

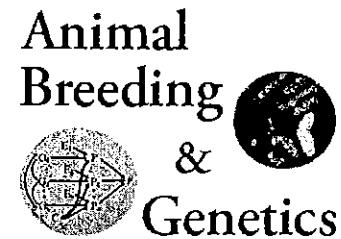
Use of High-density SNP Genotyping for Genetic Improvement of Livestock

Jack Dekkers
Dorian Garrick
Rohan Fernando



Iowa State University
Ames, IA, USA

June 1-10 2009



Module A: Basics of QTL mapping and MA-selection (June 1-4)

Provides basis for Module B

- Day 1 Multi-locus Population Genetics – Linkage & Disequilibrium
- Day 2 QTL detection – Basic Principles
- Day 3 Linkage Disequilibrium Mapping
- Day 4 Identity By Descent and Marker-assisted Genetic Evaluation

Module B: Genomic Selection (June 5-6 and 8-10)

- Statistical, quantitative, and computational aspects of genomic selection
- Strategies for implementation of genomic selection

USE AND ACKNOWLEDGEMENT OF SHORT COURSE MATERIALS

Materials provided in these notes are copyright of the authors and the Animal Breeding and Genetics group at Iowa State University but are available for use with proper acknowledgement of the author and the short course. Materials that include references to third parties should properly acknowledge the original source.



2009 ISU AB&G Short Course

"Use of High-Density SNP Genotyping for Genetic Improvement of Livestock" June 1-10, 2009

Instructors: Dr. Jack Dekkers, Dr. Rohan Fernando, and Dr. Dorian Garrick

Behnam Abasht

Heritage Breeders
10789 Stewart Neck Road
Princess Anne, MD 21853 USA
Behnam.Abasht@Perdue.com

Mohammed K. A. Abo-Ismail, Ph.D. Student
Department of Animal & Poultry Science
University of Guelph
50 Stone Road East – Bldg. #70
Guelph, Ontario CANADA
maboisma@uoguelph.ca

Everestus Chima Akanno, Graduate Student
University of Guelph
50 Stone Road East
Guelph N1G 2W1
CANADA
eakanno@uoguelph.ca

Cristina Andreeescu, Graduate Student
Department of Animal Science
229 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
cyan@iastate.edu

Jesus A. Arango, Statistical Geneticist
Hy-Line International
2583 – 240th Street
Dallas Center, IA 50063 USA
jarango@hyline.com

Jesus A. Baro, Professor
Universidad de Valladolid
Avda – De Madrid S/N
Palencia 34004
SPAIN
baro@agro.uva.es

Devori Beckman, Graduate Student
Department of Animal Science
229 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
dbeckman@iastate.edu

Stefano Biffani, Researcher
Parco Tecnologico Padano
Via A. Einstein
Lodi 26900
ITALY
stefano.biffani@tecnoparco.org

Teodor Stefan Bildea, Research Scientist
Pioneer Hi-Bred International Inc.
PO Box 1000
Johnston, IA 50131-0184 USA
stefan.bildea@pioneer.com

Nicholas Boddicker, Graduate Student
Department of Animal Science
227 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
njb@iastate.edu

Weigu Cai, Research Assistant
Department of Animal Science
231 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
cweigu@iastate.edu

Carlos Carleos, Professor Contratado Doctor
Universidad Oviedo, Facultad de Ciencias –
c/ calvo sotelo, Oviedo / Asturias 33007
SPAIN
carleos@uniovi.es

Ane Marie Closter, Ph.D. Student
Wageningen University & Research
Marijkeweg 40
Wageningen 6709 AH
THE NETHERLANDS
ane-marie.closter@wur.nl

Dr. Jack Dekkers
Department of Animal Science
239D Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
jdekkers@iastate.edu

Fernando A. Di Croce, Ph.D. Student
University of Tennessee – 102 McCold Hall
2640 Morgan Circle Drive
Knoxville, TN 37996-4574 USA
fdicroce@utk.edu

Zhiqiang Du, Post Doc
Department of Animal Science
2262 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
zhqdu@iastate.edu

Dawn Elkins, Graduate Student
Department of Animal Science
233 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
delkins@iastate.edu

Linda Engblom, Post Doc
Department of Animal Science
109 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
lengblom@iastate.edu

Bin Fan, Post Doctorate
Department of Animal Science
2255 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
bfan@iastate.edu

Dr. Rohan L. Fernando
Department of Animal Science
237A Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
rohan@iastate.edu

Dr. Mark Fife, Post Doc
Institute for Animal Health
High Street – Compton
Berkshire RG 20 7NN
UK
mark.fife@bbsrc.ac.uk

Flavio Forabosco, Researcher
Interbull SLV
Uppsala SE-750007
SWEDEN
flavio.forabosco@hgen.slu.se

A.E. Freeman, Professor Emeritus
Department of Animal Science
225 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
genef@iastate.edu

Stuart Gardner, Ph.D. Student
Department of Statistics & Microbiology
1144 Veterinary Medicine
Iowa State University
Ames, IA 50011 USA
stugard@iastate.edu

Dr. Dorian Garrick
Department of Animal Science
225D Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
dorian@iastate.edu

Danielle Gorbach, Graduate Research Assistant
Department of Animal Science
2255 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
dmbowen@iastate.edu

Andres Gordillo, Breeder
Ag Reliant Genetics LLC
972 County Road 500 East
Ivesdale, IL 61851 USA
andres.gordillo@agreliantgenetics.com

Kent Gray, Ph.D. Graduate Student
NC State University
2842 Avent Ferry Road - #201
Raleigh, NC 27606 USA
kgray@ncsu.edu

Jose Guerra, Biometristian
Keygene N.V.
Agro Business Park
90 Wageningen 6708 PW
THE NETHERLANDS
sa@keygene.com

Wei He, Research Assistant
Department of Animal Science
233 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
hewei@iastate.edu

Elliot L. Heffner, Graduate Student
Cornell University
422 Bradfield Hall
Ithaca, NY 14853 USA
elh39@cornell.edu

Dr. Gerald Herbert, Senior Geneticist
Hubbard LLC
PO Box 415 – 195 Main Street
Walpole, NH 03608 USA
gerald.herbert@hubbardbreeders.com

Julie Ho, Research Scientist
Pioneer Hi-Bred
1039 S. Milton-Shopiere Rd.
Janesville, WI 53546 USA
julie.ho@pioneer.com

Mohsen Jafarikia, Geneticist
Canadian Centre for Swine Improvement
Central Experimental Farm – Bldg. #54
Ottawa ON - K1A 0CG CANADA
inohsen@ccsi.ca

Bob Kemp
RAK Genetic Consulting Ltd.
54 Coachwood Point W.
Lethbridge, Alberta T1K6A9
CANADA
kempb@shaw.ca

Kadir Kizilkaya, Post Doc
Department of Animal Science
233 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
kadirk@iastate.edu

Konrad Kulak, Research Scientist
PIC USA
100 Bluegrass Commons Blvd.
Hendersonville, TN 37075 USA
konrad.kulak@pic.com

Honghao (Eileen) Li, Graduate Student
Alberta Bovine Genomics
Department of Agricultural, Food and
Nutritional Science
1430 College Plaza – 8215-112 Street
Edmonton, AB T6G 2C8
CANADA
honghao@ualberta.ca

Aaron Lorenz, Postdoctoral Research Associate
USDA-ARS – Robert W. Holly Center
Towers Road
Ithaca, NY 14853 USA
ajl289@cornell.edu

Olivia N. Mapholi, Researcher
Animal Breeding & Genetics
ARC - Irene
SOUTH AFRICA
ntanga@arc.agric.za

Myrthe Maurice-van Eijndhoven,
Ph.D. Student
Animal Sciences Group – Wageningen UR
Edelhertweg 15 – Lelystad 821gPH
THE NETHERLANDS
myrth.maurice-vaneijndhoven@wur.nl

Jennifer McDonald, Visiting Scientist
Department of Animal Science
233 Kildee Hall - Iowa State University
Ames, IA 50011-3150 USA
jennifer_mcdonald@newsham.com

Dr. Anupama Mukherjee
227B Kildee Hall – Iowa State University
Department of Animal Science
Ames, IA 50011-3150
mukherje@iastate.edu

Dr. Elly Ana Navajas, Researcher
Scottish Agricultural College
Sir Stephen Watson Building
Penicuik, Midlothian EH26 OPH
UK
elly.navajas@8ac.ac.uk

Mark Newell, Ph.D. Graduate Student
Department of Agronomy
1203 Agronomy - Iowa State University
Ames, IA 50011 USA
newell@iastate.edu

Hong Nguyen Nguyen, Research Scientist
The World Fish Center
Jalan Batu Maung
Batu Maung – 11960 Bayan Lepas
Penang, MALAYSIA
n.nguyen@cgiar.org

Marja Nikkilae, Graduate Student
Department of Animal Science
227 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
nikkilae@iastate.edu

Katrina Olson, Post Doc
Virginia Tech
2150 Litton-Reaves Hall (0315)
Blacksburg, VA 24060 USA
kmolson@vt.edu

Suneel K. Onteru, Post Doc
Department of Animal Science
2262 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
sonterus@iastate.edu

Sunday Peters, Grad Student/Visiting Scholar
Department of Animal Science
233 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
sopeters@iastate.edu

Laknath Peiris, Post Doc Research Associate
Department of Animal Science
239E Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
laknathp@iastate.edu

Fabiano Pita, Biologist Scientist
Dow Agrosciences
9330 Zionsville Road
Indianapolis, IN 46268 USA
fpita@dow.com

Markus Schneeberger, Lecturer
Institute of Animal Sciences
ETH Zurich – TAN D2
Zurich 8092
SWITZERLAND
markus.schneeberger@inw.agrl.ethz.ch

Dr. James Schneider, Research Geneticist
US Meat Animal Research Center
PO Box 166
State Spur 18D
Clay Center, NE 68933-0166 USA
Jim.Schneider@ars.usda.gov

Ghislaine Schopen, Ph.D.
Wageningen University & Research
Marijkeweg 40
Wageningen 6900 AH
THE NETHERLANDS
ghyslaine.schopen@wur.nl

Anouk Schurink, Ph.D.
Wageningen University & Research
Marijkeweg 40
Wageningen 6900 AH
THE NETHERLANDS
anouk3.schurink@wur.nl

J. R. Tait, Associate Scientist
Department of Animal Science
2255 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
rtait@iastate.edu

B. K. Thamasandra Narayana,
Graduate Research Assistant
Department of Agronomy
1301 Agronomy Hall
Iowa State University
Ames, IA 50011 USA
nbharath@iastate.edu

Johannes van Kaam, Researcher
Anafi Via Bergamo
2g2 Cremona 26100
ITALY
jikaam@anafi.it

Huiyu Wang, Graduate Student
University of Georgia
425 River Road
Athens, GA 30602 USA
huiyu@uga.edu

Jing Wang, SRA
Pioneer Hi-Bred, Inc.
NW 62nd Avenue
Johnston, IA 50131 USA
jing.wang@pioneer.com

Shaolin Wang, Graduate Student
Auburn University
650 N. Ross Street A8
Auburn, AL 36830 USA
wangsha@auburn.edu

Zhiqian Wang, Associate Professor
University of Alberta
4007 – 104A Street
Edmonton, AB T6J 6L3
CANADA
zhiqian@ualberta.ca

Kristina L. Weber, Graduate Student
UC Davis
644 Almondo Avenue - #115
Davis, CA 95616 USA
klweber@ucdavis.edu

Jamie Williams, Graduate Student
University of Georgia
425 River Road
Athens, GA 30602 USA
jwill@uga.edu

Huiqin Xue, Scientist
Monsanto Company
3302 SE Convenience Blvd.
Ankeny, IA 50021 USA
huiqin.xue@monsanto.com

Aimin Yan, Post Doc
Department of Animal Science
229 Kildee Hall
Iowa State University
Ames, IA 50011-3150 USA
aiminy@iastate.edu

Wenzhao Yang, Graduate Student
Department of Animal Science
1205F Anthony Hall
Michigan State University
East Lansing, MI 48824-1225 USA
yangwenz@msu.edu

Jian Zeng, Research Assistant
Department of Animal Science
233 Kildee Hall
Iowa State University
Ames, IA 50011-3150
jzeng@iastate.edu

Honghua Zhao, Research Scientist
Pioneer Hi-Bred Inc.
NW 62nd Avenue
Johnston, IA 50131
honghua.zhao@pioneer.com

Xia Zhao, Graduate Student
Department of Animal Science
2262 Kildee Hall
Iowa State University
Ames, IA 50011-3150
zhx@iastate.edu

Shengqiang Zhong
Monsanto Company
3302 SE Convenience Blvd.
Ankeny, IA 50021
shengqiang.zhong@monsanto.com

Introduction, Background, and Mathematical Foundation of Quantitative Genetics

Jack Dekkers¹ and Jean-Luc Jannink²

¹ Dept. Animal Science, Iowa State University

² Currently at Cornell University, formerly Dept. Agronomy, Iowa State University

Quantitative genetics is the study of continuous traits (such as height or weight) and its underlying mechanisms. It is based on extending the principles of Mendelian and populations genetics to quantitative traits.

Mendelian inheritance:

1. *Law of segregation*: A trait is influenced by a pair of alleles but each individual only passes a single, random allele on to its progeny.
2. *Law of independent assortment*: Alleles of different factors combine independently in the gamete.

Population Genetics is the study of the allele frequency distribution and change under the influence of the four evolutionary forces: natural selection, genetic drift, mutation, and migration.

Falconer and Mackay:

“Quantitative genetics theory consists of the deduction of the consequences of Mendelian inheritance when extended to the properties of populations and to the simultaneous segregation of genes at many loci.”

For the purposes of this class: Quantitative genetics = A set of concepts based on the theory of inheritance that help us understand and dissect the genetic basis of quantitative traits and predict what the consequences of different breeding choices will be and therefore allow us to make decisions that lead to the most desirable outcomes.

Quantitative traits

Quantitative genetics covers all traits that are determined by many genes.

- Continuous traits are quantitative traits with a continuous phenotypic range. They are usually polygenic, and may also have a significant environmental influence.
- Traits whose phenotypes are expressed in whole numbers, such as number of offspring, or number of bristles on a fruit fly. These traits can be either treated as approximately continuous traits or as threshold traits.
- Some qualitative traits can be treated as if they have an underlying quantitative basis, expressed as a threshold trait (or multiple thresholds). E.g. diseases that are controlled by multiple traits but for which phenotype is observed as healthy/diseased.

See also Lynch and Walsh Chapter 1 and “Philosophical and Historical Overview.pdf”

Mendelian Genetics

Discrete traits

- Theory of heredity at individual locus level
- Inheritance of genes (alleles) at a locus from parent to progeny
 - o *Law of segregation*
 - o *Law of indep. assortment*

Darwin's Evolutionary Genetics –

quantitative traits

- Focus on variation as spice of evolution
- Resemblance among relatives
 - o *Some heritable variation*
- Differential reproductive success
But: no clear model of inheritance
Progeny resemble parents yet need to differ from parents to maintain variation
????

Galton's Biometrical Genetics –

quantitative traits

- Progeny resemble their parents
- Regression towards mediocrity
 - o Children of tall (short) parents tend to be shorter (taller) than parents
- *But: how is variation maintained?*
 - o Johansson: $P = G + E$
 - o *G = single inherited block*
 - o *Evolution through mutation*

Population Genetics

Individual loci

- Theory of allele/genotype frequencies in populations
- Theory of changes in frequencies due to evolutionary forces: *Natural selection*
Genetic drift *Mutation*

Multifactorial model for quantitative traits

George Udny Yule (1902)

- multiple genetic factors + environment
 - continuous variation
 - regression towards mediocrity
- *Experimental evidence from corn breeders*
 - *outbreak of variation in F2*

Quantitative Genetics Theory

- Theory underlying the inheritance of quantitative traits
- Falconer and McKay: "the deduction of the consequences of Mendelian inheritance when extended to the properties of populations and to the simultaneous segregation of genes at many loci."
- Theory of population changes in quantitative trait as a result of *selection, genetic drift (inbreeding), mutation, migration* (crossing)
- A set of concepts based on the theory of inheritance that help us understand and dissect the genetic basis of quantitative traits and predict what the consequences of different breeding choices will be and therefore allow us to make decisions that lead to the most desirable outcomes

“...genetics is meant to explain two apparently antithetical observations – that organisms resemble their parents and differ from their parents. That is, genetics deals with both the problem of heredity and the problem of variation.” Lewontin, 1974.

Francis Galton (1822-1911): regression toward mediocrity – progeny of parents with extreme phenotypes tend to be closer to average.

The modern synthesis of Quantitative Genetics was founded by R.A. Fisher, Sewall Wright, and J.B.S. Haldane, based on evolutionary concepts and population genetics, and aimed to predict the response to selection given data on the phenotype and relationships of individuals.

Analysis of Quantitative trait loci, or QTL, is a more recent addition to the study of quantitative genetics. A QTL is a region in the genome that affects the trait or traits of interest.

Some Basic Quantitative Genetic Concepts and Models

Quantitative genetics dwells primarily on developing theory or mathematical models that represent our understanding of phenomena of interest, and uses that theory to make predictions about how those phenomena will behave under specific circumstances. The model that exists to explain observations of quantitative traits contains the following components:

- Loci that carry alleles that affect phenotype – so-called quantitative trait loci or QTL
- Many such quantitative trait loci
- Alleles at QTL that act in pairs (2 alleles per locus) but that are passed on to progeny individually
- Which of the parent's alleles are passed on to progeny occurs at random (i.e. a random one of the pair of alleles that a parent has at a locus is passed on to a given progeny), which introduces variability among progeny
- Loci that affect phenotype sometimes show independent assortment (unlinked loci); sometimes not (linked loci)
- Environmental factors influence the trait

In order to develop the quantitative genetic theory and models and to deduce its consequences or predictions it might make, quantitative geneticists have translated these concepts and their behavior into mathematical and statistical terms/models. The most basic model of quantitative genetics is that the phenotypic value (P) of an individual is the combined effect of the individual's genotypic value (G) and the environmental deviation (E):

$$P = \mu + G + E \quad \text{where } \mu \text{ is the trait mean}$$

G is the combined effect of all the genes that affect the trait.

E is the combined effect of all environmental effects that affect the phenotype of the individual.

The simplest model to describe inheritance of a quantitative trait (under a lot of assumptions that will be covered later), is that the genotypic value of the offspring can be expressed in terms of the genotypic values of its sire (s) and dam (d), based on the fact that half of the genes that the offspring have come from each parent:

$$G_o = \frac{1}{2} G_s + \frac{1}{2} G_d + RA_s + RA_d$$

Here the terms RA_s and RA_d are random assortment or Mendelian sampling terms, which reflect that parents pass on a random half of their alleles (i.e. a random one of two alleles at each locus).

Developing these quantitative genetic models and deducing their consequences, e.g. the consequences of natural or artificial selection on the trait and the population, then involves manipulating the mathematical terms, that is doing algebra and even a little calculus sometimes (!). Quantitative geneticists were really pioneers in this type of mathematical treatment of biological phenomena and as a result the early growth of quantitative genetics was almost synonymous with the early growth of

statistics. Indeed, R.A. Fisher is hailed as a founder of quantitative genetics but also of analysis of variance and randomization procedures in statistics. The early geneticists Galton and Pearson originated the concepts of regression and correlation. Anyway, the upshot for us here is that we will be deeply involved with the mathematical manipulation and statistical evaluation of our representations of the basic quantitative genetic model. We will review some of the rules of probability and statistics, such as variance, covariance, correlation and regression, and will give a hint at how they may relate to the quantitative genetic model.

Mathematical Foundations for Quantitative Genetics

See also Lynch and Walsh Chapters 2 and 3

Random Variables

In principle, we are interested in the random and non-random processes that determine the value of variables. If the variable of interest is which allele a heterozygous (Mm) father passed on to his daughter for a given marker locus, the rule of random segregation indicates that this is a random process. If the variable of interest is the height of the son of a tall woman, some portion of the variable will be non-random (we expect a relatively tall son) and some portion will be random (we don't know exactly what the height will be). Either way, we can identify a random variable with a symbol (say X^P to designate the paternally inherited marker allele, or Y to designate height). Common notation is to use capitals for the name of a variable (e.g. X or Y) and regular font to represent the value (or class) of that variable. E.g. $X=x$ indicates the event that variable X has value x .

Sample Space

The sample space is the set of possible values that a random variable can take. So, for example $X^P \in [M, m]$ (i.e., the progeny inherits either allele M or allele m from its heterozygous Mm father), and $1 < Y < 2.5$ if height is measured in meters. Note that these two example random variables are very different. Random variable X^P can take on just two states (one of the two alleles that the parent has), it is a *categorical* variable, while Y can take on all values between 1 and 2.5, it is a *continuous* variable. Nevertheless, many of the mathematical manipulations we will discuss below can be applied equally to either type variable.

Probability (~ frequency)

We designate the probability of an event A as $\text{Pr}(A)$. For example, if the event A is "the daughter received marker allele M from her heterozygous Mm father" then $\text{Pr}(A) = \text{Pr}(X^P = M)$. In this case $\text{Pr}(A) = \frac{1}{2}$. The probability function $\text{Pr}(\cdot)$ has certain rules assigned to it, just like, for example multiplication has rules assigned to it. For example if event A is "any possible event in the sample space of events" then $\text{Pr}(A) = 1$. Thus, the probability that $X^P = M$ or $X^P = m$ for a progeny of a heterozygous Mm father is equal to $\frac{1}{2} + \frac{1}{2} = 1$. Intuitively, though, it is most useful to think of the $\text{Pr}(A)$ as the chance that event A will happen. If you look at many events (N events, with N very big) and you count N_A , the number of times event A happens, then we can interpret $\text{Pr}(A)$ as a frequency, i.e. $\text{Pr}(A) = N_A/N$. As examples related to the random variables we gave above, if the father is a heterozygote, then Mendel's law of segregation say $\text{Pr}(X^P = M) = \text{Pr}(X^P = m) = \frac{1}{2}$. For the height Y of the son of a tall woman, we can guess that $\text{Pr}(1.5 < Y \leq 1.6) < \text{Pr}(1.8 < Y \leq 1.9)$, that is, the son is less likely to be in a short ten centimeter bracket than a relatively tall ten centimeter bracket.

Probability Density (~ frequency distribution for continuous variables)

The second example leads to the question what is $\Pr(Y = 1.8)$? And the answer, oddly, is zero. That is, given that Y can take on an infinite number of values in the range [1, 2.5], there is a probability of zero that it will take on any specific value. Intuitively, though, we want to be able to express the idea that the chance that the height will be some tall value is greater than the chance it will be some short value. To do this we define the probability density $f(y) = \Pr(y < Y \leq y+e)/e$ as e comes increasingly close to zero. This probability density will be useful to discuss random variables that vary continuously (such as the value of a quantitative trait). Using the probability density function (or pdf) and integration, we can calculate the probability that Y is contained in a certain bracket as

$$\Pr(1.5 < Y \leq 1.6) = \int_{1.5}^{1.6} f(y) dy.$$

The most prominent pdf that we will use is that of the normal distribution, i.e. the bell-shaped curve, which is illustrated in Figure 1.

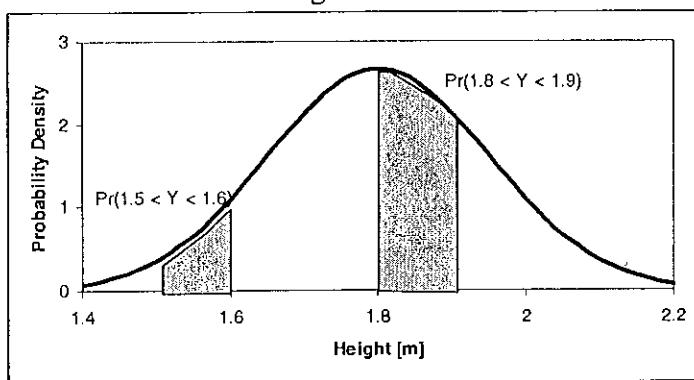


Figure 1

Expected Value (~ mean or average)

The expected value of a random variable is a measure of its location in the sample space, and can be thought of as a mean or an average. It takes slightly different forms depending on whether the variable is categorical or continuous. Consider a categorical variable X with sample space x_1, x_2, \dots, x_k . The expected value of X is essentially calculated as a weighted average of the values that X can take on,

with weights equal to the probability with which X takes on each value: $E(X) = \sum_{i=1}^k x_i \Pr(X = x_i)$.

Example 1: The number of florets per spikelet in oat (= variable X) is affected by a recessive allele that inhibits development of tertiary kernels (this example is slightly fictitious but serves its purpose). Note that the expected value of a categorical trait may not belong to any of the categories of the trait: the expected value for the number of florets per spikelet is $E(X) = 2.75$ though any given spikelet obviously has a whole number of florets.

Table 1 Example for computing expectations for a categorical variable

Genotype	Probability (frequency) $= \Pr(X=x_i)$	Number of florets per spikelet $X=x_i$	$x_i * \Pr(X=x_i)$	$x_i^2 * \Pr(X=x_i)$
t / t	0.25	3	0.75	2.25
T / t	0.50	3	1.50	4.50
T / T	0.25	2	0.50	1.00
Sum	1.00	-	$E(X) = 2.75$	$E(X^2) = 7.75$

Example 2. Now consider the continuous variable height discussed above. The sample space for Y given was $1 < Y < 2.5$, and the pdf is $f(y) = \Pr(y < Y \leq y+e)/e$ as e comes increasingly close to zero. Its expected value is $E(Y) = \int_1^{2.5} yf(y)dy$. Here, instead of multiplying the value of a category by the probability of that category as we did above, we multiply the value by its probability density and integrate over the sample space of the continuous variable. Note that integration is the continuous variable equivalent of summation for categorical variables and the pdf is the equivalent of the probability of each value occurring.

Example 3. Consider again a categorical variable X with sample space x_1, x_2, \dots, x_k . Now assume that there is a function $g(X)$, and we want the expected value of $g(X)$. This expectation is again computed as a weighted average, but now the average of $g(X)$, rather than X itself. The formula for the expectation of $g(X)$ is: $E_X[g(X)] = \sum_{i=1}^k g(x_i)\Pr(X = x_i)$.

Here, E_X means that the expectation is taken over all possible values of variable X . E.g., referring back to Example 1, the expectation of $g(X) = X^2$ is equal to 7.75, as calculated in the last column in Table 1.

Properties of Expectations

Assuming X and Y are random variables and a is a constant (e.g. $a=5$):

$$E(a) = a \quad \text{The expectation of a constant is that constant}$$

$$E(aX) = aE(X) \quad \text{The expectation of the product of a random variable by a constant is the product of the constant and the expectation of the random variable}$$

$$E(X + Y) = E(X) + E(Y) \quad \text{The expectation of a sum of two variables is the sum of their expectations.}$$

Note that $E(XY) = E(X)E(Y)$ **ONLY IF** X and Y are *independent* – see later

Joint Probability (~ joint frequency)

The joint probability is the probability for given values of two or more random variables to occur together. The joint probability that random variable $X = x$ and random variable $Y = y$ is denoted $\Pr(X = x, Y = y)$.

As an example, assume two genetic loci A and B. The genotypes of a set of individuals are obtained for both loci, resulting in two random variables (G_A and G_B). One obtains a table of the joint probability of carrying specific genotypes at each of the two loci:

Table 2 Example of joint probabilities

Genotype for locus A (G_A)	Genotype for locus B (G_B)			Marginal Prob. for G_A
	bb	Bb	BB	
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg.Prob. G_B	0.30	0.32	0.38	1.00

The entries in the body of this table are the joint probabilities. So, for example the joint probability that an individual has genotypes Aa and BB is: $\Pr(G_A = Aa, G_B = BB) = 0.16$.

Marginal probability (~ marginal frequency)

Marginal probability is used in Table 2 to show the probabilities of, for example, $G_B = bb$, as the sum down a column of joint probabilities. That is,

$$\Pr(G_B = bb) = \Pr(G_B = bb, G_A = aa) + \Pr(G_B = bb, G_A = Aa) + \Pr(G_B = bb, G_A = AA) \\ = 0.1 + 0.14 + 0.6 = 0.30$$

What works in the columns for G_B also works in the rows to get marginal probabilities for G_A .

In general if $\{E_1, E_2, \dots, E_n\}$ is a mutually exclusive and exhaustive set of events (i.e. a set of non-overlapping events that includes the complete parameter space for the variables involved), then marginal probabilities for event I can be calculated as the sum of joint probabilities of event I and events E_i : $\Pr(I) = \sum_{i=1}^n \Pr(E_i, I)$

In Table 2, for example, events $G_A = aa$, $G_A = Aa$, and $G_A = AA$ are mutually exclusive and exhaustive events and marginal probabilities for G_B can be obtained by summing the joint probabilities in a column of Table 2.

Conditional probability

Intuitively, the conditional probability is the probability of a certain event to occur when you already know that another event is true. Alternately, it is the probability of obtaining a given value for one variable (say, $X=x$), conditional on the fact that the value of another variable (say $Y=y$) has already been observed. This conditional probability is denoted $\Pr(X=x | Y=y)$. First, in order to obtain a given value for X (say $X=x$) while Y has another value (say $Y=y$), both conditions have to hold. So we need the joint probability $\Pr(X=x, Y=y)$. Second, because we know that $Y=y$, the parameter space for X is restricted to the subset of events where $Y=y$. All this to help you intuit the definition of conditional probability:

$$\Pr(X = x | Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

In words, the probability of $X=x$ given $Y=y$, is the joint probability of $X=x$ and $Y=y$ divided by the marginal probability of $Y=y$.

Referring back to Table 2, the probability of Aa cows having genotype BB is the probability of $G_B = BB$ conditional on $G_A = Aa$, which is:

$$\Pr(G_B = BB | G_A = Aa) = \frac{\Pr(G_B = BB, G_A = Aa)}{\Pr(G_A = Aa)} = \frac{0.16}{0.48} = 0.333.$$

One way to interpret this conditional probability is as follows: assuming that we have a total of 100 individuals, then on average 48 ($=0.48*100$) will be Aa and of those, on average 16 ($=0.16*100$) will be BB. Thus, the proportion of Aa cows that are BB = $16/48 = 0.333$.

Bayes' Theorem

Sometimes, the conditional probability of X given Y is more difficult to derive than the conditional probability of Y given X . We can then use conditional probabilities to convert one into the other, as follows: $\Pr(X = x | Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$.

Then, using $\Pr(Y = y | X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$,

$$\text{we can write this as: } \Pr(X = x | Y = y) = \frac{\Pr(Y = y | X = x) \Pr(X = x)}{\Pr(Y = y)}$$

This is known as Bayes' Theorem.

For example, suppose somebody tosses a coin three times and gets three heads. What is the probability that this is a double-headed coin, instead of a fair coin?

Let X represent a variable that denotes the state of the coin, i.e. $X = \text{'double'}$ or $X = \text{'fair'}$

Let Y represent the data, in our case $Y = 3$ heads in three tosses.

Thus, we are looking for the following conditional probability: $\Pr(X = \text{double} | Y = 3)$

Using Bayes' theorem, we can also write this as:

$$\Pr(X = \text{double} | Y = 3) = \frac{\Pr(Y = 3 | X = \text{double}) \Pr(X = \text{double})}{\Pr(Y = 3)}$$

Considering each of the three probabilities:

$\Pr(Y=3 | X=\text{double}) = 1$ because every toss will give heads for a double-headed coin

$\Pr(X = \text{double})$ is known as the 'prior' probability of a random coin being double-headed, rather than fair. So what proportion of all coins is double-headed. Let's say that that is 0.01.

$\Pr(Y=3)$ is the probability of getting 3 heads out of 3 tosses for a randomly chosen coin, which can be a double-headed coin with probab=0.01 and a fair coin with prob=0.99

$$\begin{aligned} \text{Thus } \Pr(Y=3) &= \Pr(Y=3 | X=\text{double}) * \Pr(X=\text{double}) + \Pr(Y=3 | X=\text{fair}) * \Pr(X=\text{fair}) \\ &= 1.0 * 0.01 + (0.5)^3 * 0.99 = 0.134 \end{aligned}$$

Filling these probabilities into the Bayes' theorem equation gives:

$$\Pr(X = \text{double} | Y = 3) = \frac{1 * 0.01}{0.134} = 0.075$$

Statistical independence

Random variable X is statistically independent of Y if the probabilities of obtaining different categories of X are the same irrespective of the value of Y .

That is, $\Pr(X = x_i | Y = y_j) = \Pr(X = x_i | Y = y_k) = \Pr(X = x_i)$ for all i, j , and k .

In other words, the conditional probabilities are equal to the marginal probabilities. It follows from the definition of conditional probability that if X is statistically independent of Y , the joint probability is equal to the product of their marginal probabilities:

$$\Pr(X = x_i, Y = y_j) = \Pr(X = x_i) \Pr(Y = y_j).$$

For the example in Table 2, G_A and G_B are NOT independent because, e.g.:

$$\Pr(G_B = \text{BB} | G_A = \text{Aa}) = 0.333 \text{ is NOT equal to } \Pr(G_B = \text{BB}) = 0.38.$$

Also, $\Pr(G_B = \text{BB}, G_A = \text{Aa}) = 0.16$ is NOT equal to the product of the marginal probabilities:

$$\Pr(G_B = \text{BB}) \Pr(G_A = \text{Aa}) = 0.38 * 0.48 = 0.1824$$

Conditional expectation (~ conditional mean or average)

The expectation (=mean) for variable X conditional on variable Y being equal to y is:

$$E(X | Y = y) = \sum_{i=1}^k x_i \Pr(X = x_i | Y = y)$$

and, for continuous variables, $E(X | Y = y) = \int_x x f(x | Y = y) dx$

So conditional expectation is also computed as a weighted average, but now with weights being equal to the conditional probabilities.

For example, in the oat example of Table 1, consider the expectation for the number of florets per spikelet, conditional on the fact that the line carries at least one T allele. From Table 1, first computing the conditional probabilities:

$$\Pr(G = T/t \mid G \text{ contains } T) = \frac{\Pr(G = T/t, G \text{ contains } T)}{\Pr(G \text{ contains } T)} = \frac{\Pr(G = T/t)}{\Pr(G = T/t) + \Pr(G = T/T)} = \frac{0.5}{0.5 + 0.25} = 2/3$$

$$\Pr(G = T/T \mid G \text{ contains } T) = \frac{\Pr(G = T/T)}{\Pr(G = T/t) + \Pr(G = T/T)} = \frac{0.25}{0.5 + 0.25} = 1/3$$

Then the conditional expectation is : $E(X \mid G \text{ contains } T) = 3*(2/3) + 2*(1/3) = 8/3 = 2.67$

Note that this expectation is slightly lower than the overall $E(X)$ (=2.75). So, if we know that the line carries one T allele, we expect the number of florets per spikelet to be slightly lower than average.

Variance

The variance of a random variable is a measure of the spread of a variable over the sample space. Intuitively, we want to know how far we can expect the value of a given variable on average to be from its expected value. That is, we want to know something about the average deviation of the random variable from its expected location. The way to obtain a variance is to find the average of the squared deviation from the mean:

$$\begin{aligned} \text{var}(Y) &= E\{[Y - \mu_Y]^2\} \quad \text{where } \mu_Y = E(Y) \\ &= E\{Y^2 - 2Y\mu_Y + \mu_Y^2\} = E(Y^2) - 2E[Y\mu_Y] + \mu_Y^2 = E(Y^2) - 2\mu_Y\mu_Y + \mu_Y^2 \end{aligned}$$

Thus: $\text{var}(Y) = E\{[Y - \mu_Y]^2\} = E(Y^2) - \mu_Y^2$

Looking back at Table 1, the number of florets per spikelet given different genotypes,

$$\text{var}(X) = 7.75 - (2.75)^2 = 0.1875$$

Note from your statistics class that when we have a sample of N observations for a random variable X (instead of frequencies of the variable attaining certain values), the variance of the sample can be

$$\text{computed as: } \text{var}(X) = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} \quad \text{or as } \text{var}(X) = \frac{\sum_{i=1}^N x_i^2}{N} - \bar{x} \quad \text{where } \bar{x} \text{ is the average of } X$$

Realizing that taking the average is sample equivalent to taking the expectation of a variable, note that these equations are similar to the equations for variances based on expectations, as given above.

Covariance

The covariance between variables X and Y quantifies the (linear) relationship or dependence between X and Y based on the extent to which they “co-vary”.

$$\begin{aligned} \text{Cov}(X, Y) &= E\{(X - \mu_X)(Y - \mu_Y)\} \\ &= E(XY) - \mu_X\mu_Y \quad \text{where } E(XY) = \sum_i \sum_j x_i y_j \Pr(X = x_i, Y = y_j) \end{aligned}$$

Example: The covariance between the genotypic value and the phenotypic value will play a big role in quantitative genetic inferences. Refer back to Table 1, the number of florets per spikelet, conditional on the oat genotype. In Table 1, the genotypic value for the number of florets per spikelet G is considered the same as the phenotypic value for the number of florets per spikelet P . In that case, the covariance between the genotypic and phenotypic values is equal to the variance of the phenotypic values (0.1875, see above). But consider a slightly more complicated situation in which the environment also contributes to determining the phenotype so that:

Table 3 Example for computing covariances

Genotype, T	Probability	Genotypic value G	Phenotypic value P	$\Pr(T) \times GP$	E
t / t	0.20	2.8	3	1.68	0.2
t / t	0.05	2.8	2	0.28	-0.8
T / t	0.30	2.6	3	2.34	0.4
T / t	0.20	2.6	2	1.04	-0.6
T / T	0.05	2.2	3	0.33	0.8
T / T	0.20	2.2	2	0.88	-0.2
Expectation:		2.55	2.55	6.55	0

With this environmental effect, the covariance between genetic and phenotypic values is:

$$\text{Cov}(G, P) = E(GP) - E(G)E(P) = 6.55 - (2.55)^2 = 0.0475.$$

Check that for this specific example, $\text{Cov}(G, P) = \text{Var}(G) = 0.0475$

The variance of phenotype is greater: $\text{Var}(P) = 0.2475$

The model that relates phenotype to genotype is: $P = G + E$ where E represents the effect of environment. So, for the first row in Table 3 the $E = 3-2.8=+0.2$. For the second row: $E=2-2.8=-0.8$. Environmental effects are in the last column of Table 3. Note that $E(E)=0$. You can also check that:

$\text{Cov}(G, E) = 0$ (i.e. environmental effects are independent of genetic effects)

$\text{Cov}(P, E) = 0.2$

$\text{Var}(E) = 0.2$

Properties of Variance and Covariance

Assuming again that a is a constant:

$\text{Var}(a) = 0$ The variance of a constant is zero

$\text{Var}(aX) = a^2 \text{Var}(X)$ The variance of the product of a variable by a constant is the product of the constant squared and the variable's variance

$\text{Cov}(X, Y) = \text{Cov}(Y, X)$

$\text{Cov}(X, aY) = a\text{Cov}(X, Y)$

$\text{Cov}(X, Y+Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ The variance of a sum is the sum of variance plus twice the covariance

$$\begin{aligned} \text{(for the Table 3 example: } \text{Var}(P) &= \text{Var}(G+E) = \text{Var}(G) + \text{Var}(E) + 2\text{Cov}(G, E) = \\ &= 0.0475 + 0.2 + 2 * 0 = 0.2475 \end{aligned}$$

Generalizing the equation for $\text{Var}(X+Y)$ to the sum of many variables:

$$\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i) + 2\sum_{i < j} \text{Cov}(X_i, X_j) \quad \text{If } X_i \text{'s are independent } \Rightarrow \text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$$

$$\text{Also: } \text{Var}(X-Y) = \text{Var}[X+(-Y)] = \text{Var}(X) + \text{Var}(-1*Y) + 2\text{Cov}[X, (-1*Y)] =$$

$$= \text{Var}(X) + (-1)^2 \text{Var}(Y) + 2*(-1)*\text{Cov}(X, Y) =$$

$$= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

$$\begin{aligned}
 \text{Cov}(X, X) &= E(XX) - E(X)E(X) \\
 &= E(X^2) - [E(X)]^2 \\
 &= \text{Var}(X) \quad \rightarrow \text{the covariance of a variable with itself is its variance}
 \end{aligned}$$

If X and Y are independent: $E(XY) = \sum_i \sum_j x_i y_j \Pr(X = x_i, Y = y_j)$

$$\begin{aligned}
 &= \sum_i \sum_j x_i y_j \Pr(X = x_i) \Pr(Y = y_j) \\
 &= [\sum_i x_i \Pr(X = x_i)] [\sum_j y_j \Pr(Y = y_j)] \\
 &= E(X)E(Y)
 \end{aligned}$$

So that $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$

Correlation

The correlation measures the (linear) relationship between two variables on a standardized scale, by dividing their covariance by the product of their standard deviations:

$$r_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad \text{Note that: } -1 \leq r_{XY} \leq 1$$

For the example of Table 3: $r_{GP} = \frac{\text{Cov}(G, P)}{\sqrt{\text{Var}(G)\text{Var}(P)}} = \frac{0.0475}{\sqrt{0.0475 * 0.2475}} = 0.438$

Based on rearrangement of the correlation equation, we get the following expression for the covariance, which we also frequently use:

$$\text{Cov}(X, Y) = r_{XY} \sqrt{\text{Var}(X)\text{Var}(Y)}$$

Regression

A repeated theme in quantitative genetics is the estimation of quantities associated with individuals or parameters associated with populations when those quantities or parameters are themselves not directly observable. The most obvious example is the desire to estimate an individual's genotypic value for a trait when the only information we have available derives from the individual's phenotype. Regression is used for this kind of estimation.

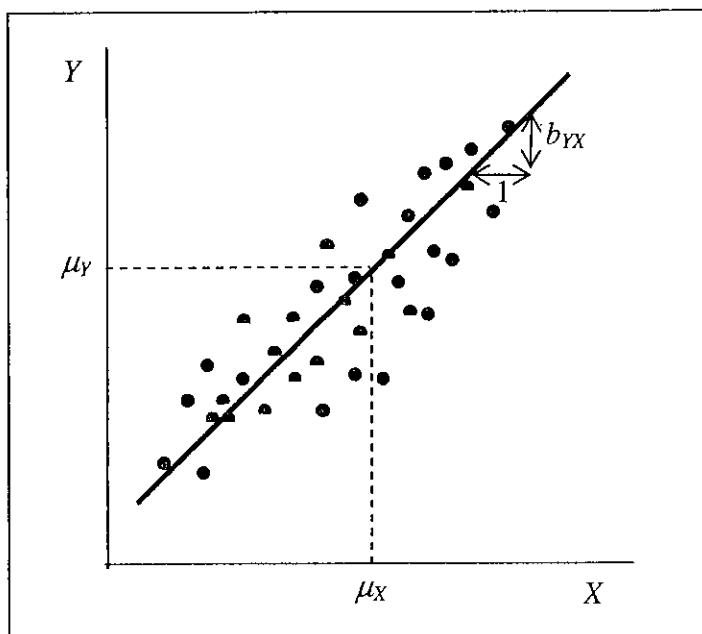
Definition: The regression of Y on X is the expected value of Y conditional on having a certain value for variable X :

$$\hat{y} = E(Y|X)$$

This is also called the *best (linear) predictor* of Y given X .

Regression can be used to define a model:

$y = \hat{y} + e$ where e is called the residual, which is the deviation of the observed value for Y from its expected value conditional on X .



For quantitative variables, the predicted value for Y can be derived using linear regression:

$$\hat{y} = E(Y|X) = \mu_Y + b_{YX}(x - \mu_X)$$

with $\mu_Y = E(Y)$ $\mu_X = E(X)$

b_{YX} = coefficient of regression of Y on X = expected change in Y per 1 unit increase in X

Given data, b_{YX} can be derived by fitting the following linear regression model:

$$y = \mu_Y + b_{YX}(x - \mu_X) + e$$

Using least squares (see Lynch & Walsh p39), b_{YX} can be derived to be equal to:

$$b_{YX} = \text{Cov}(Y, X) / \text{Var}(X)$$

Note that b_{YX} can also be expressed in terms of the correlation coefficient:

$$b_{YX} = \text{Cov}(Y, X) / \text{Var}(X) = r_{XY} \sqrt{\text{Var}(Y) \text{Var}(X)} / \text{Var}(X) = r_{XY} \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}}$$

So the important equations to remember for the regression coefficient are:

$$b_{YX} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = r_{XY} \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}}$$

Note that these only hold for simple regression with a single independent variable (X).

For the example of Table 3, suppose we want to predict the genotypic value of an individual based on its observed phenotypic value. We would use the following regression model:

$$G = \bar{G} + b_{GP}(P - \bar{P}) + e \quad \text{with } \bar{G} = E(G) = E(P) = \bar{P} = 2.55$$

The regression coefficient can be computed as:

$$b_{GP} = \text{Cov}(G, P) / \text{Var}(P) = 0.0475 / 0.2475 = 0.192$$

$$\text{or } b_{GP} = r_{GP} \sqrt{\frac{\text{Var}(G)}{\text{Var}(P)}} = 0.438 \sqrt{0.0475 / 0.2475} = 0.192$$

So the prediction model is: $\hat{G} = \bar{G} + b_{GP}(P - \bar{P}) = 2.55 + 0.192(P - 2.55)$.

Results are in Table 4. The last column in this table shows the prediction error: $\hat{e} = G - \hat{G}$

Table 4 Example prediction based on linear regression

Genotype, T	Probability	G	P	E	\hat{G}	\hat{e}
t/t	0.20	2.8	3	0.2	2.636	0.164
t/t	0.05	2.8	2	-0.8	2.444	0.356
T/t	0.30	2.6	3	0.4	2.636	-0.036
T/t	0.20	2.6	2	-0.6	2.444	0.156
T/T	0.05	2.2	3	0.8	2.636	-0.436
T/T	0.20	2.2	2	-0.2	2.444	-0.244
Expectation:		2.55	2.55	0	2.55	0.0004

Properties of Regression

1. The average of predicted values is equal to the average of Y's: $E(\hat{Y}) = E(Y) = \mu_Y$

$$E(\hat{y}) = E[\mu_Y + b_{YX}(x - \mu_X)] = E(\mu_Y) + E[b_{YX}(x - \mu_X)] = \mu_Y + b_{YX}[E(x) - \mu_X] = \mu_Y$$

This also implies that the regression line always passes through the mean of both X and Y; substituting μ_X for x into the prediction equation gives $\hat{y} = \mu_Y$

2. The average value of the residual is zero: $E(e) = 0$

$$\begin{aligned} E(e) &= E(Y - \hat{Y}) && \text{from regression model} \\ &= E(Y) - E(\hat{Y}) && \text{property of expectation} \\ &= 0 && \text{from property 1 above} \end{aligned}$$

3. The expectation of the residual is zero for all values of X: $E(e|X) = 0$

$$\begin{aligned} E(e|X) &= E(Y - \hat{Y}|X) && \text{from regression model} \\ &= E(Y|X) - E(\hat{Y}|X) && \text{property of expectation} \\ &= \hat{Y} - \hat{Y} = 0 && \text{by definition of regression} \end{aligned}$$

This implies that predictions of Y are on average equal to the true Y across the range of possible values for X.

4. **Accuracy of prediction** = $\text{Corr}(\hat{Y}, Y) = r_{\hat{y}y}$

The accuracy of the prediction equation is equal to the correlation of \hat{y} with its true value y. We can derive accuracy as:

$$\text{Accuracy} = r_{\hat{y}y} = \frac{\text{Cov}(\hat{y}, y)}{\sqrt{\text{Var}(\hat{y})\text{Var}(y)}} = \frac{\text{Cov}(\mu_Y + b_{YX}(x - \mu_X), y)}{\sqrt{\text{Var}(\mu_Y + b_{YX}(x - \mu_X))\text{Var}(y)}}$$

Since μ_Y and μ_X are constants, this simplifies to:

$$\text{Accuracy} = \frac{\text{Cov}(b_{YX}x, y)}{\sqrt{\text{Var}(b_{YX}x)\text{Var}(y)}} = \frac{b_{YX}\text{Cov}(x, y)}{\sqrt{b_{YX}^2\text{Var}(x)\text{Var}(y)}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = r_{XY}$$

So the accuracy of a prediction equation based on simple (= 1-variable) regression is equal to the correlation between the dependent and independent variables.

5. **Decomposition of variance in Y into that explained by the prediction and unexplained variance**

Using the above equation, we can also show that the variance of Y is the sum of the variance explained by the regression on X and residual variance (note that $\text{Cov}(X, e) = 0$):

$$\text{Var}(y) = \text{Var}(\mu_Y + b_{YX}(x - \mu_X) + e) = b_{YX}^2\text{Var}(x) + \text{Var}(e) = [\text{Cov}(y, x)]^2/\text{Var}(x) + \text{Var}(e)$$

Note that because $\text{Cov}(y, x) = r_{XY}\sqrt{\text{Var}(x)\text{Var}(y)}$ the first term can also be written as:

$$[\text{Cov}(y, x)]^2/\text{Var}(x) = r_{XY}^2 \text{Var}(x) \text{Var}(y) / \text{Var}(x) = r_{XY}^2 \text{Var}(y) = r_{\hat{y}y}^2 \text{Var}(y)$$

This is the variance in Y that is explained by the X through the prediction model

By subtraction we get $\text{Var}(e) = [1 - r_{XY}^2]\text{Var}(y)$. This is the unexplained/residual variance.

Thus, variance of Y can be decomposed as: $\text{Var}(y) = r_{XY}^2\text{Var}(y) + [1 - r_{XY}^2]\text{Var}(y)$

Note that the variance of predicted values is equal to the explained variance:

$$\begin{aligned}\text{Var}(\hat{y}) &= \text{Var}[\mu_Y + b_{YX}(x - \mu_X)] = b_{YX}^2 \text{Var}(x) = \left\{ \frac{\text{Cov}(y, x)}{\text{Var}(x)} \right\}^2 \text{Var}(x) = \\ &= \frac{[\text{Cov}(y, x)]^2}{\text{Var}(x)} = \frac{[\text{Cov}(y, x)]^2}{\text{Var}(x)\text{Var}(y)} \text{Var}(y) = r_{YX}^2 \text{Var}(y)\end{aligned}$$

So the variance of predicted values is equal to the variance explained by the model, which depends on the correlation between Y and X .

The above equations apply when prediction is based on one variable (x), in which case $r_{yy} = r_{xy}$.

In general, prediction can be based on multiple x 's = multiple regression. In that case the partitioning of variance is: $\text{Var}(y) = \text{Var}(\hat{y}) + \text{Var}(e) = r_{yy}^2 \text{Var}(y) + [1 - r_{yy}^2] \text{Var}(y)$

6. Residuals are uncorrelated with the predictor variable, X :

$$\text{Cov}(X, e) = 0$$

$$\begin{aligned}\text{Cov}(x, e) &= \text{Cov}[x, y - \hat{y}] = \text{Cov}[x, y - (\mu_Y + b_{YX}(x - \mu_X))] = \\ &= \text{Cov}(x, y) - \text{Cov}(x, \mu_Y) - b_{YX} \text{Cov}(x, x) - b_{YX} \text{Cov}(x, \mu_X) \\ &= \text{Cov}(x, y) - 0 - b_{YX} \text{Var}(x) - 0 \\ &= \text{Cov}(x, y) - \frac{\text{Cov}(y, x)}{\text{Var}(x)} \text{Var}(x) = 0\end{aligned}$$

7. Residuals are uncorrelated with the predictions:

$$\text{Cov}(\hat{y}, e) = 0$$

$$\begin{aligned}\text{Cov}(\hat{y}, e) &= \text{Cov}[\hat{y}, y - \hat{y}] = \text{Cov}(\hat{y}, y) - \text{Var}(\hat{y}) = \\ &= \text{Cov}(x, y) - \text{Cov}(x, \mu_Y) - b_{YX} \text{Cov}(x, x) - b_{YX} \text{Cov}(x, \mu_X) \\ &= \text{Cov}(x, y) - 0 - b_{YX} \text{Var}(x) - 0 \\ &= \text{Cov}(x, y) - \frac{\text{Cov}(y, x)}{\text{Var}(x)} \text{Var}(x) = 0\end{aligned}$$

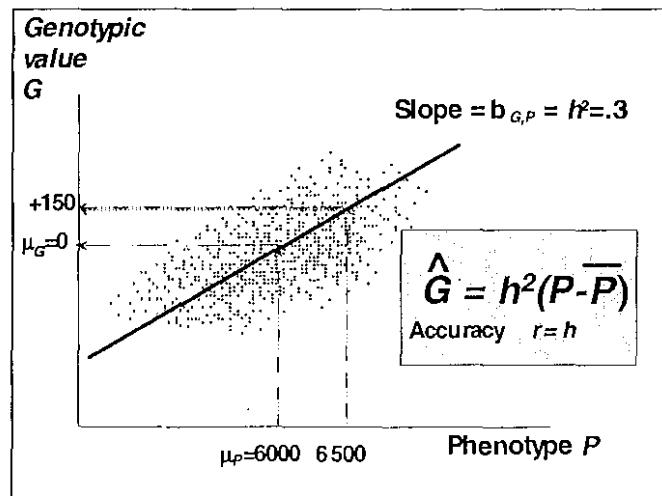
Properties 6 and 7 imply that all information on Y that is contained in X is captured in the predicted values, as the residual is uncorrelated to both X and the predicted values.

Some Applications to Quantitative Genetic Theory

The standard quantitative genetics model equation for the observed phenotype of an individual i for a quantitative trait (P_i) is that it is the sum of the effect of genetics (the genotypic value G_i) and the effect of environment (E_i): $P_i = \mu_P + G_i + E_i$

In practice, we only observe phenotype and cannot directly observe G_i or E_i . However, if we could observe both P_i and G_i for a group of individuals, we could plot genotypic against phenotypic values, as in the figure below.

Using such a hypothetical plot, or model, and statistics such as correlation, covariance, variance, and regression, we can specify important population parameters such as heritability (h^2) and make a number of inferences or predictions, such as predicting an individual's genotypic value or 'breeding value' from its observed phenotype:



1) Covariance and correlation between phenotypic and genotypic values:

Based on $P_i = \mu_P + G_i + E_i$

$$\text{Cov}(P, G) = \text{Cov}(\mu + G + E, G) =$$

$$= \text{Cov}(G, G) + \text{Cov}(G, E) = \text{Var}(G)$$

The last step assumes that $\text{Cov}(G, E) = 0$, i.e. that the environment that an individual receives is independent of its genotypic value. The result of this covariance, $\text{Var}(G)$, which is often denoted σ_G^2 , is the **genetic variance** in the population, i.e. the variance of genotypic values of individuals in a population. This in contrast to the phenotypic variance, $\text{Var}(P)$, often denoted σ_P^2 , which is the variance of phenotypic values of individuals in a population.

Then, the correlation between phenotypic and genotypic values can be derived as:

$$r_{P,G} = \frac{\text{Cov}(G, P)}{\sqrt{\text{Var}(G) \text{Var}(P)}} = \frac{\sigma_G^2}{\sqrt{\sigma_G^2 \sigma_P^2}} = \frac{\sigma_G}{\sigma_P}$$

Thus, the correlation between genotypic and phenotypic values of individuals in a population is equal to the ratio of the genetic and phenotypic standard deviations for the trait.

The square of this correlation, therefore, is equal to the ratio of the genetic and phenotypic variances, or to the proportion of phenotypic variance that is genetic. This proportion is also

defined as the **heritability** of the trait ($= h^2$). Thus: $(r_{P,G})^2 = \frac{\sigma_G^2}{\sigma_P^2} = h^2$

2) Regression of genotypic on phenotypic values:

Using the above model and referring to the figure, we can also set up a regression equation between the genotypic and phenotypic values to predict G :

$$G_i = \mu_G + b_{G,P}(P_i - \mu_P) + e_i \quad \text{where } b_{G,P} \text{ is the coefficient of regression of } G \text{ on } P.$$

This regression coefficient can be derived as: $b_{G,P} = \frac{\text{Cov}(G, P)}{\text{Var}(P)} = \frac{\sigma_G^2}{\sigma_P^2} = h^2$

Thus, the slope of the regression of genotypic on phenotypic values is equal to heritability

3) Prediction of genotypic values:

The above regression model can be used to predict an individual's genotypic value based on its observed phenotype, using the following prediction equation:

$$\hat{G}_i = \mu_G + h^2(P_i - \mu_P)$$

In practice, we often set μ_G to zero, because we're primarily interested in ranking individuals in a population. Thus: $\hat{G}_i = h^2(P_i - \mu_P)$

As an **example** (see figure), assume a dairy cow produces 6500 kg milk, which is its phenotypic value (P_i). The mean production of the herd she is in is 6000 kg ($= \mu_P$).

Milk production is a trait with an (assumed known) heritability of 0.3, a phenotypic standard deviation of 1200 kg ($\sigma_P = 1200$). Using $h^2 = \frac{\sigma_G^2}{\sigma_P^2}$ and, thus, $\sigma_G^2 = h^2 \sigma_P^2$, the genetic standard

deviation for milk yield is equal to $\sigma_G = h\sigma_P = \sqrt{0.3} * 1200 = 657.3$ kg

Then, this cow's genotypic value can be predicted to be:

$$\hat{G}_i = h^2(P_i - \mu_P) = 0.3 (6500 - 6000) = +150 \text{ kg}$$

So this cow's genotypic value is expected to be 150 kg greater than the average in this herd.

We can also attach an accuracy to this prediction, based on the previously derived result that the correlation between predicted and true values based on linear regression is equal to the correlation to the dependent (Y) and independent (X) variables: $r_{\hat{G},G} = r_{G,P} = h$

Thus, when predicting an individual's genotypic value based on its phenotypic value, the accuracy of this prediction will be equal to the square-root of heritability of the trait.

When we predict genotypic values for all individuals in a population in this manner, and take the variance of these predicted values, we expect this variance to be equal to (based on property 5):

$$\text{Var}(\hat{y}) = r_{YX}^2 \text{Var}(y)$$

which in this case simplifies to: $\text{Var}(\hat{G}) = h^2 \sigma_G^2 = h^4 \sigma_p^2$

And, using property 5 above, the variance of prediction errors ($e_i = G - \hat{G}$) is equal to:

$$\text{Var}(e) = [1 - r_{XY}^2] \text{Var}(y)$$

which in this case simplifies to: $\text{Var}(e) = (1-h^2) \sigma_G^2$

For the example, the variance of predicted values is: $\text{Var}(\hat{G}) = h^2 \sigma_G^2 = 0.3 * 657.3^2 = 129600$

and the variance of prediction errors is: $\text{Var}(e) = (1-h^2) \sigma_G^2 = 0.7 * 657.3^2 = 302430$

Note that these two variances sum to the genetic variance: $129600 + 302430 = 43203 = 657.3^2$

Based on $\text{Var}(e) = 302430 \text{ kg}^2$ we can also add a confidence interval to our prediction (see later).

4) Regression of offspring phenotype on parent phenotype

One of the problems with predicting genotypic values, as described above, is that it requires you to know the heritability of the trait. Luckily, we can also get estimates of heritability for a trait from phenotypic data. We do this by observing how similar the phenotype of offspring is to that of their parents; if these are very similar, we expect the trait to be more heritable.

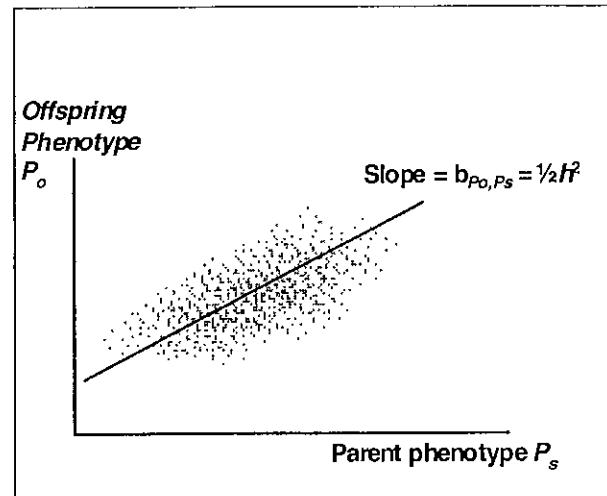
When we have phenotypes observed on offspring and their sires, we can estimate heritability by regressing the phenotype of the offspring on that of their parents, as illustrated below:

The regression model is:

$$P_o = \mu_o + b_{PoPs}(P_s - \mu_s) + e$$

The regression coefficient can be derived as:

$$\begin{aligned} b_{PoPs} &= \frac{\text{Cov}(P_o, P_s)}{\text{Var}(P_s)} = \frac{\text{Cov}(G_o + E_o, G_s + E_s)}{\sigma_p^2} \\ &= \frac{\text{Cov}(G_o, G_s) + \text{Cov}(G_o, E_s) + \text{Cov}(E_o, G_s) + \text{Cov}(E_o, E_s)}{\sigma_p^2} = \frac{\text{Cov}(G_o, G_s)}{\sigma_p^2} \end{aligned}$$



The last step assumes that the environment that the offspring progeny received is independent of the phenotype of the sire (a sometimes strong assumption), making the last 3 covariance terms 0.

To derive the covariance between the genotypic value of offspring and that of their sire, we can express the genotypic value of the offspring in terms of the genotypic values of its sire (s) and dam (d), based on the fact that half of the genes that the offspring have come from each parent:

$$G_o = \frac{1}{2} G_s + \frac{1}{2} G_d + RA_s + RA_d$$

Here the terms RA_s and RA_d are random assortment or Mendelian sampling terms, which reflect that parents pass on a random half of their alleles (i.e. a random one of two alleles at each locus). Using this genetic model (which has quite a number of assumptions, which well be covered later), we can continue our derivation as:

$$\begin{aligned}\text{Cov}(G_o, G_s) &= \text{Cov}(\frac{1}{2}G_s + \frac{1}{2}G_d + RA_s + RA_d, G_s) = \\ &= \text{Cov}(\frac{1}{2}G_s, G_s) + \text{Cov}(\frac{1}{2}G_d, G_s) + \text{Cov}(RA_s, G_s) + \text{Cov}(RA_d, G_s)\end{aligned}$$

Assuming random mating and the fact that Mendelian sampling terms are independent (see later), the last three covariance terms are zero, resulting in:

$$b_{Pops} = \frac{\text{Cov}(G_o, G_s)}{\sigma_p^2} = \frac{\text{Cov}(\frac{1}{2}G_s, G_s)}{\sigma_p^2} = \frac{\frac{1}{2}\text{Cov}(G_s, G_s)}{\sigma_p^2} = \frac{\frac{1}{2}\sigma_G^2}{\sigma_p^2} = \frac{1}{2}h^2$$

Thus, heritability of a trait can be estimated based on phenotypes of relatives, by measuring the degree of resemblance between relatives, using statistics such as linear regression. More on this later.

Some Distributions useful in Population and Quantitative Genetics

Bernoulli distribution.

Named after the mathematician Daniel Bernoulli, 1700-1782. A Bernoulli random variable is characterized by one parameter, that is typically designated p and is sometimes called the “probability of success”. The random variable can have one of two values: 1 with probability p and 0 with probability $1 - p$.

If Y is a Bernoulli random variable with probability p , its expectation is:

$$E(Y) = \sum_{i=1}^2 y_i \Pr(Y = y_i) = 0(1 - p) + 1(p) = p$$

Its variance is

$$\begin{aligned}\text{var}(Y) &= E(Y^2) - E(Y)^2 \\ &= [0^2(1 - p) + 1^2(p)] - p^2 \\ &= p - p^2 \\ &= p(1 - p)\end{aligned}$$

The Bernoulli distribution is used in population and quantitative genetics in relation to the presence or inheritance of alleles at a locus. For example, for a locus with two possible alleles, A and a , and with the frequency of allele A in the population denoted by p , then the process of drawing one allele at this locus from a population can be specified by a Bernoulli distribution by specifying a variable Y that is equal to 1 if allele A is drawn and equal to 0 if allele a is drawn.

Binomial Distribution

The Binomial distribution is based on the Bernoulli distribution. A binomial random variable is the sum of k independent Bernoulli random variables all with parameter p . The binomial is therefore characterized by two parameters, k and p and can have integer values from 0 to k . If X is binomially distributed with k trials and p probability of success: $X \sim \text{Binomial}(k, p)$, then:

From the properties of expectation of a sum, the **expected value** of X is kp : $E(X) = kp$.

From the properties of variance of a sum of independent variables, the **variance** of X is

$$kp(1 - p): \text{var}(X) = kp(1 - p)$$

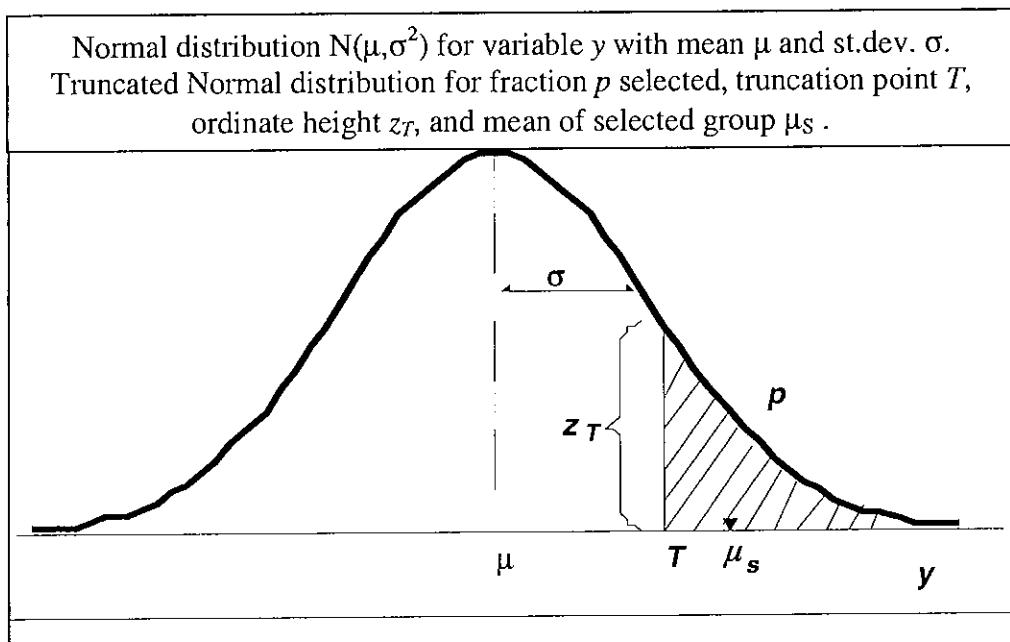
The probability density function $\Pr(X = x)$ is

$$\Pr(X = x) = \binom{k}{x} p^x (1-p)^{k-x} \text{ where } \binom{k}{x} = \frac{k!}{x!(k-x)!} \text{ and } a! = 1*2*3*\dots*a$$

When considering population or quantitative genetics, the Binomial Distribution could correspond to the process of randomly drawing k alleles at a locus from a population.

Normal or Gaussian distribution.

This is perhaps the most important distribution in quantitative genetics, as phenotypes for most quantitative traits approximately follow a normal distribution, or can be transformed to follow a normal distribution. This is a property of the fact that phenotype is the sum of many genetic factors and of many environmental factors. Following the Central Limit Theorem of statistics, this is expected to result in a Normal distribution, even if the distribution of variables that are included in the sum is not Normal. See also Falconer and MacKay Chapter 6.



The probability distribution function for a variable y that has a Normal distribution with mean μ and

standard deviation σ , denoted by $y \sim N(\mu, \sigma^2)$ is:

$$\Pr(y) = z = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

It is often useful to work with the *Standard Normal* distribution, which has mean zero and standard deviation 1: $N(0,1)$ Any Normally distributed variable can be ‘standardized’ to a variable that follows $N(0,1)$ by subtracting the mean and dividing by the standard deviation:

$$\text{If } y \sim N(\mu, \sigma^2) \quad \text{then} \quad y' = (y - \mu)/\sigma \quad \text{follows } N(0,1)$$

Truncated Normal distribution.

In plant and animal breeding, we often are interested in using individuals with the highest phenotype for breeding. If phenotype (y) is Normally distributed ($y \sim N(\mu, \sigma^2)$) then it is of interest to know something about the distribution of phenotypes of the selected individuals. This is the Truncated Normal distribution, as illustrated in the figure above:

Selecting a proportion p of individuals from a population based on phenotype (y) is equivalent to truncating the Normal distribution at a truncation point T , such that a fraction p falls above the truncation point.

The mean phenotype for the selected individuals is denoted by μ_S (see Figure).

The difference between the mean of the selected individuals over that of all individuals is called the selection differential: $S = \mu_S - \mu$

Maximum Likelihood Estimation

Maximum Likelihood (ML) is a procedure for estimating parameters from an observed set of data. It was introduced by Fisher and is widely used in population and quantitative genetics.

The basic idea of ML estimation is to find the value of the parameter(s) that is ‘most likely’ to have produced the data that is observed, i.e. that maximizes the likelihood of getting the data that you got.

As a simple example to illustrate the concept of ML estimation, consider the following observed genotype frequencies.

Table 1. Falconer and Mackay, p. 1, blood group categories in Iceland:

Blood Group	Counts	Probabilities
MM	233	$P = 233 / 747 = 0.312$
MN	385	$H = 385 / 747 = 0.515$
NN	129	$Q = 129 / 747 = 0.173$
Total	747	$747 / 747 = 1.000$

P , H , and Q are the estimated genotype frequencies – obtained by counting

To estimate allele/gene frequencies, we could obtain these simply by counting: 2 * 747 alleles were sampled; the number of M alleles is $(2P + H) * 747$. Thus, the allele/gene frequency of allele M is

$$p = \frac{(2P + H) * 747}{2 * 747} = P + \frac{1}{2}H$$

So $p = 0.312 + 0.515 / 2 = 0.57$ and $q = 0.173 + 0.515 / 2 = 0.43$.

This estimates of allele frequency obtained by counting is actually an ML estimate: for the example of Table 1, if 57% of all alleles in the sample is M (vs. N, as is observed in the sample), then the ML estimate of p , the frequency of M in the population that the sample came from, is 0.57, because that is the value of p that is most likely to have produced a sample with 57% of alleles being M.

A more formal derivation of this estimate uses the Binomial distribution to specify the **Likelihood** of the data as a function of the parameter (= **Likelihood function**): if out of n alleles sampled n_M are M, then the likelihood to get these counts given the population frequency of M is equal to the probability that the value of a Binomial variable with parameters n and p is equal to n_M :

$$\text{Likelihood}(\text{data} | p) = Pr(\text{data} | p) = \binom{n}{n_M} p^{n_M} (1-p)^{n-n_M}$$

For the data in Table 1 $n = 2 * 747 = 1494$ and $n_M = 2 * 233 + 385 = 851$

$$\text{So: } \text{Likelihood}(\text{data} | p) = \binom{1494}{851} p^{851} (1-p)^{643}$$

Now the ML estimate of p is the value of p that maximizes the above function. To find this value we can take the first derivative of the *Likelihood* and set it equal to zero. However, it is often easier to first take the natural log of the *Likelihood* and to maximize it for p :

$$L(\text{data} | p) = \ln \left\{ \binom{n}{n_M} p^{n_M} (1-p)^{n-n_M} \right\}$$

Then, using some algebra, this can be ‘simplified’ to:

$$\begin{aligned} L(\text{data} | p) &= \ln \left(\binom{n}{n_M} \right) + \ln(p^{n_M}) + \ln((1-p)^{n-n_M}) = \\ &= \ln \left(\binom{n}{n_M} \right) + n_M \ln(p) + (n - n_M) \ln(1-p) \end{aligned}$$

The first derivative of the LogLikelihood with respect to p is: $n_M \frac{1}{p} + (n - n_M) \frac{-1}{1-p}$

Setting this to zero to find the maximum and solving for the ML estimate of p , \hat{p} , gives:

$$\begin{aligned} n_M \frac{1}{\hat{p}} + (n - n_M) \frac{-1}{1-\hat{p}} &= 0 \rightarrow n_M \frac{1}{\hat{p}} = (n - n_M) \frac{1}{1-\hat{p}} \rightarrow \frac{1-\hat{p}}{\hat{p}} = \frac{n - n_M}{n_M} \\ \rightarrow \frac{1}{\hat{p}} - 1 &= \frac{n}{n_M} - 1 \quad \rightarrow \frac{1}{\hat{p}} - 1 = \frac{n}{n_M} - 1 \quad \rightarrow \hat{p} = \frac{n_M}{n}, \text{ i.e. count estimate.} \end{aligned}$$

This is obviously a simple example, where we don’t need ML estimation to obtain a good estimate (we can just count).

Another (obvious) example is the following: Suppose n values, y_1, y_2, \dots, y_n , are sampled independently from an underlying Normal distribution with unknown mean μ and variance 1. What is the MLE for μ given the data?

Let’s denote the data by a vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$. Using the probability density function of the Normal distribution with mean μ and standard deviation 1, the likelihood for a given data point y_i

given the mean, μ , is: $Likelihood(y_i | \mu) = \Pr(y_i | \mu) \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{(y_i - \mu)^2}{2} \right]}$ Because each

observation is independent, the likelihood function for all observation \mathbf{y} is the product of n normal density functions:

$$Likelihood(\mathbf{y} | \mu) = \Pr(\mathbf{y} | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{(y_i - \mu)^2}{2} \right]} = (2\pi)^{-n/2} \sum_{i=1}^n e^{\left[-\frac{(y_i - \mu)^2}{2} \right]}$$

Again, taking the natural log of the likelihood: $L(\mathbf{y} | \mu) = -\left(\frac{n}{2}\right) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2$

Maximizing by taking the first derivative gives:

$$\frac{\partial L(\mu | y)}{\partial \mu} = \sum_{i=1}^n (y_i - \mu) = n(\bar{y} - \mu) \quad \text{where } \bar{y} \text{ is the average of the observations}$$

Setting this equal to zero gives: $n(\bar{y} - \mu) = 0 \rightarrow$ the MLE of μ is: $\hat{\mu} = \bar{y}$

Again, this is obvious but it does illustrate the principle behind the use of ML to estimate parameters in more complex situation. For example, if we want to estimate a parameter such as heritability from data (y) we have observed in a pedigree population, we can formally state the problem by that of finding the MLE of heritability, given the observed data; i.e. what is the most likely value of heritability that would have given rise to the data that we observed. To do this, we need to formulate the Likelihood function, or the log of the likelihood, and maximize it.

$$\text{Likelihood}(\text{data} | h^2) = \Pr(\text{data} | h^2)$$

This is the basis of ML procedures for estimation of genetic parameters.

A Review of Elementary Matrix Algebra

Notes developed by John Gibson for Economics of Animal Breeding Strategies notes
(Dekkers, Gibson, van Arendonk)

Dr. B.W. Kennedy originally prepared this review for use alongside his course in Linear Models in Animal Breeding. His permission to use these notes is gratefully acknowledged. Not all the operations outlined here are necessary for this course, but most would be necessary for some applications in animal breeding.

A much more complete treatment of matrix algebra can be found in "Matrix Algebra Useful for Statistics" by S.R. Searle. See also Chapter 8 of Lynch and Walsh.

A.1 Definitions

A matrix is an ordered array of numbers. For example, an experimenter might have observations on a total of 35 animals assigned to three treatments over two trials as follows:

Treatment	Trial	
	1	2
1	6	4
2	3	9
3	8	5

The array of numbers of observations can be written as a matrix as

$$\mathbf{M} = \begin{bmatrix} 6 & 4 \\ 3 & 9 \\ 8 & 5 \end{bmatrix}$$

with rows representing treatments (1,2,3) and columns representing trials (1,2).

The numbers of observations then represent the elements of matrix \mathbf{M} . The order of a matrix is the number of rows and columns it consists of. \mathbf{M} has order 3×2 .

A vector is a matrix consisting of a single row or column. For example, observations on 3 animals of 3, 4 and 1, respectively, can be represented as column or row vectors as follows:

A column vector: $\mathbf{x} = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix}$

A row vector: $\mathbf{x}' = [3 \ 4 \ 1]$

A scalar is a single number such as 1, 6 or -9.

A.2 Matrix Operations

A.2.1 Addition

If matrices are of the same order, they are conformable for addition. The sum of two conformable matrices, is the matrix of sums element by element of the two matrices. For example, suppose \mathbf{A} represents observations on the first replicate of a 2×2 factorial experiment, \mathbf{B} represents observations on a second replicate and we want the sum of each treatment over replicates. This is given by matrix $\mathbf{S} = \mathbf{A} + \mathbf{B}$.

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 1 & 9 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} -4 & 6 \\ 5 & 2 \end{bmatrix},$$
$$\mathbf{S} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 - 4 & 5 + 6 \\ 1 + 5 & 9 - 2 \end{bmatrix} = \begin{bmatrix} -2 & 11 \\ 6 & 7 \end{bmatrix}.$$

A.2.2 Subtraction

The difference between two conformable matrices is the matrix of differences element by element of the two matrices. For example, suppose now we want the difference between replicate 1 and replicate 2 for each treatment combination, i.e. $\mathbf{D} = \mathbf{A} - \mathbf{B}$,

$$\mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} 2 + 4 & 5 - 6 \\ 1 - 5 & 9 + 2 \end{bmatrix} = \begin{bmatrix} 6 & -1 \\ -4 & 11 \end{bmatrix}.$$

A.2.3 Multiplication

Scalar Multiplication

A matrix multiplied by a scalar is the matrix with every element multiplied by the scalar. For example, suppose \mathbf{A} represents a collection of measurements taken on one scale which we would like to convert to an alternative scale, and the conversion factor is 3.

For a scalar $\lambda = 3$. $\lambda\mathbf{A} = 3 \begin{bmatrix} 2 & 5 \\ 1 & 9 \end{bmatrix} = \begin{bmatrix} 6 & 15 \\ 3 & 27 \end{bmatrix}.$

Vector Multiplication

The product of a row vector with a column vector is a scalar obtained from the sum of the products of corresponding elements of the vectors. For example, suppose \mathbf{v} represents the number of observations taken on each of 3 animals and that \mathbf{y} represents the mean of these observations on each of the 3 animals and we want the totals for each animal.

$$\mathbf{v}' = [3 \quad 4 \quad 1] \quad \mathbf{y} = \begin{bmatrix} 1 \\ 5 \\ 2 \end{bmatrix},$$

$$t = \mathbf{v}'\mathbf{y} = [3 \quad 4 \quad 1] \begin{bmatrix} 1 \\ 5 \\ 2 \end{bmatrix} = 3(1) + 4(5) + 1(2) = 25.$$

Matrix Multiplication

Vector multiplication can be extended to the multiplication of a vector with a matrix, which is simply a collection of vectors. The product of a vector and a matrix is a vector and is obtained as follows:

$$\text{e.g. } \mathbf{v}' = [3 \quad 4 \quad 1] \quad \mathbf{M} = \begin{bmatrix} 6 & 4 \\ 3 & 9 \\ 8 & 5 \end{bmatrix}$$

$$\mathbf{v}'\mathbf{M} = [3 \quad 4 \quad 1] \begin{bmatrix} 6 & 4 \\ 3 & 9 \\ 8 & 5 \end{bmatrix}$$

$$= [3(6) + 4(3) + 1(8) \quad 3(4) + 4(9) + 1(5)]$$

$$= [38 \quad 53]$$

That is, each column (or row) of the matrix is treated as a vector multiplication.

This can be extended further to the multiplication of matrices. The product of two conformable matrices is illustrated by the following example:

$$\mathbf{A} \times \mathbf{B} = \begin{bmatrix} 2 & 5 \\ 1 & 9 \end{bmatrix} \begin{bmatrix} 4 & -6 \\ -5 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 2(4) + 5(-5) & 2(-6) + 5(2) \\ 1(4) + 9(-5) & 1(-6) + 9(2) \end{bmatrix}$$

$$= \begin{bmatrix} -17 & -2 \\ -41 & 12 \end{bmatrix}.$$

For matrix multiplication to be conformable, the number of columns of the first matrix must equal the number of rows of the second matrix.

A.2.4 Transpose

The transpose of a matrix is obtained by replacing rows with corresponding columns and vice-versa,

e.g.

$$\mathbf{M}' = \begin{bmatrix} 6 & 4 \\ 3 & 9 \\ 8 & 5 \end{bmatrix}' = \begin{bmatrix} 6 & 3 & 8 \\ 4 & 9 & 5 \end{bmatrix}.$$

The transpose of the product of two matrices is the product of the transposes of the matrices taken in reverse order, e.g.

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$

A.2.5 Determinants

The determinant of a matrix is a scalar and exists only for square matrices. Knowledge of the determinant of a matrix is useful for obtaining the inverse of the matrix, which in matrix algebra is analogous to the reciprocal of scalar algebra. If \mathbf{A} is a square matrix, its determinant can be symbolized as $|\mathbf{A}|$. Procedures for evaluating the determinant of various order matrices follow.

The determinant of a scalar (1×1 matrix) is the scalar itself, e.g. for $\mathbf{A} = 6$, $|\mathbf{A}| = 6$. The determinant of a 2×2 matrix is the difference between the product of the diagonal elements and the product of the off-diagonal elements, e.g. for

$$\mathbf{A} = \begin{bmatrix} 5 & 2 \\ 6 & 3 \end{bmatrix}$$

$$|\mathbf{A}| = 5(3) - 6(2) = 3.$$

The determinant of a 3×3 matrix can be obtained from the expansion of three 2×2 matrices obtained from it. Each of the second order determinants is preceded by a coefficient of +1 or -1, e.g. for

$$\mathbf{A} = \begin{bmatrix} 5 & 2 & 4 \\ 6 & 3 & 1 \\ 8 & 7 & 9 \end{bmatrix}$$

Based on elements of the first row,

$$\begin{aligned} |\mathbf{A}| &= 5(+1) \begin{vmatrix} 3 & 1 \\ 7 & 9 \end{vmatrix} + 2(-1) \begin{vmatrix} 6 & 1 \\ 8 & 9 \end{vmatrix} + 4(+1) \begin{vmatrix} 6 & 3 \\ 8 & 7 \end{vmatrix} \\ &= 5(27 - 7) - 2(54 - 8) + 4(42 - 24) \\ &= 5(20) - 2(46) + 4(18) \\ &= 100 - 92 + 72 = 80 \end{aligned}$$

The determinant was derived by taking in turn each element of the first row, crossing out the row and column corresponding to the element, obtaining the determinant of the resulting 2×2 matrix, multiplying this determinant by +1 or -1 and the element concerned, and summing the resulting products for each of the three first row elements. The (+1) or (-1) coefficients for the ij^{th} element were obtained according to $(-1)^{i+j}$. For example, the coefficient for the 12 element is $(-1)^{1+2} = (-1)^3 = -1$. The coefficient for the 13 element is $(-1)^{1+3} = (-1)^4 = 1$. The determinants of each of the 2×2 sub-matrices are called minors. For example, the minor of first row element 2 is

$$\begin{bmatrix} 6 & 1 \\ 8 & 9 \end{bmatrix} = 46$$

When multiplied by its coefficient of (-1), the product is called the co-factor of element 12. For example, the co-factor of elements 11, 12 and 13 are 20, -46 and 18.

Expansion by the elements of the second row yields the same determinant, e.g.

$$\begin{aligned} |A| &= 6(-1) \begin{bmatrix} 2 & 4 \\ 7 & 9 \end{bmatrix} + 3(+1) \begin{bmatrix} 5 & 4 \\ 8 & 9 \end{bmatrix} + 1(-1) \begin{bmatrix} 5 & 2 \\ 8 & 7 \end{bmatrix} \\ &= -6(18 - 28) - 3(45 - 32) + 1(35 - 16) \\ &= 60 + 39 - 19 = 80 \end{aligned}$$

Similarly, expansion by elements of the third row again yields the same determinant, etc.

$$\begin{aligned} |A| &= 8(+1) \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} + 7(-1) \begin{bmatrix} 5 & 4 \\ 6 & 1 \end{bmatrix} + 9(+1) \begin{bmatrix} 5 & 2 \\ 6 & 3 \end{bmatrix} \\ &= 8(2 - 12) - 7(5 - 24) + 9(15 - 12) \\ &= -80 + 133 + 27 = 80 \end{aligned}$$

In general, multiplying the elements of any row by their co-factors yields the determinant. Also, multiplying the elements of a row by the co-factors of the elements of another row yields zero, e. g. the elements of the first row by the co-factors of the second row gives

$$\begin{aligned} &5(-1) \begin{bmatrix} 2 & 4 \\ 7 & 9 \end{bmatrix} + 2(+1) \begin{bmatrix} 5 & 4 \\ 8 & 9 \end{bmatrix} + 4(-1) \begin{bmatrix} 5 & 2 \\ 8 & 7 \end{bmatrix} \\ &= -5(18 - 28) + 2(45 - 32) + 4(35 - 16) \\ &= 50 + 26 - 76 = 0 \end{aligned}$$

Expansion for larger order matrices follows according to $|A| = \sum_{j=1}^n a_{ij} (-1)^{i+j} |M_{ij}|$

for any i where n is the order of the matrix, $i = 1, \dots, n$ and $j = 1, \dots, n$, a_{ij} is the ij^{th} element, and $|M_{ij}|$ is the minor of the ij^{th} element.

A2.6 Inverse

As suggested earlier, the inverse of a matrix is analogous to the reciprocal in scalar algebra and performs an equivalent operation to division. The inverse of matrix A is symbolized as A^{-1} . The multiplication of a matrix by its inverse gives an identity matrix (I), which is composed of all diagonal elements of one and all off-diagonal elements of zero, i.e. $A \times A^{-1} = I$. For the inverse of a matrix to exist, it must be square and have a non-zero determinant.

The inverse of a matrix can be obtained from the co-factors of the elements and the determinant.

The following example illustrates the derivation of the inverse.

$$A = \begin{bmatrix} 5 & 2 & 4 \\ 6 & 3 & 1 \\ 8 & 7 & 9 \end{bmatrix}$$

- i) Calculate the co-factors of each element of the matrix, e.g. the co-factors of the elements of the first row are $(+1)\begin{bmatrix} 3 & 1 \\ 7 & 9 \end{bmatrix}$, $(-1)\begin{bmatrix} 6 & 1 \\ 8 & 9 \end{bmatrix}$, and $(+1)\begin{bmatrix} 6 & 3 \\ 8 & 7 \end{bmatrix} = 20, -46$ and 18 .

Similarly, the co-factors of the elements of the second row are $= 10, 13$ and -19

and the co-factors of the elements of the third row are $= -10, 19$ and 3 .

- ii) Replace the elements of the matrix by their co-factors, e.g.

$$A = \begin{bmatrix} 5 & 2 & 4 \\ 6 & 3 & 1 \\ 8 & 7 & 9 \end{bmatrix} \text{ yields } C = \begin{bmatrix} 20 & -46 & 18 \\ 10 & 13 & -19 \\ -10 & 19 & 3 \end{bmatrix}$$

- iii) Transpose the matrix of co-factors, e.g.

$$C' = \begin{bmatrix} 20 & -46 & 18 \\ 10 & 13 & -19 \\ -10 & 19 & 3 \end{bmatrix}' = \begin{bmatrix} 20 & 10 & -10 \\ -46 & 13 & 19 \\ 18 & -19 & 3 \end{bmatrix}$$

- iv) Multiply the transpose matrix of co-factors by the reciprocal of the determinant to yield the inverse, e.g.

$$|A| = 80, 1/|A| = 1/80$$

$$A^{-1} = \frac{1}{80} \begin{bmatrix} 20 & 10 & -10 \\ -46 & 13 & 19 \\ 18 & -19 & 3 \end{bmatrix}.$$

- v) As a check, the inverse multiplied by the original matrix should yield an identity matrix, i.e. $A^{-1}A = I$, e.g.

$$\frac{1}{80} \begin{bmatrix} 20 & 10 & -10 \\ -46 & 13 & 19 \\ 18 & -19 & 3 \end{bmatrix} \begin{bmatrix} 5 & 2 & 4 \\ 6 & 3 & 1 \\ 8 & 7 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The inverse of a 2×2 matrix is: $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

A.2.7 Linear Independence and Rank

As indicated, if the determinant of a matrix is zero, a unique inverse of the matrix does not exist. The determinant of a matrix is zero if any of its rows or columns are linear combinations of other rows or columns. In other words, a determinant is zero if the rows or columns do not form a set of linearly

independent vectors. For example, in the following matrix $\begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}$

rows 2 and 3 sum to row 1 and the determinant of the matrix is zero.

The rank of a matrix is the number of linearly independent rows or columns. For example, the rank of the above matrix is 2. If the rank of matrix A is less than its order n , then the determinant is zero and the inverse of A does not exist, i.e. if $r(A) < n$ then A^{-1} does not exist.

A.2.8 Generalized Inverse

Although a unique inverse does not exist for a matrix of less than full rank, generalized inverses do exist. If A^* is a generalized inverse of A , it satisfies $AA^*A = A$. Generalized or g -inverses are not unique and there are many A^* which satisfy $AA^*A = A$. There are also many ways to obtain a g -inverse, but one of the simplest ways is to follow these steps:

- Obtain a full rank subset of A and call it M .
- Invert M to yield M^{-1} .
- Replace each element in A with the corresponding element of M^{-1} .
- Replace all other elements of A with zeros.
- The result is A_* , a generalized inverse of A .

Example $\mathbf{A} = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

a) \mathbf{M} , a full rank subset, is $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

b) $\mathbf{M}^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$

c) Replacing elements of \mathbf{A} with corresponding elements of \mathbf{M}^{-1} and all other elements with 0's gives

d) $\mathbf{A}^- = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$

A.2.9 Special Matrices

In many applications of statistics we deal with matrices that are the product of a matrix and its transpose, e.g.

$$\mathbf{A} = \mathbf{X}'\mathbf{X}$$

Such matrices are always symmetric, that is every off-diagonal element above the diagonal equals its counterpart below the diagonal. For such matrices

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \mathbf{X}$$

and $\mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'$ is invariant to $(\mathbf{X}'\mathbf{X})^-$, that is, although there are many possible g -inverses of $\mathbf{X}'\mathbf{X}$, any g -inverse pre-multiplied by \mathbf{X} and post-multiplied by $\mathbf{X}'\mathbf{X}$ yields the same matrix \mathbf{X} .

A.2.10 Trace

The trace of a matrix is the sum of the diagonal elements. For the matrix \mathbf{A} of order n with elements (a_{ij}) , the trace is defined as

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$$

As an example, the trace of

$$\begin{bmatrix} 3 & 1 & 4 \\ 1 & 6 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

is $3 + 6 + 5 = 14$

For products of matrices, $\text{tr}(AB) = \text{tr}(BA)$ if the products are conformable. This can be extended to the product of three or more matrices, e.g.

$$\text{Tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

A.3 Quadratic Forms

All sums of squares can be expressed as quadratic forms that is a $y'Ay$. If $y \sim (\mu, V)$, then

$$E(y'Ay) = \mu'A\mu$$

Exercises

1. For $A = \begin{bmatrix} 6 & 3 \\ 0 & 5 \\ -5 & 1 \end{bmatrix}$ $B = \begin{bmatrix} 3 & 8 \\ 2 & -4 \\ 5 & -1 \end{bmatrix}$

Find the sum of $A + B$.

Find the difference of $A - B$.

2. For A and B above and $v' = [1 \ 3 \ -1]$, find $v'A$ and $v'B$.

3. For $B' = \begin{bmatrix} 3 & 2 & 5 \\ 8 & -4 & -1 \end{bmatrix}$ and A as above. Find $B'A$. Find AB' .

4. For A and B above, find AB .

5. Obtain determinants of the following matrices

$$\begin{bmatrix} 3 & 8 \\ 2 & -4 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 3 \\ 1 & 5 \\ -5 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 3 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 6 & 4 & 3 \\ 2 & 8 & 5 & 4 \\ 3 & 8 & 7 & 5 \\ 4 & 9 & 7 & 7 \end{bmatrix}$$

6. Show that the solution to $\mathbf{Ax} = \mathbf{y}$ is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$.

7. Derive the inverses of the following matrices:

$$\begin{bmatrix} 4 & 2 \\ 6 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 3 \\ -5 & 1 & 0 \\ 1 & 4 & -2 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

8. For $\mathbf{A} = \begin{bmatrix} 1 & 1 & 3 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$,

show that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

Day 1

Multi-locus Population Genetics – Linkage & Disequilibrium Objective

Present population genetic principles of allele, haplotype and genotype frequencies, and of linkage and linkage disequilibrium

1. Single locus allele and genotype frequencies
2. Multi-locus haplotype and genotype frequencies
3. Measures of Linkage Disequilibrium (LD)
4. Estimating LD from genotype data
5. Linkage maps and recombination
6. Mechanisms that generate and erode LD
7. LD balance between drift and recombination
8. Persistence of LD across breeds
9. Erosion of LD in crosses vs. outbred population
10. LD always exists within families

1

1. Single locus allele and genotype frequencies

Consider a single locus in a random mating outbred population.

The locus has alleles A_1 and A_2 with allele (or gene) frequencies p and q

Under random mating (Hardy Weinberg Equilibrium), the allele received from one parent is *independent* of the allele received from the other parent, resulting in the following relationship between allele and genotype frequencies:

Table 1: Genotype probabilities, single locus two-allele case

<i>Paternal allele</i>	<i>Maternal allele</i>		
	$Pr(A_1) = p$	$Pr(A_2) = q$	<i>Marginal prob</i>
$Pr(A_1) = p$	p^2	pq	$p^2 + pq = p(p + q) = p$
$Pr(A_2) = q$	pq	q^2	$pq + q^2 = q(p + q) = q$
<i>Marginal prob.</i>	$p^2 + pq = p(p + q) = p$	$pq + q^2 = q(p + q) = q$	

This results in the HWE genotype frequencies: p^2 , $2pq$, q^2

2

1

2. Multi-locus haplotype and genotype frequencies

With *multiple* loci we also need to consider *haplotypes* and their frequencies, and relationships between *allele*, *haplotype*, and *genotype* frequencies.

Haplotype = the combination of alleles at >1 locus that an individual inherited from a parent
E.g. an individual with (unordered) genotype A_1A_2 and B_1B_2 at loci A and B, can have the following combinations of haplotype pairs (separated by /):

$A_1B_1/A_2B_2 \rightarrow$ alleles A_1 and B_1 received from one parent and A_2 and B_2 from the other
 $A_1B_2/A_1B_2 \rightarrow$ alleles A_1 and B_2 received from one parent and A_2 and B_1 from the other

Haplotype frequency = frequency of a given haplotype in a population

What is the relationship between *haplotype frequencies* and the *frequencies of alleles* that make up each haplotype?

This depends on whether the alleles at the two loci are *dependent* or *independent*:

The term 'linkage' in Linkage disequilibrium is actually not quite correct and a bit misleading because disequilibrium can occur between *unlinked* loci, although it is more likely to be present (and persist) between linked loci (see later). Thus, 'Gametic phase' disequilibrium is a better term; gametic phase refers to the haploid phase of chromosomes and disequilibrium refers to dependence between alleles that make up the haplotypes that are present in the current generation and which originated from the haploid gametes produced by their parents.



3

Haplotype probabilities / frequencies

What is the probability of a progeny to receive from a parent: allele A_i at locus A and allele B_j at locus B ?

i) if the alleles at the two loci are *independent* from each other
→ joint probability = product of marginal probabilities

Locus B allele freq's	Locus A – allele frequencies		Marginal prob
	$Pr(A_1) = p_A$	$Pr(A_2) = q_A$	
$Pr(B_1) = p_B$	$Pr(A_1B_1) = p_Ap_B$	$Pr(A_2B_1) = q_Ap_B$	$p_Ap_B + q_Ap_B$ = $p_B(p_A + q_A) = p_B$
$Pr(B_2) = q_B$	$Pr(A_1B_2) = p_Aq_B$	$Pr(A_2B_2) = q_Aq_B$	$p_Aq_B + q_Aq_B$ = $q_B(p_A + q_A) = q_B$
Marginal prob	$p_Ap_B + p_Aq_B$ = $p_A(p_B + q_B) = p_A$	$q_Ap_B + q_Aq_B$ = $q_A(p_B + q_B) = q_A$	Haplotype frequency

Locus B allele freq's	Locus A – allele frequencies		Marginal prob
	$Pr(A_1) = 0.5$	$Pr(A_2) = 0.5$	
$Pr(B_1) = 0.5$	$Pr(A_1B_1) = 0.25$	$Pr(A_2B_1) = 0.25$	$0.25 + 0.25$ = 0.5
$Pr(B_2) = 0.5$	$Pr(A_1B_2) = 0.25$	$Pr(A_2B_2) = 0.25$	$0.25 + 0.25$ = 0.5
Marginal prob	$0.25 + 0.25$ = 0.5	$0.25 + 0.25$ = 0.5	

4

Haplotype probabilities / frequencies

What is the probability of a progeny to receive from a parent: allele A_i at locus A
and allele B_j at locus B ?

ii) What if the alleles at the two loci are NOT independent?

→ joint probabilities deviate from product of marginal probabilities (by $+D$)

Locus B	Locus A		Marginal prob
	$Pr(A_1) = p_A$	$Pr(A_2) = q_A$	
$Pr(B_1) = p_B$	$Pr(A_1B_1) = r$ $= p_Ap_B + D$	$Pr(A_2B_1) = t$ $= q_Ap_B - D$	$p_Ap_B + D + q_Ap_B - D$ $= p_B(p_A + q_A) = p_B$
$Pr(B_2) = q_B$	$Pr(A_1B_2) = s$ $= p_Aq_B - D$	$Pr(A_2B_2) = u$ $= q_Aq_B + D$	$p_Aq_B - D + q_Aq_B + D$ $= q_B(p_A + q_A) = q_B$
Marginal prob	$p_Ap_B + D + p_Aq_B - D$ $= p_A(p_B + q_B) = p_A$	$q_Ap_B - D + q_Aq_B + D$ $= q_A(p_B + q_B) = q_A$	$D = r - p_Ap_B$ \uparrow $Pr(A_1B_1) - Pr(A_1)Pr(B_1)$

D = measure of disequilibrium

Value of $|D|$ is the same irrespective of the haplotype used

Locus B allele freq's	Locus A – allele frequencies		Marginal prob
	$Pr(A_1) = 0.5$	$Pr(A_2) = 0.5$	
$Pr(B_1) = 0.5$	$Pr(A_1B_1) = 0.4$	$Pr(A_2B_1) = 0.1$	$0.4 + 0.1 = 0.5$
$Pr(B_2) = 0.5$	$Pr(A_1B_2) = 0.1$	$Pr(A_2B_2) = 0.4$	$0.1 + 0.4 = 0.5$
Marginal prob	$0.4 + 0.1 = 0.5$	$0.1 + 0.4 = 0.5$	$D = 0.4 - 0.5 * 0.5 = 0.15$ 5

3. Measures of Linkage Disequilibrium (LD)

$$D = Pr(A_1B_1) - p_Ap_B$$

D' = D standardized to make it less dependent on allele frequencies

$$D' = D/D_{max} \quad \text{where } D_{max} = \text{Min}(p_Ap_B, q_Aq_B) \text{ if } D < 0$$

$$D_{max} = \text{Min}(p_Aq_B, q_Ap_B) \text{ if } D > 0 \quad \text{See F&M Ex 1.6 p17}$$

r^2 = squared correlation between allele at locus A and allele at locus B
- also measures ability (R^2) to predict allele at locus A from allele at locus B

$$r^2 = \frac{D^2}{p_Aq_Ap_Bq_B}$$

$$D = 0.4 - 0.5 * 0.5 = 0.15$$

$$D' = 0.15/0.25 = 0.6$$

$$D_{max} = \text{Min}(0.5 * 0.5, 0.5 * 0.5) = 0.25$$

$$r^2 = \frac{0.15^2}{0.5 * 0.5 * 0.5 * 0.5} = 0.36$$

$|D'|$ and r^2 range between 0 and 1

$|D'|$ is strongly inflated if one haplotype has low frequency

r^2 is the preferred measure of LD for most uses

To derive r^2 : Let $X = 1$ when allele A_1 present, $X = 0$ if A_2 present (= Bernoulli var.)
 $Y = 1$ when allele B_1 present, $Y = 0$ if B_2 present (= Bernoulli var.)

Then: $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$
 $= r - p_A p_B = D$

$$\Rightarrow \text{Corr.} = r_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{D}{\sqrt{(p_A q_A)(p_B q_B)}}$$

$$\Rightarrow r^2 = r_{XY}^2 = \frac{D^2}{p_A q_A p_B q_B} \quad (\text{Note: this } r \text{ is different than } r \text{ in the table in previous slide})$$

	$A_1 X=1$	$A_2 X=0$
$B_1 Y=1$	$\Pr(A_1 B_1) = r$ $XY = 1$	$\Pr(A_2 B_1) = t$ $XY = 0$
$B_2 Y=0$	$\Pr(A_1 B_2) = s$ $XY = 0$	$\Pr(A_2 B_2) = u$ $XY = 0$

If A is a marker and B a QTL $\Rightarrow r^2 = \text{proportion of QTL variance observed at marker}$

- eg if QTL variance = 200 kg^2 , and $r^2 = 0.2 \Rightarrow$ variation observed at marker = 40 kg^2

r^2 is a key parameter determining the power of LD mapping to detect QTL

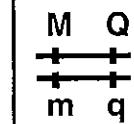
- Experiment sample size must be increased by $1/r^2$
to have the same power as an experiment observing the QTL directly

For multi-allelic markers, see Zhao et al. 2005 and 2007. Genetical Research

7

Why is LD important?

Use of linked markers relies on
association of markers with phenotype
QTL detection



Marker Genotype	Mean Phenotype
-----------------	----------------

MM

20

Mm

18

mm

14

Allele M is associated with favorable QTL allele

MAS

Select MM or individuals that inherited allele M

Requires Linkage Disequilibrium between marker and QTL

8

4. Estimating LD from Genotype Data

Disequilibrium is quantified by comparing haplotype frequencies to their expected frequencies based on independence ($D = \Pr(A_1B_1) - \Pr(A_1)\Pr(B_1)$).

The problem is that genotyping data is in the form of unordered genotypes, not haplotypes, requiring special methods to estimate haplotype frequencies.

With 2 loci with 2 alleles, there are 4 possible haplotypes, 16 ordered genotypes (ordered based on haplotypes), and 9 unordered genotypes (see tables 2,3)

Table 2: Haplotype frequencies and genotype frequencies under random mating (HWE)

Haplotype - freq		Maternal haplotype						
		A_1B_1	r	A_1B_2	s	A_2B_1	t	A_2B_2
<i>Paternal haplotype</i>	A_1B_1	r	A_1B_1/A_1B_1	r^2	A_1B_1/A_1B_2	rs	A_1B_1/A_2B_1	rt
	A_1B_2	s	A_1B_2/A_1B_1	sr	A_1B_2/A_1B_2	s^2	A_1B_2/A_2B_1	st
	A_2B_1	t	A_2B_1/A_1B_1	tr	A_2B_1/A_1B_2	ts	A_2B_1/A_2B_1	t^2
	A_2B_2	u	A_2B_2/A_1B_1	ur	A_2B_2/A_1B_2	us	A_2B_2/A_2B_1	ut

9

2 loci with 2 alleles \rightarrow 4 haplotypes \rightarrow 16 ordered genotypes \rightarrow 9 unordered genotypes

Table 2: Haplotype frequencies and genotype frequencies under random mating (HWE)

Haplotype - freq		Maternal haplotype						
		A_1B_1	r	A_1B_2	s	A_2B_1	t	A_2B_2
<i>Paternal haplotype</i>	A_1B_1	r	A_1B_1/A_1B_1	r^2	A_1B_1/A_1B_2	rs	A_1B_1/A_2B_1	rt
	A_1B_2	s	A_1B_2/A_1B_1	sr	A_1B_2/A_1B_2	s^2	A_1B_2/A_2B_1	st
	A_2B_1	t	A_2B_1/A_1B_1	tr	A_2B_1/A_1B_2	ts	A_2B_1/A_2B_1	t^2
	A_2B_2	u	A_2B_2/A_1B_1	ur	A_2B_2/A_1B_2	us	A_2B_2/A_2B_1	ut

Table 3: Unordered and ordered genotypes and their frequencies under random mating

Unordered genotypes	Frequency =sum of ordered frequencies	Possible ordered genotypes and their frequencies (from Table 2) 'ordered' based on parental origin (paternal haplotype/maternal haplotype)			
$A_1A_1B_1B_1$	r^2	A_1B_1/A_1B_1	r^2		
$A_1A_1B_1B_2$	$2rs$	A_1B_1/A_1B_2	rs	A_1B_2/A_1B_1	sr
$A_1A_1B_2B_2$	s^2	A_1B_2/A_1B_2	s^2		
$A_1A_2B_1B_1$	$2rt$	A_1B_1/A_2B_1	rt	A_2B_1/A_1B_1	tr
$A_1A_2B_1B_2$	$2ru+2st$	A_1B_1/A_2B_2	ru	A_1B_2/A_2B_1	st
$A_1A_2B_2B_2$	$2su$	A_1B_2/A_2B_2	su	A_2B_2/A_1B_2	ts
$A_2A_2B_1B_1$	t^2	A_2B_1/A_2B_1	t^2		
$A_2A_2B_1B_2$	$2tu$	A_2B_1/A_2B_2	tu	A_2B_2/A_2B_1	ut
$A_2A_2B_2B_2$	u^2	A_2B_2/A_2B_2	u^2		

The unordered genotype is what is obtained from genotyping, i.e. the genotype at each locus

10

Simple method for estimating haplotype frequencies

The **simple method** to estimate haplotype frequencies (r, s, t, u) is to assume that double heterozygotes are equally likely to have either haplotype configuration.

Simple method to estimate haplotype frequencies and LD

Observed data			Exp. Freq	Haplotype counts			
Genotype	Counts	Frequency		A1B1	A1B2	A2B1	A2B2
A1A1B1B1	19	0.475	r^2	38			
A1A1B1B2	5	0.125	$2rs$	5	5		
A1A1B2B2	0	0	s^2		0		
A1A2B1B1	8	0.2	$2rt$	8		8	
A1A2B1B2	8	0.2	$2ru+2st$	4	4	4	4
A1A2B2B2	0	0	$2su$		0		0
A2A2B1B1	0	0	t^2			0	
A2A2B1B2	0	0	$2tu$			0	0
A2A2B2B2	0	0	u^2			0	
	40			0.6875	0.1125	0.15	0.05
				= r	= s	= t	= u
			D = ru-st =		0.0175		

8 individuals are observed to be double heterozygotes, so it is assumed that of the $2*8 = 16$ haplotypes that are carried by these individuals, 4 are A_1B_1 , 4 are A_1B_2 , 4 are A_2B_1 , and 4 are A_2B_2 .

The problem with this method is that the 4 haplotypes are NOT equally likely. In fact, even based on the simple method, the frequency of the A_1B_1 haplotype is 0.69 and that of A_1B_2 is 0.11. 11

In fact, using the frequencies obtained from the simple method, we can calculate the probability that a double heterozygote will have the one versus the other haplotype configuration

$$\Pr\left(\frac{A_1B_1}{A_2B_2} \mid A_1A_2B_1B_2\right) = \frac{2ru}{2ru + 2st} = \frac{ru}{ru + st} \quad \text{and} \quad \Pr\left(\frac{A_1B_2}{A_2B_1} \mid A_1A_2B_1B_2\right) = \frac{2st}{2ru + 2st} = \frac{st}{ru + st}$$

For the example data, these probabilities will equal:

$$\Pr\left(\frac{A_1B_1}{A_2B_2} \mid A_1A_2B_1B_2\right) = \frac{0.6875 * 0.05}{0.6875 * 0.05 + 0.1125 * 0.15} = 0.67$$

$$\Pr\left(\frac{A_1B_2}{A_2B_1} \mid A_1A_2B_1B_2\right) = \frac{0.1125 * 0.15}{0.6875 * 0.05 + 0.1125 * 0.15} = 0.33$$

Observed data			Exp. Freq	Haplotype counts			
Genotype	Counts	Frequency		A1B1	A1B2	A2B1	A2B2
A1A1B1B1	19	0.475	r^2	38			
A1A1B1B2	5	0.125	$2rs$	5	5		
A1A1B2B2	0	0	s^2		0		
A1A2B1B1	8	0.2	$2rt$	8		8	
A1A2B1B2	8	0.2	$2ru+2st$	4	4	4	4
A1A2B2B2	0	0	$2su$		0		0
A2A2B1B1	0	0	t^2			0	
A2A2B1B2	0	0	$2tu$			0	0
A2A2B2B2	0	0	u^2			0	
	40			0.6875	0.1125	0.15	0.05
			D = ru-st =		0.0175		

Thus, based on these haplotype frequencies, the A_1B_1/A_2B_2 haplotype configuration is twice as likely as A_1B_2/A_2B_1 .

So now we can use these to adjust haplotype counts for double heterozygotes to 5.33, 2.67, 2.67, 5.33 and use these to re-estimate the haplotype frequencies.

We can then repeat this procedure until the haplotype frequencies don't change anymore (have converged).

Observed data			Exp. Freq	Haplotype counts			
Genotype	Counts	Frequency		A1B1	A1B2	A2B1	A2B2
A1A1B1B1	19	0.475	r^2	38			
A1A1B1B2	5	0.125	$2rs$	5	5		
A1A1B2B2	0	0	s^2		0		
A1A2B1B1	8	0.2	$2rt$	8		8	
A1A2B1B2	8	0.2	$2ru+2st$	5.33	2.67	2.67	5.33
A1A2B2B2	0	0	$2su$		0		0
A2A2B1B1	0	0	t^2			0	
A2A2B1B2	0	0	$2tu$			0	0
A2A2B2B2	0	0	u^2			0	
	40			0.7045	0.0954	0.1329	0.0671
			D = ru-st =		0.0346		

This is the Expectation Maximization algorithm for Maximum Likelihood haplotype frequency estimation.

12

EM algorithm for ML haplotype frequency estimation

Based on (assumed) unrelated individuals

Implemented in the 'EM for haplotype frequencies' sheet in the EM_estimation.xls file

$$P(A1B1//A2B2) = \\ = P(A1B1//A2B2 | A1A2B1B2) * P(A1A2B1B2) = ru/(ru+st) * P(A1A2B1B2) = st/(ru+st) * P(A1A2B1B2)$$

EM algorithm

Genotype	Counts	Freq	Iterations of EM algorithm -->								
			E(Freq)1	E(Freq)2	E(Freq)3	E(Freq)4	E(Freq)5	E(Freq)6	E(Freq)7		
A1A1B1B1	19	0.475	A1B1//A2B2	0.100	0.134	0.158	0.170	0.174	0.176	0.177	
A1A1B1B2	5	0.128	A1B2//A2B1	0.100	0.066	0.042	0.030	0.026	0.024	0.023	
A1A1B2B2	0	0	Freq	M(Freq)1	M(Freq)2	M(Freq)3	M(Freq)4	M(Freq)5	M(Freq)6	M(Freq)7	
A1A2B1B1	6	0.2	pA =	0.8000	0.8000	0.8000	0.8000	0.8000	0.8000	0.8000	
A1A2B1B2	8	0.2	pB =	0.8375	0.8375	0.8375	0.8375	0.8375	0.8375	0.8375	
A1A2B2B2	0	0	r = P(A1B1)	0.6700	0.6875	0.7046	0.7163	0.7223	0.7247	0.7257	0.7260
A2A2B1B1	0	0	s = P(A1B2)	0.1300	0.1125	0.0954	0.0837	0.0777	0.0753	0.0743	0.0740
A2A2B1B2	0	0	t = P(A2B1)	0.1675	0.1500	0.1329	0.1212	0.1152	0.1128	0.1118	0.1115
A2A2B2B2	0	0	u = P(A2B2)	0.0325	0.0500	0.0671	0.0788	0.0848	0.0872	0.0882	0.0885
	40		D	0.0000	0.0175	0.0346	0.0463	0.0523	0.0547	0.0557	0.0560
			Change in frequencies	0.0350	0.0341	0.0235	0.0119	0.0049	0.0018	0.0007	

Implemented in Haploview <http://www.broad.mit.edu/mpg/haploview/>

Other software: FastPhase http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/fastPHASE.php

MCMC methods Phase http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/PHASEv2.php
- also estimates recombination rates

Also used for - computing missing genotypes
- assigning haplotype probabilities to individuals

13

What if individuals are not unrelated?

Does the assumption of unrelatedness bias LD results?

Little if sample size large enough (e.g. De Roos et al. Genetics 2009)

But unrelatedness assumption DOES affect haplotype probabilities of individuals

- Marker genotypes of relatives help determine haplotypes of individual

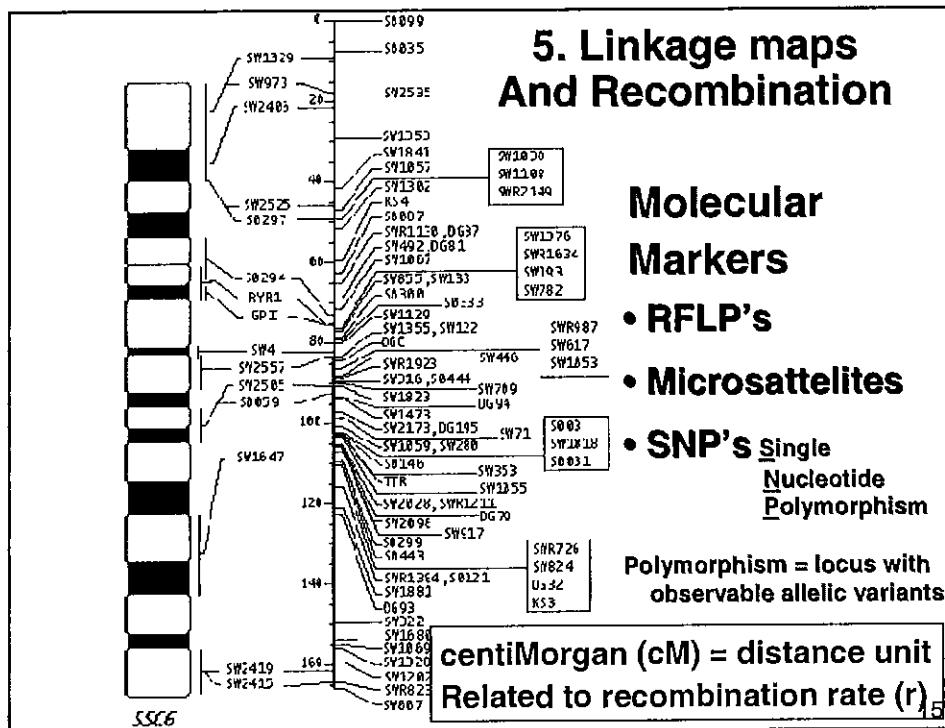
In large paternal half sib families (dairy cattle)

- haplotype phase of a sire can be inferred based on which sire alleles co-occur most often in progeny
- Maternal haplotypes received by progeny then obtained by subtracting sire haplotype from progeny genotype

In complex pedigrees - much more difficult

- SimWalk
- GenoProb
- Iterative peeling
- MCMC methods

14



RECOMBINATION FREQUENCY VERSUS MAP DISTANCE					
Recombination rate (A,B) = c = proportion of recombinants generated by meiosis. = measure of distance between loci in terms of recombination rate					
We like distances to be additive: if A, B, and C are on a string and $\text{dist}(A, B) = c_{AB}$ and $\text{dist}(B, C) = c_{BC}$, then $\text{dist}(A, C)$ is $c_{AB} + c_{BC}$?					
$c_{AC} = c_{AB} + c_{BC} ?$					
Are recombination frequencies additive? NO					
Given c_{AB} and c_{BC} , what is c_{AC} ?					
Four possibilities:					
A-B interval	B-C interval	→	A-C interval	probability	
A ← no rec →	B ← no rec →	C →	A ← no rec → C	$(1 - c_{AB})(1 - c_{BC})$	
A no rec	B rec	C →	A rec C	$(1 - c_{AB}) c_{BC}$	
A rec	B no rec	C →	A rec C	$c_{AB} (1 - c_{BC})$	
A rec	B rec	C →	A no rec C	$c_{AB} c_{BC}$	
$\rightarrow A-C \text{ rec.rate} = c_{AC} = \Pr(2.) + \Pr(3.) = (1 - c_{AB})c_{BC} + c_{AB}(1 - c_{BC}) = c_{AB} + c_{BC} - 2c_{AB}c_{BC}$					
Example $c_{AB} = 0.2$ $c_{BC} = 0.3$ $\rightarrow c_{AC} = 0.2 + 0.3 - 2*0.2*0.3 = 0.38$					
Note: $c_{AC} < 0.5 = 0.2 + 0.3$, because 12% of all gametes ($2*0.2*0.3$) are double recombinant, thus, despite recombination, a parental configuration of alleles will exist between A and C (A_1C_1 or A_2C_2).					
An important assumption in this calculation is <i>independence</i> of recombination events – no interference					

Recombination \leftrightarrow Crossover

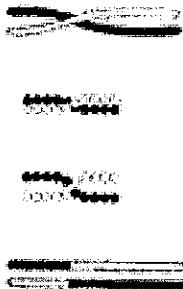


Fig. 8.1. Behavior of chromosomes in meiosis after crossing over at the synaptonemal complex.



Fig. 8.2. Behavior of chromosomes in meiosis after double crossing over.

Thomas Hunt Morgan's illustration of crossing over (1916)

A double crossing over

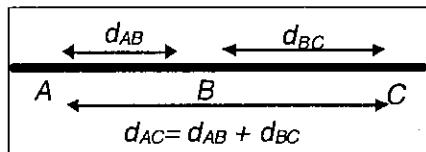
Recombinations result from crossovers.

A crossover occurs when segments of homologous chromosomes of a pair (i.e. the maternal and paternal chromosomes) are exchanged during meiosis.

Multiple crossovers can occur between loci A and B but only an **uneven number** of crossovers results in a recombination event between A and B.

17

Map distance $(A-B) = d_{AB} = E(\# \text{ crossovers in } A-B) = \text{expected #cross-overs generated during a meiosis in } A-B \text{ interval}$



Map distances ARE additive because expectations are additive (even with dependence):

$$(\# \text{ crossovers in } A-C) = (\# \text{ crossovers in } A-B) + (\# \text{ crossovers in } B-C)$$

$$\Rightarrow E(\# \text{ crossovers in } A-C) = E\{(\# \text{ crossovers in } A-B) + (\# \text{ crossovers in } B-C)\} \\ = E(\# \text{ crossovers in } A-B) + E(\# \text{ crossovers in } B-C)$$

$$\Rightarrow d_{AC} = d_{AB} + d_{BC}$$

Recombination rate, c , = proportion between 0 (completely linked) and $\frac{1}{2}$ (unlinked)

Map distance, d , is measured in Morgans: if $d_{AB} = 1$ Morgan \rightarrow on average 1 cross-over event will occur per meiosis

- if $d_{AB} = 1$ Morgan \rightarrow on average 1 cross-over event occurs between A and B per meiosis
- 1 Morgan = 100 cM
- For cattle, genetic map length ~ 30 M (3000 cM) $\rightarrow \sim 30$ crossovers per meiosis.

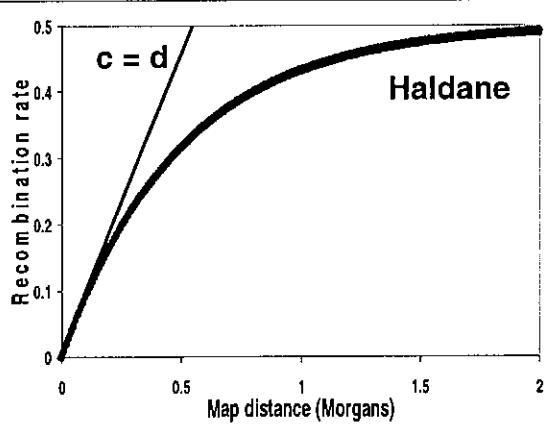
$c \leq d$ because an even number of cross-overs results in a non-recombinant gamete

18

Mapping Function

Provides the relationship between recombination rate (c) and map distance (d)

- Complete interference $\rightarrow c = d$
- No interference \rightarrow Haldane mapping function: $c = (1 - e^{-2d})/2$ $d = -\ln(1-2c)/2$
- Some interference \approx Kosambi mapping function $c = (e^{4d} - 1)/2(e^{4d} + 1)$
 $d = \frac{1}{4} \ln[(1+2c)/(1-2c)]$



E.g. $d = 0.2$
 $\rightarrow c_{\text{Haldane}} = (1 - e^{-2*0.2})/2 = 0.165$
 $\rightarrow c_{\text{Kosambi}} = (e^{4*0.2} - 1)/2(e^{4*0.2} + 1) = 0.190$
 $\rightarrow c_{\text{complete int}} = 0.200$

19

6. Mechanisms that Generate and Erode LD

A variety of mechanisms generate linkage disequilibrium, and several of these can operate simultaneously. They can be separated into:

1. Recurrent factors – operate to create LD each generation
 - a. *Drift* (inbreeding) in small populations – by chance or sampling, haplotypes passed on to the next generation are not in LE frequencies
 - b. *Recurrent migration* – continuous mixing of populations in which haplotypes occur in different frequencies (e.g. $\Pr(A_1B_1)=1$ for pop. 1 and $=0$ for pop. 2)
 - c. *Selection* – certain haplotypes may be selected upon and increase in frequency
 - selection also creates LD between loci that are selected upon
 (= Bulmer effect – see later)
 - selection with epistasis (certain combinations of alleles are favorable) also creates LD between loci involved.
2. Punctual factors – operate only sporadically over time to create LD
 - a. *Mutation* – occurs in a specific haplotype, which is then the only haplotype that contains that mutation, resulting it to be in LD with the mutation.
 - b. *One-time admixture/migration/crossing* (e.g. producing F_1/F_2) – results in mixing populations with different haplotype frequencies
 - c. *Population bottleneck / founder effects* – severe drift from 1-time sampling effects

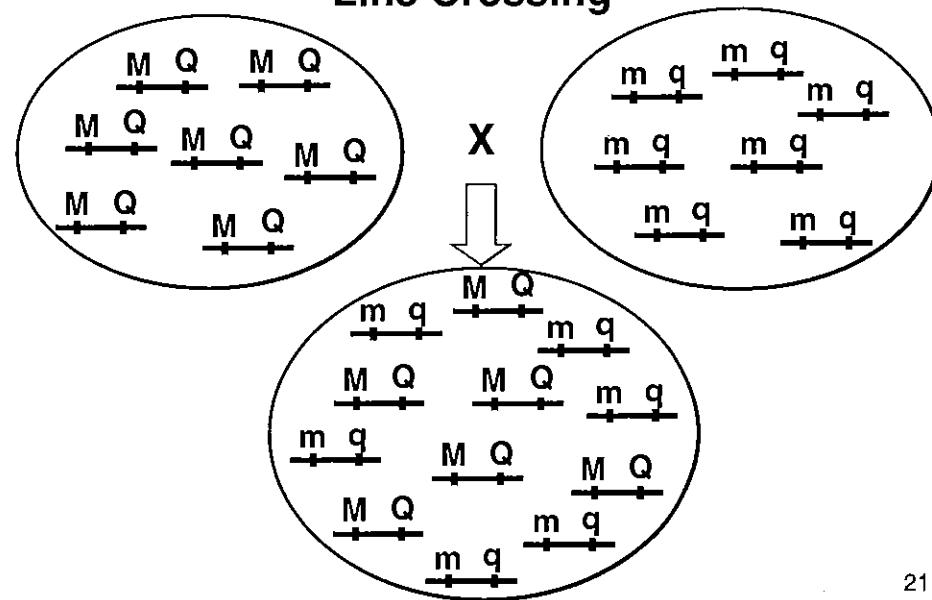
20

Processes that create LD

Inbred line I

Line Crossing

Inbred line II

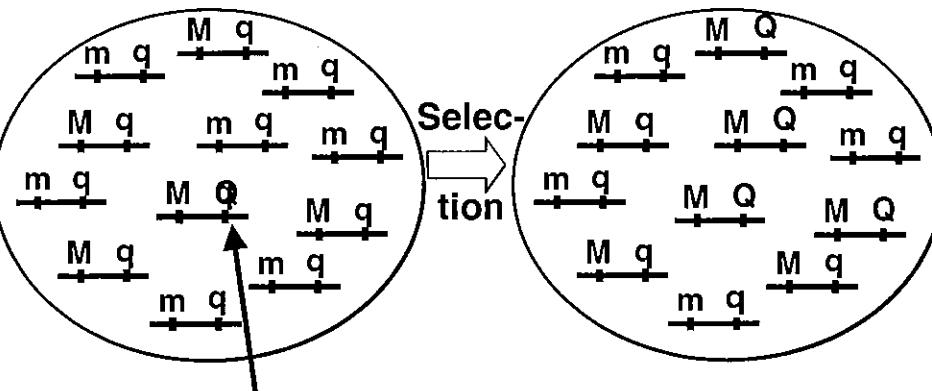


21

Processes that create LD

Mutation and Selection

Selection



QTL allele on M chromosome mutates from q to Q
and then increases in frequency because of

- random drift
- or selection on Q → selective sweep = LD block around Q

22

Selective Sweep

Original mutation ($q \rightarrow Q$) occurred in marker haplotype:

001110010Q01001110110

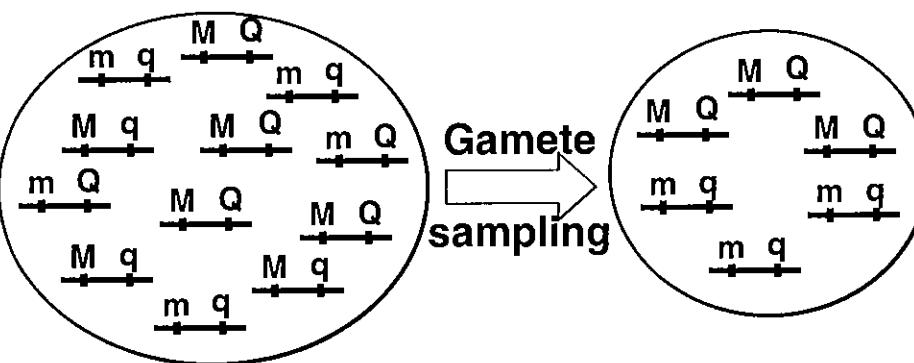
Many generations ↓ of recombination

100110010Q01100110100
011110010Q01001011010
001001110Q01000010111
001110110Q01101101110
011010010Q01001100010
000110010Q01001000111
111010010Q01011101111
010110010Q01001101010

Unique haplotype
associated with Q

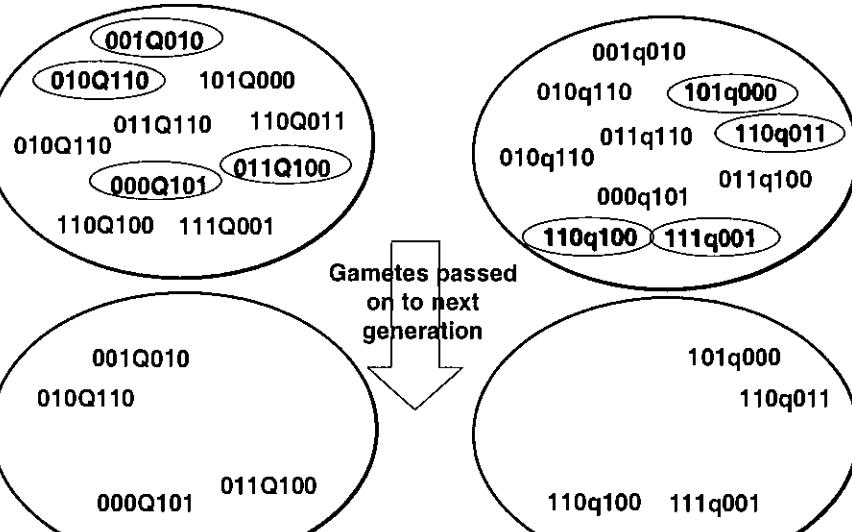
23

Processes that create LD Random drift/inbreeding



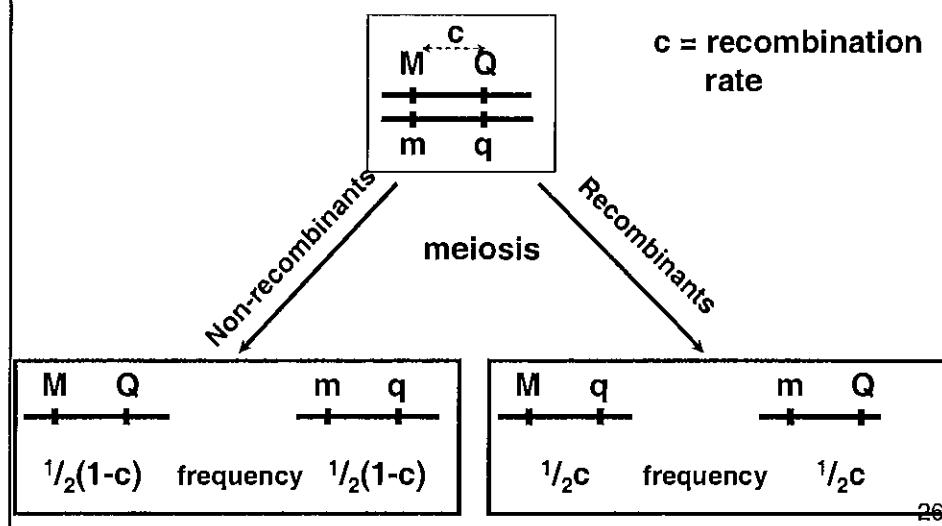
24

LD created by Drift



25

LD is continuously eroded by recombination



26

LD continuously eroded by recombination: how does D change over time?

Let r_0 = frequency of A_1B_1 haplotypes in generation 0 $\rightarrow D_0 = r_0 \cdot p_A p_B$

What is the frequency of A_1B_1 haplotypes in generation 1?

In the following derivation, we will consider parental origin of haplotypes and let \bullet indicate 'any' allele, so $A_1B_1/A_\bullet B_\bullet$ indicates an individual that received the A_1B_1 from its father and any haplotype (A_1B_1 or A_1B_2 or A_2B_1 or A_2B_2) from its mother)

There are four ways that parents from generation 0 can generate gametes that carry the A_1B_1 haplotype and that will produce generation 1:

1. non-recombinant A_1B_1 haplotype produced by a $A_1B_1/A_\bullet B_\bullet$ parent
2. non-recombinant A_1B_1 haplotype produced by a $A_\bullet B_\bullet/A_1B_1$ parent
3. recombinant A_1B_1 haplotype produced by a $A_1B_\bullet/A_\bullet B_1$ parent
4. recombinant A_1B_1 haplotype produced by a $A_\bullet B_1/A_1B_\bullet$ parent

Case 1: the frequency of $A_1B_1/A_\bullet B_\bullet$ parents is r_0 and the frequency of non-recombinant A_1B_1 haplotypes produced by these parents is $\frac{1}{2}(1-c)$. Since these two events are independent, the frequency of A_1B_1 haplotype produced by $A_1B_1/A_\bullet B_\bullet$ parents = $\text{Prob}(1) = \frac{1}{2}(1-c)r_0$.

Case 2. results in the same frequency: $\text{Prob}(2.) = \frac{1}{2}(1-c)r_0$

Case 3: the frequency of $A_1B_\bullet/A_\bullet B_1$ parents is $p_A p_B$ because the frequency of generation 0 individuals that received a A_1B_\bullet haplotype from their father = frequency of individuals that received the A_1 allele from their father = frequency of the A_1 allele = p_A . Similarly, the frequency generation 0 individuals that received a $A_\bullet B_1$ haplotype from their mother = p_B . Then, the frequency of recombinant A_1B_1 haplotypes produced by these parents is $\frac{1}{2}c$, so the overall frequency = $\frac{1}{2}cp_A p_B$.

Case 4. results in the same frequency: $\text{Prob}(4.) = \frac{1}{2}cp_A p_B$.

27

LD continuously eroded by recombination: how does D change over time?

Let r_0 = frequency of A_1B_1 haplotypes in generation 0 $\rightarrow D_0 = r_0 \cdot p_A p_B$

What is the frequency of A_1B_1 haplotypes in generation 1?

Thus, the overall frequency of A_1B_1 gametes produced by generation 0 is the sum of these four mutually exclusive cases:

$$\rightarrow r_1 = r_0(1-c) + p_A p_B c$$

$$\begin{aligned} \rightarrow D_1 &= r_1 \cdot p_A p_B = r_0(1-c) + p_A p_B c - p_A p_B = \\ &= r_0(1-c) \cdot p_A p_B (1-c) = (r_0 + p_A p_B)(1-c) \\ &\approx D_0(1-c) \end{aligned}$$

$$\begin{aligned} \rightarrow D_2 &= D_1(1-c) = \{D_0(1-c)\} (1-c) = \\ &= D_0(1-c)^2 \end{aligned}$$

$$\rightarrow D_t = D_0(1-c)^t \quad \rightarrow D_\infty = 0$$

\rightarrow Erosion of LD by recombination occurs faster when loci are further apart.

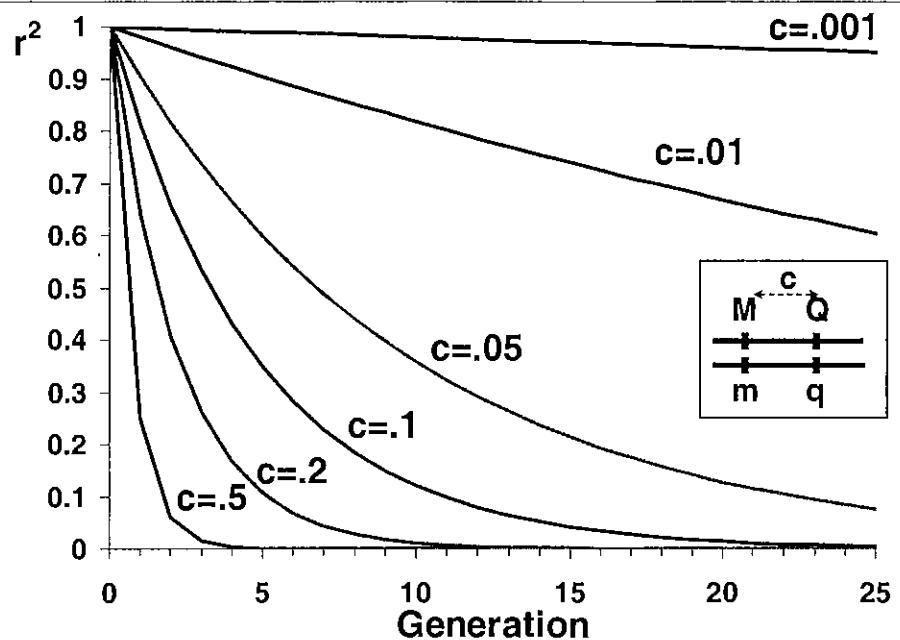
LD is halved each generation if loci are unlinked ($c = \frac{1}{2}$).

Since $r^2 = \frac{D}{p_A q_A p_B q_B}$, LD measured by r^2 will decline at a rate of $(1-c)^2$ per generation:

$$r_t^2 = r_0^2 (1-c)^{2t}$$

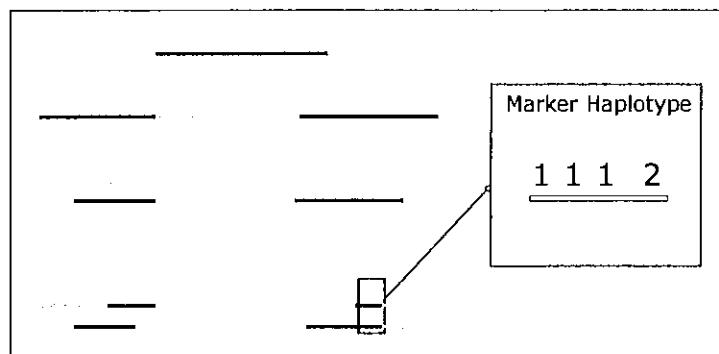
28

Break-up of LD by recombination



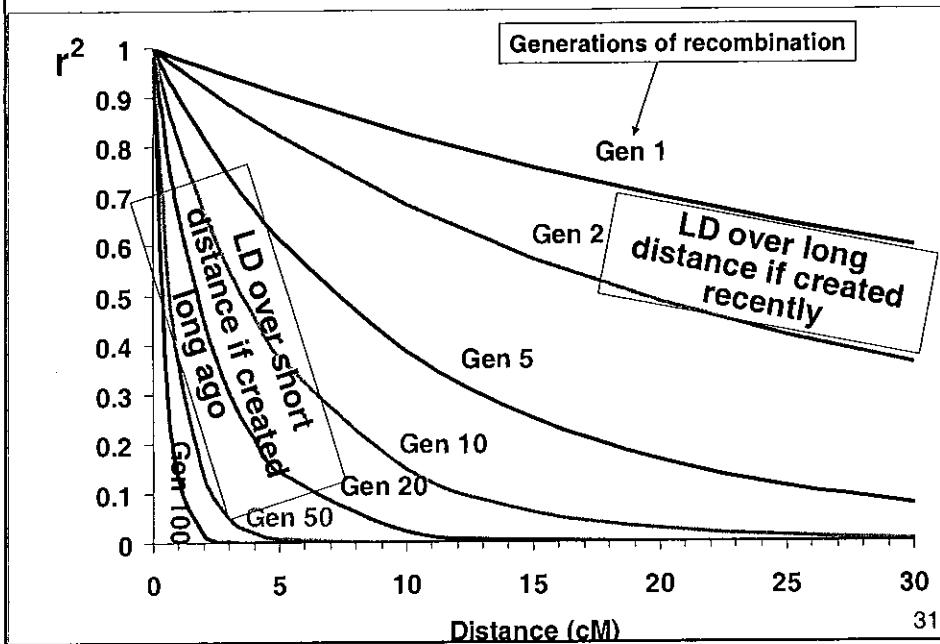
Another way of looking at LD

Conservation of chunks of ancestral chromosomes



Size of conserved chunks depends on how
Long ago LD was created – longer if N_e larger

Recent LD extends over large distances

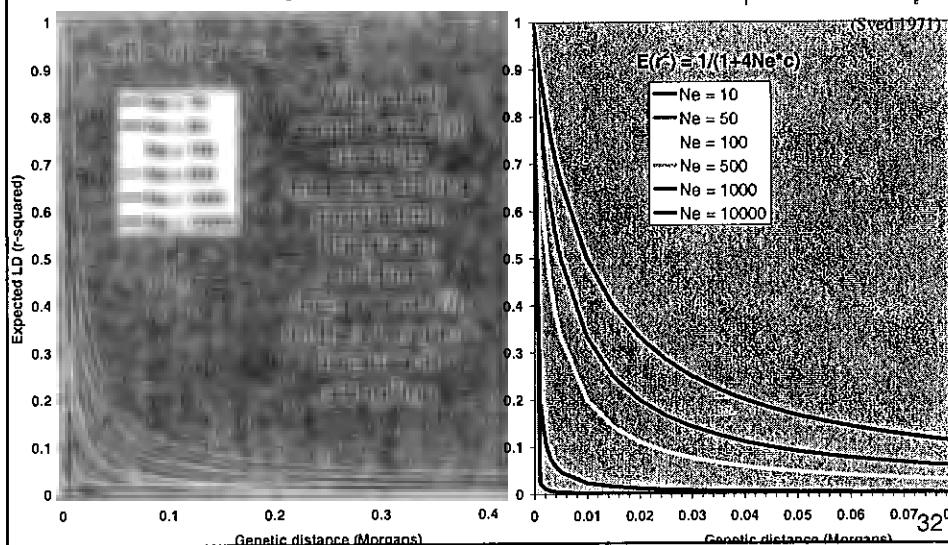


7. Balance between Drift and Recombination

In small(er) closed populations

- LD is continuously created by drift – more with smaller effective pop. size, N_e
- LD is continuously eroded by recombination – faster at longer distances

This results in a balance/equilibrium of average LD at a given distance: $E(r^2_{\infty,c}) = \frac{1}{1 + 4N_e c}$



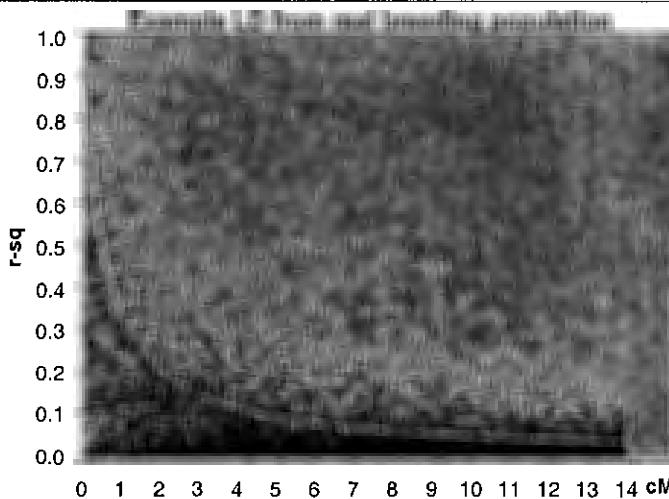
7. Balance between Drift and Recombination

In small(er) closed populations

- LD is continuously created by drift – more with smaller effective pop. size, N_e
- LD is continuously eroded by recombination – faster at longer distances

This results in a balance/equilibrium of average LD at a given distance: $E(r^2_{\infty,c}) = \frac{1}{1+4N_e c}$

(Sved 1971)



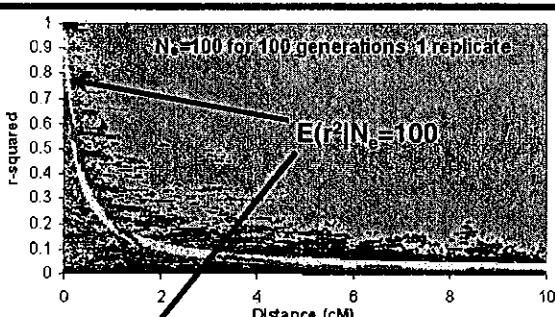
LD is very variable

LD at short distances is often lower than expected based on a given N_e

Because LD reflects historical N_e and this has not been constant

33

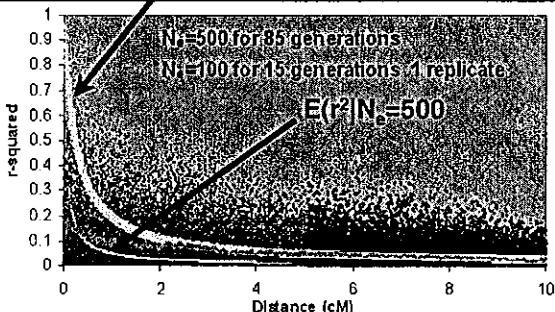
Effect of historical N_e on LD



LD at distance c (M) :

$$E(r_c^2) = \frac{1}{1+4N_{e,t} c}$$

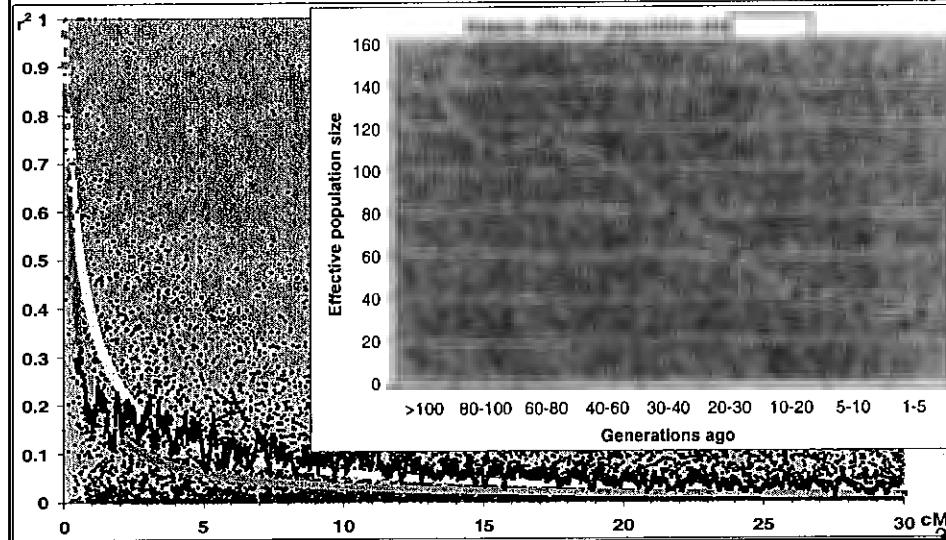
- Where $t = 1/(2c)$ generations ago
 - markers 0.1M (10cM) apart reflect N_e 5 generations ago
 - markers 0.001 (0.1cM) apart reflect N_e 500 generations ago
- LD at short distances reflects historical effective population size
- LD at longer distances reflects more recent population history



34

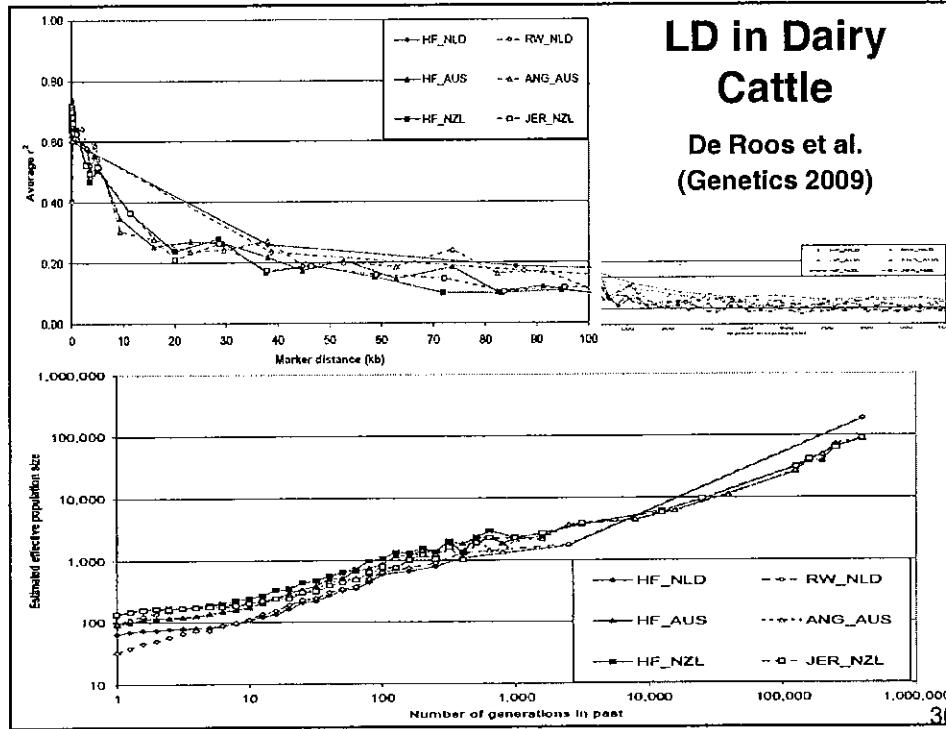
Estimating historical N_e from average LD at a given distance

$$E(r_c^2) = \frac{1}{1 + 4N_{e,t} c} \implies \hat{N}_{e,t} = \frac{1}{4c} \left(\frac{1}{r_c^2} - 1 \right)$$



LD in Dairy Cattle

De Roos et al.
(Genetics 2009)



8. Persistence of LD across breeds

- Can the same marker be used across breeds?
 - Yes if marker-QTL LD is similar in both breeds
- This can be assessed by evaluating the consistency of LD between SNPs across breeds
 - Could compare r^2 between same pair of SNPs across breeds
 - However, the r^2 statistic between two SNPs can be same value even if phases of haplotypes are reversed
 - ➔ Use comparison of r instead = correlation between SNP alleles, instead of square of correlation

37

Persistence of LD across breeds Use r instead of r^2

Breed 1

Marker B	Marker A		Frequency
	A1	A2	
B1	0.4	0.1	0.5
B2	0.1	0.4	0.5
Frequency	0.5	0.5	

$$r = \frac{D}{\sqrt{P_{A1}P_{A2}P_{B1}P_{B2}}} = \frac{(0.4 - 0.5 * 0.5)}{\sqrt{0.5 * 0.5 * 0.5 * 0.5}} = 0.6$$

Breed 2

Marker B	Marker A		Frequency
	A1	A2	
B1	0.3	0.2	0.5
B2	0.2	0.3	0.5
Frequency	0.5	0.5	

$$r = \frac{(0.3 - 0.5 * 0.5)}{\sqrt{0.5 * 0.5 * 0.5 * 0.5}} = 0.2$$

Breed 3

Marker B	Marker A		Frequency
	A1	A2	
B1	0.2	0.3	0.5
B2	0.3	0.2	0.5
Frequency	0.5	0.5	

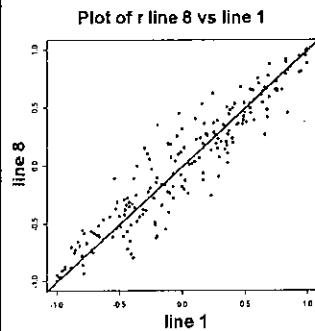
$$r = \frac{(0.2 - 0.5 * 0.5)}{\sqrt{0.5 * 0.5 * 0.5 * 0.5}} = -0.2$$

Hayes '07 38

Consistency of LD in commercial broiler breeding lines

correlation of r within 1 cM

Andreeescu et al. Genetics, 2007



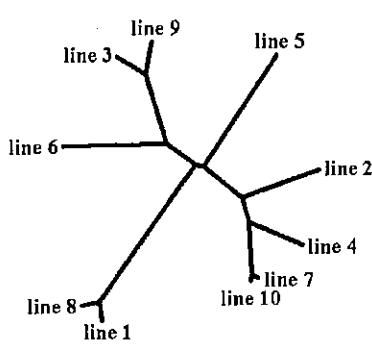
chr1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8	Line 9	Line 10
Line 1	.15	.40	.52	.46	.54	.45	.94	.57	.45
Line 2		.36	.68	.35	.53	.69	.44	.49	.67
Line 3			.42	.28	.62	.49	.50	.90	.50
Line 4				.44	.51	.72	.36	.55	.72
Line 5					.37	.43	.42	.42	.41
Line 6						.57	.50	.69	.52
Line 7							.50	.58	.97
Line 8								.51	.52
Line 9									.58

A high correlation between r means that SNP effects are expected to persist across lines – assuming QTL effects are consistent across populations

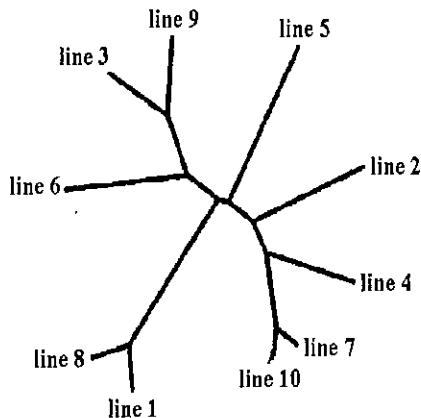
39

Phylogenetic trees

LD Correlation-based



Allele frequency-based



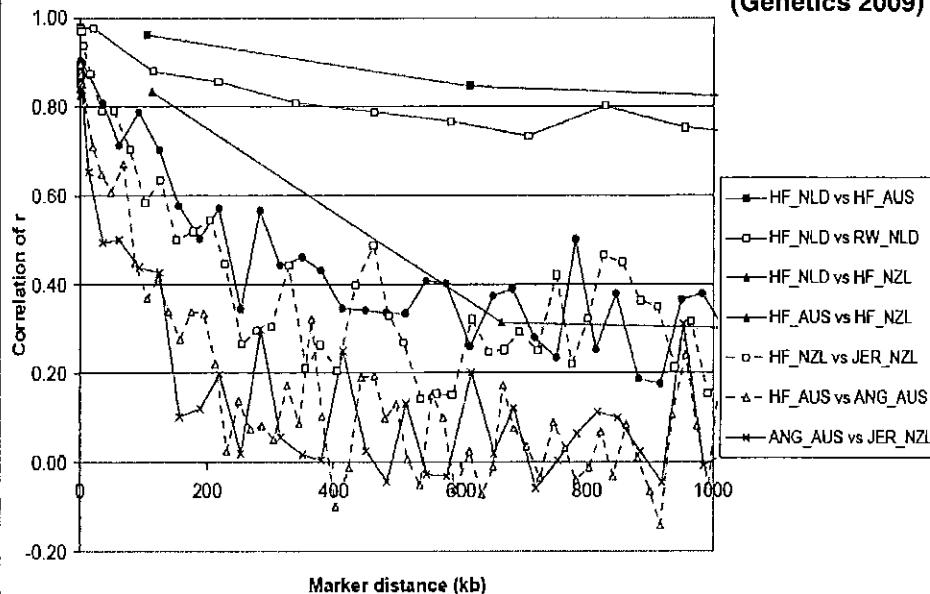
LD correlations are expected to be related to the number of generations the lines / breeds have been separated

More generations of separation → more erosion of LD by recombination
→ less consistent LD

40

LD across cattle breeds

De Roos et al.
(Genetics 2009)



LD correlations expected to decline with distance → more recombination

41

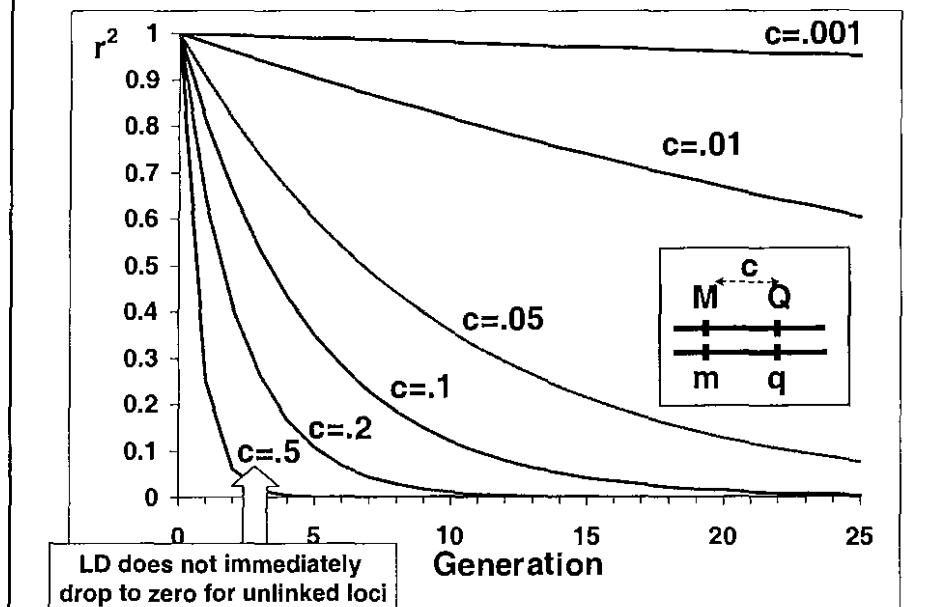
Persistence of LD across breeds

- Recently diverged breeds / lines have good prospects of using a marker found in one line in the other line
- More distantly related breeds, will need very dense marker maps to find markers which can be used across breeds
- Important in multi breed populations
 - eg. beef, sheep, pigs

42

9. Erosion of LD in crosses vs. outbred pop.

Break-up of LD by recombination in outbred population:



Erosion LD for unlinked loci in F2 cross vs. outbred pop.

In random mating outbred populations, D does not drop immediately to zero for unlinked loci but LD is halved each generation:

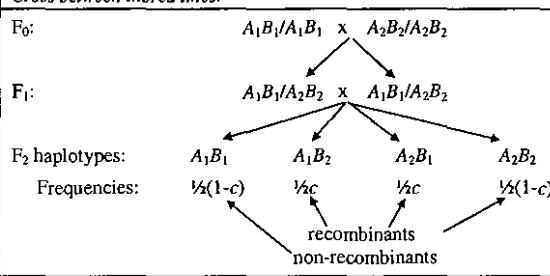
$$D_t = D_0(1-c)^t, \text{ which with } c=\frac{1}{2} \text{ gives: } D_t = D_0(1-\frac{1}{2})^t = \frac{1}{2}^t D_0$$

This is different when crossing breeds or lines, e.g. when producing an F₂ population for QTL mapping: in an F₂, LD=0 for unlinked loci.

Reason is that $D_t = D_0(1-c)^t$ only holds only if parents were produced by random mating.

In F₂ cross, parents (=F₁'s) were NOT produced by random mating (F₁'s are A_1B_1/A_2B_2):

Cross between inbred lines:



Disequilibrium in the F₂ is:

$$D_{F2} = \Pr(A_1B_1) - \Pr(A_1)\Pr(B_1) = \frac{1}{2}(1-c) - \frac{1}{2} \frac{1}{2} = \frac{1}{4}(1-2c)$$

→ for unlinked loci

$$c = \frac{1}{2} \rightarrow D_{F2} = 0$$

→ for completely linked loci

$$c = 0 \rightarrow D_{F2} = \frac{1}{4}$$

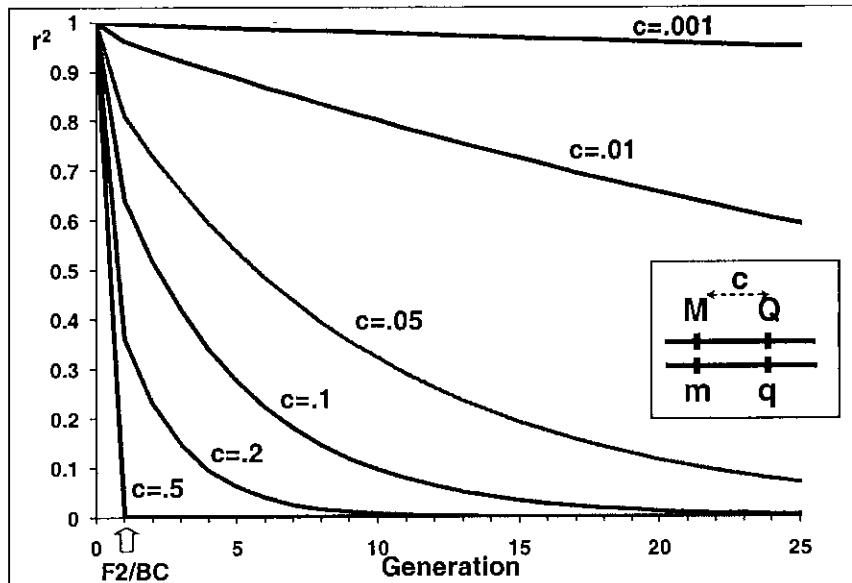
• But $D_{max} = \min(p_Aq_B, q_Ap_B) = \frac{1}{4} \rightarrow D'_{F2} = \frac{1}{4} / \frac{1}{4} = 1$

• Also: $r^2_{F2} = (D_{F2})^2 / (p_Aq_Ap_Bq_B) = (\frac{1}{4})^2 / (\frac{1}{2})^4 = 1 \rightarrow$ LD between linked loci is maximum in F₂

In F₃, etc, the standard equation does apply (random mating): $D_{F2+1} = D_{F2}(1-c)^t$

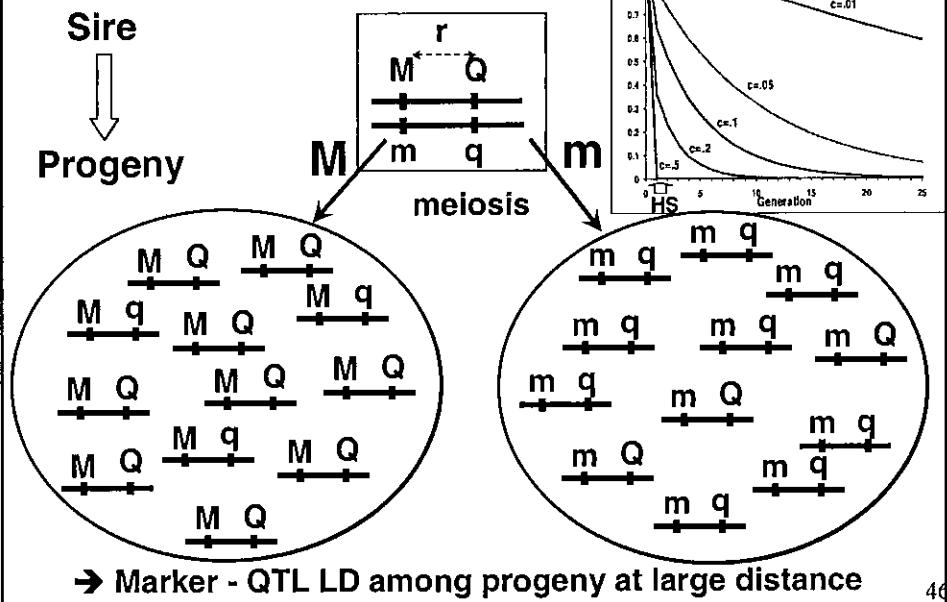
44

Break-up of LD by recombination in Advanced Intercross Line



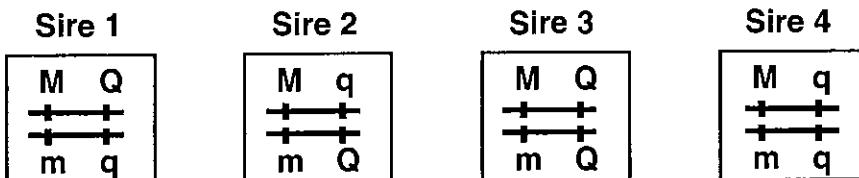
10. LD always exists within families

LD behavior similar to BC/F2



BUT Within-family LD is not consistent across families

Except for (close) markers with population-wide LD



Different marker-QTL linkage phases within each family

Linkage phase = assortment of alleles into haplotypes

Sire 1 has genotypes Mm and Qq; haplotypes MQ and mq

Alleles M and Q are in coupling linkage phase

Alleles M and q are in repulsion linkage phase

47

Day 1

Multi-locus Population Genetics – Linkage & Disequilibrium

Objective

Present population genetic principles of allele, haplotype and genotype frequencies, and of linkage and linkage disequilibrium

1. Single locus allele and genotype frequencies
2. Multi-locus haplotype and genotype frequencies
3. Measures of Linkage Disequilibrium (LD)
4. Estimating LD from genotype data
5. Linkage maps and recombination
6. Mechanisms that generate and erode LD
7. LD balance between drift and recombination
8. Persistence of LD across breeds
9. Erosion of LD in crosses vs. outbred population
10. LD always exists within families

4

Day 2 QTL Detection

Objective

Present principles for detection of genes affecting quantitative traits (QTL) using genetic markers
in 'simple' experimental designs
Concepts covered relevant to issues in 'genomic selection'

1. Single locus quantitative genetic model
2. Principle of use of LD to detect QTL using markers
3. Overview of strategies for QTL detection
4. QTL detection using line crosses
5. QTL interval mapping in line crosses
6. QTL detection in line crosses – additional topics
 - a. Significance testing
 - b. Accuracy of position estimates
 - c. Breed crosses (vs inbred line crosses)
7. QTL detection in outbred populations – linkage analysis
8. Summary and limitations
9. Software for QTL mapping

1

1. Single locus Quantitative Genetic Model

- Partition phenotype into genetic and environmental components:
$$P = \text{mean} + G + E$$
- G = collective effect of many genes
$$= \text{quantitative trait loci (QTL)}$$
- Genotypes for QTL have an associated genotypic value:

$$G_T = E(P | T)$$

G_T = phenotype you expect to get
from an individual with genotype T

G_T = Average phenotype over all individuals
with genotype T

G_T is often deviated from the mean → overall average G_T is zero

2

1

Falconer Model for effects of QTL

Genotype T	A_2A_2	A_1A_2	A_1A_1
Genotypic value G_T	$\mu - a$	μ	$\mu + d$

μ is NOT the population mean - it is the "mid-homozygote" value.
 - it is often standardized to zero (by subtraction)

Under HWE:	T	Frequency, $f(T)$	Genotypic value, G_T	$f(T) \times G_T$
	A_1A_1	p^2	a	p^2a
	A_1A_2	$2pq$	d	$2pqd$
	A_2A_2	q^2	$-a$	$-q^2a$

$$\begin{aligned}\text{Population mean} &= E(G_T) = M = p^2a + 2pqd + -q^2a \\ &= a(p - q) + 2pqd\end{aligned}$$

3

Example

The pygmy gene in mice

Allele frequency: $\Pr(+)$ = $p = 0.7$ $q = 0.3$

Genotype:	$++$	$+pg$	$pg\ pg$
Average weight (gr):	14	12	6 $\rightarrow \mu = \frac{14+6}{2} = 10$
Genotypic value G_T:	$a = 4$	$d = 2$	$-a = -4$
Expected freq. under HWE:	$p^2 = 0.49$	$2pq = 0.42$	$q^2 = 0.09$

$$\begin{aligned}\text{Mean } G_T &= E(G_T) = M = 0.49*4 + 0.42*2 + 0.09*(-4) = 2.44 \\ &= a(p - q) + 2pqd = 4(0.7 - 0.3) + 2*0.7*0.3*2 = 2.44\end{aligned}$$

$$\text{Expected population mean} = 0.49*14 + 0.42*12 + 0.09*6 = 12.44 = \mu + E(G_T)$$

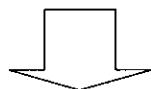
Most QTL have much smaller effects than the mouse pygmy gene
 and cannot be observed directly

4

How can we find these QTL?

Since we cannot observe the QTL directly,
we want to use (or create) an association
between the QTL and something we CAN
observe:

A genetic marker...

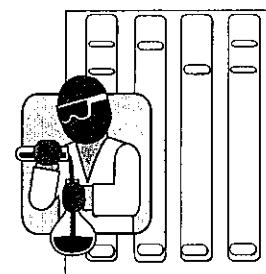
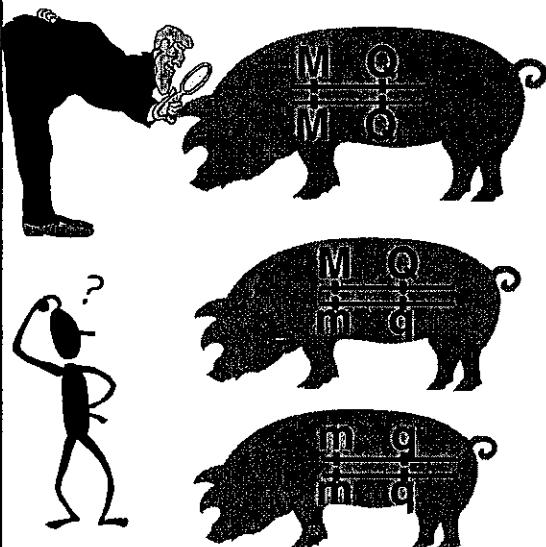


2. Principles of the use of LD to detect QTL using markers

5

Molecular Genetics

“In Search of the Holy Grail”



Major genes
Quantitative
Trait
Loci (QTL)

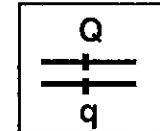
= position (locus) on
genome associated
with genetic
differences for a
quantitative trait

6

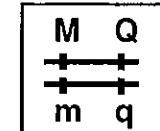
Most QTL cannot be observed at DNA level

Two types of observable molecular genetic loci

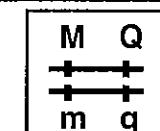
- Functional mutations - known genes
 - Most beneficial and easy to use
 - Difficult to find



- Anonymous markers linked to QTL
 - Easier to find
 - More restrictive and difficult to use



Use of markers for QTL detection and MAS relies on association of markers with phenotype



QTL detection

<u>Marker Genotype</u>	<u>Mean Phenotype</u>	
MM	20	
Mm	18	
mm	14	Allele M is associated with favorable QTL allele

MAS

Select MM or individuals that inherited allele M

Requires Linkage Disequilibrium between marker and QTL

**Illustration that marker genotype means don't differ
if marker and QTL are in Linkage Equilibrium**

Allele frequencies: $P(M)=p_M$ $P(m)=q_M$ $P(Q)=p$ $P(q)=q$ $D=0$

Genotypic value	Frequency		
	M	Q	
QTL has effect on phenotype, marker does not	M	Q	
	M	Q	$p_M^2 p^2$
	M	Q	
	M	q	$p_M^2 pq$
	M	q	
	M	Q	$p_M^2 pq$
	M	q	
	M	q	$p_M^2 q^2$
	M	Q	
	m	Q	$2p_M q_M p^2$
	m	Q	
	m	q	$2p_M q_M pq$
	m	q	
	m	Q	$2p_M q_M pq$
	m	q	
	m	Q	$q_M^2 p^2$
	m	q	
	m	q	$q_M^2 pq$
	m	Q	
	m	q	$q_M^2 q^2$
Average	$\mu+a(p-q)+2pqd$		$\mu+a(p-q)+2pqd$
			$\mu+a(p-q)+2pqd_9$

**Illustration that marker genotype means don't differ
if marker and QTL are in Linkage Equilibrium**

Allele frequencies: $P(M)=p_M$ $P(m)=q_M$ $P(Q)=0.7$ $P(q)=0.3$ $D=0$

Genotypic value	Example		
	M	Q	
10	M	Q	$p_M^2 (.49)$
	M	Q	
	M	Q	$2p_M q_M (.49)$
	M	Q	
	M	q	$p_M^2 (.21)$
	M	q	
	M	Q	$p_M^2 (.21)$
	M	q	
	M	Q	$2p_M q_M (.21)$
	M	q	
8	m	Q	$q_M^2 (.49)$
	m	Q	
	m	q	$2p_M q_M (.21)$
	m	q	
	m	Q	$q_M^2 (.21)$
	m	q	
	m	Q	$q_M^2 (.21)$
	m	q	
	m	Q	$q_M^2 (.09)$
	m	q	
Average	$.49*10+.21*8+.21*8+.09*5=8.71$		$.49*10+.21*8+.21*8+.09*5=8.71$
			$.49*10+.21*8+.21*8+.09*5=8.7$

**Detection of QTL based on markers requires
Linkage Disequilibrium between marker and QTL**
Relative frequency of Q must differ between marker genotypes

Example (arbitrary)

$$\text{Allele frequencies: } P(M) = p_M = 0.4 \quad P(m) = q_M = 0.6 \\ P(Q) = p = 0.7 \quad P(q) = q = 0.3$$

**Assumed
Haplotype frequencies**

$$\begin{array}{c} M \quad Q \\ \text{---} \quad \text{---} \end{array} \quad 0.38 = p_M p + D$$

$$\begin{array}{c} M \quad q \\ \text{---} \quad \text{---} \end{array} \quad 0.02 = p_M q - D$$

$$\begin{array}{c} m \quad Q \\ \text{---} \quad \text{---} \end{array} \quad 0.32 = q_M p - D \quad \text{Disequilibrium} = D \\ \begin{array}{c} m \quad q \\ \text{---} \quad \text{---} \end{array} \quad 0.28 = q_M q + D \quad = P(MQ) - p_M p \\ = 0.38 - (0.4)(0.7) = +0.10$$

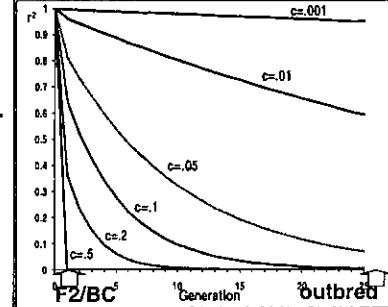
Genotypic value	Example D=+0.10		
	Random mating of parents		
10	$M \quad Q$ Frequency $(.38)(.38) = .1444$	$M \quad Q$ $2(.38)(.32) = .2432$	$m \quad Q$ $(.32)(.32) = .1024$
8	$M \quad Q$ $(.38)(.02) = .0076$	$m \quad q$ $2(.38)(.28) = .2128$	$m \quad q$ $(.32)(.28) = .0896$
8	$M \quad q$ $(.02)(.38) = .0076$	$m \quad Q$ $2(.02)(.32) = .0128$	$m \quad Q$ $(.28)(.32) = .0896$
5	$M \quad q$ $(.02)(.02) = .0004$	$m \quad q$ $2(.02)(.28) = .0112$	$m \quad q$ $(.28)(.28) = .0784$
Average		9.80	8.94
		7.92	
12			

3. Overview of Strategies for QTL Detection

Depend on the type of LD between markers and QTL
that you want to exploit

- LD you create by a cross
 - F2 cross
 - Backcross
 - Advanced Intercross Line – AIL
 - Recombinant Inbred Line – RIL
- LD that exists within families
 - Within half-sib families
 - In extended pedigree
- LD that is already present in an outbred population
 - LD created in past by drift, mutation, selection, migration

Strategies differ in the # of rounds of recombination that occurred since creation of LD and, therefore, in how close a marker needs to be to be in sufficient LD with a QTL



Type of LD used affects marker density required, type of analysis needed, and how results are to be interpreted

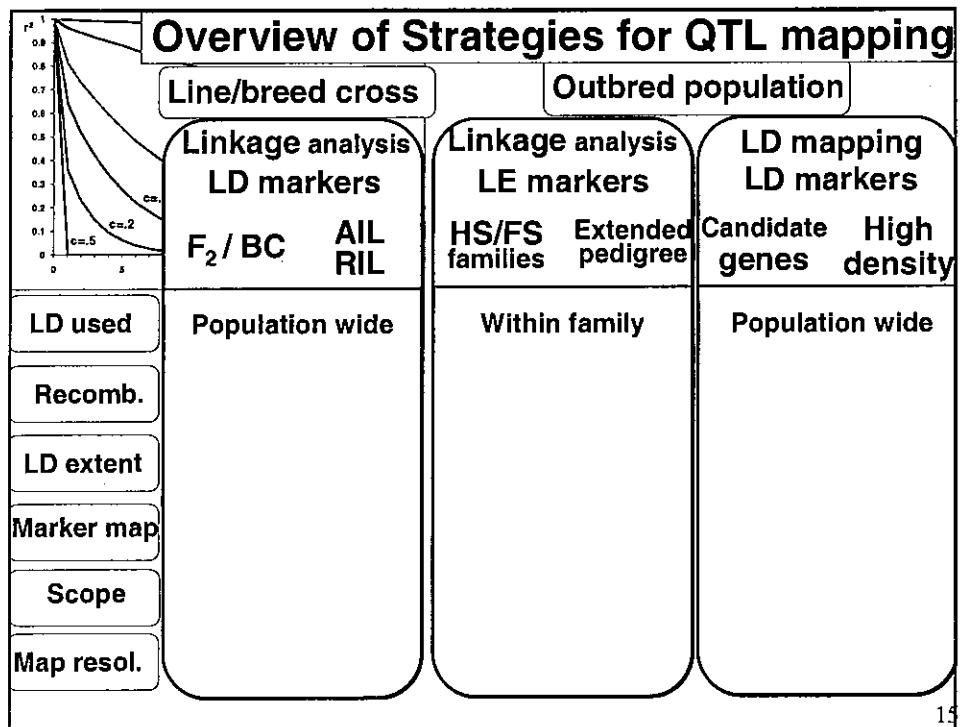
13

Scope of QTL Detection Strategy

- Targeted – e.g. candidate gene approach
 - Look for QTL in targeted region of the genome
- Genome-wide – genome scan approach
 - Place markers across the genome
 - Look for associations of markers with trait phenotype across the genome
 - Identify QTL across the genome

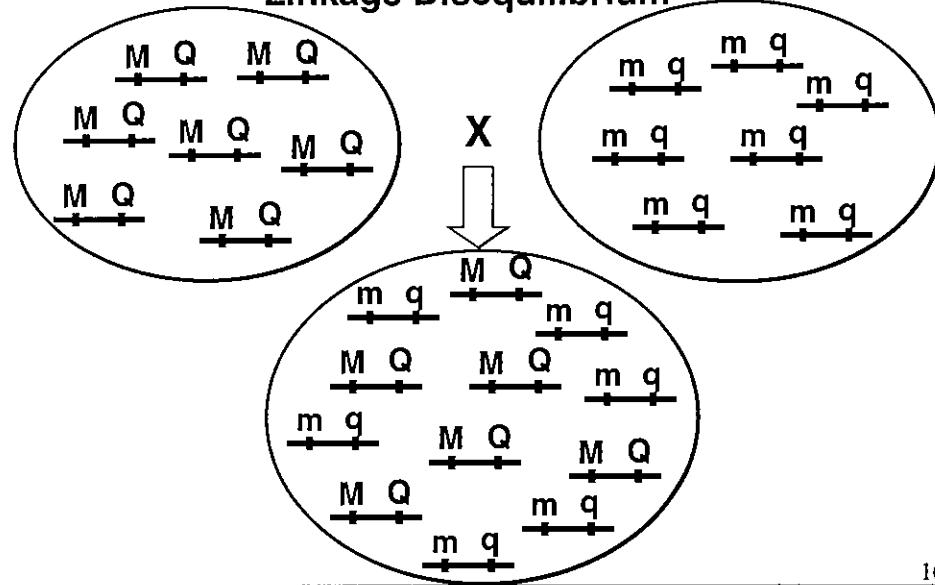
M_1	M_2	M_3	Q	M_4	M_5	M_6
m_1	m_2	m_3	q	m_4	m_5	m_6

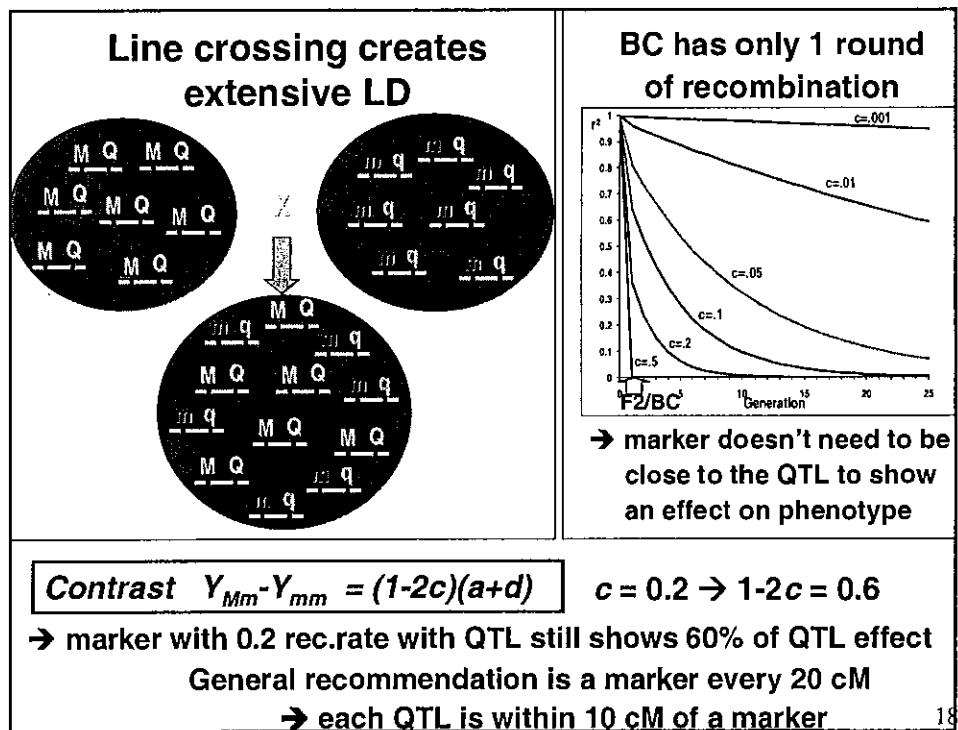
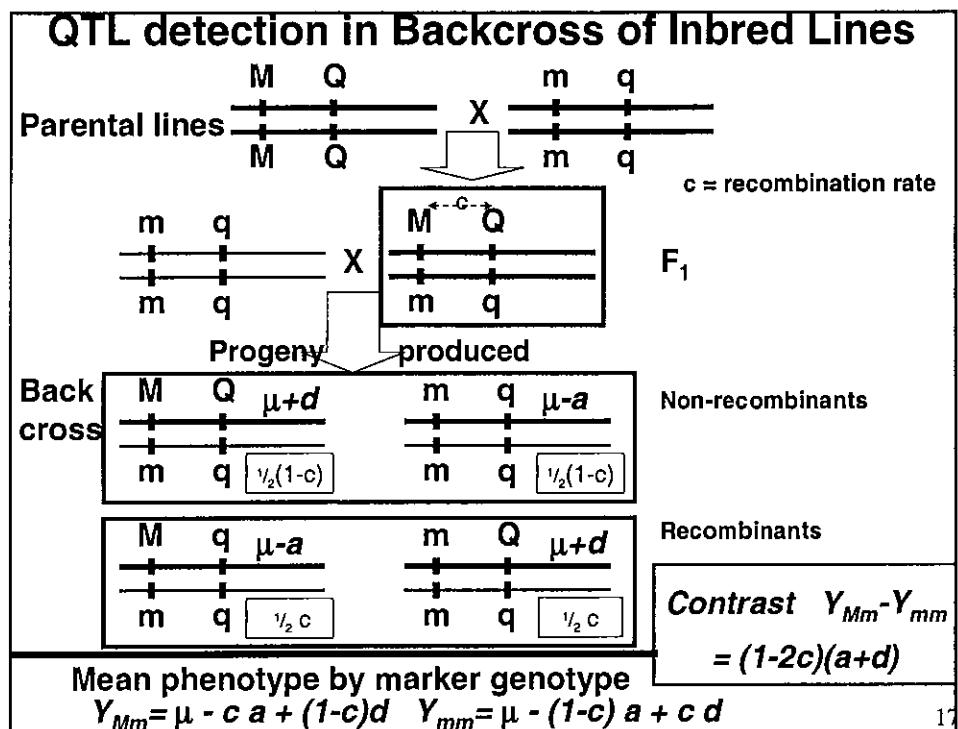
14



4. QTL detection in Line Crosses

Line crossing creates extensive
Linkage Disequilibrium





F₂ Cross between Inbred Lines

F ₂			
M Q $\mu+a$	m Q $\mu+a$		
M Q $\frac{1}{4}(1-c)(1-c)$	m Q $\frac{1}{4}cc$		
M q $\mu+d$	m q $\mu+d$		
M Q $\frac{1}{4}c(1-c)$	m Q $\frac{1}{4}(1-c)c$		
M Q $\mu+d$	m Q $\mu+d$		
M q $\frac{1}{4}(1-c)c$	m q $\frac{1}{4}c(1-c)$		
M q $\mu-a$	m q $\mu-a$		
M q $\frac{1}{4}cc$	m q $\frac{1}{4}(1-c)(1-c)$		

Expected mean of marker genotypes

$$Y_{MM} = \mu + (1-c)^2 a + 2c(1-c)d - c^2 a \quad Y_{mm} = \mu + c^2 a + 2c(1-c)d - (1-c)^2 a$$

19

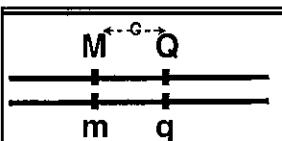
F₂ Cross between Inbred Lines

	M Q $\mu+a$	m Q $\mu+a$	m Q $\mu+a$
M Q $\frac{1}{4}(1-c)(1-c)$	m Q $\frac{1}{4}(1-c)c$	M Q $\frac{1}{4}c(1-c)$	m Q $\frac{1}{4}cc$
M q $\mu+d$	m q $\mu+d$	m q $\mu+d$	m q $\mu+d$
M Q $\frac{1}{4}c(1-c)$	m Q $\frac{1}{4}cc$	M Q $\frac{1}{4}(1-c)(1-c)$	m Q $\frac{1}{4}(1-c)c$
M Q $\mu+d$	m Q $\mu+d$	m Q $\mu+d$	m Q $\mu+d$
M q $\frac{1}{4}(1-c)c$	m q $\frac{1}{4}(1-c)$	m q $\frac{1}{4}cc$	m q $\frac{1}{4}c(1-c)$
M q $\mu-a$	m q $\mu-a$	m q $\mu-a$	m q $\mu-a$
M q $\frac{1}{4}cc$	m q $\frac{1}{4}(1-c)$	M q $\frac{1}{4}(1-c)c$	m q $\frac{1}{4}(1-c)$

$Y_{MM} = \mu + (1-c)^2 a + 2c(1-c)d - c^2 a$ $Y_{Mm} = \mu + c^2 d + (1-c)^2 d$ $Y_{mm} = \mu + c^2 a + 2c(1-c)d - (1-c)^2 a$

$a(1-2c) = (Y_{MM} - Y_{mm})/2$ $d(1-2c)^2 = Y_{Mm} - 1/2(Y_{MM} + Y_{mm})$

20



Summary

Expectation if $c = 0.5$

Backcross: $(1-2c)(a+d) = Y_{Mm} - Y_{mm} = 0$

F2 cross: $(1-2c)a = (Y_{MM} - Y_{mm})/2 = 0$

$$(1-2c)^2 d = Y_{Mm} - \frac{1}{2}(Y_{MM} + Y_{mm}) = 0$$

Estimates confound QTL position and effect

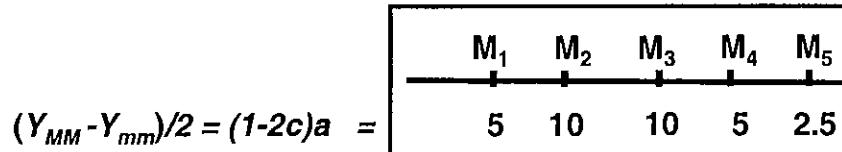
E.g. if $(Y_{MM} - Y_{mm})/2 = 10 \text{ kg}$ (F2 cross)

- QTL could be near M with $a = 10$ (if $c=0$)
 - QTL could be distant ($c=0.25$) with $a = 20$
 - or any other possibility
 - QTL can be on either side of the marker
- $\left. \right\} \text{Marker-associated effect} = 10$

21

But, if we test multiple markers

and find the following marker-associated effects:



there is evidence that the QTL is between M_2 and M_3

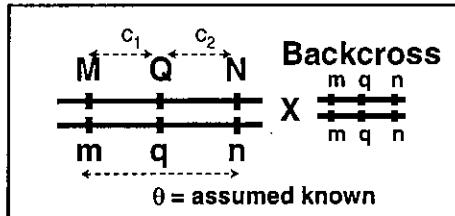
(although we cannot exclude presence of multiple QTL)

22

5. QTL Interval Mapping in Line Crosses

Use of flanking markers

To estimate QTL position and effect separately



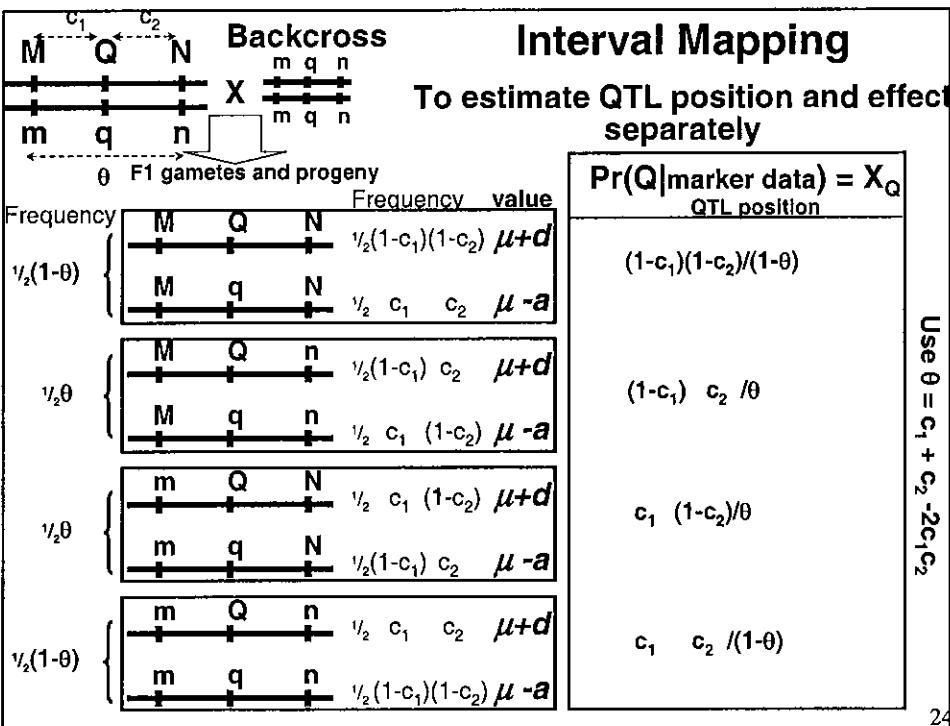
$$\text{Contrast } Y_{Mm} - Y_{mm} = (1-2c_1)(a+d)$$

$$\text{Contrast } Y_{Nn} - Y_{nn} = (1-2c_2)(a+d)$$

$$\text{No interference} \rightarrow \theta = c_1 + c_2 - 2c_1c_2$$

} 3 equations
3 unknowns $c_1, c_2, (a+d)$

23



$E(Y_i | \text{Marker Genotype})$

- Two possible QTL genotypes: Qq or qq
 - If Qq , $E(Y_i | Qq) = \mu + d$
 - If qq , $E(Y_i | qq) = \mu - a$
- Put those two together with $P(Qq | g_{\text{marker}}) = X_{Qi}$
and $P(qq | g_{\text{marker}}) = 1 - X_{Qi}$
- $$\begin{aligned} E(Y_i | M) &= (\mu + d)X_{Qi} + (\mu - a)(1 - X_{Qi}) \\ &= (\mu - a) + (a + d)X_{Qi} \\ &= m + b_Q X_{Qi} \end{aligned}$$
- Regression model: $Y_i = m + b_Q X_{Qi} + e$

25

Regression Interval Mapping

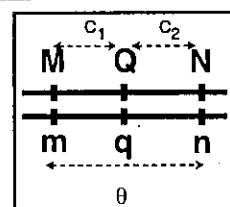
Estimate QTL position and effect separately

Haley and Knott (1992)
Heredity 69: 315

Backcross regression model

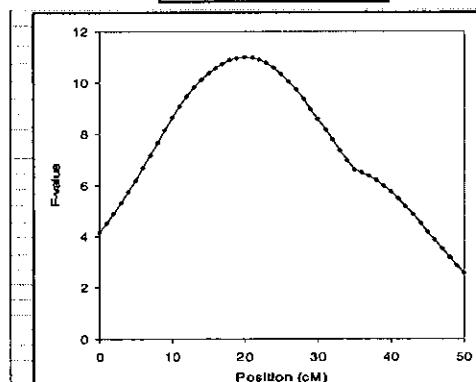
$$Y_i = m + b_Q X_{Qi} + e_i$$

$$E(b_Q) = a+d$$



Fit Model for various positions of QTL
(e.g. in steps of 1 cM)

Position with lowest RSS
or highest F-test gives best estimate of c_1 and $b_Q (=a+d)$



F₂ Cross between Inbred Lines

g _{markers}	Pr(QQ g _{markers})	Pr(Qq g _{markers})	Pr(qq g _{markers})
MM NN	$f(c_1, c_2, \theta)$		
MM Nn		$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$
Mm nn			

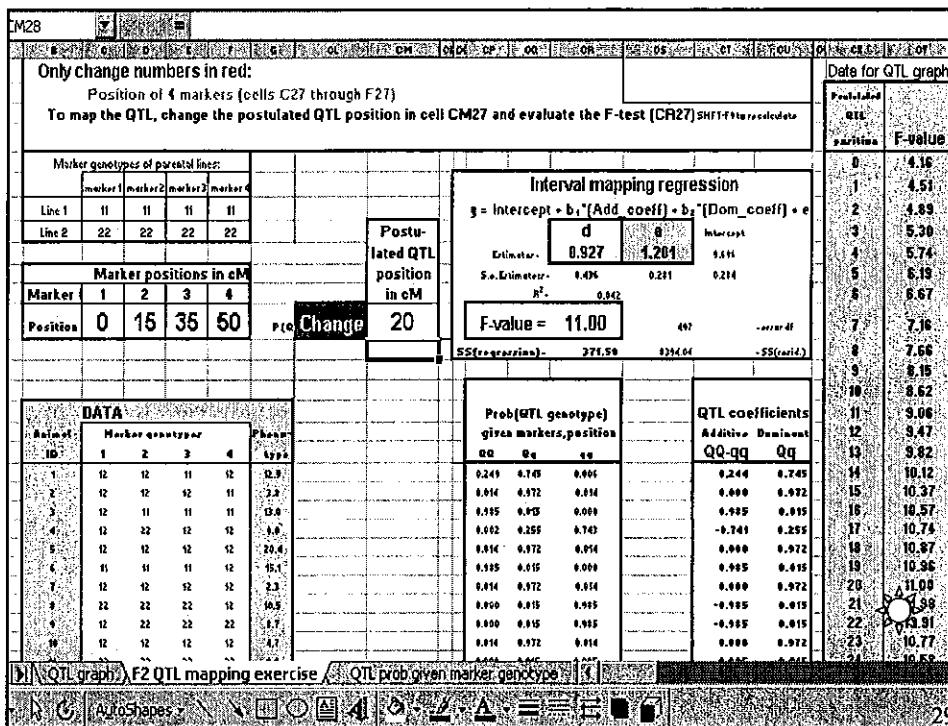
F₂ Cross between Inbred Lines Haley and Knott (1992)
Heredity 69: 315

Markers	Pr(QQ)	Pr(Qq)	Pr(qq)	Additive coef. X_{add} Pr(QQ)-Pr(qq)	Dom. Coef. X_{dom} Pr(Qq)
MM NN					
MM Nn	$f(c_1, c_2, \theta)$				
MM nn					
Mm NN	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$
Mm Nn					
Mm nn					
mm NN					
mm Nn					
mm nn					

$Y_i = \mu + b_a X_{add,i} + b_d X_{dom,i} + e_i$ at QTL position

$E(b_a) = a$ $E(b_d) = d$

Fitted at each 1 cM position on chromosome
Position with highest F-test \rightarrow QTL (if significant)



6. QTL detection in line crosses

Additional Topics

(see also Lynch and Walsh ch 15)

- a. Significance test for presence of QTL
 - b. Accuracy of position estimates
 - Advanced intercross lines
 - c. Breed Crosses (vs inbred line crosses)

6a. How to decide if you've detected a QTL?

Test statistic (e.g. F or LR) > threshold T

Set T to control the Type I error rate (False Positives)

- Comparison-wise test at 5% : set threshold T such that:

- $\text{Prob}(\text{test} > T \mid \text{no QTL}) < .05$ allow 5% FP tests

Possible outcomes for test for QTL at a given position:

Result of significance test		
True state	Accept H_0	Reject H_0
H_0 is true (no QTL)	True negative	False positive Type I error
H_0 is false (QTL)	False negative Type II error	True positive

31

Expected result for tests at 100 positions on chromosome with NO QTL at 5% comparison-wise test level:

Result of significance test		
True state	Accept H_0	Reject H_0
H_0 is true (no QTL)	95	5 Type I error
H_0 is false (QTL)	0 Type II error	0

→ Significance testing complicated by:

- Large # tests performed (many markers, QTL positions)
 - At $\alpha = 0.05$, 5% of tests significant even if no QTL exist
- Tests on the same chromosome are dependent
 - Bonferroni adjustment ($\alpha^* = \alpha / (\# \text{ tests})$) is too stringent

32

Strategies to control % false positives (%FP)

(Lander & Kruglyak, 1995, Nature Genetics 11: 241-247)

- Chromosome-wise test - control % FP at chrom. level
 - Account for multiple (correlated) tests on chrom.
 - # FP/chromosome ≥ 1 on 5% of chromosomes
- Experiment-wise test - control %FP within experiment
 - Account for all tests conducted in experiment
 - # FP/experiment ≥ 1 on 5% of experiments
- Genome-wise test - control % FP at genome level
 - Account for all tests conducted on the genome
 - # FP/genome ≥ 1 on 5% of genomes tested
- Significance Levels (Lander & Kruglyak, 1995)
 - Significant Linkage at $p < .05$: $\text{Prob}(\geq 1 \text{ FP}) < .05$
 - Suggestive Linkage : at least 1 false positive test

33

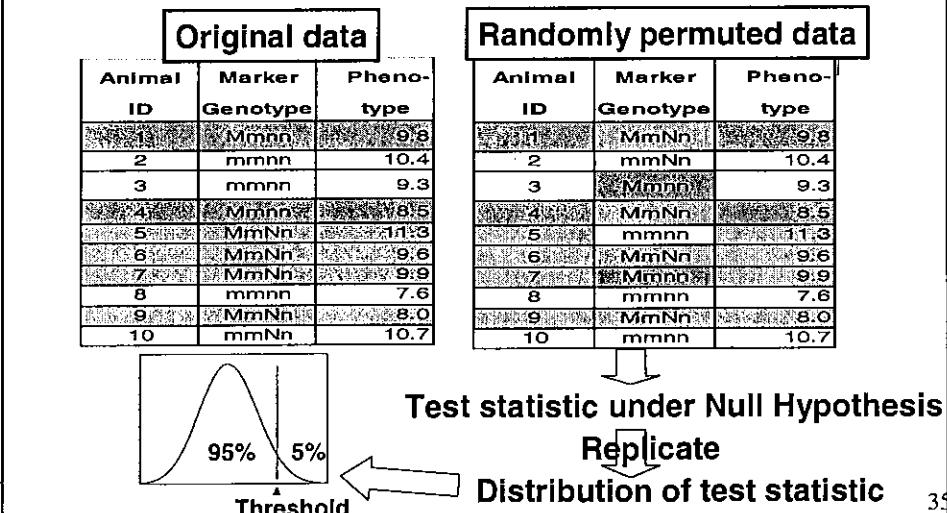
Computing significance thresholds

- Adjust Table test statistic values by equation of Lander & Kruglyak (1995)
 - Assumes high-density marker map
- Develop empirical threshold based on permutation test
(Churchil and Doerge, 1994, Genetics 138:963)
 - Simulate data under the Null Hypothesis (=no QTL)
 - Compute test statistic (F-test / LR)
 - Replicate many times
 - Determine 95 % level of tests statistic (for 5% test)

34

Significance thresholds by Permutation test (Churchill&Doerge, 1994 Genetics 138:963)

- Simulate data under the Null Hypothesis (=no QTL)
- Compute test statistic (F-test / LR)
- Replicate many times
- Determine 95 % level of tests statistic (for 5% test)



35

Control of False Discovery Rate (FDR)

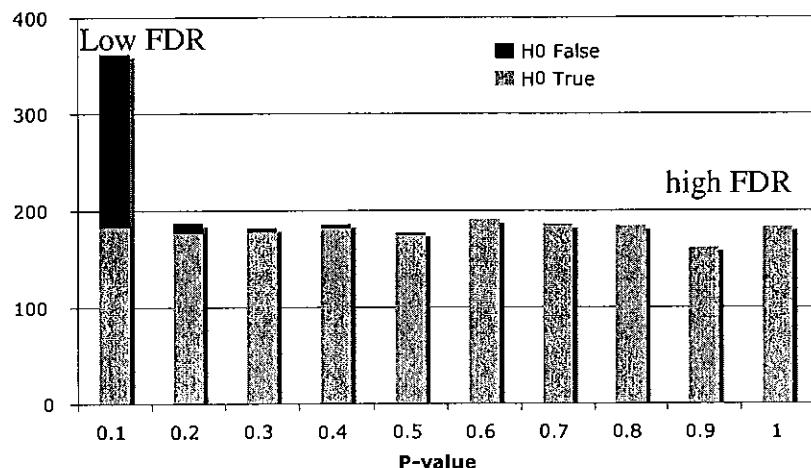
		Result of significance test	
True state		Accept H_0	Reject H_0
H_0 is true		U	V Type I error
H_0 is false		T Type II error	S

FDR - Control the expected proportion of significant tests that are false positives

- Control $E(V/(V+S))$

36

Frequency Distribution of p-values across many tests



See notes “False discovery rate.doc” for further details

37

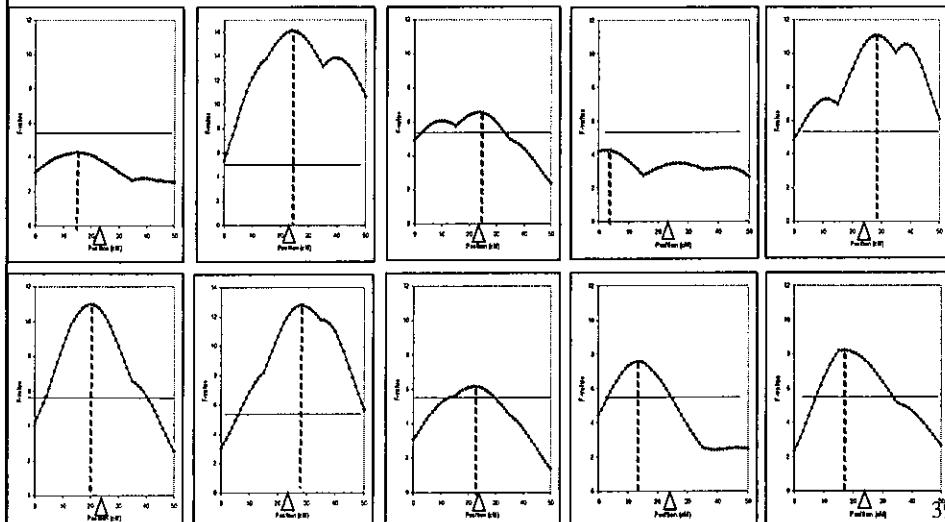
6. QTL detection in line crosses

Additional Topics

- a. Significance test for presence of QTL
- b. Accuracy of position estimates
 - Advanced intercross lines
- c. Breed Crosses (vs inbred line crosses)

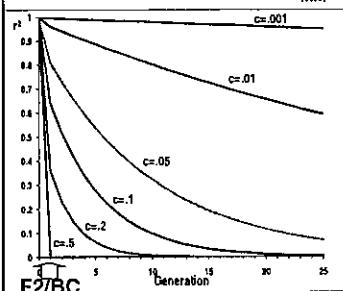
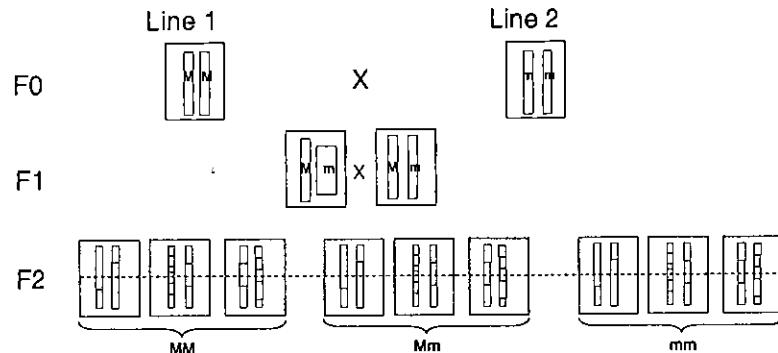
38

Replicate Genome Scan results for F2
N=500 6 markers Trait with SD=2
QTL at 23 cM a=1 d=0.5 --> 14% of variance



Replicate	Position	a	d
1	15	0.791	0.19
2	24	1.56	0.19
3	23	1.03	0.3
4	27	0.771	0.24
5	20	1.201	0.93
6	28	1.35	0.94
7	22	0.96	0.14
8	13	0.991	0.64
9	17	0.94	0.52
10	29	0.924	1.44
Average	21.8	1.052	0.55
St.dev.	5.231	0.236	0.41
TRUE	23	1	0.5
			40

F2 cross design

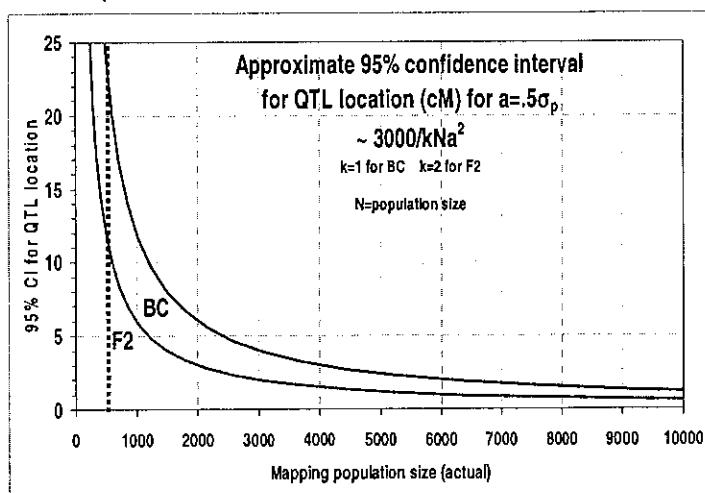


- Large chunks
- High LD
- Only 1 round of recombination
→ low accuracy of QTL position

4

Resolving Power of QTL Mapping

(Darvasi & Soller 1997. Behavior Genetics)



Increase resolution with advanced intercross lines

- Recombination breaks genome up in smaller pieces
 - reduces LD except at short distance

(Darvasi & Soller 1995 Genetics)

Strategies to increase accuracy of estimates of QTL position in line crosses

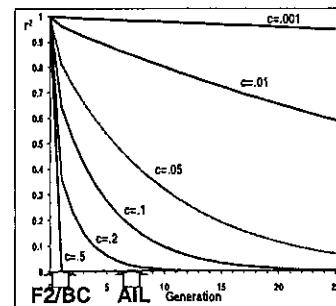
F2/BC:

- Increasing marker density limited effect
- Increase population size

Advanced intercross lines

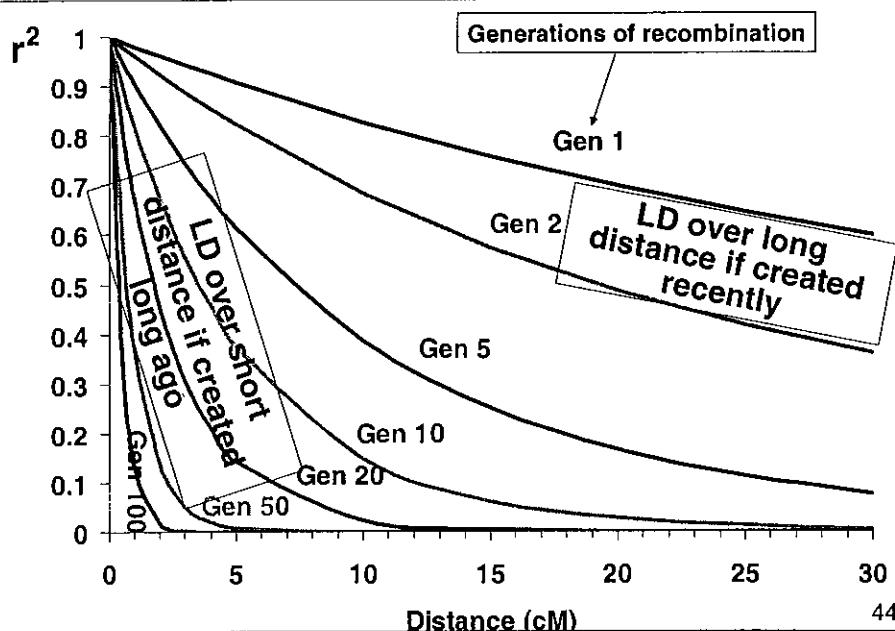
→ Higher accuracy of QTL position

- Requires more markers to maintain power to detect QTL (lower LD)



43

Recent LD extends over large distances



44

Overview of Strategies for QTL mapping

Line/Breed cross		Outbred population	
Linkage analysis LD markers		Linkage analysis LE markers	LD mapping LE markers
F_2 / BC AIL		HS/FS families Ext. pedigree	Cand. genes High density
LD used	Population wide		
Recomb.	1 rnd >1 rnd		
LD extent	Long Smaller		
Marker map	Sparse Denser		
Coverage	Genome wide		
Map resol.	Poor Better		

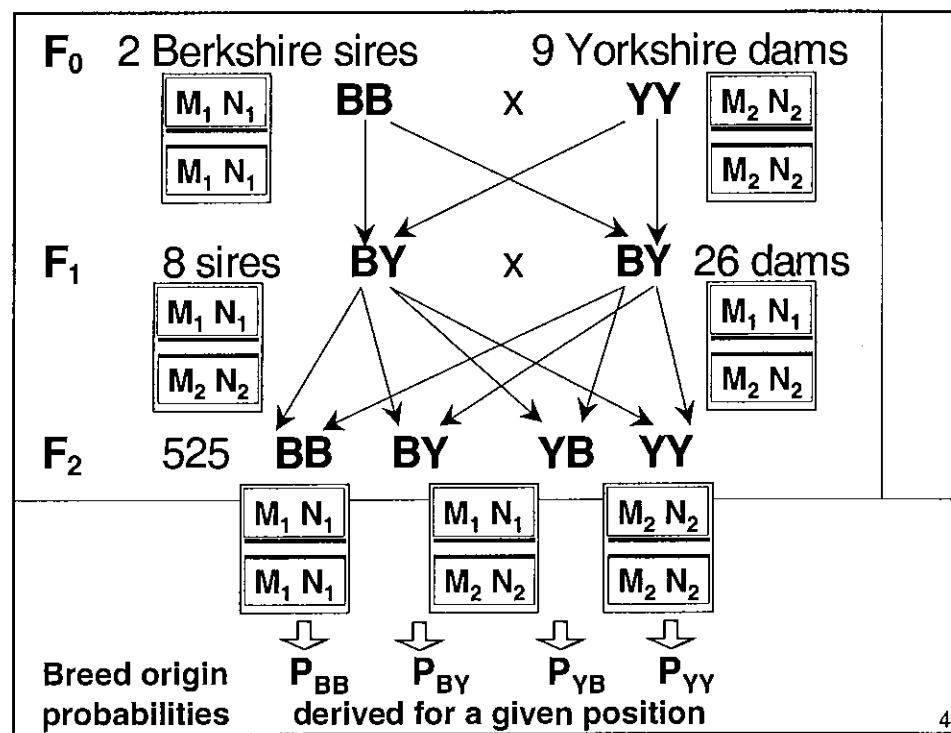
45

6. QTL detection in line crosses

Additional Topics

- a. Significance test for presence of QTL
- b. Accuracy of position estimates
 - Advanced intercross lines
- c. Breed Crosses (vs inbred line crosses)

46



48

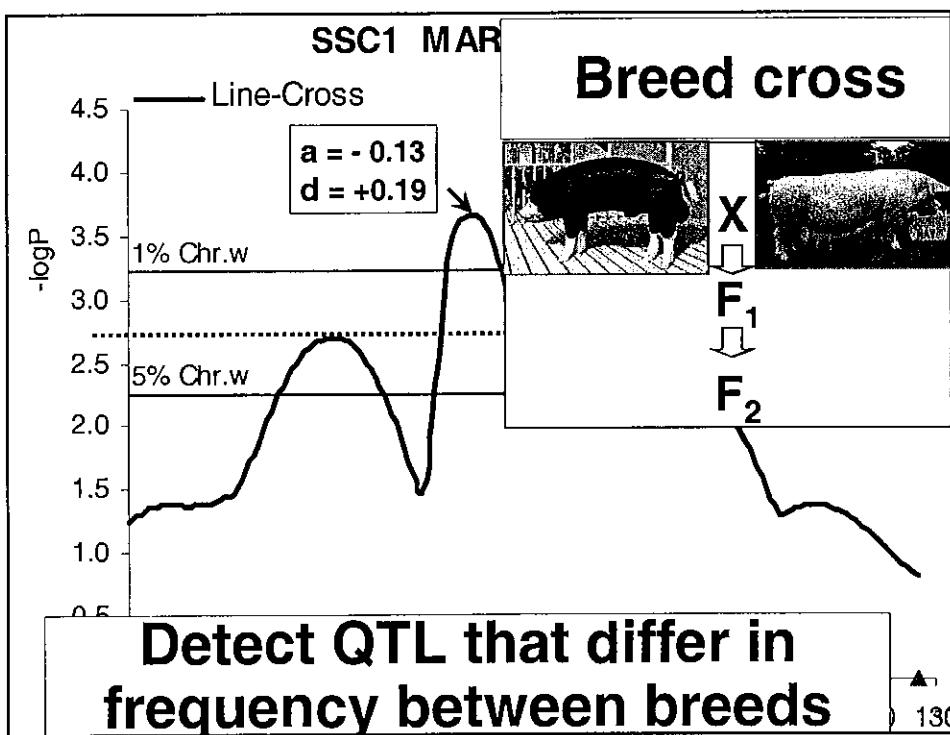
F₂ Cross between breeds					Haley and Knott (1992) Heredity 69: 315
Identical to cross of inbreds but follow B vs. Y alleles				Additive coef. X_{add} Pr(BB)-Pr(YY)	Dom. Coef. X_{dom} Pr(BY)
Markers	Pr(BB)	Pr(BY)	Pr(YY)		
MM NN					
MM Nn					
MM nn					
Mm NN	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$	$f(c_1, c_2, \theta)$
Mm Nn					
Mm nn					
mm NN					
mm Nn					
mm nn					

$Y_i = \mu + b_a X_{add,i} + b_d X_{dom,i} + e_i$ at QTL position

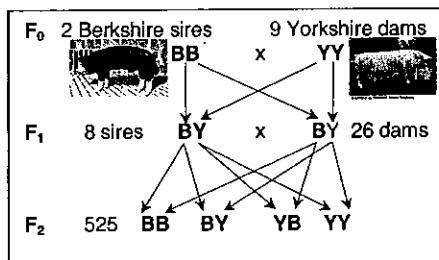
$E(b_a) = a$ $E(b_d) = d$

Fitted at each 1 cM position on chromosome
Position with highest F-test → QTL (if significant)

49



Breed cross interval mapping



Compares average Berk allele to average York allele

→ QTL only detected if breeds differ in frequency

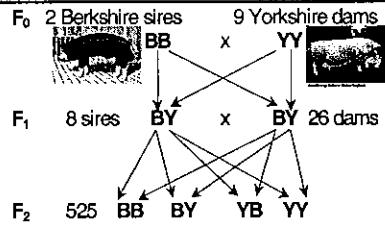
	<u>Berk</u>	<u>X</u>	<u>York</u>	
Frequency of Q	p_B		p_Y	
Line cross additive effect	=	$(p_B - p_Y)a$		
Line cross dominance effect	=	$(p_B - p_Y)d$		QTL effect
				QQ +a Qq d qq -a

Summary of QTL mapping in Line/ Breed Crosses

- QTL detection requires LD between markers and QTL
- Cross → extensive LD
→ genome scan with markers @ 20 cM
- Regression interval mapping
→ estimate QTL position, effect
- Estimates have limited accuracy
→ 10 – 30 cM confidence intervals
- Fine mapping not limited by # markers but requires
 - larger populations
 - crosses that accumulate recombinations
 - Recombinant Inbred Lines
 - Advanced Intercross Lines
- Only detects QTL that differ between breeds

52

Breed cross QTL scan



- QTL that differ in frequency between breeds
- Wide QTL region (20-50 cM)

Within-breed MAS requires QTL that segregate within breeds

Follow-up within-breed research in QTL region:

→ Linkage mapping → see next

Evans et al. (2003 Genetics:621) - confirmed QTL in 10 commercial lines

→ LD mapping → day 3

53

Overview of Strategies for QTL mapping					
Line/Breed cross			Outbred population		
	Linkage analysis LD markers F ₂ / BC AIL		Linkage analysis LE markers HS/FS families Ext. pedigree		LD mapping LE markers Cand. genes High density
LD used	Population wide				
Recomb.	1 rnd >1 rnd				
LD extent	Long Smaller				
Marker map	Sparse Denser				
Coverage	Genome wide				
Map resol.	Poor Better				

54

7. QTL detection in outbred populations – linkage analysis

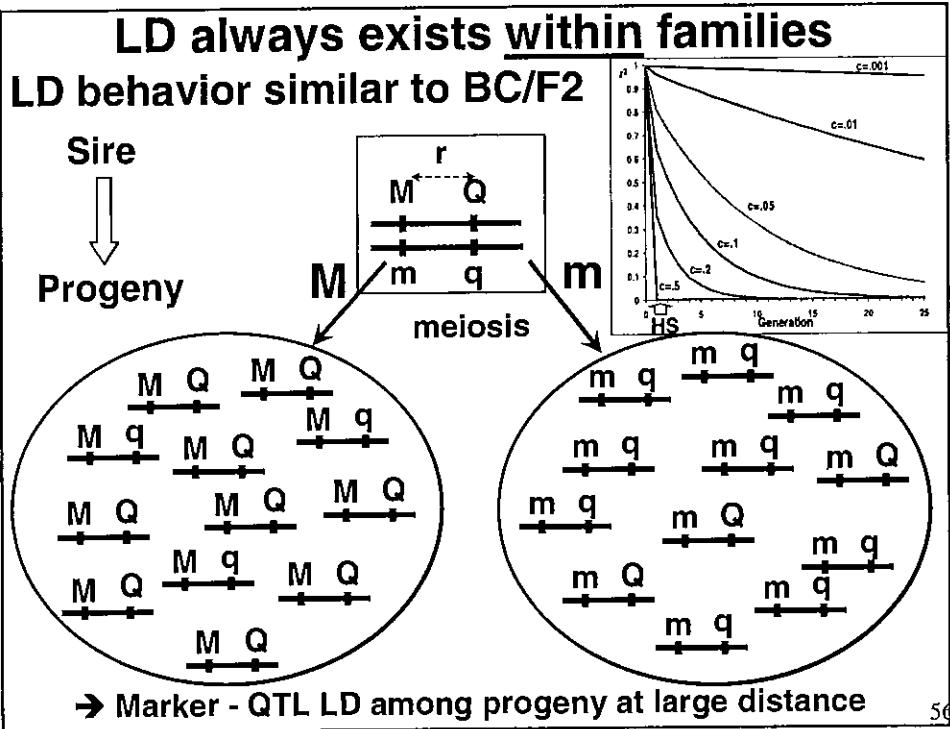
e.g. livestock, wildlife, human

Reading

Dekkers and van der Werf (2007) Chapter 10 at

<http://www.fao.org/docrep/010/a1120e/a1120e00.htm>

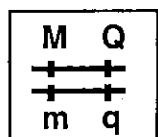
55



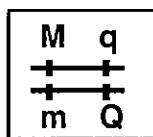
QTL mapping in half-sib family design

Within-family LD not consistent across families

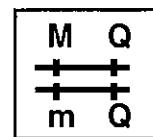
Sire 1



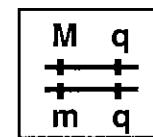
Sire 2



Sire 3



Sire 4



→ Analysis must allow for

different marker-QTL linkage phases within each family

QTL effects must be fitted w/in family:

$$Y_{ij} = \mu_i + \alpha_{Q,i} P_{Q,ij} + e_{ij}$$

$P_{Q,ij} = \text{Prob}(Q_{Mi} | \text{marker genotype, QTL position})$

$\alpha_{Q,i}$ = QTL allele substitution effect for sire i

See e.g. Knott et al. Theor.Appl.Genet. 1996. 93: 71-80

57

Power of alternative QTL mapping designs

For given number of animals genotyped

F2 > BC > Fullsib > Halfsib

Typical size used
animals

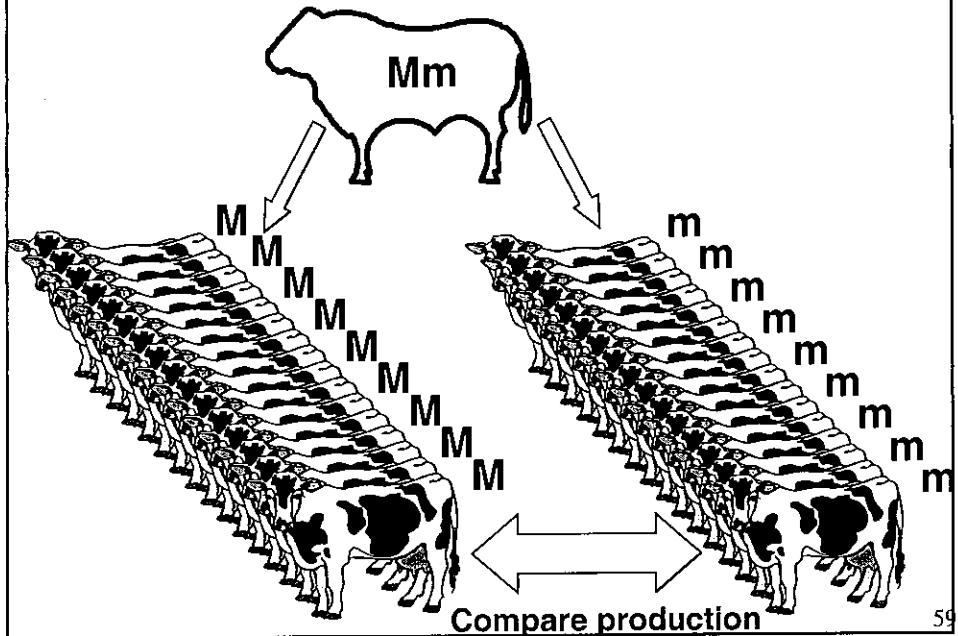
> 500 animals

>1000

Outbred designs: Fraction p^2+q^2 of parents are homozygous for QTL
= non-informative

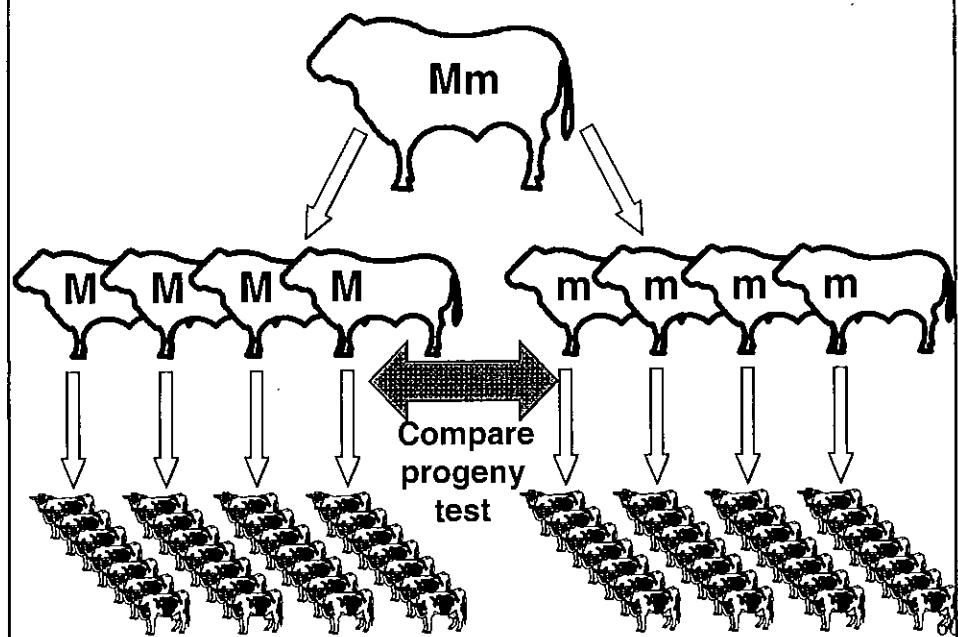
58

Daughter design for QTL detection and MAS

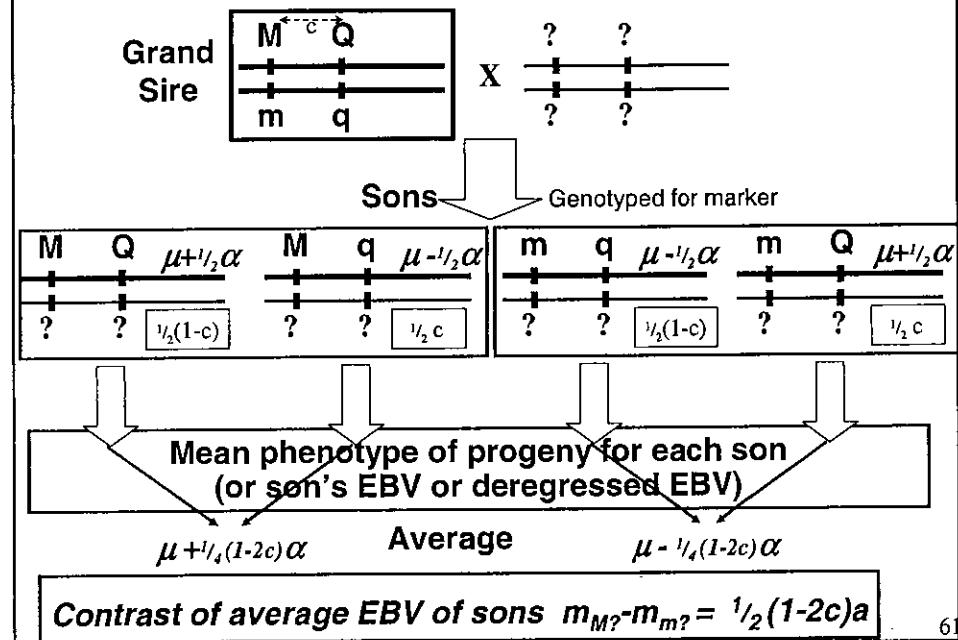


59

Grand daughter design



Grand-daughter Design (Weller et al. 1990)



61

Overview of Strategies for QTL mapping					
Line/Breed cross		Outbred population			
	Linkage analysis LD markers F_2 / BC AIL	Linkage analysis LE markers HS/FS families	LD mapping LD markers Cand. genes		
LD used	Population wide	Within family			
Recomb.	1 rnd >1 rnd	1 rnd >1 rnd			
LD extent	Long Smaller	Long Smaller			
Marker map	Sparse Denser	Sparse Denser			
Coverage	Genome wide	Genome wide			
Map resol.	Poor Better	Poor Better			
Linkage Analysis in extended pedigrees by random QTL effects - see later					

62

8. Summary and limitations of QTL mapping in outbred populations using sparse markers

- Within family → extensive LD
 - genome scan with markers @ 20 cM
- Regression interval mapping
 - estimate QTL position, effect
- Estimates of marker/QTL effects differ by family
 - complicates MAS
- Estimates have limited accuracy
 - 10 – 30 cM confidence intervals
- Fine mapping not limited by # markers but requires
 - larger populations
 - Populations that accumulate recombinations
 - Linkage analysis in deep pedigrees
 - Historical recombination → LD mapping

63

Software for QTL mapping by linkage analysis

Many programs available (with tutorials)

See: <http://linkage.rockefeller.edu/soft/list.html>

- For inbred line crosses: Mapmaker QTL
 - http://www.broad.mit.edu/genome_software/other/qtl.html
 - <http://darwin.eeb.uconn.edu/notes/qtl-mapmaker.pdf>
- For breed crosses and outbred populations: QTL Express
 - <http://qtl.cap.ed.ac.uk/>

64

Day 2 QTL Detection

Objective

Present principles for detection of genes affecting quantitative traits (QTL) using genetic markers in 'simple' experimental designs

Concepts covered relevant to issues in 'genomic selection'

- 1. Single locus quantitative genetic model**
- 2. Principle of use of LD to detect QTL using markers**
- 3. Overview of strategies for QTL detection**
- 4. QTL detection using line crosses**
- 5. QTL interval mapping in line crosses**
- 6. QTL detection in line crosses – additional topics**
 - a. Significance testing**
 - b. Accuracy of position estimates**
 - c. Breed crosses (vs inbred line crosses)**
- 7. QTL detection in outbred populations – linkage analysis**
- 8. Summary and limitations → need for LD mapping**
- 9. Software for QTL mapping**

65

