Module 7 - Multilocus Model and Map: Towards Linkage and Association

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline

Expanding Chapter 2.3 - The Biology Underlying Mendelian Inheritance

- From Mendel's first law to second law of segregation
- Mendel's pea study
- Morgan's fruit study
- Meiosis, crossover and recombination

Part of Chapter 5 - The General Concepts of Gene Mapping

- What is a map? Why do we need a map? How do we use a map?
- Measuring distance: Genetic distance, recombination fraction and linkage
- Map function
- Missing information and linkage
- Haplotype, phase and inheritance process

Overview of Genetic Map and Mapping Studies I

- A **genetic map orders and provides distance** for a set of genetic markers on the genome.

- Why are maps important: a set of ordered genetic markers **serves as landmarks on the genome**, which help us to perform genetic mapping studies.

- Given a set of markers whose locations are already known, a gene mapping study can then **identify the marker that is in 'close' proximity to the gene** that is responsible for the trait of interest, and specify the distance between the marker and the gene as well, using a suitable statistical analysis method.

- **Design issues**:
  - How many landmarks do we put on the genome? How many do we need? (as well as how many can we afford?)
  - What is the measure of closeness/dependency?
    within linkage/dependence: linkage equilibrium/independence vs. linkage disequilibrium/dependence.
  - What is the method to detect and estimate the dependency?

Overview of Genetic Map and Mapping Studies II

- Linkage analysis and association studies are two most commonly used statistical methods, on which we will spend a few lectures.
  - **Linkage analysis relies on the concept of linkage/dependence between two loci within families** from one generation to the next.
  - Two loci from the same chromosome would be in linkage; linkage analysis is a long-range mapping approach!
  - **Association analysis relies on the concept of linkage disequilibrium (LD)/dependence between two loci at population level**.
  - Only loci are very close to each other can be in LD; association analysis is a fine mapping approach!

A Toy Example I

How to find a house/gene in a city/genome?

- Location of the house/gene is completely unknown.
- Build **a rough map** of the city/genome using known houses/markers e.g. using one house as a landmark to represent each neighbourhood, say 1000 houses/markers for the city/genome.
- Evidence for the unknown gene/house being close to a particular landmark house (e.g. in the same neighbourhood) is **linkage evidence** obtained using linkage analysis.
- Now the **coarse search/linkage analysis** have identified a candidate neighbourhood/genomic region for us perform **fine-mapping/association study.**

A Toy Example II

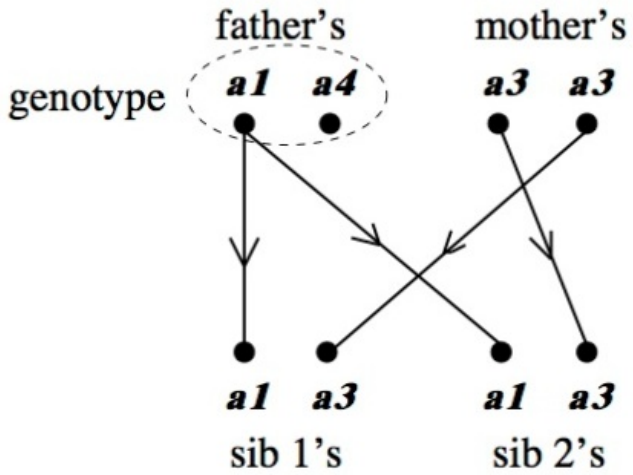
- Build **a more refined map**, surrounding the identified landmark house by using more landmark houses/markers
  e.g. using one landmark house to represent each street in the neighbourhood identified, say 100 markers/houses for the neighbourhood/genomic region.
- Evidence for gene/house very close to the actual house (e.g. on the same street) is **association evidence** obtained using association analysis.

- If we had enough resources/money to have a more refined map of the city/genome with many landmark houses/markers to start with (e.g. 100,000 houses/markers to benchmark all streets in the city/genome), then we can skip the linkage step and go directly with the association analysis, hence the genome-wide association studies (GWAS) or the next generation sequencing (NGS) studies.
- BUT, there are many statistical and computational issues related to this study design which we discuss later.

Now we need models to handle multiple loci jointly
and measure distance between loci!

Revisit Mendel's First Law of Segregation I

- It is a model for **single locus inheritance**: *underlies the concept of Mendelian transmissions of alleles from one generation to the next.*
- *One allele of each parent is* **randomly and independently** *selected, with probability 1/2, for transmission to the offspring; the alleles unite randomly to form the offspring's genotype.*



Revisit Mendel's First Law of Segregation II

- e.g. using our previous notation

$$P(X_1 = (a1, a3)|P_1 = (a1, a4), P_2 = (a3, a3)) = \frac{1}{2} \cdot 1.$$

- Note that in the example, we do not know the origin of $a3$ of the offsprings, so in fact, we applied the Mendel's first law of segregation twice!

$$\begin{aligned} &P(X_1 = (a1, a3)|P_1 = (a1, a4), P_2 = (a3_1, a3_2)) \\ &= P(X_1 = (a1, a3_1)|P_1 = (a1, a4), P_2 = (a3_1, a3_2)) \\ &+P(X_1 = (a1, a3_2)|P_1 = (a1, a4), P_2 = (a3_1, a3_2)) \\ &= \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}. \end{aligned}$$

Mendel's Second Law of Segregation

- It is a model for **multilocus inheritance**.
- It assumes **independent segregation of multiple loci**. That is segregation at one locus is not affected by segregation at another locus.

$$\begin{aligned} &P(a1 \text{ and } b2 \text{ were transmitted } | Parents = (a1, a4) \text{ and } (b2, b3)) \\ &= P(a1 \text{ was transmitted } | (a1, a4)) \times P(b2 \text{ was transmitted } | (b2, b3)) \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

- Well, it turns out the law of independent segregation is **only partially true**!

Testing Mendel's 2nd Law - Mendel's Garden Pea Study I

- Consider two traits (two loci).
- Traits of interest: colour of unripe pods and form of ripe seeds.
  - yellow (AA Aa), or green (aa).
  - round (BB Bb), or wrinkled (bb).
- Design and data:

Testing Mendel's 2nd Law - Mendel's Garden Pea Study II

```
F1:     Aa           x   Aa          <- first law:  AA(1/4)  Aa(1/2)  aa(1/4)
        Bb(bB)           Bb(bB)      <- first law: BB(1/4)  Bb(1/2)  bb(1/4)
       (yellow,round)   (yellow,round)
                     |
                     V
F2:  AA(1/4)  AA(1/4)  Aa(1/2)  Aa(1/2)
     BB(1/4)  Bb(1/2)  BB(1/4)  Bb(1/2)
       (1/16)   (1/8)    (1/8)    (1/4)  <- Mendel's second law
     ---------------------------------
     (9/16: yellow,round)

     AA(1/4)  Aa(1/2)
     bb(1/4)  bb(1/4)
       (1/16)   (1/8)                   <- Mendel's second law
     ----------------
     (3/16: yellow,wrinkled)

     aa(1/4)  aa(1/4)
     BB(1/4)  Bb(1/2)
       (1/16)   (1/8)                   <- Mendel's second law
     ----------------
     (3/16: green,round)

     aa(1/4)
     bb(1/4)
       (1/16)                           <- Mendel's second law
     --------
     (1/16: green,wrinkled)


Obs.  315: yellow,round       (559 x 9/16 ~ 314)
      101: yellow,wrinkled    (559 x 3/16 ~ 105)
      108: green,round        (559 x 3/16 ~ 105)
       35: green,wrinkled      (559 x 1/16 ~ 35)
```

Testing Mendel's 2nd Law - Mendel's Garden Pea Study III

- Assume the total number of peas is fixed $n = 315 + 101 + 108 + 35 = 559$.
- Let $X = (n_{yr}, n_{yw}, n_{gr}, n_{gw})$ denote the number of peas in each of the four categories.
- What is a reasonable distribution for X?

$$X \sim Multinomial(n, \theta),$$

$$\theta = (p_{yr}, p_{yw}, p_{gr}), \;\; p_{gw} = 1 - p_{yr} - p_{yw} - p_{gr}.$$

- Translate Mendels' Second Law into a null hypothesis

$$H_0 : \theta = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}).$$

Testing Mendel's 2nd Law - Mendel's Garden Pea Study IV

- Apply the Pearson χ^2 test (we can also apply the LRT etc):

$$T = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_3$$

$$T_{obs} = \frac{(315 - 559 \cdot \frac{9}{16})^2}{559 \cdot \frac{9}{16}} + \frac{(101 - 559 \cdot \frac{3}{16})^2}{559 \cdot \frac{3}{16}}$$

$$+ \frac{(108 - 559 \cdot \frac{3}{16})^2}{559 \cdot \frac{3}{16}} + \frac{(35 - 559 \cdot \frac{1}{16})^2}{559 \cdot \frac{1}{16}} = 0.237$$

$$\Longrightarrow \text{p-value} = P(T \geq T_{obs} | \chi^2_3) = 0.97.$$

- In this case, the data support Mendel's second law. However, this is **true only if the two loci are on different chromosomes (unlinked)**!

Testing Mendel's 2nd Law - Morgan's Fruit Flies Study I

- Traits of interest: eye color and wing length.
  - red (CC Cc) or purple (cc).
  - normal (DD Dd) or vestigial (dd).

- Design and data:

- Note that in this case, genotypes are ordered at the two loci, e.g. in the F1 generation, C and D are on the same chromosome (also known as haplotype; more on this later).
- We typically use (Cc, Dd) to denote unordered genotypes at two loci, and (CD, cd) or (Cd, cD) to denote the ordered/phased genotypes at two loci (In the graph below, we used a vertical bar to this effect).

Testing Mendel's 2nd Law - Morgan's Fruit Flies Study II

```
F1:     C|c     x   c|c       <- first law:  Cc(1/2)  cc(1/2)
        D|d         d|d       <- first law:  Dd(1/2)  dd(1/2)
        red         purple
        normal      vestigial

                |  Based on Mendel's first and second law
                V  we would expect

F2:  C|c(1/2)  C|c(1/2)   c|c(1/2)  c|c(1/2)
     D|d(1/2)  d|d(1/2)   D|d(1/2)  d|d(1/2)
        (1/4)     (1/4)      (1/4)     (1/4)  <- second law
     --------  --------   --------  --------
     red       red        purple    purple
     normal    vestigial  normal    vestigial

Obs.  1339: red and normal              (2839 x 1/4 ~ 710)
       151: red and vestigial           (2839 x 1/4 ~ 710)
       154: purple and normal           (2839 x 1/4 ~ 710)
      1195: purple and vestigial        (2839 x 1/4 ~ 710)


Explanation: 'linkage' between the two traits/genes/loci!

Among F2,  C  and c  are more likely than C and c
           D      d                       d     D
```

Testing Mendel's 2nd Law - Morgan's Fruit Flies Study III

- The testing would be again related to multinomial distribution with

$$H_0 : \theta = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}).$$

- We would have strong evidence against the null hypothesis!
  - Pearson's χ^2 test.
  - LRT.
  - etc.

Meiosis and Crossover

- Meiosis is a division process (reproduction process) during which an egg or a sperm cell is formed.
  - For a pair of autosomal chromosomes, each chromosome first **duplicates** to form a pair of **sister chromatids** (4 chromatid strands)
  - The pair of chromosomes mingle together to exchange genetic material according to **crossover process**.
  - Crossovers occur among the 4 chromatid strands, but only **between non-sister pairs** - **chiasma process** (defined on the 4 chromatid strands). (There are different hypothesized stochastic models to describe the chiasma process.
  - Then the 4 chromatid strands separate and transmit only one strand to each egg or sperm for the next generation
  - Note that **the transmitted strand is now a mixture of the original two**, containing genetic material from both chromosomes.
  - Different meioses are independent of each other.

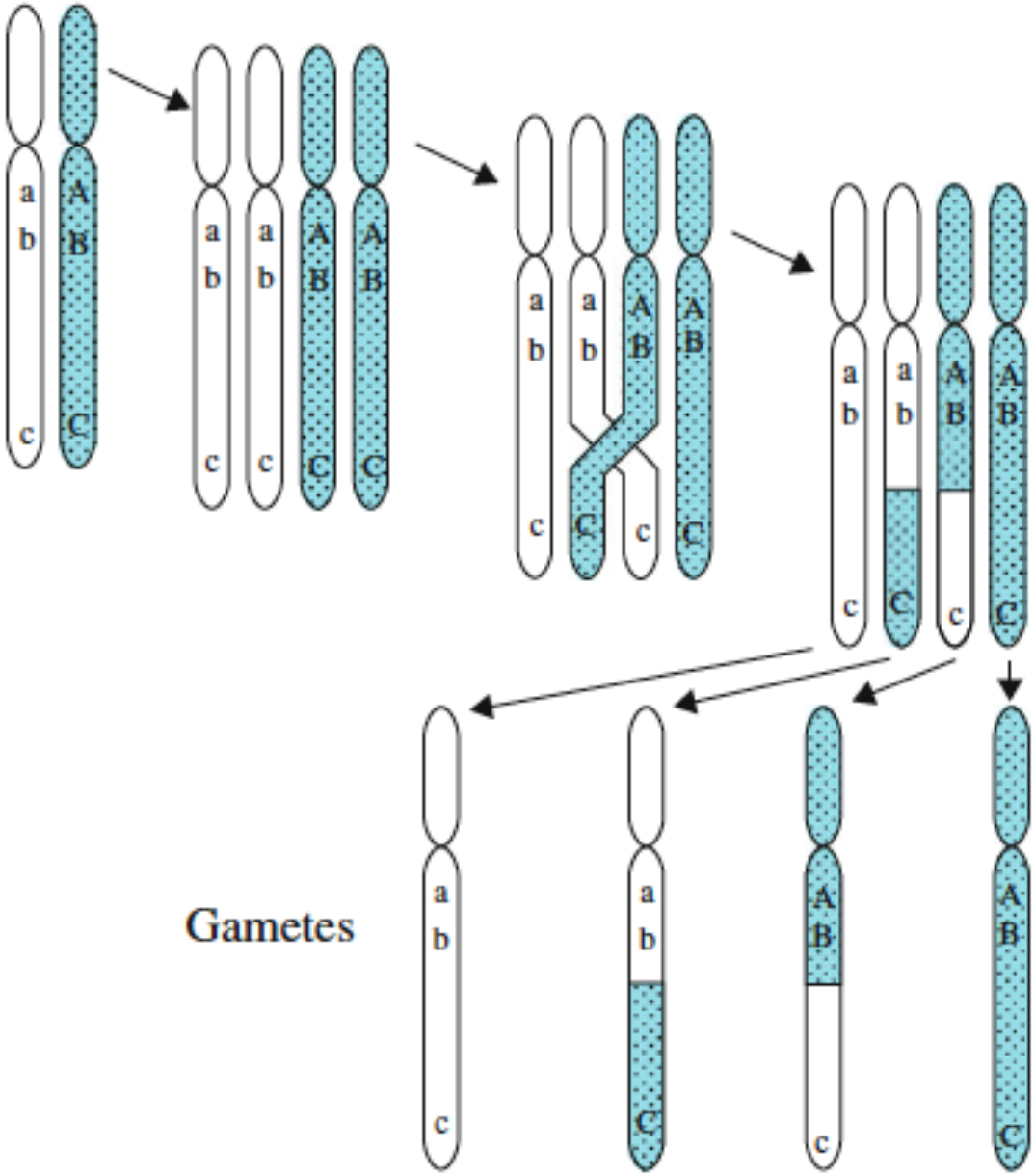
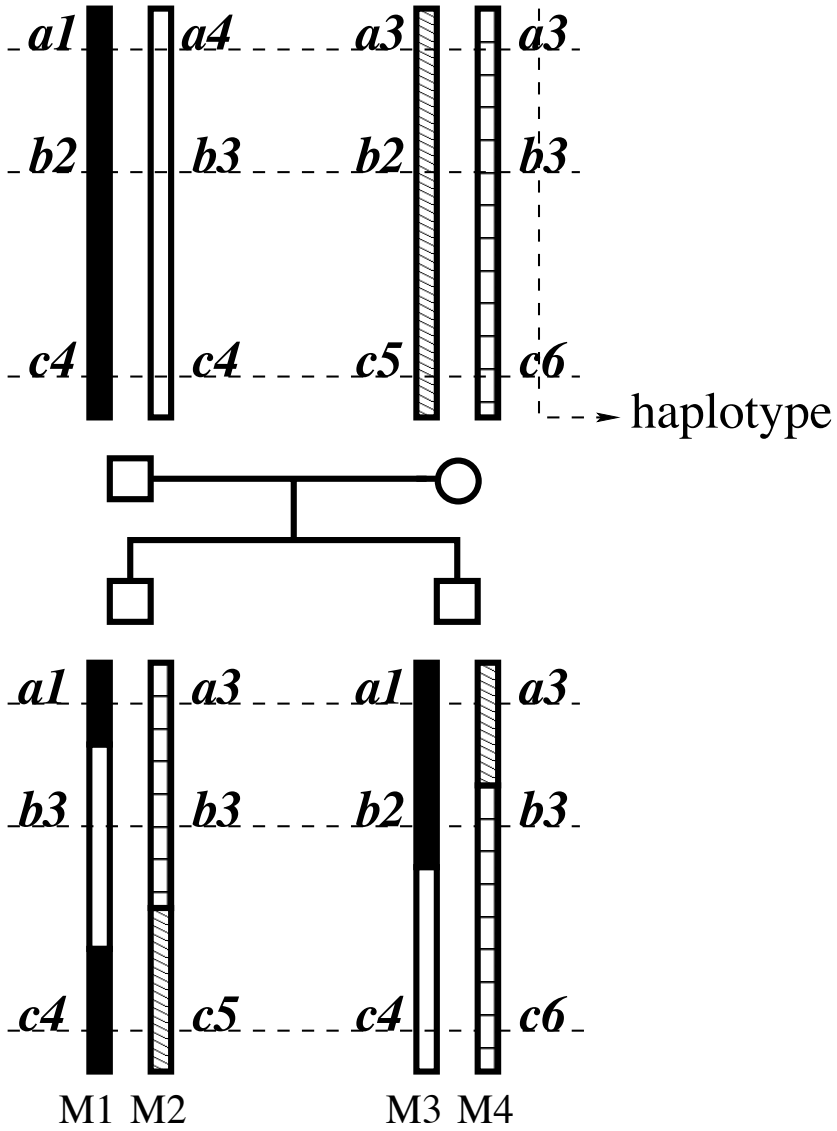


Fig. 2.5 Crossing-over and recombination during the formation of gametes (germ cells) or meiosis

Recombination

- *Crossovers are inherently unobservable, so we use the concept of recombination to describe crossovers.*
- **Recombination** occurs between two loci if the genetic material at two loci are from different chromosomes.
- e.g. Consider the second gamete, abC, of Figure 2.5 of the Textbook: *no recombination has occurred between A and B, but a recombination occurred between B and C.*
- *There is not a one-to-one relationship between recombination events and crossing over*, because a **recombination** occurs between two loci whenever there are **odd number of crossovers** between the two loci.
- Nevertheless, *recombination refers only to what can be observed between the two specific loci, whereas crossing over refers to events that can occur anywhere in the interval.*



- This figure illustrates the realizations of four independent meioses, M1, M2, M3 and M4, present in a simple nuclear family/pedigree.
- The duplication is behind the scene not depicted here. In fact, the whole process and crossovers are behind the scene and not observable.
- Here in M1, we say there is a recombination between locus A and locus B, but there is no recombination between locus A and locus C.
- *The set of alleles lying on the same chromosome is called the* **haplotype**. Alleles in a haplotype are said to be in **phase**.

Measuring Distance - Genetic Distance I

- Defined by Sturtevant (1913), a student of Morgan.
- Genetic distance t between two loci is the **expected number of crossover events** between the two loci, **per meiosis** on a **single chromatid** strand.
- Unit: Morgan, or **centiMorgan** (cM), one-hundredth of a Morgan, more often used in practice as the unit.

- A few notes on genetic distance.
  - Genetic distance t is additive.
  - Human genome is about 3000 cM or 30 Morgans long.

    (We expect 60 crossover events to happen during one meiosis or the chiasma process! Remember that genetic distance is defined on one chromatid strand, while the meiosis chiasma process involves all 4 chromatid strands but among non-sister strands.)

Measuring Distance - Genetic Distance II

- There is in fact sex difference in chromosomal length.
  - The timing of meiosis in males and females differs.
  - The total number of crossovers occurring in a sperm or egg differs.
  - Male and female recomb. rates differ markedly across chromosomes.
  - In general, female rate is higher than male rate. Male autosomes: $\approx$26.5 Morgans, Female autosomes: $\approx$39 Morgans.
  - However, most analyses assume equal recombination rate, but utilizing non-equal rates may be critical for some studies!
- Genetic distance $\neq$ physical distance (in bp).
  It's know that the crossover process varies in intensity along a chromosome: crossover/recombination hot spots or desert.
- But in general, 1cM $\approx$ 1 million bp (l,000 kb).
- In practice, crossovers thus genetic distance t cannot be observed directly in human data.

Measuring Distance - Genetic Distance III

Table 5.1 Approximate lengths of human chromosomes measured in cM and in Mb. *Source*: Yang (2000)

Chromosome #	1	2	3	4	5	6	7	8
Length (Mb)	236	255	214	203	194	183	171	155
Length (cM)	293	277	233	212	198	201	184	166
Chromosome #	9	10	11	12	13	14	15	16
Length (Mb)	145	144	144	143	114	109	106	98
Length (cM)	167	182	156	169	118	129	110	131
Chromosome #	17	18	19	20	21	22	X	Y
Length (Mb)	92	85	67	72	50	56	164	59
Length (cM)	129	124	110	97	60	58	198	–
Total (with Y)								
Length (Mb)	3200							
Length (cM)	3702							

Measuring Distance - Recombination Fraction I

- A **recombination** occurs between two loci when the genetic material at the two loci were inherited from different chromosomes
(or there are odd number of crossovers between the two loci on that gamete).

- Unlike crossover events, recombination events may be observed or estimated through study design.

- The **recombination fraction** θ between two loci is the **probability** that there is recombination between the two loci during a single meiosis.

Measuring Distance - Recombination Fraction II

- **Property of recombination fraction θ.**
  - Under some mild assumption, it can be shown that θ has boundaries

$$0 \leq \theta \leq 1/2.$$

  - $\theta = 0$: complete or perfect linkage; genetic material/alleles at the two loci from a chromosome are always transmitted together.
  - $\theta = 1/2$: unlinked; alleles at the two loci segregate/transmit independently (this is Mendel's second law).
  - Note that the upper bound of θ is NOT 1.
    Being 1 implies there is always recombination therefore NOT independence.
  - $\theta < 1/2$: **linkage**.
  - θ is a measure of distance, but it's **not additive**!

Map Function I

- A map functions specifies the relationship between genetic distance t and recombination fraction θ.
- Most commonly used map function is called **Haldane map function** (t is in unit of Morgan):

$$\theta = \frac{1 - e^{-2t}}{2},$$

$$t = -(log(1 - 2\theta))/2.$$

- Derivation required more advanced stat gene study.
  - Need the no-interference assumptions for the chiasma process.
  - Use Poisson process to describe the crossover events.

$$P(N(t) = n) = e^{-2t}\frac{(2t)^n}{n!}.$$

  - Mather's formula

$$\theta = \frac{1}{2}\{1 - P(N(t) = 0)\}.$$

Map Function II

- There are other types of map functions, e.g.
- Kosambi map function:

$$\theta = \frac{1}{2} \cdot \frac{e^{4t} - 1}{e^{4t} + 1}$$

- Gamma map function ($m = 4$ for humans):

$$\theta = \frac{1}{2}\{1 - \sum_{i=0}^{m-1} \frac{e^{-2t}(2t)^i}{i!}(1 - \frac{i}{m})\}$$

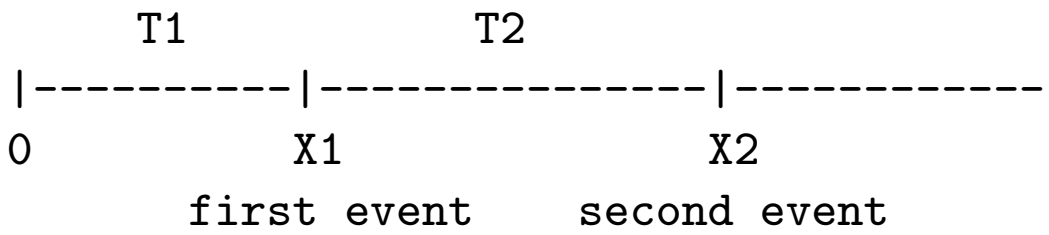
Map Function Details (Optional) I

No-interference assumptions for the chiasma process.

- Assumption I
  - Each pair of non-sister chromatids is equally likely to be involved in a given crossover, independent of which was involved in other crossovers.
  - Violation of assumption I is called chromatid interference.
- Assumption II
  - The locations of crossovers along the bundle of chromosomes follow a Poisson Process.
  - Violation of assumption II is called chiasma interference or crossover position interference.
- In practice
  - Data seem to violate the above two assumptions, especially the second one.
  - Probability of having two crossovers in close proximity next to each other is smaller than the Haldane prediction.
  - Biological reason for interference is not clear. However, observations of interference are well documented.
  - Statistical genetics analysis typically assume no-interference model.

Map Function Details (Optional) II

Poisson process (memoryless process)



- T_i: *iid* $exp(\lambda)$:

$$f(t) = \lambda e^{-\lambda t},\ F(t) = 1 - e^{-\lambda t},$$

  - λ: the rate of the process.
  - $E[T] = 1/\lambda$: expected waiting time for an event to happen.

Map Function Details (Optional) III

- Memoryless property:

$$P(T > t + s | T > s) = P(T > t)$$

  - The probability that there is no event from time 0 to time $t + s$, given that there is no event from time 0 to time s, is the same as the probability that there is no event from time 0 to time t.
  - Continuous case: only exponential distribution.
  - Discrete case: only geometric distribution, $p(1-p)^{n-1}$.

- Let $N(t)$ be the number of events observed up to time/distance t, then $N(t)$ is Poisson distributed with parameter λt:

$$P(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Map Function Details (Optional) IV

No-interference model for the chiasma process

- $N(t)$: the number of crossovers observed between two loci, t Morgan apart, along the bundle of 4 chromatid strands during a meiosis.

- $N(t)$ is a Poisson process with parameter $\lambda = 2$:

$$P(N(t) = n) = e^{-2t}\frac{(2t)^n}{n!}.$$

- Alternatively, the distribution of t between two crossover events is exponential with rate $\lambda = 2$:

$$f(t) = 2e^{-2t}.$$

- Why rate $\lambda = 2$ and has no unit?

Map Function Details (Optional) V

- Genetic distance t is in the unit of Morgan, which is defined as the expected number of crossovers between two loci on a single chromatid strand during a meiosis.
- However, chiasma process is defined on the 4 chromatid strands.
- Because each pair of non-sister chromatids is equally likely to be involved in a given crossover, the rate of the crossover process on a single chromatid strand would be $\lambda/2 = 1$.
- This means the expected time (genetic distance) for an event (crossover event) to happen on a single chromatid strand is $1/(\lambda/2) = 1$, which is precisely the definition of genetic distance.

- Mather's formula
  - Assuming no chromatid interference, Mather derived (1935):

$$\theta = \frac{1}{2}\{1 - P(N(t) = 0)\}.$$

  - Why $\frac{1}{2}$?

Map Function Details (Optional) VI

- If there are $n > 0$ crossovers in the chiasma process, having i crossovers on a given chromatids is

$$\binom{n}{i}(\frac{1}{2})^i(\frac{1}{2})^{n-i} = \binom{n}{i}(\frac{1}{2})^n$$

- Recombination involves odd # crossovers

$$\sum_{j=0}^{\lceil \frac{n-1}{2} \rceil} \binom{n}{2j+1}(\frac{1}{2})^n = \frac{1}{2}.$$

➠ Under no interference model

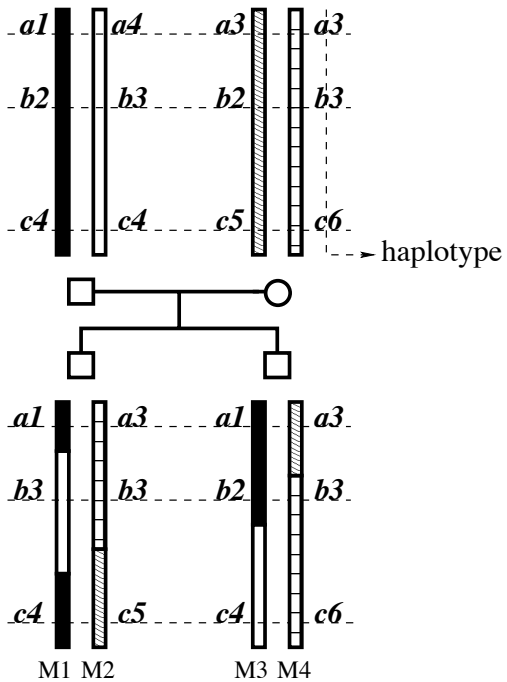
$$P(N(t) = n) = e^{-2t}\frac{(2t)^n}{n!}.$$

➠ Therefore, we have the following Haldane map function

$$\theta = \frac{1}{2}(1 - P(N(t) = 0)) = \frac{1}{2}(1 - e^{-2t}\frac{(2t)^0}{0!}) = \frac{1 - e^{-2t}}{2}.$$

Recombination Fraction I

- How do we translate "The **recombination fraction** θ between two loci is the **probability** that there is recombination between the two loci during a single meiosis." into a probability statement?



Recombination Fraction II

- For a given parent (say the father in the previous figure), if we denote the two chromosomes (black and white) as 0 and 1, and denote the two loci of interest as A and B, then

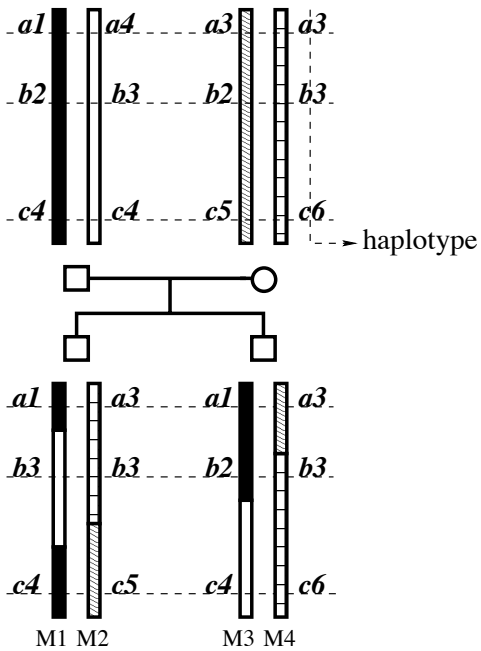
$$\begin{aligned}\theta_{AB} &= P(\text{offspring at locus B} = 1 | \text{offspring at locus A} = 0) \\ &= P(\text{offspring at locus B} = 0 | \text{offspring at locus A} = 1).\end{aligned}$$

- Note that joint distribution is (Also see Figure 6.1 of the Textbook).

$$\begin{aligned}&P(\text{offspring at locus B} = 1, \text{offspring at locus A} = 0) \\ &= P(\text{offspring at locus B} = 1 | \text{offspring at locus A} = 0) \cdot P(\text{offspring at locus A} = 0) \\ &= \frac{1}{2} \cdot \theta_{AB}.\end{aligned}$$

Recombination Fraction III

- Now consider three loci A, B, C jointly:



- Consider meiosis 1 (M1):

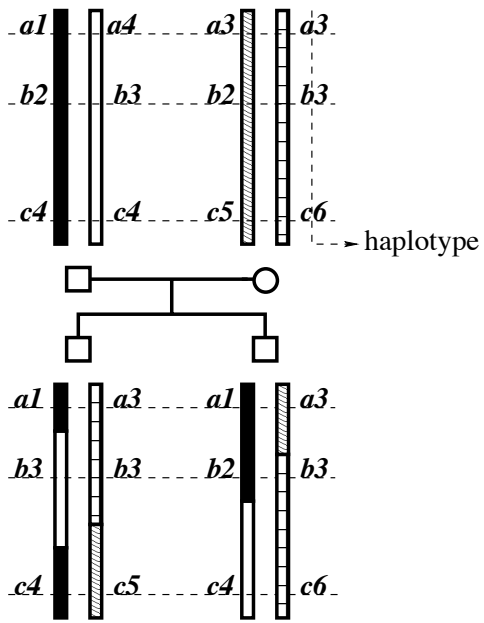
$$P(\text{locus } C = 0, \text{locus } B = 1, \text{locus } A = 0)$$
$$= P(C = 0|B = 1, A = 0) \cdot P(B = 1|A = 0) \cdot P(A = 0)$$
$$= P(C = 0|B = 1) \cdot P(B = 1|A = 0) \cdot P(A = 0)$$
$$= \theta_{BC} \cdot \theta_{AB} \cdot \frac{1}{2}.$$

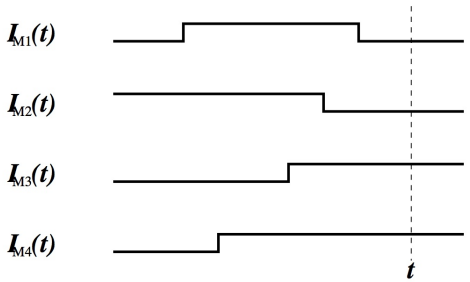
- Note that there are several advanced stat gene topics here, e.g. using the 0, 1 notation to define the inheritance process.
- And we used a Hidden Markov Model (HMM) to model the inheritance process and obtain

$$P(C = 0|B = 1, A = 0) = P(C = 0|B = 1)$$

- Joint consideration of multiple meioses (e.g. M1, M2, M3, M) describe the outcomes of the inheritance pattern in a pedigree, providing the foundation for the likelihood calculation for multiple loci and for general pedigrees.

Inheritance Process and Vector (Optional) I





Inheritance processes along the chromosome

- $I(t)$ is an indicator of, at location t along the chromosome, whether an offspring inherited a given parent's paternal allele (e.g., $I(t) = 0$) or maternal allele (say, $I(t) = 1$).
- Recall that r.f. θ between two loci t_1 and t_2 is

$$\theta = P(I(t_1) = 0 | I(t_2) = 1) = P(I(t_1) = 1 | I(t_2) = 0)$$

- Consider the four meiosis shown in the Figure, four corresponding inheritance processes, $\{I_{M1}(t)\}$, $\{I_{M2}(t)\}$, $\{I_{M3}(t)\}$ and $\{I_{M4}(t)\}$ can be defined.
- The joint process $\{\mathbf{I}(\mathbf{t})\}$,

$$\mathbf{I}(\mathbf{t}) = (I_1(t), I_2(t), I_3(t), I_4(t)),$$

describes the outcomes of all meioses in the pedigree, and it contains complete information on the inheritance pattern in this pedigree.

Inheritance Process and Vector (Optional) II

- **Inheritance process in general**
  - Consider a pedigree with f founders and n non-founders.
  - $2n$ meioses present in the pedigree: two for each of the non-founders.
  - For the realization of the k_{th} meiosis, a corresponding crossover process $\{I_k(t)\}$ can be defined.
  - The joint process

$$\{\mathbf{I(t)}\} = \{(I_1(t), I_2(t), ..., I_{2n-1}(t), I_{2n}(t))\}$$

    describes inheritance pattern in that pedigree.
  - Under the Haldane no-interference model, each $I_k(t)$ is a Markov process with 2 states, and $\{\mathbf{I(t)}\}$ would be a continuous-time Markov random walk on the vertices of a $2n-$dimensional hypercube (Donnelley 1983).
  - IBD sharing information along a chromosome is completely determined by the inheritance process.

Inheritance Process and Vector (Optional) III

- **Inheritance vector $\mathbf{I}(\mathbf{t}^*)$**, for a particular location t^*,

$$\mathbf{I}(\mathbf{t}^*) = (I_1(t^*), I_2(t^*), ..., I_{2n-1}(t^*), I_{2n}(t^*))$$

  is the inheritance vector defined by Lander and Green (1987).

- For the example discussed before:
  - Joint inheritance vectors at the three marker loci for the two sibs:

		$\mathbf{I}(\mathbf{A})$	$\mathbf{I}(\mathbf{B})$	$\mathbf{I}(\mathbf{C})$
sib 1	$M1$	0	1	0
	$M2$	1	1	0
sib 2	$M3$	0	0	1
	$M4$	0	1	1

  - Rows are independent of each other because meioses are independent of each other.
  - Columns correspond inheritance vectors at different loci and are not independent of each.

Inheritance Process and Vector (Optional) IV

- Based on the above inheritance matrix, the number of alleles shared IBD by the sib pair at each of the three loci can be inferred:

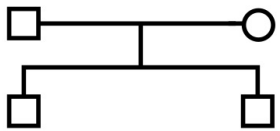
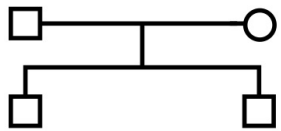
$IBD(A)$	$IBD(B)$	$IBD(C)$
1	1	0

Missing Information and Linkage

- In practice, we do not observe the underlying gametes nor the colour/pattern coded chromosomes as shown in the previous two meiosis figures (i.e. the 0 1 index for the parental origins of offspring's genotypes are uknonwn). Instead we have various sources of missing information

- **Process unobserved**: only discrete genotype data at a set of markers are observed.
- **Genotype unordered**: paternally inherited allele and maternally inherited allele cannot be distinguished based on the available genotype data at each locus.
- **Phase/haplotype unknown**: In practice, it is usually difficult to determine the phase of alleles at different loci or infer the haplotypes.
- **Missing genotype data**: genotype data may be missing for some individual(s) at some marker(s).

- However, **linkage analysis essentially is inference of missing data**, in particular haplotype and its recombination status (hence the recombination fraction θ)!

a1 a4 a3 a3	*0 0 a3 a3*
b2 b3 b2 b3	*0 0 0 0*
c4 c4 c5 c6	*0 0 c5 c6*
	
a1 a3 a1 a3	*a1 a3 a1 a3*
b3 b3 b2 b3	*b3 b3 b2 b3*
c4 c5 c4 c6	*c4 c5 c4 c6*

Haplotype and Recombination Inference I

- Example 1: 2 loci, 2 generation pedigree data.

```
father       mother
  Aa      x    Aa
  Bb(bB)       Bb(bB)
         |
         v
sib1  sib2  sib3  sib4
 Aa    aa    Aa    aa
 Bb    bb    bb    Bb
```

  - Haplotype of the sibs and parents can not be determined.
  - Recombination status (distance) cannot be inferred.

Haplotype and Recombination Inference II

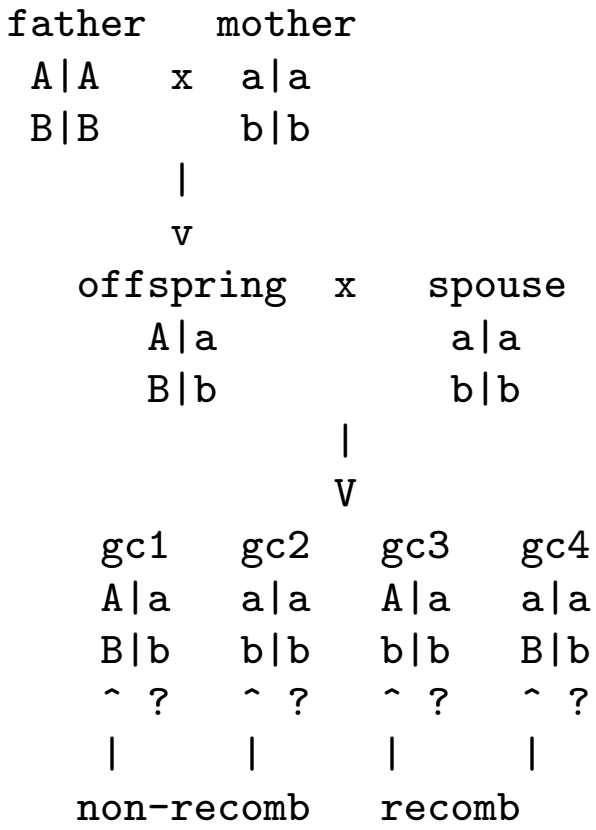
- Example 2: 2 loci, 2 generation pedigree data.

```
 father       mother
   Aa      x    a|a
   Bb(bB)       b|b
          |
          v
sib1  sib2  sib3  sib4
A|a   a|a   A|a   a|a
B|b   b|b   b|b   B|b
```

- Haplotype of the four sibs can be determined, because the mother has to transmit *ab*.
- However, recombination status cannot be inferred, because the haplotype of the father is unknown, and the mother is homozygous at both loci.

Haplotype and Recombination Inference III

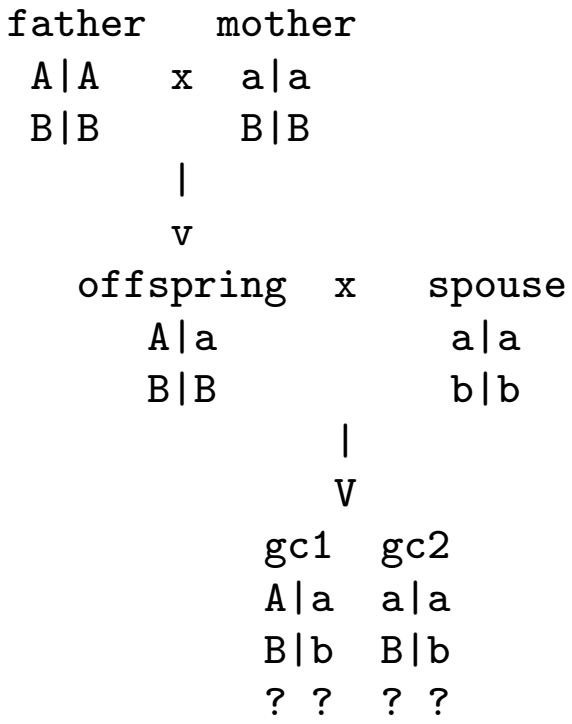
- Example 3: 2 loci, 3 generation pedigree data.



  - All the 16 haplotypes are known.
  - However, recombination status is only known for 4 haplotypes.

Haplotype and Recombination Inference IV

- Example 4: 2 loci, 3 generation pedigree data.



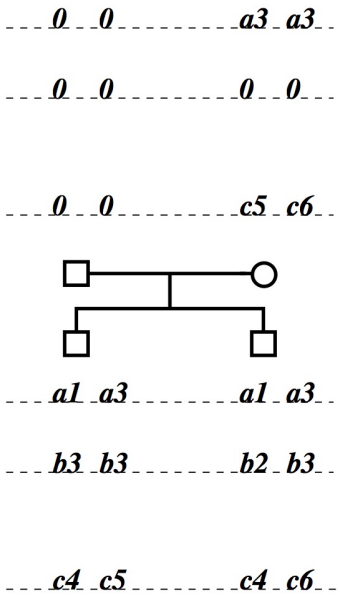
  - All the haplotypes are known.
  - However, recombination status cannot be determined for gc1 and gc2, because the parent (denoted as offspring in the graph) who transmits the haplotype is not heterozygous at both loci.

Haplotype and Recombination Inference V

- So, conditions necessary in order for recombination status of a haplotype to be determined unambiguously:
  - Genotype data of the two parents must be known in sufficient details to determine haplotype of the offspring (e.g. examples 2 and 3).
  - Genotype data of the grandparents must be known in sufficient detail to infer the haplotype of the parents (e.g. example 3).
    - Note that animal/plant genetics can use backcross design (example 3)
    - Two different strains of animal, and use inbred (e.g. brother-sister matings) for > 40 generations to obtain F_1 generation that is homozygous across the whole genome.
    - Mate F_2 generation (offspring of the F_1 generation) which is heterozygous across the genome with $F1_1$ generation.
  - It's necessary that the parent who transmits the haplotype is heterozygous at both loci (i.e. doubly heterozygous), so that recombination can be determined (e.g. example 3 but not example 4).

Haplotype and Recombination Inference VI

- In humans, informative data (e.g. example 3) are rarely available (unlike in be collected by experimental design).



- Joint analysis of many markers, with missing genotype data, for general pedigree structures can be computationally intensive.
- Needs the knowledge of likelihood calculation over general pedigrees.
- Elston-Stewart Peeling algorithm, designed for large pedigrees, but not for data with large number of markers.
- Lander-Green Hidden Markov Model (HMM) algorithm, based on a Hidden Markov formulation of the pattern of inheritance at different marker loci; limited by the size of the pedigree, but computational time is linear w.r.t. the number loci.

Haplotype and Recombination Inference VII

- An example of constructing haplotype or calculate posterior IBD probabilities in human pedigree data: essentially calculate the posterior distribution of inheritance vectors, conditional on observed genotype data.

```
father  x   mother
(0) (1)    (0) (1)    <- indicator for origin
a1  a4     a3  a3     <- obs. genotype
       |
       V
 sib1        sib2
a1 a3       a4 a3     <- obs.genotype
```

Inheritance vector	Prior	Posterior	
0 0 0 0	1/16	0	
0 0 0 1	1/16	0	
0 0 1 0	1/16	1/4	
0 0 1 1	1/16	1/4	
0 1 0 0	1/16	0	
0 1 0 1	1/16	0	
0 1 1 0	1/16	1/4	<- true one if we were
0 1 1 1	1/16	1/4	given the complete
1 0 0 0	1/16	0	inheritance information
1 0 0 1	1/16	0	as shown in the graph.
1 0 1 0	1/16	0	
1 0 1 1	1/16	0	
1 1 0 0	1/16	0	
1 1 0 1	1/16	0	
1 1 1 0	1/16	0	
1 1 1 1	1/16	0	

Exercises

- Chapter 5 Exercise 2.
- Chapter 5 Exercise 8.
- Chapter 5 Exercise 10.

What's Next

- Chapter 6 - Basic Concepts of Linkage Analysis