PHS 555

Statistical Methods for Public Health II

3 credits

Statistical Genetics and Genomics

Instructor

Rongling Wu, PhD
Department of Public Health Sciences
Pennsylvania State University
Hershey, PA 17033
Phone: (717)531-2037
Email: rwu@phs.psu.edu
Webpage: https://sites.psu.edu/statisticalgenetics/
Office hour: 3 – 4 pm Thursday

Time and location

W 6:00 – 8:55 PM (ASB 2008)

This is a research-driven course

Pure Statistics

Theory → Applications

Applied Statistics

Questions → Applications

This course

Questions → Theory → Applications

Many genetic questions beyond the scope of traditional statistics

Grading Criteria

4 Homework

- Linkage analysis
- Linkage disequilibrium analysis
- GWAS
- Functional mapping

No exam

Attendance
Attendance is required unless you obtain permission from the instructor.

Basics

•This course is intended to provide fundamental statistical concepts and tools for modeling and analysis of genetic data arising from genetic mapping, genetic association studies, population genetic studies, human genetics, and systems biology.

•Elementary knowledge of statistical theory and methods at the level of a first course in biostatistics is assumed.

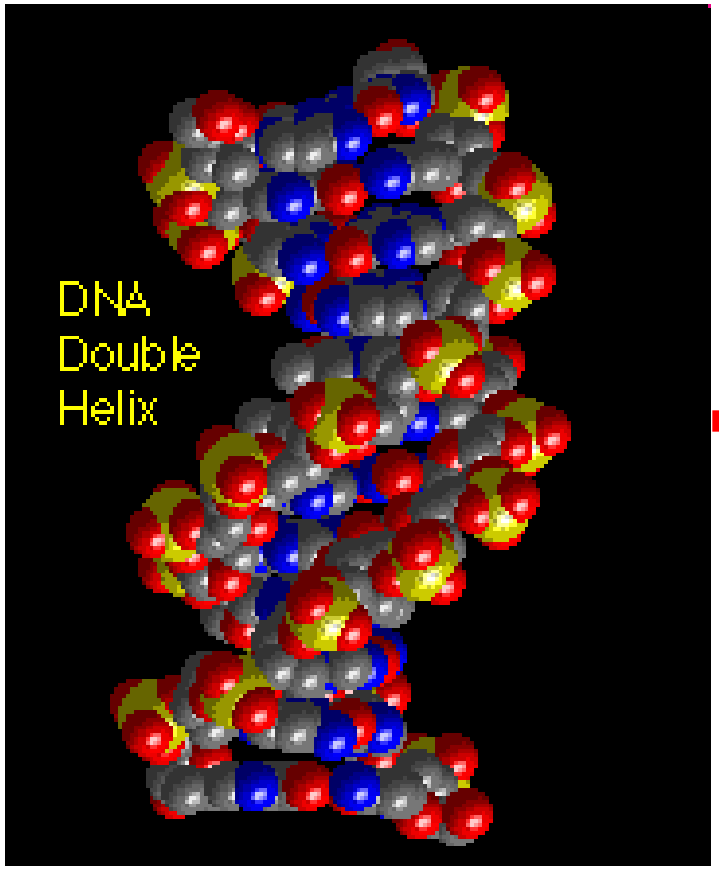
•A basic skill of computer programing with MatLab, R or any other languages is preferred.

Why Statistical Genetics?

- Understand evolution and speciation. Where do we origin from?
- Improve plant and animal breeding efficiency. How can we increase agricultural production by altering plants' and animals' genes?
- Control human diseases. How can we control diseases by developing personalized medicine?

Traditional Statistical Genetics

Genotype



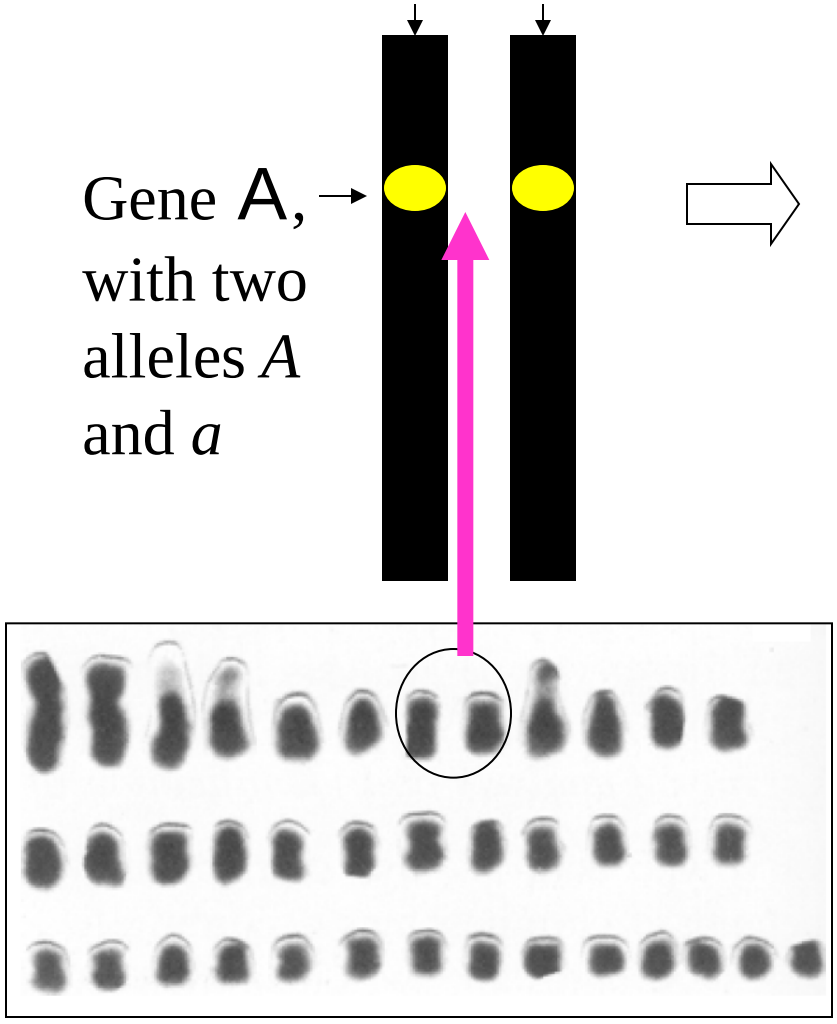
Black

Box

Phenotype

Direct association analysis between DNA and phenotypic variation

Gene, Allele, Genotype, Phenotype



Genotype	Phenotype	
	Diameter	Height
AA	185	100
AA	182	104
Aa	175	103
Aa	171	102
aa	155	101
aa	152	103

Regression model for estimating the genotypic effect

Phenotype = Genotype + Error

$$y_i = \sum_{j=1}^{3} x_i \mu_j + e_i$$

x_i is the indicator for the genotype of subject i

μ_j is the mean for genotype j

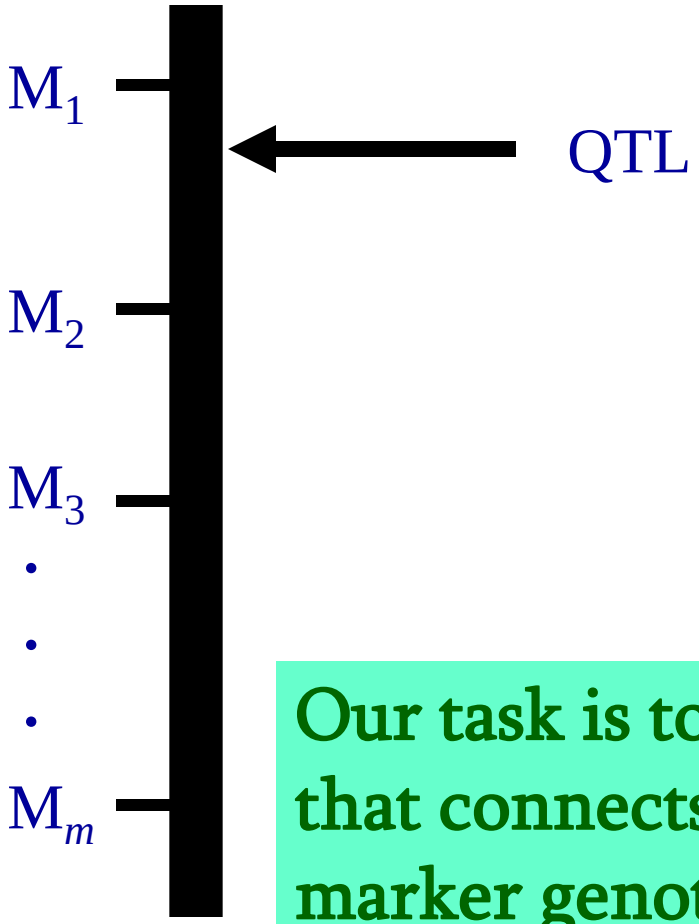
j =1 for AA, 2 for Aa, 3 for aa

$e_i \sim N(0, \sigma^2)$

The gene that is associated significantly with the trait is called quantitative trait locus (QTL)

A question is: We cannot observe such a QTL directly.

Genetic Mapping

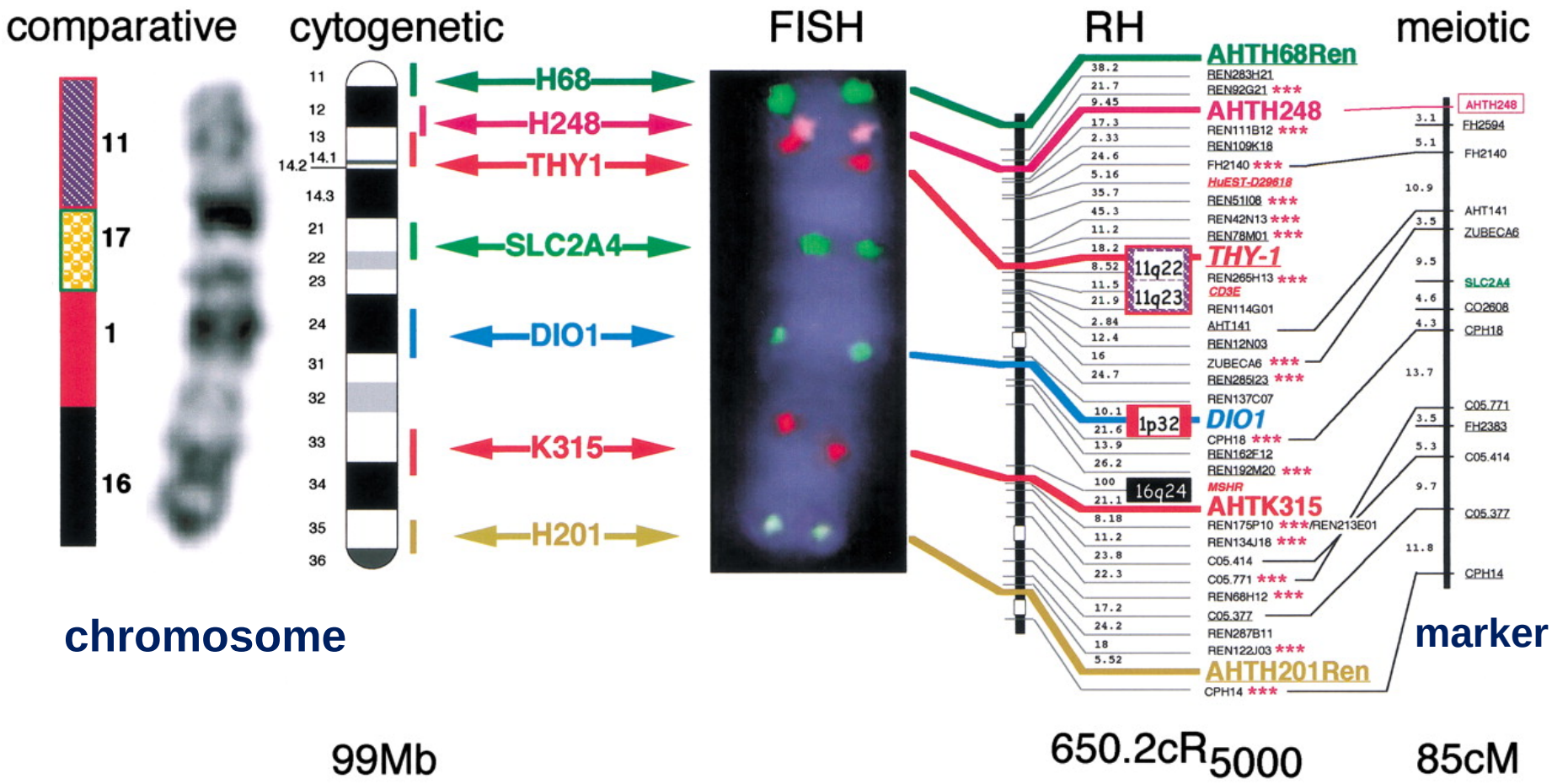


The genotypes for the trait are not observable and should be predicted from linked neutral molecular markers (M)

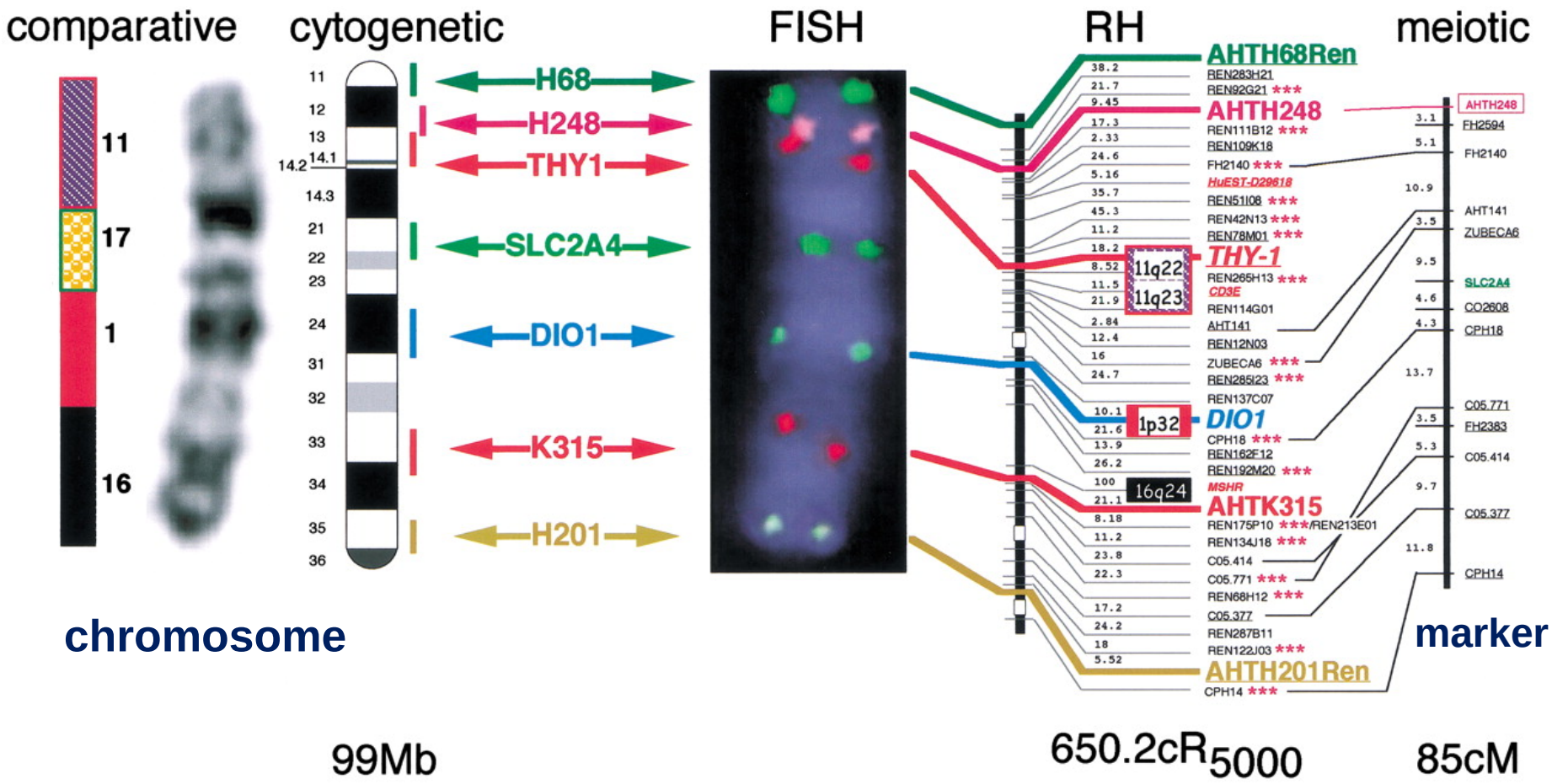
The genes that lead to the phenotypic variation are called Quantitative Trait Loci (QTL)

Our task is to construct a statistical model that connects the QTL genotypes and marker genotypes through observed phenotypes

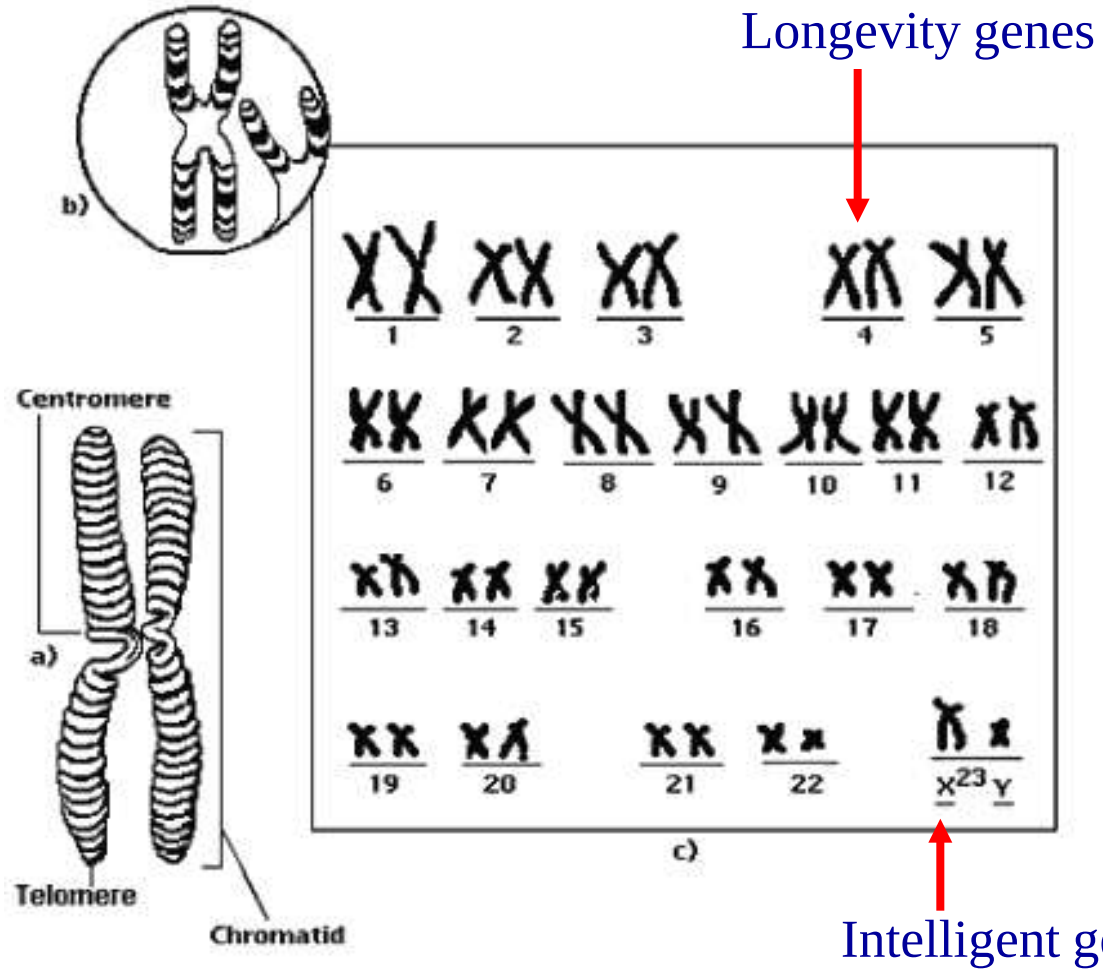
From chromosome to linkage map: an example



From chromosome to linkage map: an example



Human Chromosomes



Male ***Xy***

X ***y***

Female

XX — ***X***

XX	***Xy***
Daughter	**Son**

Contemporary Statistical Genetics

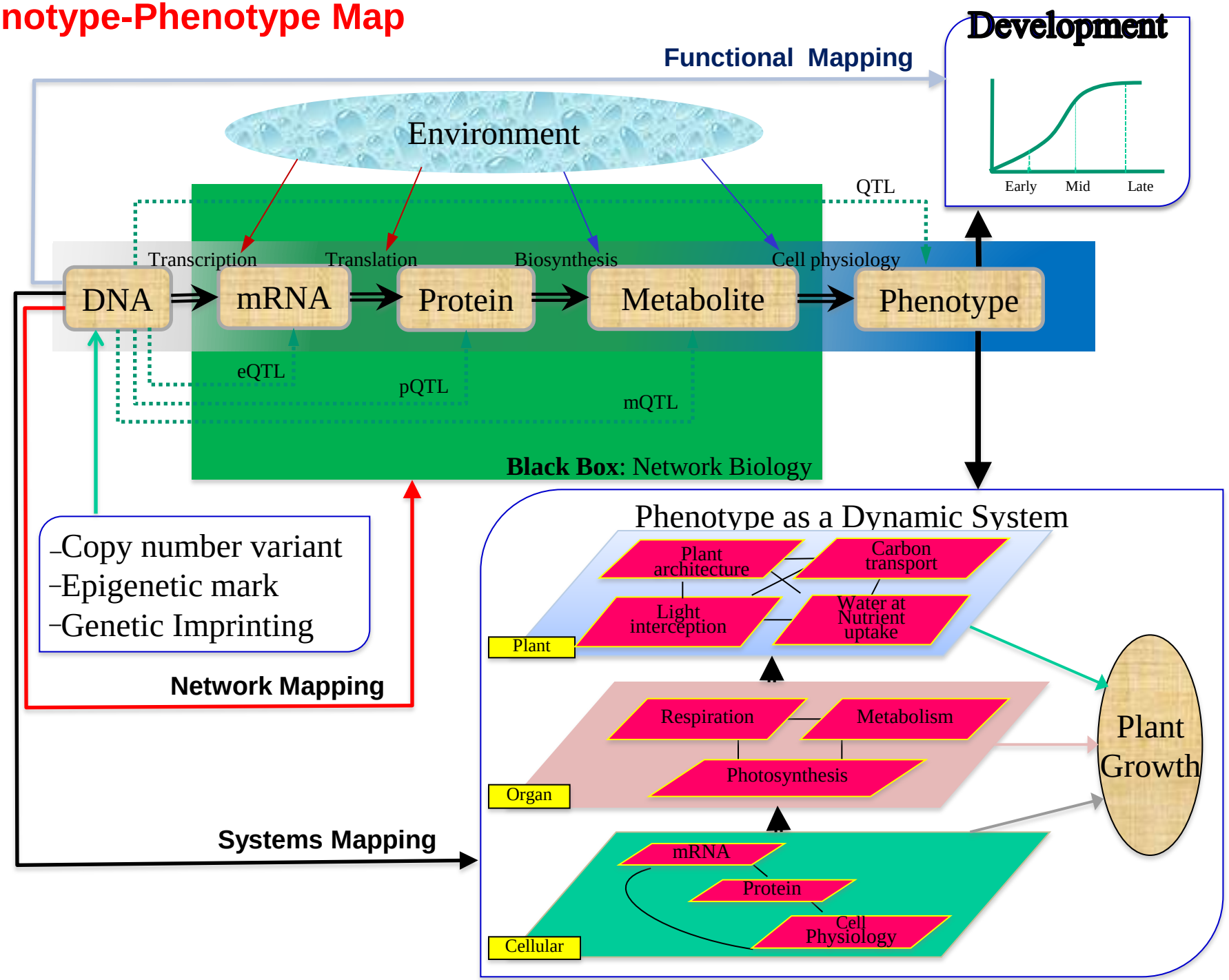
Traditional statistical genetics: Direct association between DNA variants and phenotypes, with limitations

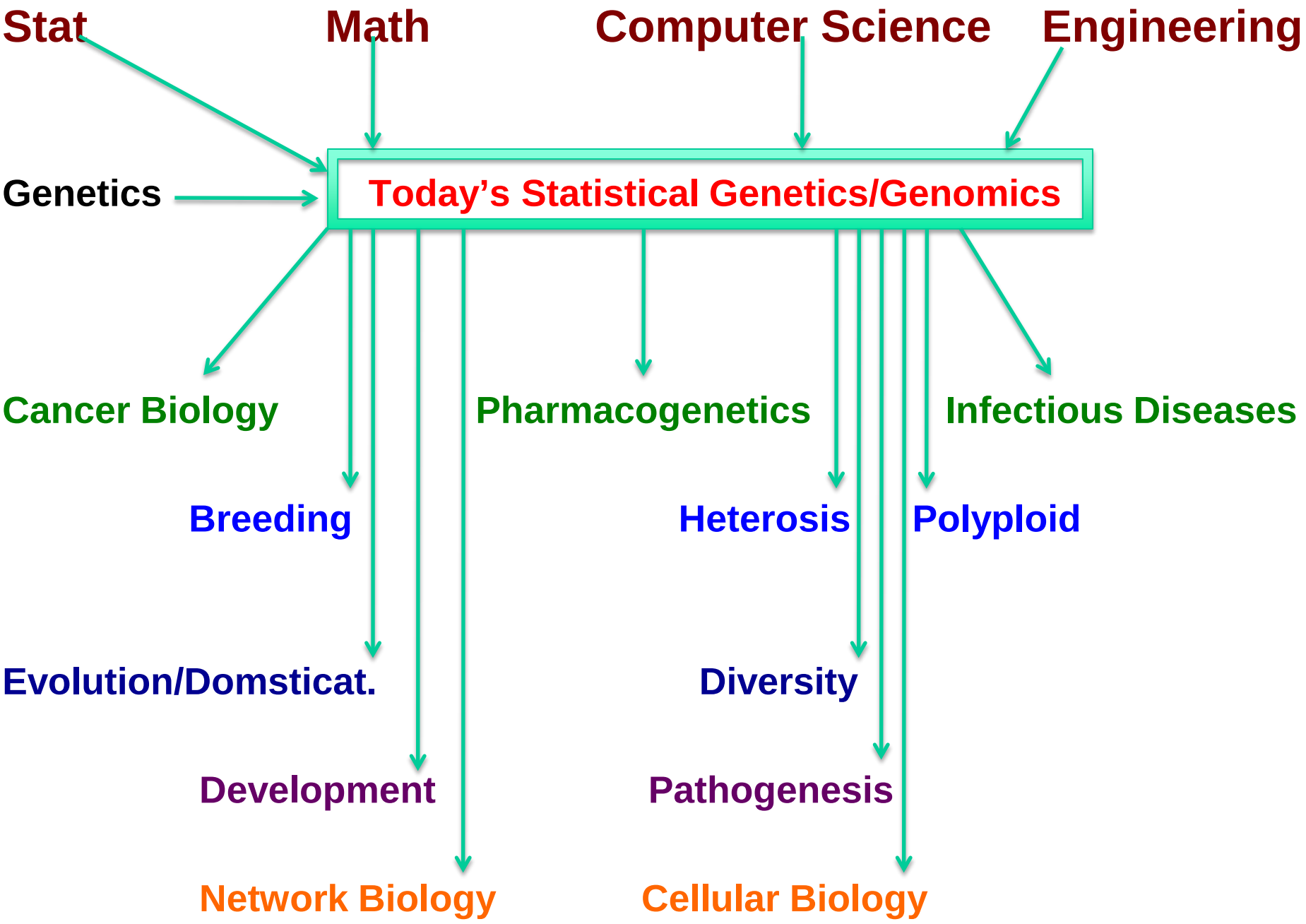
1. **Not only DNA sequences cause genetic variation (epigenetic marks, copy number variants, genetic imprinting…)**
2. **Phenotypic formation is a dynamic process that experiences various stages of development)**
3. **There is a long pathway from DNA to end-point phenotype (central dogma of biology)**

   Recent molecular genetic studies have found many types of mRNA (microRNA, small RNA, long non-coding RNA) that play an important role in guiding biological processes

All these latest discoveries should be implemented into statistical genetic research. However, no existing statistical methods can directly tackle such implementation.

Genotype-Phenotype Map




Stat
Math
Computer Science
Engineering
Genetics
Today's Statistical Genetics/Genomics
Cancer Biology
Pharmacogenetics
Infectious Diseases
Breeding
Heterosis
Polyploid
Evolution/Domsticat.
Diversity
Development
Pathogenesis
Network Biology
Cellular Biology

Nobel Prize in Chemistry 2013

Computer

Chemical reaction Systems

Biological system

Statistical Geneticists

The topics this course covers:

Traditional Statistical Genetics

- DNA marker segregation
- Linkage analysis
- Hardy-Weinberg equilibrium test
- Linkage disequilibrium estimation
- Haplotype phasing
- Quantitative genetic theory of phenotypic traits
- Genetic mapping of quantitative trait loci (QTLs)
- Genome-wide association studies (GWAS)

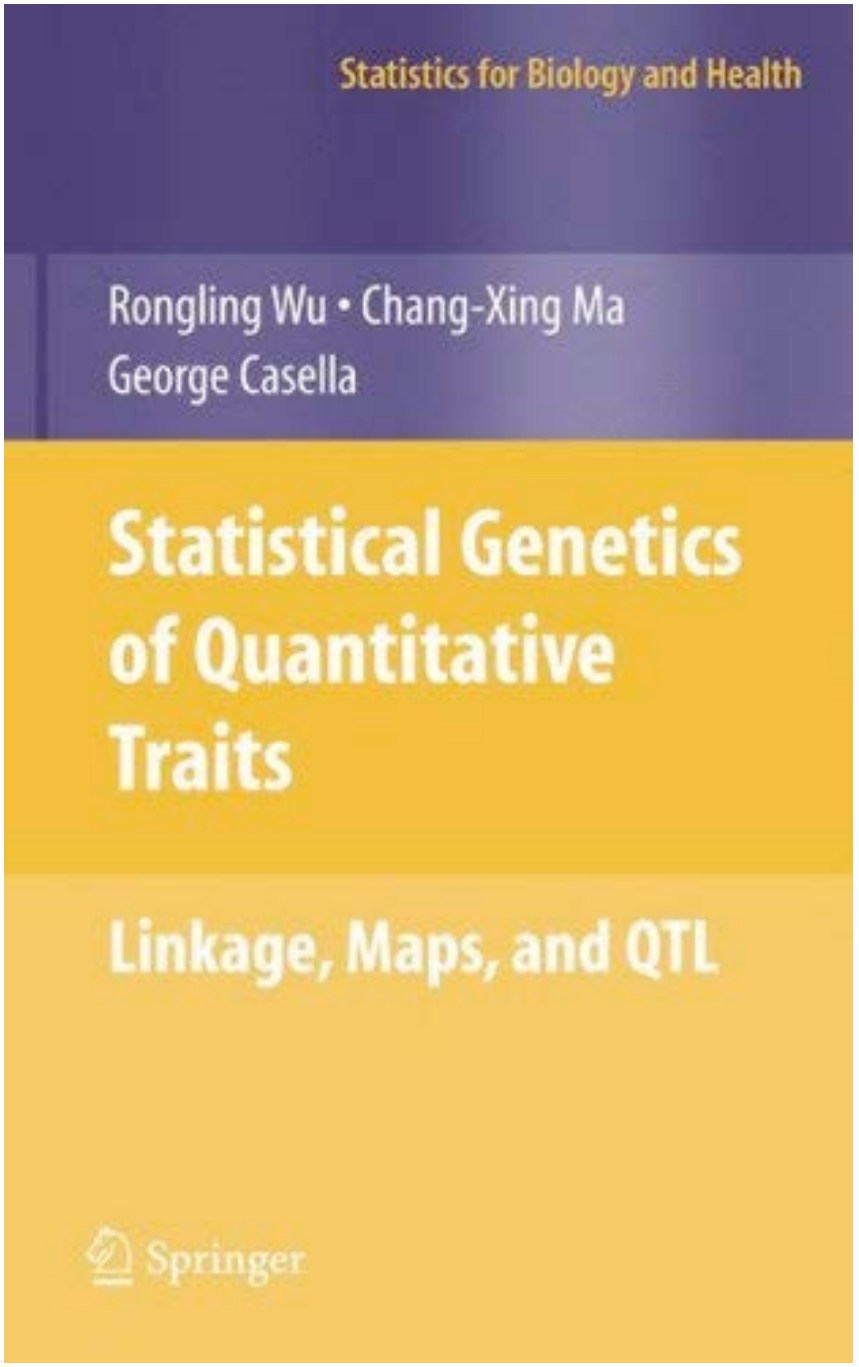
Contemporary Statistical Genetics

- Functional mapping of developmental processes
- Epigenetic detection from population-based association studies
- QTL mapping of gene expression, proteomic profiles and metabolic pathways
- Gene regulatory network construction
- Systems mapping
- eQTL mapping across multiple tissues
- Cancer genomics

Textbook

Linkage analysis
Map construction
QTL mapping

Beyond
Population genetics
Linkage disequilibrium
Functional mapping
Systems mapping
Gene expression modeling
GWAS



Other Readings

Basic Genetics

(1) Mendelian genetics

How does a gene transmit from a parent to its progeny (individual)?

(2) Population genetics

How is a gene segregating in a population (a group of individuals)?

(3) Quantitative genetics

How is gene segregation related with the phenotype of a character?

(4) Molecular genetics

What is the molecular basis of gene segregation and transmission?

(5) Developmental genetics

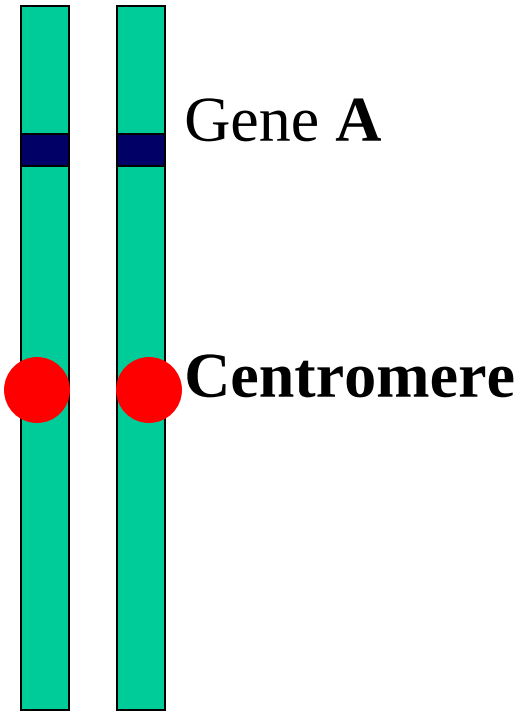
(6) Epigenetics

(7) Systems genetics

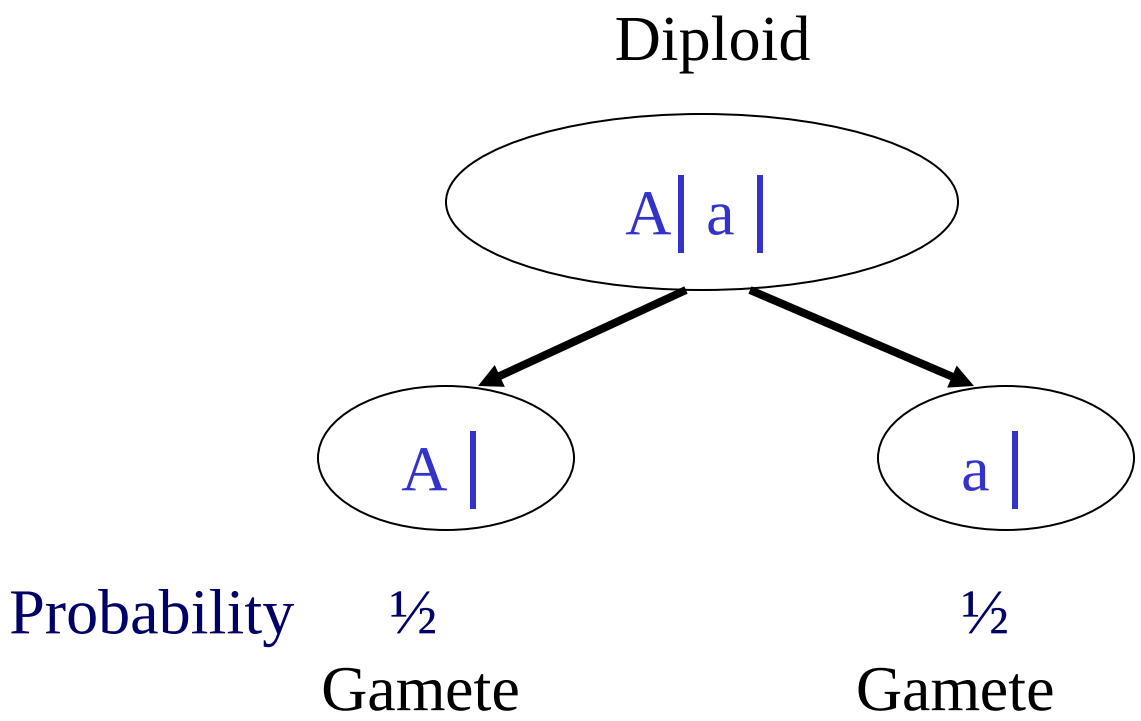
Mendel's Laws

Mendel's first law

- There is a gene with two alleles on a chromosome location (locus)
- These alleles segregate during the formation of the reproductive cells, thus passing into different gametes

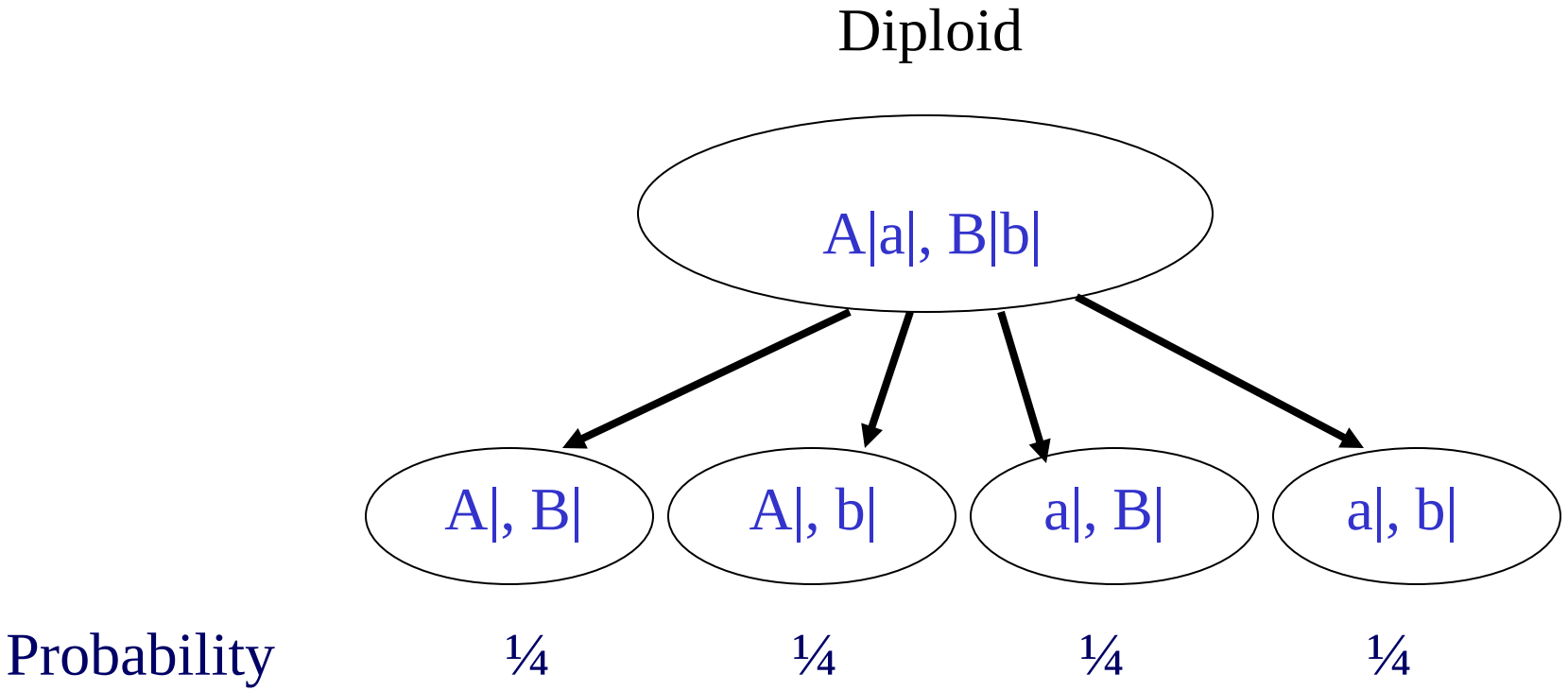


A pair of chromosomes



Mendel's second law

- There are two or more pairs of genes on different chromosomes
- They segregate independently (partially correct)

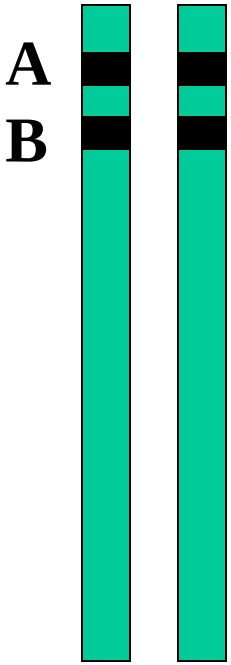


What about three genes?

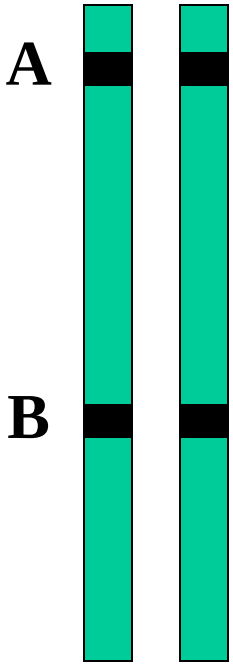
Linkage (exception to Mendel's second law)

- There are two or more pairs of genes located on the same chromosome
- They can be linked or associated (the degree of association is described by the recombination fraction)

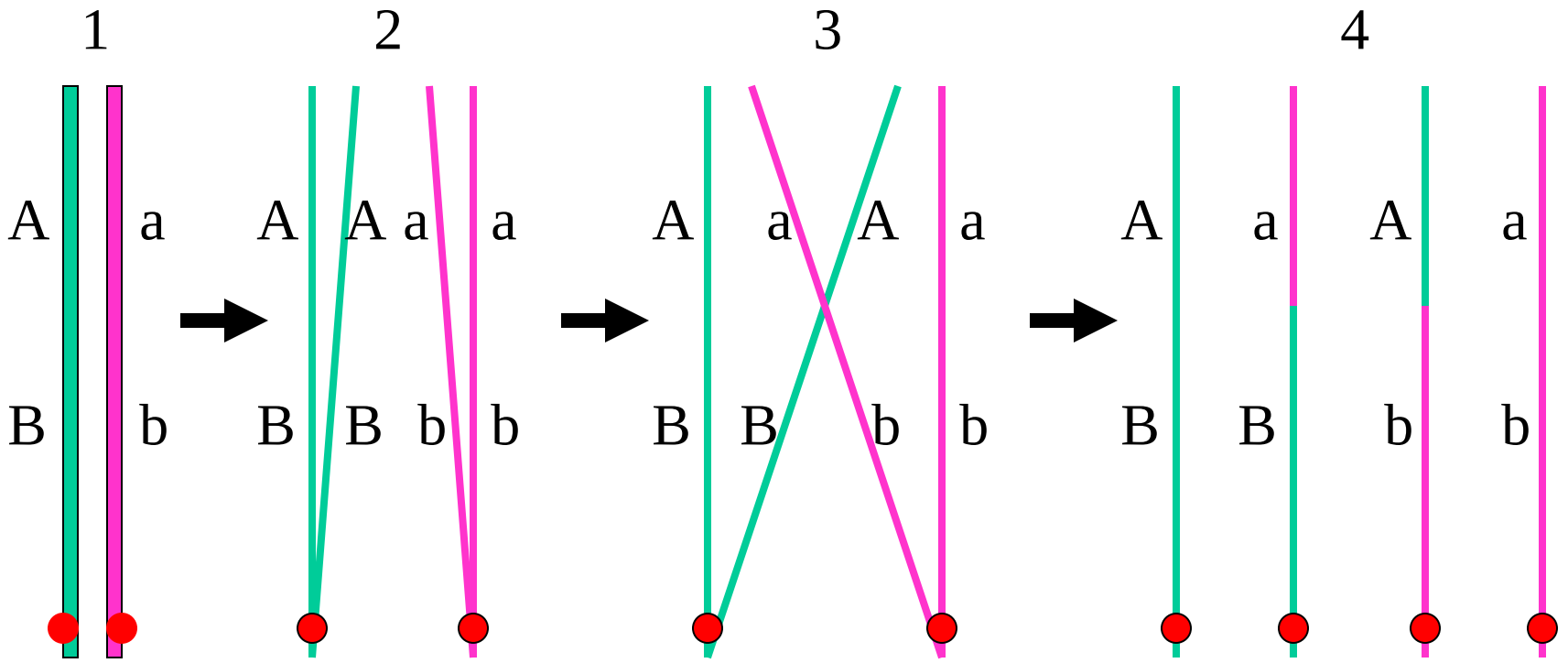
High linkage



Low linkage

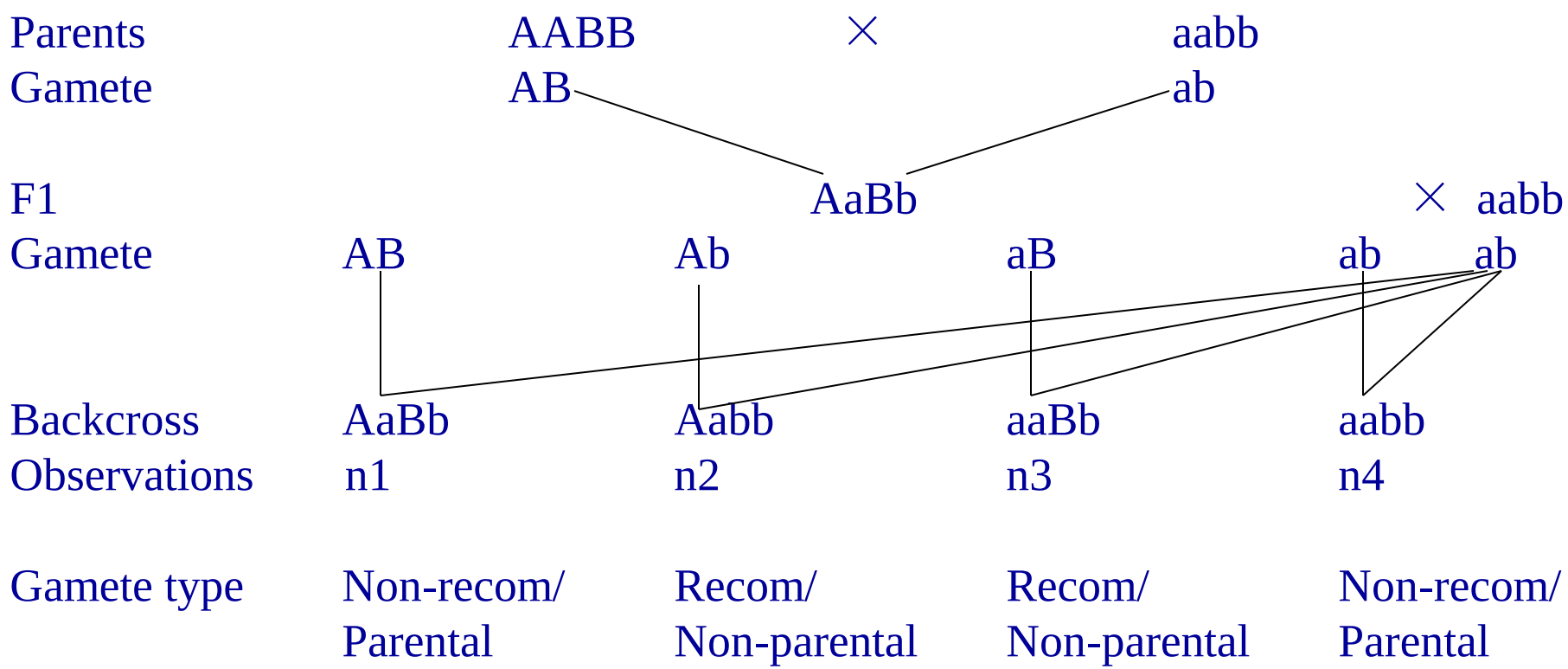


How the linkage occurs? – consider two genes A and B



Stage 1: A pair of chromosomes, one from the father and the other from the mother
Stage 2: Each chromosome is divided into two sister chromatids
Stage 3: Non-sister chromatids crossover
Stage 4: Meiosis generates four gametes AB, aB, Ab and ab –
Nonrecombinants (AB and ab) and
Recombinants (aB and Ab)

How to measure the linkage? – based on a design



Define the proportion of the recombinant gametes over the total gametes as the recombination fraction (r) between two genes A and B

$r = (n2+n3)/(n1+n2+n3+n4)$

Several concepts

Genotype and Phenotype

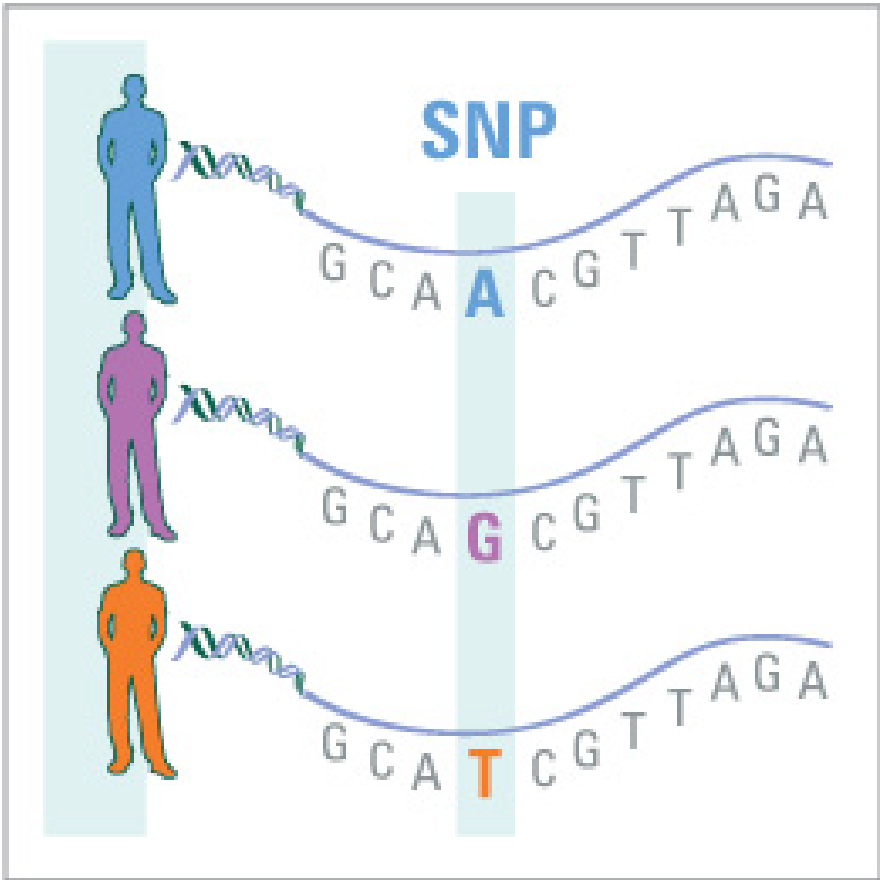
- Locus (loci), chromosomal location of a gene
- Allele (A, a), a copy of gene
- Dominant allele, one allele whose expression inhibits the expression of its alternative allele
- Recessive allele (relative to dominant allele)
- Dominant gene (AA and Aa are not distinguishable, denoted by A_)
- Codominant gene (AA, Aa and aa are mutually distinguishable)
- Genotype (AA, Aa or aa)
- Homozygote (AA or aa)
- Heterozygote (Aa)
- Phenotype: trait value

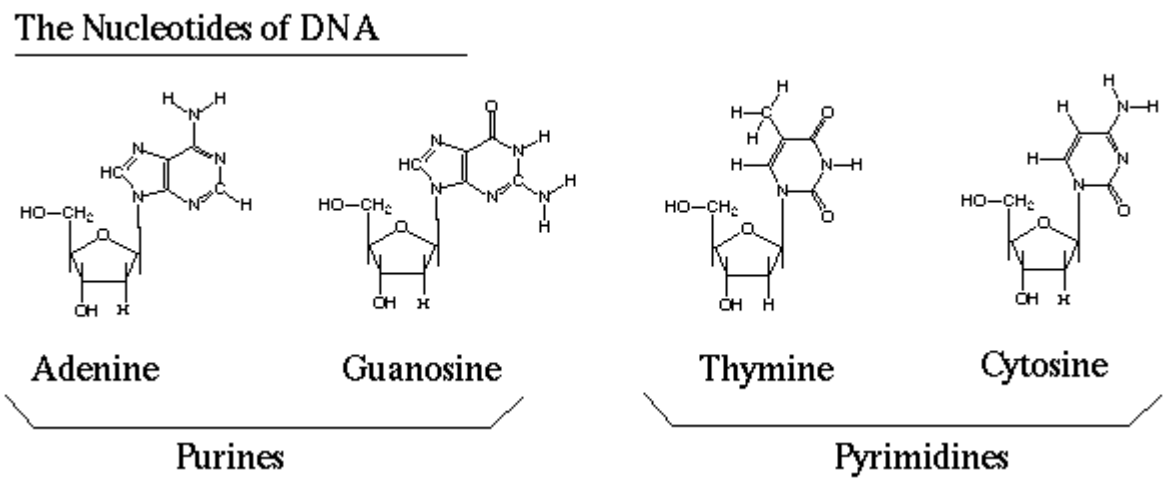
Chromosome and Meiosis

- Chromosome: Rod-shaped structure made of DNA
- Diploid (2n): An organism or cell having two sets of chromosomes or twice the haploid number
- Haploid (n): An organism or cell having only one complete set of chromosomes
- Gamete: Reproductive cells involved in fertilization. The ovum is the female gamete; the spermatozoon is the male gamete.
- Meiosis: A process for cell division from diploid to haploid ($2n \rightarrow n$) (two biological advantages: maintaining chromosome number unchanged and crossing over between different genes)
- Crossover: The interchange of sections between pairing homologous chromosomes during meiosis
- Recombination, recombinant, recombination fraction (rate, frequency): The natural formation in offspring of genetic combinations not present in parents, by the processes of crossing over or independent assortment.

Molecular markers

- Genetic markers are DNA sequence polymorphisms that show Mendelian inheritance
- Marker types
  - Restriction fragment length polymorphism (RFLP)
  - Amplified fragment length polymorphism (AFLP)
  - Simple sequence repeat (SSR)
  - Single nucleotide polymorphism (SNP)





Summary: Mendel's Laws

Mendel's first law

- There is a gene with two alleles on a chromosome location (locus)
- These alleles segregate during the formation of the reproductive cells, thus passing into different gametes

Mendel's second law

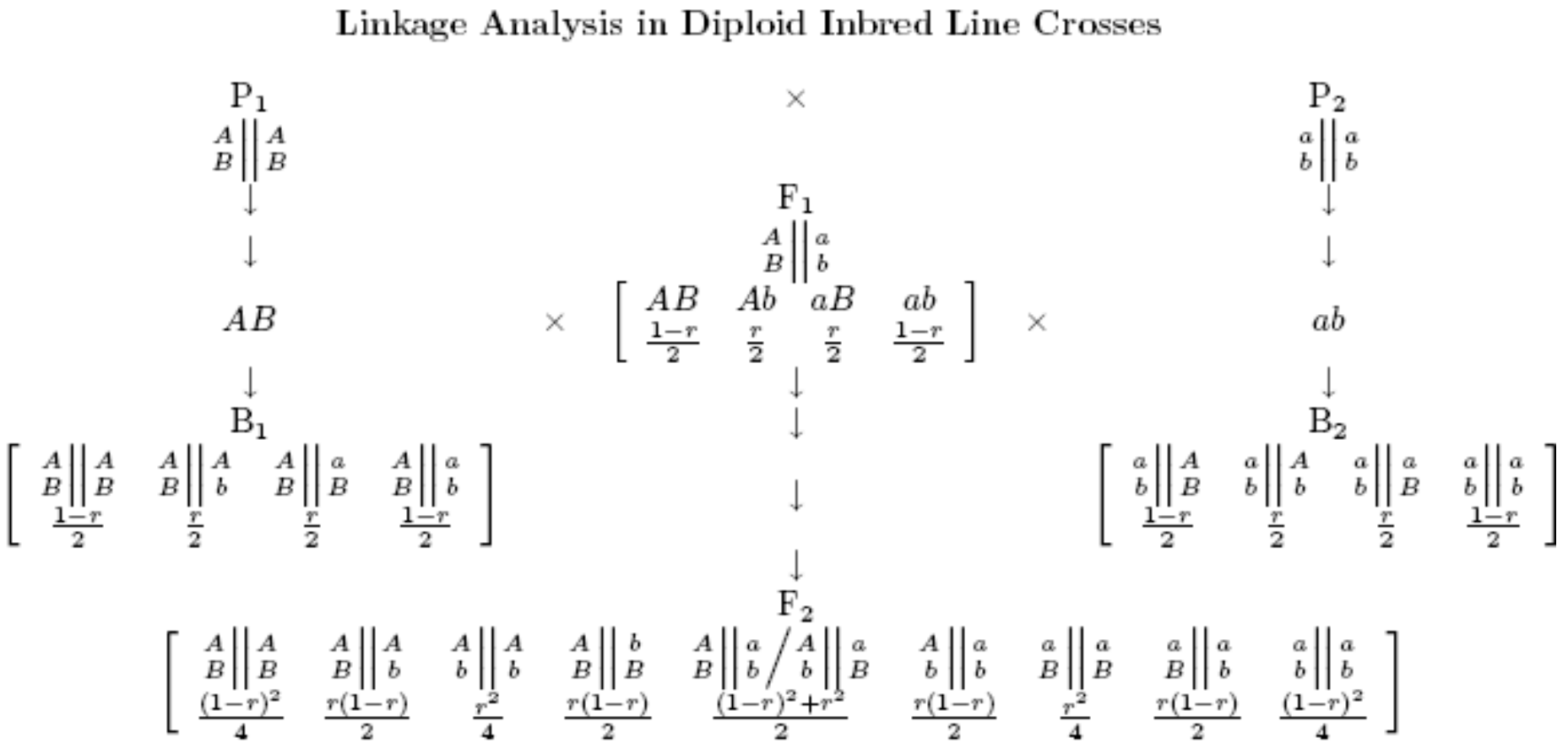
- There are two or more pairs of genes on different chromosomes
- They segregate independently (partially correct)

Linkage (exception to Mendel's second law)

- There are two or more pairs of genes located on the same chromosome
- They can be linked or associated (the degree of association is described by the recombination fraction)

Linkage Analysis and Map Construction

Genetic design



Testing Mendelian segregation

Consider marker **A** with two alleles A and a

	Backcross		F_2		
	Aa	aa	AA	Aa	aa
Observation	n_1	n_0	n_2	n_1	n_0
Expected frequency	½	½	¼	½	¼
Expected number	n/2	n/2	n/4	n/2	n/4

The x^2 test statistic is calculated by

$x^2 = \Sigma\ (obs - exp)^2 / exp$

$= (n_1\text{-}n/2)^2/(n/2) + (n_0\text{-}n/2)^2/(n/2) = (n_1\text{-}n_0)^2/n \sim x^2_{df=1}$, for BC,

$(n_2\text{-}n/4)^2/(n/4)+(n_1\text{-}n/2)^2/(n/2)+(n_0\text{-}n/4)^2/(n/4) \sim x^2_{df=2}$, for F_2

Examples

	Backcross		F2		
	Aa	aa	AA	Aa	aa
Observation	44	59	43	86	42
Expected frequency	½	½	¼	½	¼
Expected number	51.5	51.5	42.75	85.5	42.75

The x^2 test statistic is calculated by

$x^2 = \Sigma\ (obs - exp)^2 / exp$

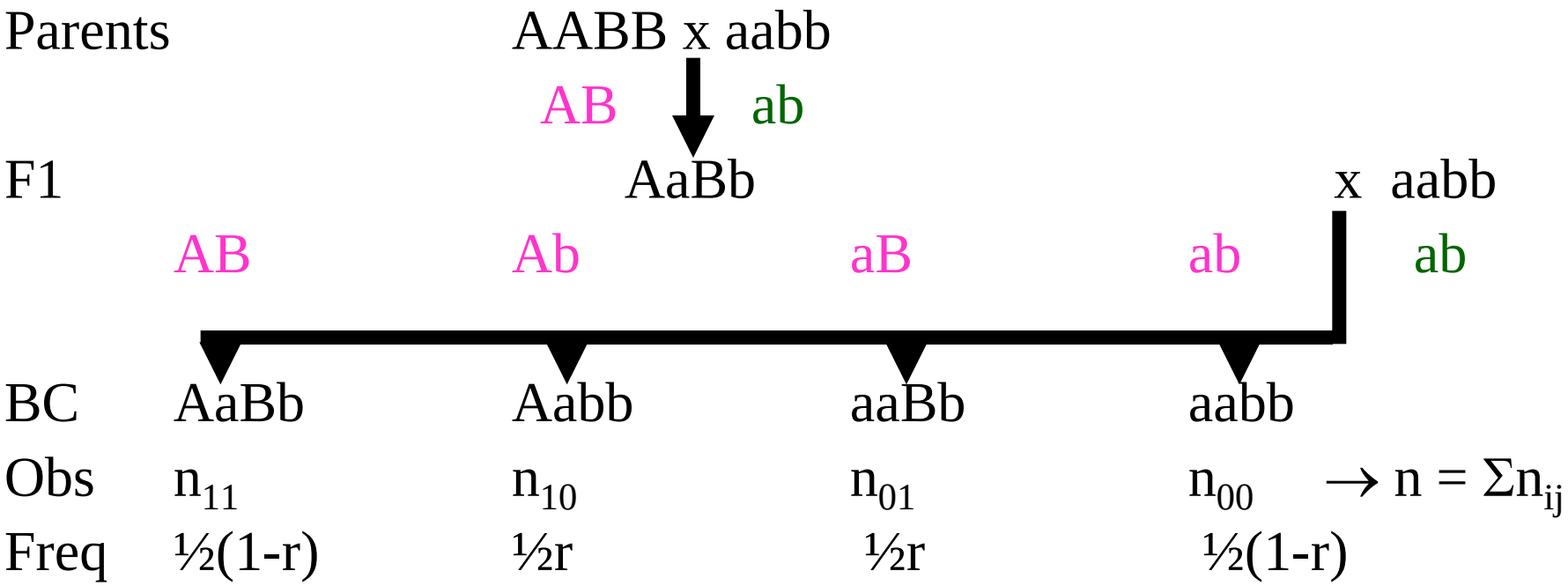
$= (44\text{-}59)^2/103 = 2.184 < x^2_{df=1} = 3.841$, for BC,

$(43\text{-}42.75)^2/42.75+(86\text{-}85.5)^2/85.5+(42\text{-}42.75)^2/42.75=0.018 < x^2_{df=2} =5.991$, for F2

The marker under study does not deviate from Mendelian segregation in both the BC and F2.

Linkage analysis

Backcross



r is the recombination fraction between two markers **A** and **B.**

The maximum likelihood estimate (MLE) of r is

$r\hat{} = (n_{10}+n_{01})/n.$ r has interval [0,0.5]: r=0 complete linkage, r=0.5, no linkage

Proof of $r\hat{} = (n_{10}+n_{01})/n$

The likelihood function of r given the observations:

$$L(r|n_{ij}) = n!/(n_{11}!n_{10}!n_{01}!n_{00}!) \times [\tfrac{1}{2}(1-r)]^{n11}[\tfrac{1}{2}r]^{n10}[\tfrac{1}{2}r]^{n01}[\tfrac{1}{2}(1-r)]^{n00}$$

$$= n!/(n_{11}!n_{10}!n_{01}!n_{00}!) \times [\tfrac{1}{2}(1-r)]^{n11+n00}[\tfrac{1}{2}r]^{n10+n01}$$

$$\log L(r|n_{ij}) = C+(n_{11}+n_{00})\log[\tfrac{1}{2}(1-r)] +(n_{10}+n_{01})\log[\tfrac{1}{2}r]$$

$$= C + (n_{11}+n_{00})\log(1-r) + (n_{10}+n_{01})\log r + n\log(\tfrac{1}{2})$$

Let the score

$$\partial \log L(r|n_{ij})/\partial r = (n_{11}+n_{00})[-1/(1-r)] +(n_{10}+n_{01})(1/r) = 0,$$

we have $(n_{11}+n_{00})[1/(1-r)]=(n_{10}+n_{01})(1/r) \rightarrow r\hat{} = (n_{10}+n_{01})/n$

Testing for linkage

BC	AaBb	aabb	Aabb	aaBb
Obs	n_{11}	n_{00}	n_{10}	n_{01} $\rightarrow n=\Sigma n_{ij}$
Freq	½(1-r)	½(1-r)	½r	½r
Gamete type	$n_{NR}= n_{11}+n_{00}$		$n_R= n_{10}+n_{01}$	
Freq with no linkage	½		½	
Exp	½n		½n	

$$\chi^2 = \Sigma(obs - exp)^2/exp$$

$$= (n_{NR} - n_R)^2/n \sim \chi^2_{df=1}$$

Example	AaBb	aabb	Aabb	aaBb
	49	47	3	4
	$n_{NR}= 49+47=96$		$n_R= 3 + 4 = 7$	

$n=96+7=103$

$$\chi^2 = \Sigma(obs - exp)^2/exp = (96-7)2/103 = 76.903 > \chi^2_{df=1} = 3.841$$

These two markers are statistically linked. r^ = 7/103 = 0.068

Gamete combinations in the F2

	AB ½(1-r)	Ab ½r	aB ½r	ab ½(1-r)
AB ½(1-r)	1	6	7	5
Ab ½r	6	2	5	8
aB ½r	7	5	3	9
ab ½(1-r)	5	8	9	4

Nine genotypes in the F2

		BB		Bb		bb	
AA	Obs	n_{22}	1	n_{21}	6	n_{20}	2
	Freq	$\frac{1}{4}(1-r)^2$		$\frac{1}{2}r(1-r)$		$\frac{1}{4}r^2$	
Aa	Obs	n_{12}	7	n_{11}	5	n_{10}	8
	Freq	$\frac{1}{2}r(1-r)$		$\frac{1}{2}(1-r)^2+\frac{1}{2}r^2$		$\frac{1}{2}r(1-r)$	
aa	Obs	n_{02}	3	n_{01}	9	n_{00}	4
	Freq	$\frac{1}{4}r^2$		$\frac{1}{2}r(1-r)$		$\frac{1}{4}(1-r)^2$	

Likelihood function

$$L(r|n_{ij}) = n!/(n_{22}!...n_{00}!)$$
$$\times[\tfrac{1}{4}(1-r)^2]^{n22+n00}[\tfrac{1}{4}r^2]^{n20+n02}[\tfrac{1}{2}r(1-r)]^{n21+n12+n10+n01}$$
$$\times[\tfrac{1}{2}(1-r)^2+\tfrac{1}{2}r^2]^{n11}$$

Let the score = 0 so as to obtain the MLE of r, but this will be difficult because AaBb contains a mix of two genotype formation types (in the dominator we will have $\frac{1}{2}(1-r)^2+\frac{1}{2}r^2$).

I will propose a shortcut EM algorithm for obtain the MLE of r

		BB	Bb	bb
AA	Obs	n_{22}	n_{21}	n_{20}
	Freq	$\frac{1}{4}(1-r)^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$
	Recombinant	0	1	2
Aa	Obs	n_{12}	n_{11}	n_{10}
	Freq	$\frac{1}{2}r(1-r)$	$\frac{1}{2}(1-r)^2+\frac{1}{2}r^2$	$\frac{1}{2}r(1-r)$
	Recombinant	1	$2r^2/[(1-r)^2+r^2]$	1
aa	Obs	n_{02}	n_{01}	n_{00}
	Freq	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}(1-r)^2$
	Recombinant	2	1	0

Based on the distribution of the recombinants (i.e., r), we have

$$r = 1/(2n)[2(n_{20}+n_{02})+(n_{21}+n_{12}+n_{10}+n_{01})+2r^2/[(1-r)^2+r^2]n_{11} \quad (1)$$

$$= 1/(2n)(2n_{2R} + n_{1R} + 2\phi n_{11})$$

where $n_{2R} = n_{20}+n_{02}$, $n_{1R} = n_{21}+n_{12}+n_{10}+n_{01}$, $n_{0R} = n_{22}+n_{00}$.

The EM algorithm is formulated as follows

E step: Calculate $2\phi = 2r^2/[(1-r)^2+r^2]$ (expected the number of recombination events for the double heterozygote AaBb)

M step: Calculate r^ by substituting the calculated ϕ from the E step into Equation 1

Repeat the E and M step until the estimate of r is stable

Example

	BB	Bb	bb
AA	$n_{22}=20$	$n_{21}=17$	$n_{20}=3$
Aa	$n_{12}=20$	$n_{11}=49$	$n_{10}=19$
aa	$n_{02}=3$	$n_{01}=21$	$n_{00}=19$

Calculating steps:

1. Give an initiate value for r, $r^{(1)}=0.1$,
2. Calculate $\phi^{(1)}=(r^{(1)})^2/[(1- r^{(1)})^2+(r^{(1)})^2] = 0.1^2/[(1-0.1)^2+0.1^2] = x$;
3. Estimate r using Equation 1, $r^{(2)} = y$;
4. Repeat steps 2 and 3 until the estimate of r is stable (converges). The MLE of r = 0.31.

How to determine that r has converged?

$|r^{(t+1)} - r^{(r)}|$ < a very small number, e.g., e^{-8}

Testing the linkage in the F2

		BB	Bb	bb
AA	Obs	$n_{22}=20$	$n_{21}=17$	$n_{20}=3$
	Exp with no linkage	1/16n	1/8n	1/16n
Aa	Obs	$n_{12}=20$	$n_{11}=49$	$n_{10}=19$
	Exp with no linkage	1/8n	¼n	1/8n
aa	Obs	$n_{02}=3$	$n_{01}=21$	$n_{00}=19$
	Exp with no linkage	1/16n	1/8n	1/16n

$$n = \Sigma n_{ij} = 191$$

$$\chi^2 = \Sigma(obs - exp)^2/exp \sim \chi^2_{df=1}$$

$$= (20\text{-}1/16\times 191)/(1/16\times 191) + \ldots = a > \chi^2_{df=1}=3.381$$

Therefore, the two markers are significantly linked.

Log-likelihood ratio test statistic

Two alternative hypotheses

$$H0: r = 0.5 \text{ vs. } H1: r \neq 0.5$$

Likelihood value under H1

$$L_1(r|n_{ij}) = n!/(n_{22}!...n_{00}!) \times [¼(1-r)^2]^{n22+n00}[¼r^2]^{n20+n02}[½r(1-r)]^{n21+n12+n10+n01}[½(1-r)^2+½r^2]^{n11}$$

Likelihood value under H0

$$L_0(r=0.5|n_{ij}) = n!/(n_{22}!...n_{00}!) \times [¼(1-0.5)^2]^{n22+n00}[¼0.5^2]^{n20+n02}[½0.5(1-0.5)]^{n21+n12+n10+n01}[½(1-0.5)^2+½0.5^2]^{n11}$$

$$LOD = \log_{10}[L_1(r|n_{ij})/L_0(r=0.5|n_{ij})]$$

$$= \{(n_{22}+n_{00})2[\log_{10}(1-r)-\log_{10}(1-0.5)+\ldots\} = 6.08 > \text{critical LOD}=3$$

Three-point analysis

- Determine a most likely gene order;
- Make full use of information about gene segregation and recombination

Consider three genes **A**, **B** and **C**.

Three possible orders **A-B-C**, **A-C-B**, or **B-A-C**

AaBbCc produces 8 types of gametes (haplotypes) which are classified into four groups

	Recombinant # between		Observation	Frequency
	A and **B**	**B** and **C**		
ABC and abc	0	0	$n_{00}=n_{ABC}+n_{abc}$	g_{00}
ABc and abC	0	1	$n_{01}=n_{Abc}+n_{abC}$	g_{01}
aBC and Abc	1	0	$n_{10}=n_{aBC}+n_{Abc}$	g_{10}
AbC and aBc	1	1	$n_{11}=n_{AbC}+n_{aBc}$	g_{11}

Note that the first subscript of n or g denotes the number of recombinant between **A** and **B**, whereas the second subscript of n or g denotes the number of recombinant between **B** and **C** (assuming order **A-B-C**)

Matrix notation

Markers **A** and **B**	Markers **B** and **C**: Recombinant	Non-recombinant	Total
Recombinant	n_{11}	n_{10}	
Non-recombinant	n_{01}	n_{00}	
Total			n
Recombinant	g_{11}	g_{10}	r_{AB}
Non-recombinant	g_{01}	g_{00}	$1-r_{AB}$
Total	r_{BC}	$1-r_{BC}$	1

What is the recombination fraction between A and C?

$r_{AC} = g_{01} + g_{10}$

Thus, we have

$r_{AB} = g_{11} + g_{10}$

$r_{BC} = g_{11} + g_{01}$

$r_{AC} = g_{01} + g_{10}$

The data log-likelihood

(complete data, it is easy to derive the MLEs of g_{ij}'s)

$$\log L(g_{00}, g_{01}, g_{10}, g_{11} | n_{00}, n_{01}, n_{10}, n_{11}, n)$$

$$= \log n! - (\log n_{00}! + \log n_{01}! + \log n_{10}! + \log n_{11}!) + n_{00} \log g_{00} + n_{01} \log g_{01} + n_{10} \log g_{10} + n_{11} \log g_{11}$$

The MLE of g_{ij} is: $\hat{g}_{ij} = n_{ij}/n$

Based on the invariance property of the MLE, we obtain the MLE of r_{AB}, r_{AC} and r_{BC}.

A relation:

$0 \le g_{11} = \frac{1}{2}(r_{AB} + r_{BC} - r_{AC}) \rightarrow r_{AC} \le r_{AB} + r_{BC}$

$0 \le g_{10} = \frac{1}{2}(r_{AB} - r_{BC} + r_{AC}) \rightarrow r_{BC} \le r_{AB} + r_{AC}$

$0 \le g_{01} = \frac{1}{2}(-r_{AB} + r_{BC} + r_{AC}) \rightarrow r_{AB} \le r_{AC} + r_{BC}$

Advantages of three-point (and generally multi-point) analysis

- Determine the gene order,
- Increase the estimation precision of the recombination fractions (for partially informative markers).

Real-life example – AoC/oBo × ABC/ooo

Eight groups of offspring genotypes

	A_B_C_	A_B_cc	A_bbC_	A_bbcc	aaB_C_	aaV_cc	aabbC_	aabbcc
Obs.	28	4	12	3	1	8	2	2

Order	**A**	**-**	**B**	**-**	**C**
Two-point analysis		0.38±0.386		0.39±0.418	
			0.18±0.056		
Three-point analysis		0.20±0.130		0.20±0.130	
			0.20±0.059		

Multilocus likelihood – determination of a most likely gene order

- Consider three markers **A, B, C,** with no particular order assumed.
- A triply heterozygous F1 ABC/abc backcrossed to a pure parent abc/abc

Genotype	ABC or abc	ABc or abC	Abc or aBC	AbC or aBc
Obs.	n_{00}	n_{01}	n_{10}	n_{11}
Frequency under				
Order **A-B-C**	$(1-r_{AB})(1-r_{BC})$	$(1-r_{AB})\, r_{BC}$	$r_{AB}(1-r_{BC})$	$r_{AB}\, r_{BC}$
Order **A-C-B**	$(1-r_{AC})(1-r_{BC})$	$r_{AC}\, r_{BC}$	$r_{AC}(1-r_{BC})$	$(1-r_{AC})r_{BC}$
Order **B-A-C**	$(1-r_{AB})(1-r_{AC})$	$(1-r_{AB})\, r_{AC}$	$r_{AB}r_{AC}$	$r_{AB}(1-r_{AC})$

r_{AB} = the recombination fraction between **A** and **B**
r_{BC} = the recombination fraction between **B** and **C**
r_{AC} = the recombination fraction between **A** and **C**

It is obvious that

$$r_{AB} = (n_{10} + n_{11})/n$$
$$r_{BC} = (n_{01} + n_{11})/n$$
$$r_{AC} = (n_{01} + n_{10})/n$$

What order is the mostly likely?

$$L_{\mathbf{ABC}} \propto (1-r_{AB})^{n00+n01} (1-r_{BC})^{n00+n10} (r_{AB})^{n10+n11} (r_{BC})^{n01+n11}$$
$$L_{\mathbf{ACB}} \propto (1-r_{AC})^{n00+n11} (1-r_{BC})^{n00+n10} (r_{AC})^{n01+n10} (r_{BC})^{n01+n11}$$
$$L_{\mathbf{BAC}} \propto (1-r_{AB})^{n00+n01} (1-r_{AC})^{n00+n11} (r_{AB})^{n10+n11} (r_{AC})^{n01+n10}$$

According to the maximum likelihood principle, the linkage order that gives the maximum likelihood for a data set is the best linkage order supported by the data. This can be extended to include many markers for searching for the best linkage order.

Map function

- Transfer the recombination fraction (non-additivity) between two genes into their corresponding genetic map distance (additivity)
- Map distance is defined as the mean number of crossovers
- The unit of map distance is Morgan (in honor of T. H. Morgan who obtained the Novel prize in 1930s)
- 1 Morgan or M = 100 centiMorgan or cM

The Haldane map function (Haldane 1919)

Assumptions:

- No interference (the occurrence of one crossover is independent of that of next)
- Crossover events follow the Poisson distribution.

Consider three markers with an order **A-B-C**

A triply heterozygous F1 ABC/abc backcrossed to a pure parent abc/abc

Event	Gamete	Frequency
No crossover	ABC or abc	$(1-r_{AB})(1-r_{BC})$
Crossover between **B**&**C**	ABc or abC	$(1-r_{AB})r_{BC}$
Crossover between **A**&**B**	Abc or aBC	$r_{AB}(1-r_{BC})$
Crossovers between **A**&**B** and **B**&**C**	AbC or aBc	$r_{AB}r_{BC}$

The recombination fraction between **A** and **C** is expected to be

$r_{AC} = (1-r_{AB})r_{BC} + r_{AB}(1-r_{BC}) = r_{AB}+r_{BC}-2r_{AB}r_{BC}$

$\rightarrow (1-2r_{AC})=(1-2r_{AB})(1-2r_{BC})$

Map distance:

A genetic length (map distance) x of a chromosome is defined as the mean number of crossovers.

Poisson distribution (x = genetic length):

Crossover event	0	1	2	3	…	t	…
Probability	e^{-x}	xe^{-x}	$\frac{x^2e^{-x}}{2!}$	$\frac{x^3e^{-x}}{3!}$	…	$\frac{x^te^{-x}}{t!}$	…

The value of r (recombination fraction) for a genetic length of x is the sum of the probabilities of all odd numbers of crossovers:

$$r = e^{-x}(x^1/1! + x^3/3! + x^5/5! + x^7/7! + \ldots)$$

$$= \tfrac{1}{2}(1 - e^{-2x})$$

$$x = -\tfrac{1}{2}\ln(1-2r)$$

We have $x_{AC} = x_{AB} + x_{BC}$ for a given order **A-B-C,** but generally, $r_{AC} \neq r_{AB} + r_{BC}$

Proof of $x_{AC} = x_{AB} + x_{BC}$

For order **A-B-C**, we have

$r_{AB} = ½(1- e^{-2xAB})$, $r_{BC} = ½(1- e^{-2xBC})$, $r_{AC} = ½(1- e^{-2xAC})$

$r_{AC} = r_{AB} + r_{BC} – 2r_{AB}r_{BC}$

$= ½(1- e^{-2xAB}) + ½(1- e^{-2xBC})$

$- 2\ ½(1- e^{-2xAB})\ ½(1- e^{-2xBC})$

$= ½[1- e^{-2xAB} +1- e^{-2xBC} -1+ e^{-2xAB} + e^{-2xBC} - e^{-2xAB}\ e^{-2xBC}$

$= ½(1- e^{-2(xAB+xBC)})$

$= ½(1- e^{-2xAC})$, which means $x_{AC} = x_{AB} + x_{BC}$

The Kosambi map function (Kosambi 1943)

The Kosambi map function is an extension of the Haldane map function

For gene order **A-B-C**

[1] $r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$

[2] $r_{AC} \approx r_{AB} + r_{BC}$, for small r's

[3] $r_{AC} \approx r_{AB} + r_{BC} - r_{AB}r_{BC}$, for intermediate r's

The Kosambi map function attempts to find a general expression that covers all the above relationships

Map Function

	Haldane	Kosambi
$x =$	$-\frac{1}{2}\ln(1-2r)$	$\frac{1}{4}\ln(1+2r)/(1-2r)$
$r =$	$\frac{1}{2}(1-e^{-2x})$	$\frac{1}{2}(e^{2x}-e^{-2x})/(e^{2x}+e^{-2x})$
$r_{AC} =$	$r_{AB}+r_{BC}-2r_{AB}r_{BC}$	$(r_{AB}+r_{BC})/(1+4r_{AB}r_{BC})$

Reference

Ott, J, 1991 *Analysis of Human Genetic Linkage*.

The Johns Hopkins University Press, Baltimore and London

Construction of genetic maps

- The Lander-Green algorithm -- a hidden Markov chain
- Genetic algorithm

- Lander ES, Green P (1987) **Construction of multilocus genetic linkage maps in humans.** Proc Natl Acad Sci U S A. 84(8): 2363–2367.
- Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA (1987) **MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations.** Genomics 1:174-81.