Statistical Methods in Genetic Epidemiology

Michael Knapp

Institute for Medical Biometry, Informatics and Epidemiology

University of Bonn

Organizational issues

Web site:

`http://imbie.meb.uni-bonn.de/~knapp/`

Written exam:

date to be announced

Literature:

“Statistics in Genetics” by Peter Almgren et al.

(`www.maths.lth.se/matstat/kurser/statgen/book`)

What is Genetic Epidemiology?

King et al., Ann Rev Public Health 5: 1–52 (1984):

Page 1 (Definition of Genetic Epidemiology):

Genetic epidemiology is the study of how and why diseases cluster in families and ethnic groups. Genetic Epidemiology addresses three questions:

1. Do diseases cluster in families?
2. Is familial clustering caused by common environmental exposure, biologically inherited susceptibility, or culturally inherited risk factors?
3. How is genetic susceptibility inherited?

Page 43 (Goal of Genetic Epidemiology):

Identification of specific susceptibility genes whose existence is inferred by statistical evidence.

Overview

- Introduction to genetics (very brief)
- Basic concepts from probability theory
- Basic concepts from inference theory
- Parametric linkage analysis
- Nonparametric linkage analysis
- Linkage analysis and imprinting
- Association analysis
- Haplotype frequency estimation
- Power calculations for linkage/association studies
- Association analysis and logistic regression
- Genome-wide association studies (GWAS)

DNA

Deoxyribonucleic acid (DNA):

double-stranded molecule, where each strand consists of a linear arrangement of four types of nucleotides: adenine (A), guanine (G), cytosine (C), or thymine (T).

```
... - - - - - - - - - - - - - - - - - - - - - - - - - - - ...
    A C C G T A T A A C G A T C C T C T G A
    ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮
    T G G C A T A T T G C T A G G A G A C T
... - - - - - - - - - - - - - - - - - - - - - - - - - - - ...
```

Each of the two strands contains all of the information present in the other strand, because adenine pairs only with thymine, and guanine pairs only with cytosine.

Chromosomes and genes

DNA is organized into *chromosomes* (humans: 23 chromosomes).

Meiosis is the process by which *haploid* cells (gametes) are produced (male gamete: sperm, female gamete: egg). *Haploid* means that only a single copy of each chromosome is present in a gamete.

Fusion of two haploid gametes forms a *diploid* zygote (with two copies of each chromosome, one of which was received from the mother and the other from the father), which grows by subsequent *mitosis* (i.e., cell division resulting in two diploid daughter cells).

A *gene* is a segment of the DNA which specifies an amino acid sequence, which in turn specifies a subunit of a protein:

$$\text{DNA} \overset{\text{transcription}}{\longrightarrow} \text{pre mRNA} \overset{\text{splicing}}{\longrightarrow} \text{mRNA} \overset{\text{translation}}{\longrightarrow} \text{protein}$$

Variation in DNA sequence

The total length of the haploid genome is $\approx 3.3 \cdot 10^9$ base pairs (bp).

More than 99.5% of the genome of any two unrelated individuals is identical.

Most important classes of sequence variation:

- microsatellite:

  different numbers of repeats of a short sequence (e. g. CA)

- single nucleotide polymorphism (SNP):

  variation in a single nucleotide

Locus, allele and genotype

A *locus* L is a well-defined position along a chromosome.

In the population, different variants can exist at a locus. These variants are called *alleles*. A locus L is described by the set of variants:

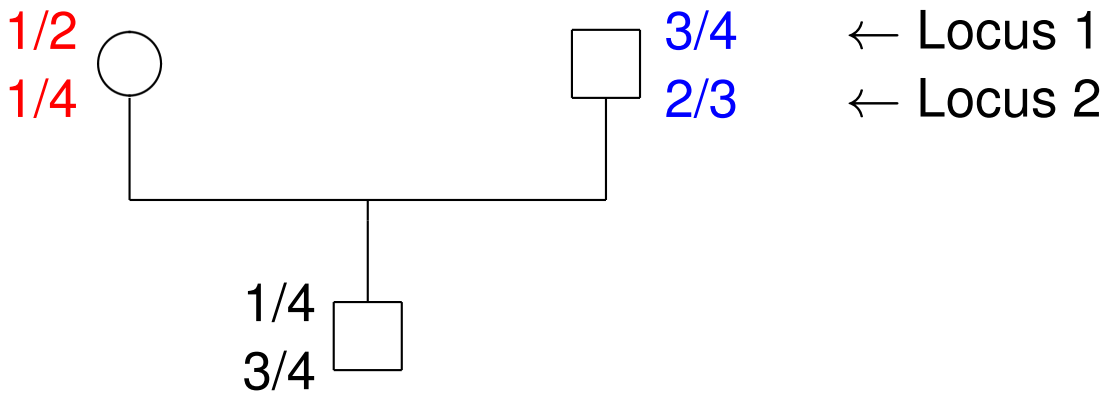
$L = \{a_1, \ldots, a_k\}$.

A *genotype* at a locus consists of a pair of alleles, one inherited from the father and one from the mother. A person is said to be *homozygous* at locus L if both alleles of the genotype at this locus are the same (i.e., a_1/a_1,...,a_k/a_k) and is said to be *heterozygous* if the alleles are different (i.e., a_i/a_j with $i \neq j$).

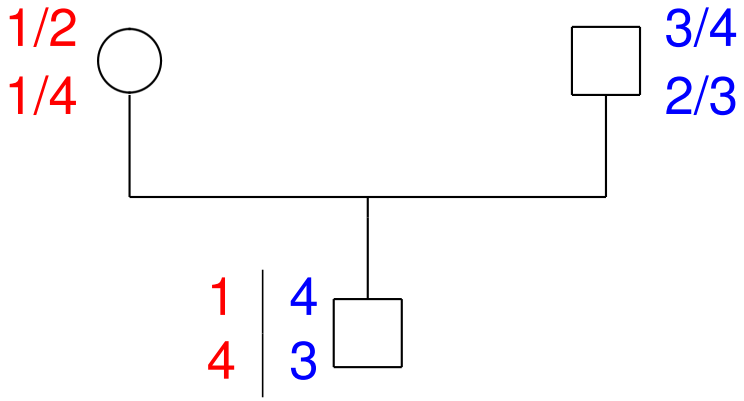
Haplotype

A sequence of alleles from different loci received from the same parent is called a *haplotype*.

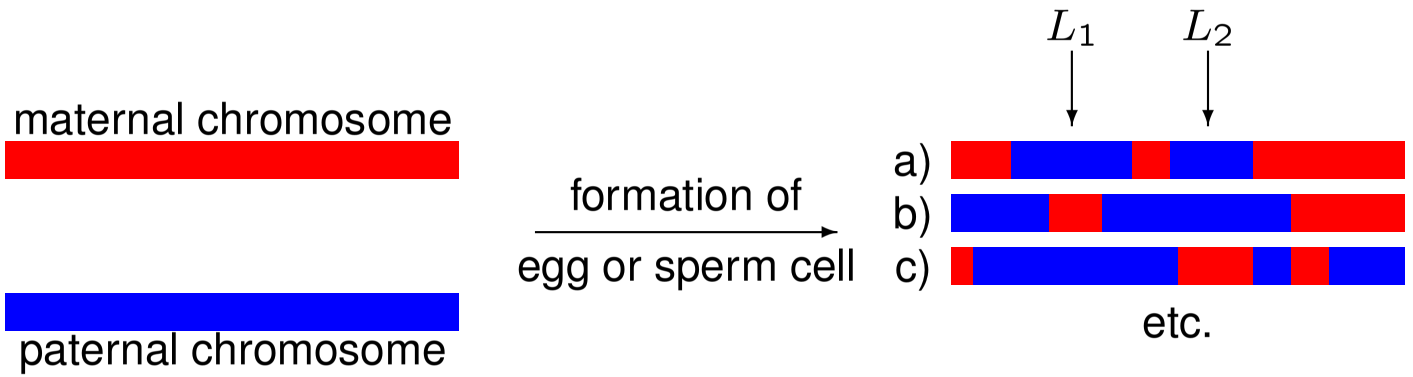
Example: Genotypes at two loci:



Haplotypes in the child:



Crossover and recombination



Crossover:
switch between paternal and maternal chromosome at meiosis before formation of gametes

Recombination between two loci L_1 and L_2:
odd number of crossovers between L_1 and L_2

Physical and genetic distance between two loci

The *physical distance* between two loci L_1 and L_2 is measured in kb (1,000 base pairs) or Mb (million base pairs).

The *genetic distance* between two loci is the expected number of crossovers occurring in a gamete between the two loci
(unit: Morgan = 100 centiMorgans).

Relationship between physical and genetic distance:
strongly depends on the chromosomal region, on average 1Mb $\sim$ 1cM

Map functions

Consider two loci L_1 and L_2. Let x denote the genetic distance between L_1 and L_2 (in Morgans) and let θ denote the recombination fraction (i.e., the probability that an odd number of crossovers occurs between L_1 and L_2).

Map functions describe the relation between x and θ:

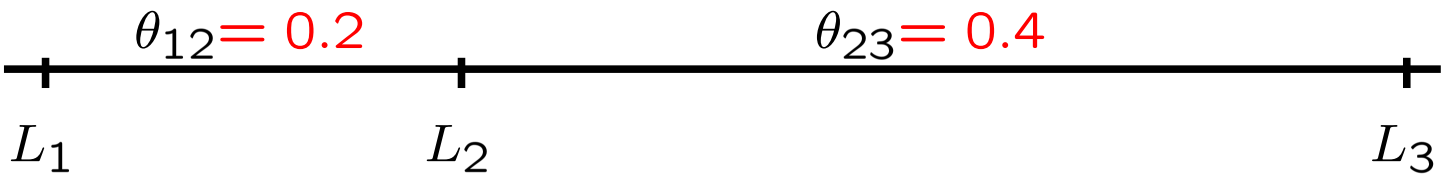
1. Haldane map function (no interference between crossovers):

   $$x = -\tfrac{1}{2}\ln(1-2\theta), \qquad \theta = \tfrac{1}{2}(1-\exp(-2\mid x \mid))$$

2. Kosambi map function (positive interference):

   $$x = \tfrac{1}{4}\ln\tfrac{1+2\theta}{1-2\theta}, \qquad \theta = \tfrac{1}{2}\cdot\tfrac{\exp(4x)-1}{\exp(4x)+1}$$

Map functions



$$\theta_{13} = ?$$

Haldane:

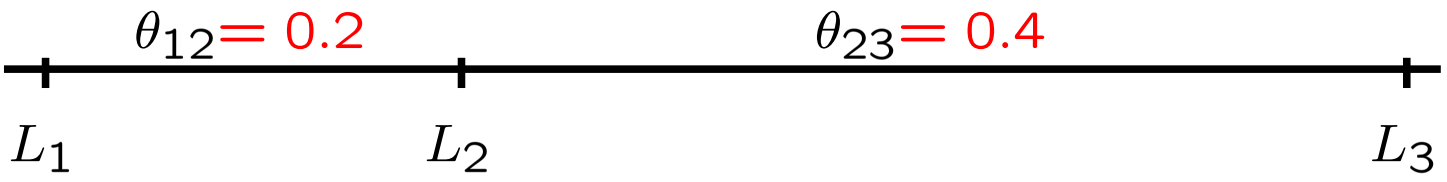
$$x_{12} = -\frac{1}{2}\ln(1-2\theta_{12}) = 0.2554M$$

$$x_{23} = -\frac{1}{2}\ln(1-2\theta_{23}) = 0.8047M$$

$$x_{13} = x_{12} + x_{23} = -\frac{1}{2}\ln(1-2\theta_{12})(1-2\theta_{23}) = 1.0601M$$

$$\theta_{13} = \frac{1}{2}(1-\exp(-2\mid x_{13}\mid)) = \theta_{12}(1-\theta_{23}) + (1-\theta_{12})\theta_{23} = 0.44$$

Map functions



$$\theta_{13} = ?$$

Exercise:

Show that, with the Kosambi map function,

$$\theta_{13} = \frac{\theta_{12} + \theta_{23}}{1 + 4\theta_{12}\theta_{23}} = 0.4545$$

Map functions