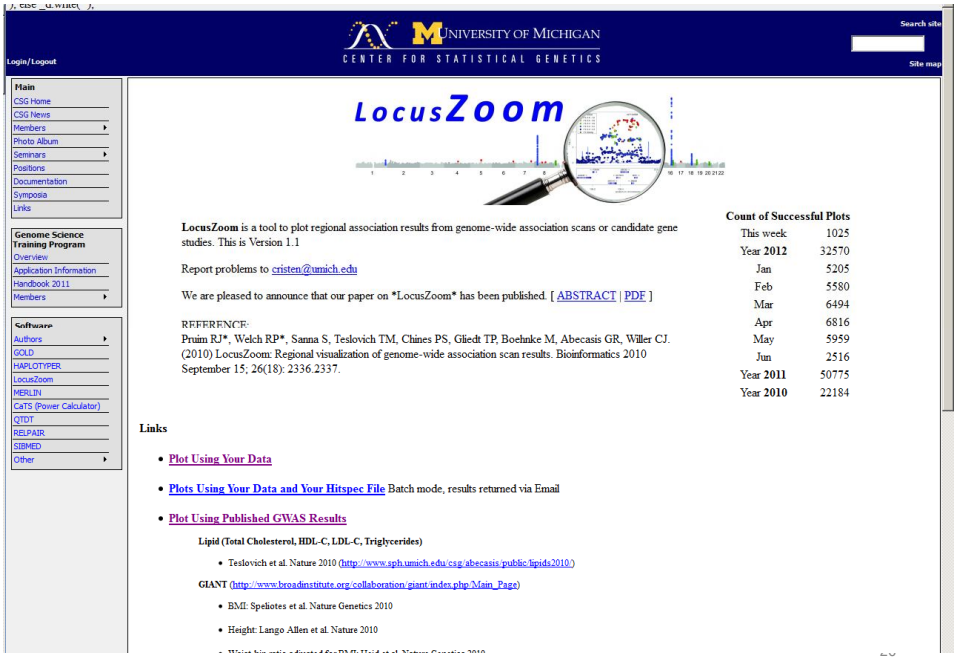




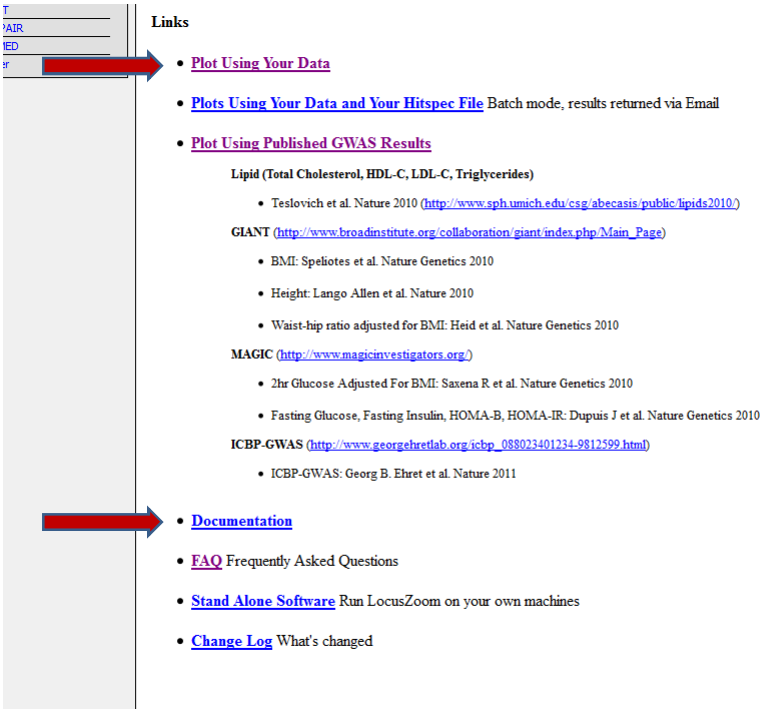
Regional Association Plots
LocusZoom

http://csg.sph.umich.edu/locuszoom/

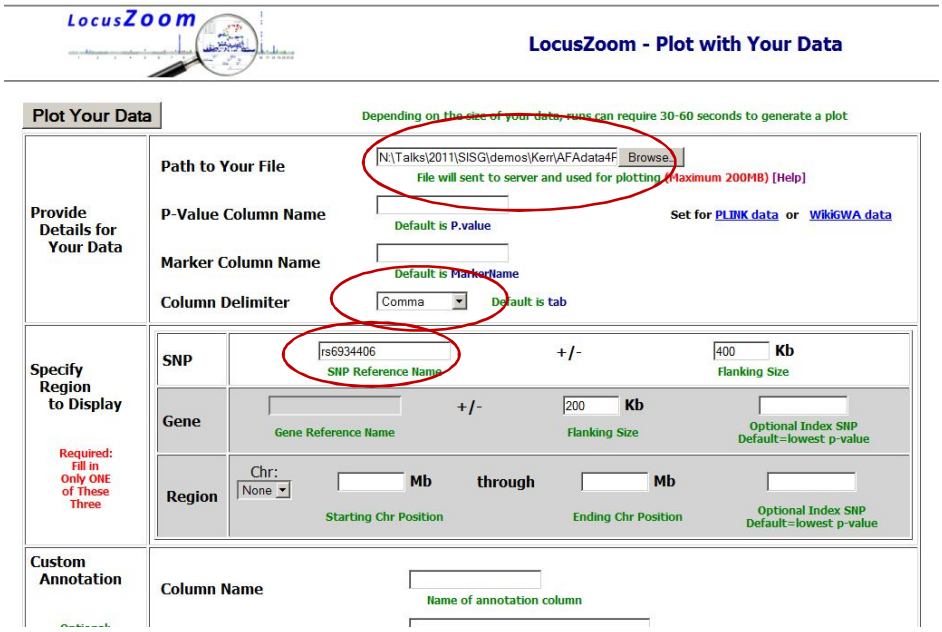
Software Demo



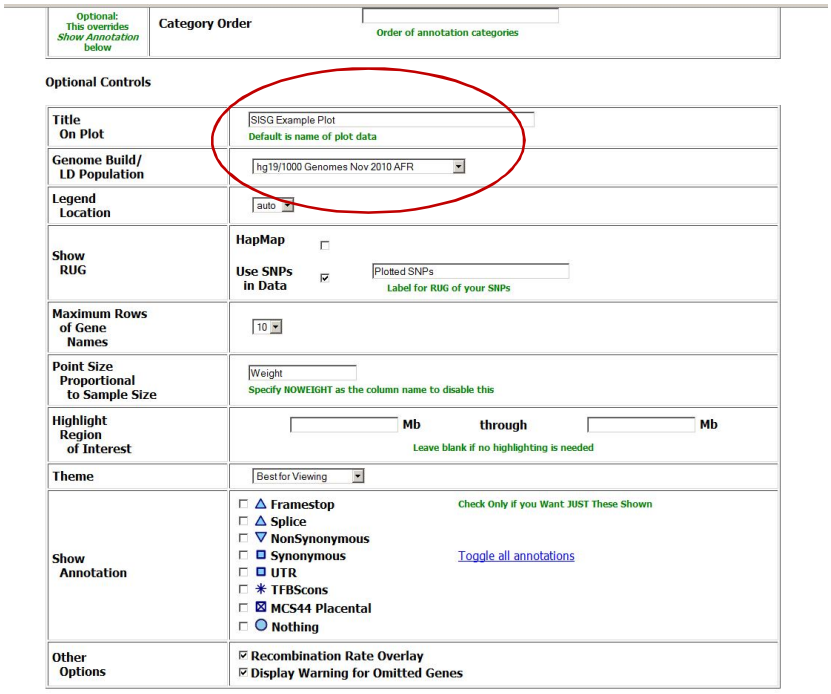
HaploView and LocusZoom

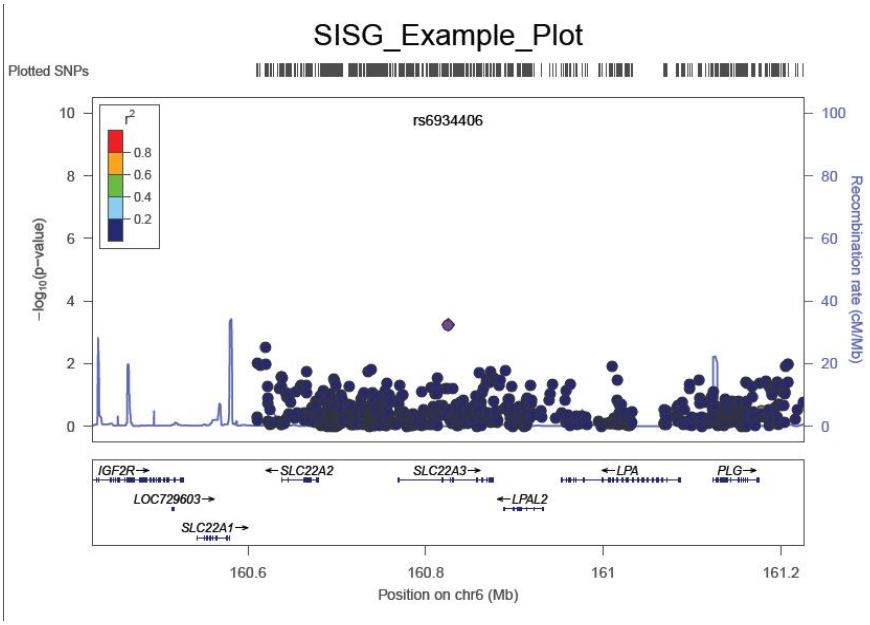


Input file example: AFAdata4RegionalAssocPlot1.csv



HaploView and LocusZoom





Statistical and Software Resources for Genetic Epidemiology Studies

Timothy Thornton and Katie Kerr

Summer Institute in Statistical Genetics 2014
Module 10
Lecture 12

Introduction

- In the last 15-20 years there has been a tremendous amount of software released for the analysis of genetic data
- There are thousands of software packages available for genetic analysis
- Finding the software that suits your needs can be challenging
- Fortunately there are resources available that can be used to aide in this process
- This guide is designed to introduce you to some of the available resources for genetic analysis software as well as highlight the usefulness for a few of them

R Software

- R is a programming language and software environment for statistical computing and graphics.
- There are a number of R packages implementing statistical methods and algorithms for the analysis of genetic data and for related population genetics studies.
- A brief description of these packages are given at the following website:
  `http://cran.r-project.org/web/views/Genetics.html`

R Software

- R packages for the following topics:
  - Population Genetics
  - Phylogenetics
  - Linkage
  - QTL mapping
  - Association
  - Linkage Disequilibrium and haplotype mapping
  - Genome-Wide Association Studies (GWAS)
  - Multiple testing
  - Sequence Data

PLINK

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner: `pngu.mgh.harvard.edu/~purcell/plink/`
- PLINK has numerous useful features for managing and analyzing genetic data
- Data management
  - Read data in a variety of formats
  - Recode and reorder files
  - Merge two or more files
  - Extracts subsets (SNPs or individuals)
  - Flip strand of SNPs
  - Compress data in a binary file format

PLINK

- Summary statistics for quality control
  - Allele, genotypes frequencies, HWE tests
  - Missing genotype rates
  - Inbreeding, IBS and IBD statistics for individuals and pairs of individuals
  - non-Mendelian transmission in family data
  - Sex checks based on X chromosome SNPs
  - Tests of non-random genotyping failure

PLINK

- Basic association testing
  - Case/control
    - Standard allelic test
    - Fisher's exact test
    - Cochran-Armitage trend test
    - Mantel-Haenszel and Breslow-Day tests for stratified samples
    - Dominant/recessive and general models
    - Model comparison tests (e.g. general versus multiplicative)

PLINK

- Family-based association (TDT, sibship tests)
- Quantitative traits, association and interaction
- Association conditional on one or more SNPs
- Asymptotic and empirical p-values
- Flexible clustered permutation scheme
- Analysis of genotype probability data and fractional allele counts (post-imputation)

PLINK

- Multimarker predictors, haplotypic tests
  - Suite of flexible, conditional haplotype tests
  - Case/control and TDT association on the probabilistic haplotype phase
  - A set of proxy association" methods to study single SNP associations in their local haplotypic context
  - Imputation heuristic, to test untyped SNPs given a reference panel
- Copy number variant analysis
  - Joint SNP and CNV tests for common copy number variants
  - Filtering and summary procedures for segmental (rare) CNV data
  - Case/control comparison tests for global CNV properties
  - Permutation-based association procedure for identifying specific loci

PLINK

- Gene-based tests of association
- Screen for epistasis
- Gene-environment interaction with continuous and dichotomous environments
- Meta-analysis
  - Automatically combine several generically-formatted summary files, for millions of SNPs

Software for Population Stratification

- EIGENSOFT/EIGENSTRAT:
  `http://genepath.med.harvard.edu/~reich/Software.htm`
  - Identify and adjust for population structure by using principal components analysis (PCA)
  - The top principal components are viewed as continuous axes of variation that reflect subpopulation genetic variation in the sample.
  - Computationally feasible for high density genome scans

Software for Ancestry Estimation

- STRUCTURE:
  http://pritch.bsd.uchicago.edu/software.html
  - This method uses MCMC to cluster genetically similar individuals into "subpopulations."
  - Not computationally feasible for GWAS
  - Assignments of individuals are highly sensitive to the number of subpopulations in the sample, which is often unknown when there is cryptic structure
- ADMIXTURE:
  http://www.genetics.ucla.edu/software/admixture
  - The ADMIXTURE software incorporates an improved algorithm from what is implemented in STRUCTURE so that the method is computationally feasible for large-scale GWAS.
- *frappe*:
  `http://med.stanford.edu/tanglab/software/frappe.html`
  - *frappe* uses a frequentist approach for estimating individual ancestry admixture proportions

Software for Structured Samples with Related Individuals

- Many genetic studies have samples with related individuals
- ROADTRIPS: RObust Association Dectection Test for Related Individuals and Population Structure
  `http://galton.uchicago.edu/~mcpeek/software/index.html`
  - Software for partially or completely unknown population and pedigree structure.
  - Uses an empirical covariance matrix to capture both population and pedigree structure
- EMMAX: Efficient Mixed-Model Association eXpedited
  `http://genetics.cs.ucla.edu/emmax/`
  - Uses a Variance Components model to account for structure in GWAS

Software for Rare Variant Association Testing

- The analysis of rare variants from sequencing studies has become an area of growing importance in genetic epidemiology
- It is believed that rare variants may explain a significant proportion of this missing heritability that have not been capture by GWAS
- Although the development of statistical methods for the analysis of rare variants is still in its infancy, there has been significant effort dedicated towards the development of powerful and computationally efficient statistical methods in the last few years.
- Many of these methods have been implemented in software.

Software for Rare Variant Association Testing

- The sequence kernel association test (SKAT) http://www.hsph.harvard.edu/xlin/software.html
  - The SKAT software is widely used to test for association between a set of rare variants and dichotomous or quantitative phenotypes
  - Aggregates individual score test statistics of SNPs in a set, while adjusting for covariates, such as principal components to account for population stratification.
- Burden rare variant association tests obtain a cumulative value for the effects of the rare variants in a region to test for association with a phenotype. Software implementing burden tests include
  - VT software: http://genetics.bwh.harvard.edu/rare_variants/
  - The CCRaVAT (Case-Control Rare Variant Analysis Tool) & QuTie (Quantitative Trait) software: http://www.sanger.ac.uk/resources/software/rarevariant/

QUANTO: Software for Computing Power in Association Studies

- Quanto is a program that computes sample size and/or power for association studies of genes, environmental factors, gene-environment interaction, or gene-gene interaction: `http://hydra.usc.edu/GxE`
- Available study designs for a disease (binary) outcome include the unmatched case-control, matched case-control, case-sibling, case-parent, and case-only designs.
- Study designs for a quantitative trait include independent individuals and case parent designs.

The Genetic Power Calculator

- The Genetic Power Calculator (GPC) software is another useful program for power calculations: `http://pngu.mgh.harvard.edu/~purcell/gpc/`
  - GPC power calculations are based on variance components methods for quantitative traits and can be used for both linkage and association mapping studies:
  - GPC can calculate power for sibships of arbitrary size.
  - The software also provides power calculations for testing association for dichotomous traits with the TDT test for trio-based study designs, as well power for a case-control design.

Software for the Graphical Display of Genetic Data

- Haploview is designed to simplify and expedite the process of haplotype analysis by providing a common interface to several tasks relating to such analyses.
  `http://www.broad.mit.edu/personal/jcbarret/haploview/`
  - Haploview currently allows users to examine block structures, generate haplotypes in these blocks, run association tests, and save the data in a number of formats.
- Online Tools
  - LocusZoom: a tool to plot regional association results from genome-wide association scans or candidate gene studies: `http://csg.sph.umich.edu/locuszoom/`
  - SNAP: a tool to plot regional association plots or regional LD plots: `http://www.broadinstitute.org/mpg/snap/`
- The R software package!

Reference Panels

- Hapmap: International Haplotype Map Project
  http://hapmap.ncbi.nlm.nih.gov/
  - Started in 2002
  - High Throughput genotyping chips
  - Latest release (Release 3) has $\sim$ 3 million SNPs
  - 11 populations
  - 1400 individuals
- The 1000 Genomes
  http://www.1000genomes.org
  - Started in 2008
  - Low coverage and exome next generation sequencing
  - Latest release (Phase 2) contains 36.6 millon snps, 3.8 millon indels and 14K deletions
  - 19 Populations
  - 1715 individuals

Reference Panels

- HGDP: Human Genetic Diversity Panel
  http://hagsc.org/hgdp/index.html
  - A group of scientists at Stanford University collaborated on a large study to understand genetic diversity in human populations
  - High throughput genotyping data
  - More than 650,000 SNPs
  - 51 populations from around the world: Africa, Europe, the Middle East, South and Central Asia, East Asia, Oceania and the Americas.
  - 1,043 individuals

HGDP: Human Genetic Diversity Panel

Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Nature Reviews | Genetics

Genotype and Phenotype Repositories: dbGaP

- The database of Genotypes and Phenotypes (dbGaP): http://www.ncbi.nlm.nih.gov/gap
  - Developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype
  - Studies include: GWAS, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.
  - Publicly available data as well as controlled access data (must apply for access).

Genotype and Phenotype Repositories: EGA

- The European Genome-Phenome Archive (EGA): http://www.ebi.ac.uk/ega/
  - EGA a repository containing genotype and phenotype information for a number of genetic studies conducted in Europe.
  - This repository holds the well known Wellcome Trust Case Control Consortium (WTCCC) genotype and phenotype data.