Bioinformatics and Statistical Genetics

Elective specialization for MDS/MIRI/MAI students

Marta Castellano

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya Barcelona, Spain

marta.castellano@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics
2. Hardy-Weinberg equilibrium
3. Linkage disequilibrium and haplotype estimation
4. Population substructure
5. Genetic association analysis
6. Relatedness analysis (allele sharing)

21 November 2023

Tuesdays 5-8pm

Hardy-Weinberg Equilibrium (HWE)

RECAP

When a population is in Hardy-Weinberg equilibrium for a gene, it is not evolving, and allele frequencies will stay the same across generations….in the absence of **disturbing factors**.

Statistical testing allows to decide whether the sample genotype frequencies adhere to the frequencies observed in HWE. The null hypothesis states that $f(AA) = p^2, f(AB) = 2pq, f(BB) = q^2$. Most frequent statistical tests:

- Pearson's test.
- Fisher's exact test (based on).
- Permutation test.

Measures of disequilibrium (like the inbreeding coefficient f) quantify how far from HWE the sample is.

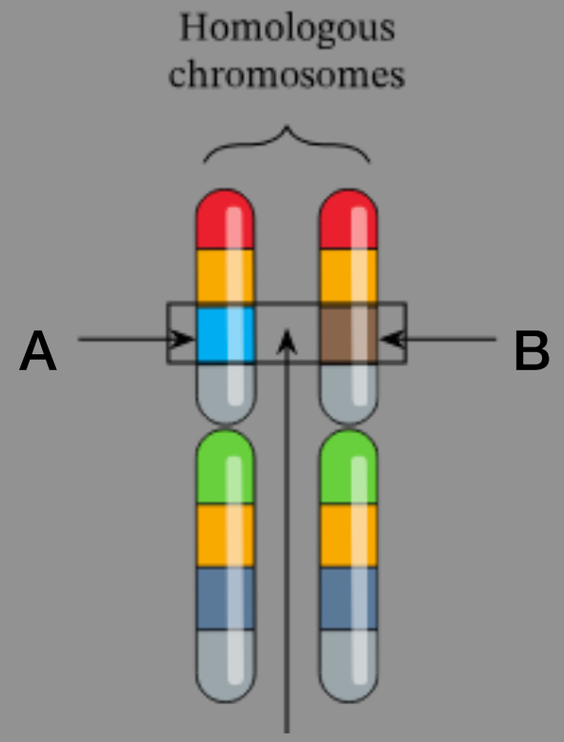
Generalizations of the HWE are required for multiple alleles, the variant has multiple copies, the organism studied is tetraploid or the variant is found on the X-chromosome.

		Females	
		A (p)	B (q)
Males	A (p)	AA (p^2)	AB (pq)
	B (q)	AB (qp)	BB (q^2)

frequency of homozygous dominant genotype — frequency of heterozygous recessive genotype

$$p^2 + 2pq + q^2 = 1$$

frequency of heterozygous genotype



Locus of interest

Content

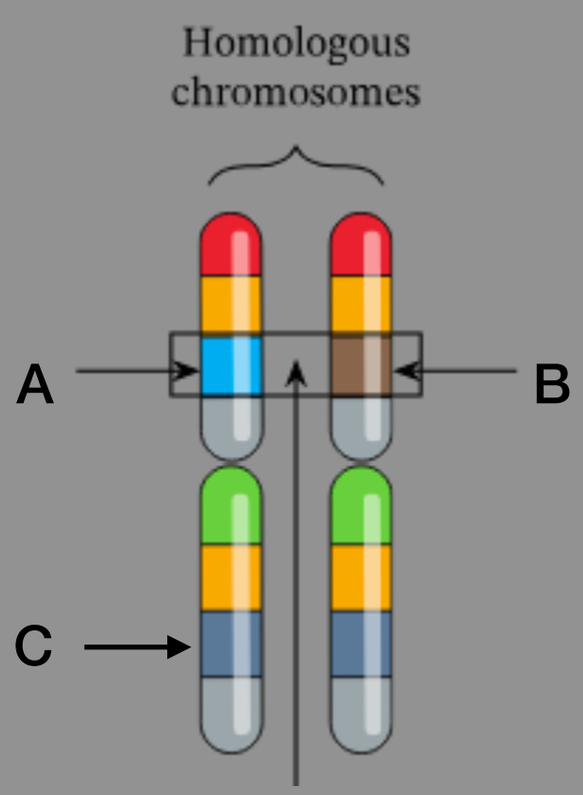
Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in population genetics: Linkage disequilibrium



Locus of interest

Both concepts refer to association between alleles but….

HWE refers to an association between alleles at the same locus

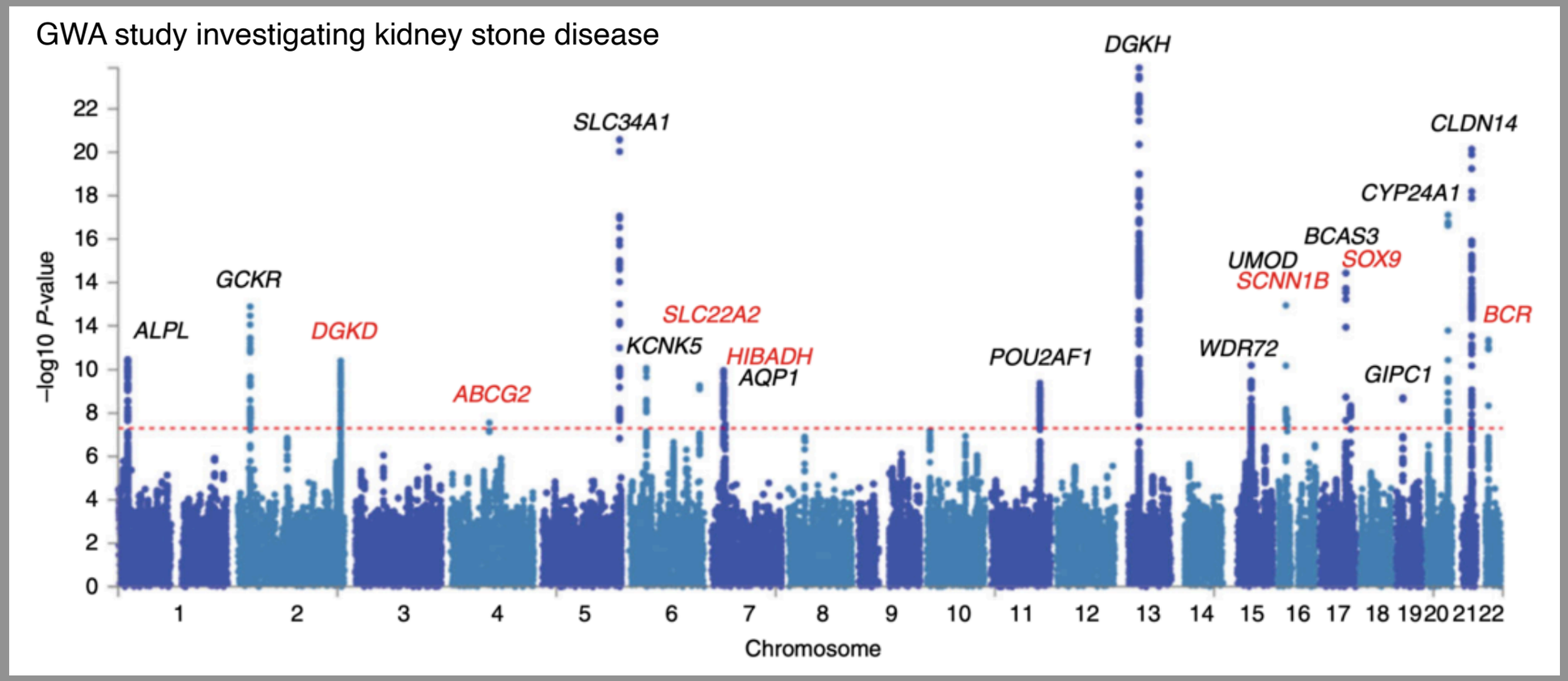
LD refers to associations between alleles at different loci

Linkage disequilibrium

Why? (Still ELI5)

Disease-marker association studies

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in ~~population~~ genetics: Linkage disequilibrium



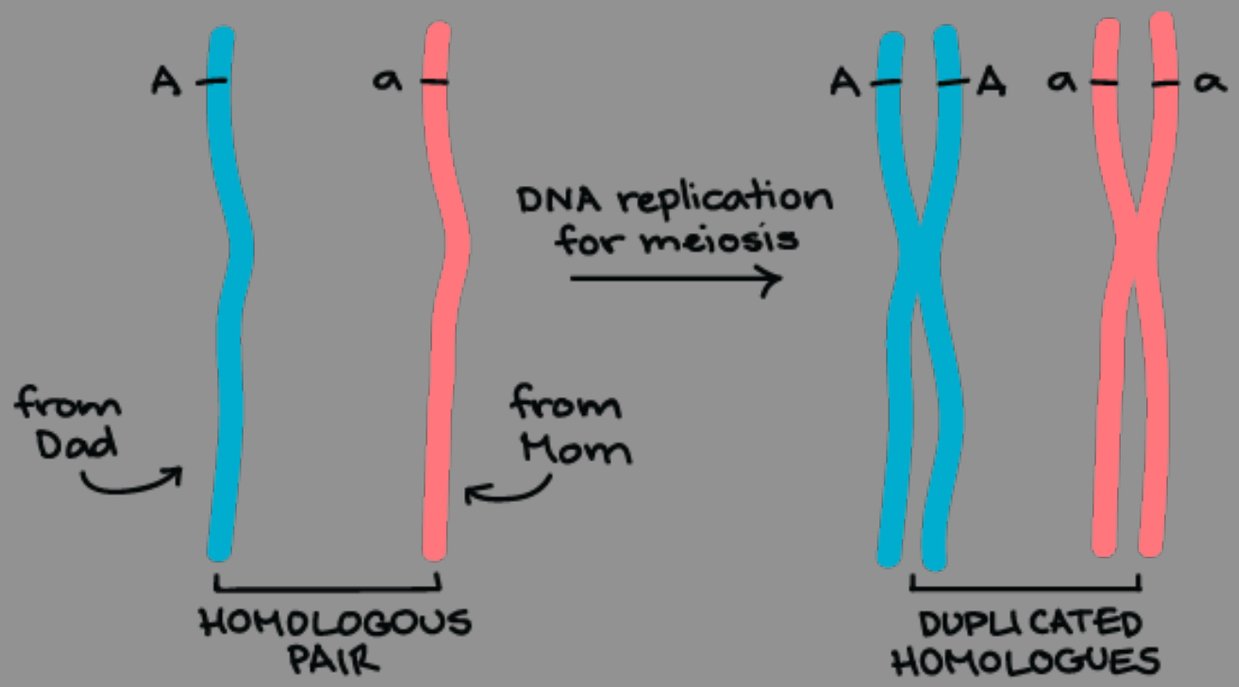
Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. The peaks indicate genetic variants that are found more often in individuals with kidney stones.

When we are looking for regions of the genome or SNP that is causal for a gene, we often find that a whole bunch of SNPs are associated with the disease. Its not that they all cause disease, it is just that a whole bunch are correlated with the causal SNP (passenger mutations). Thus it is our job to identify the causal needle in the haystack.

Linkage disequilibrium

Why? (Still ELI5)

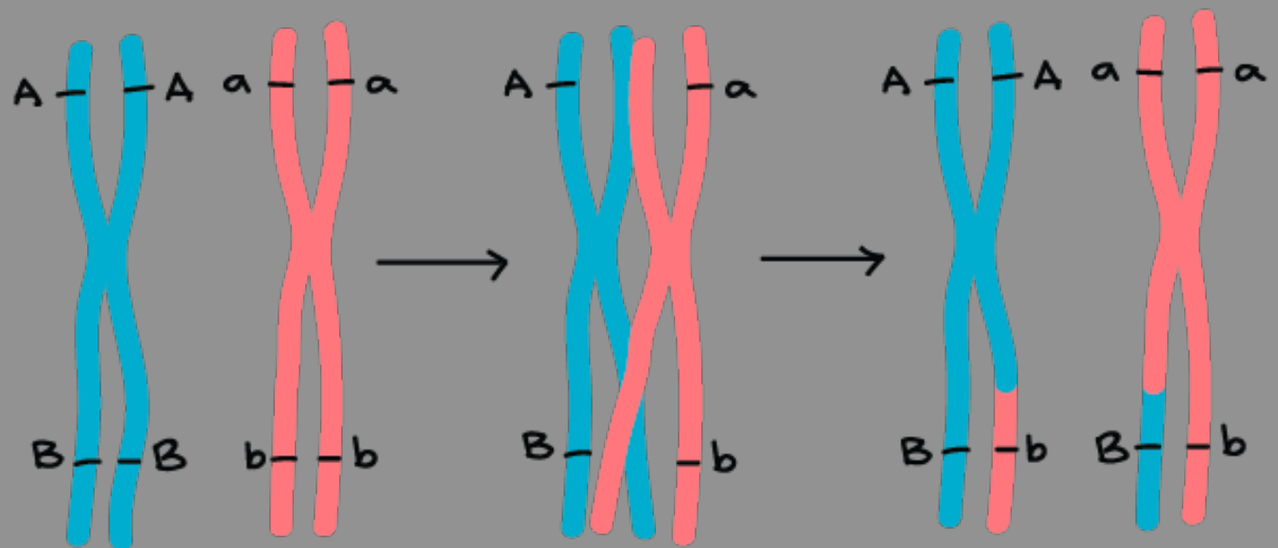
LD is not (only) a intra-chromosome concept but helps to understand it if we start talking about the first stage in cell division (meiosis)…

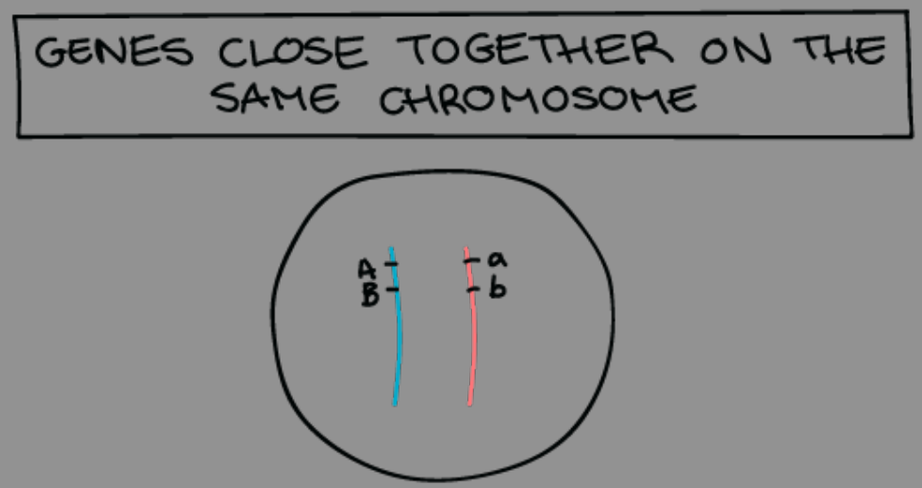


Linkage disequilibrium

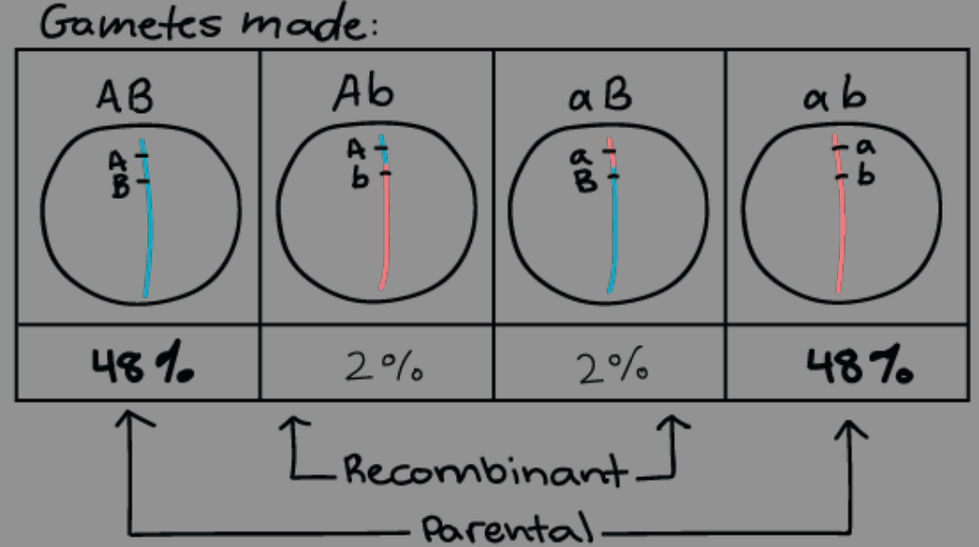
Why? (Still ELI5)

LD is not (only) a intra-chromosome concept but helps to understand it if we start talking about meiosis…





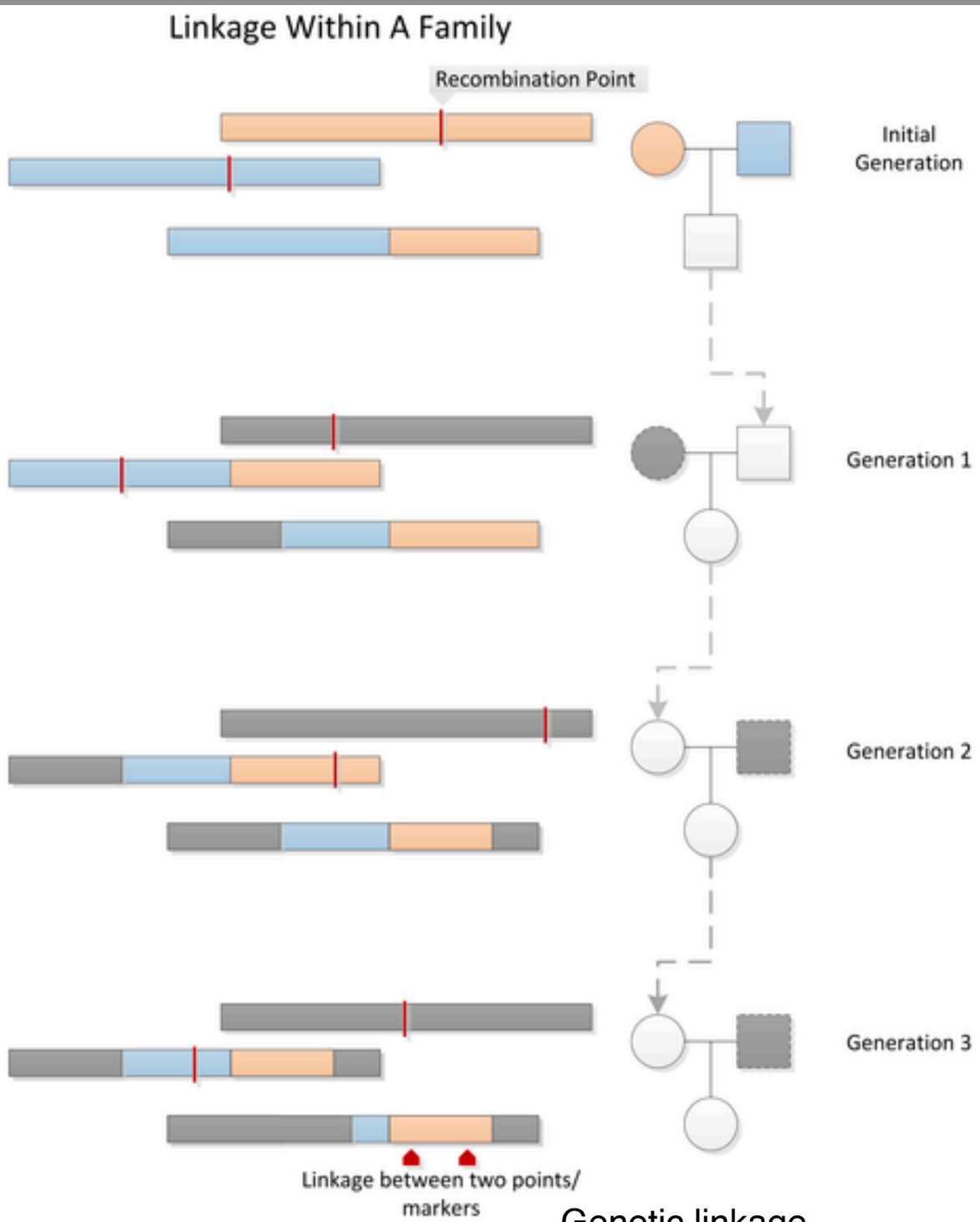
- At the beginning of meiosis, homologous chromosomes randomly exchange matching fragments (genetic recombination).



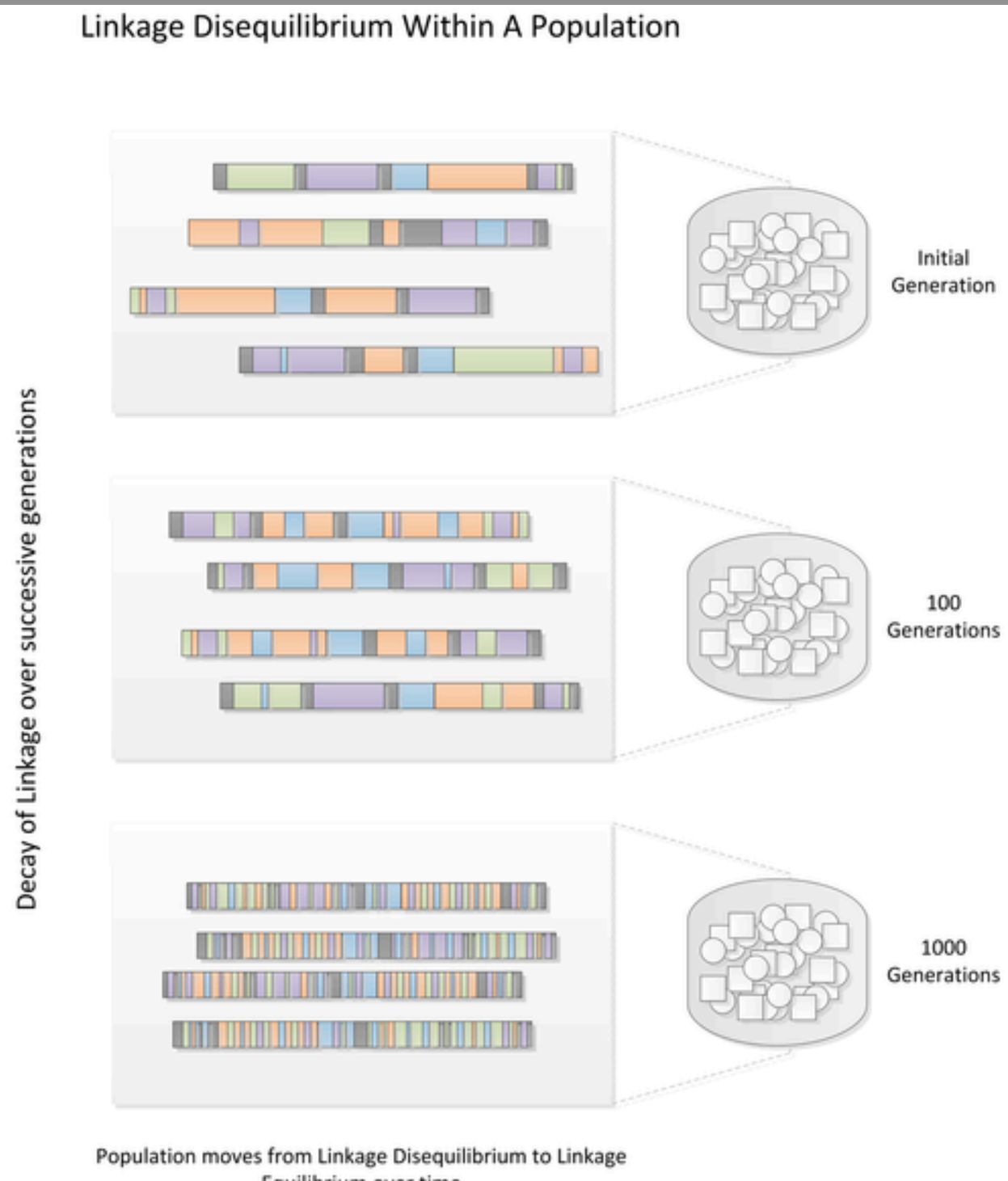
Linkage disequilibrium

Why? (Still ELI5)

Stronger-than-expected linkage can reflect selection for functional relationships between genes, such as combinations of alleles that are advantageous when inherited together, or shared regulatory mechanisms (positive selection)



Genetic linkage



Linkage disequilibrium

Introduction to LD

- **LD:** in population genetics, linkage disequilibrium (LD) is the non-random association of alleles at different loci in a given population.

- The terms suggests this to be a consequence of the physical closeness of the sites (genetic linkage), but this is not necessarily so.

- LD is an important concept in disease-marker association studies.

- There are many factors affecting the linkage disequilibrium. At a population level, consider:

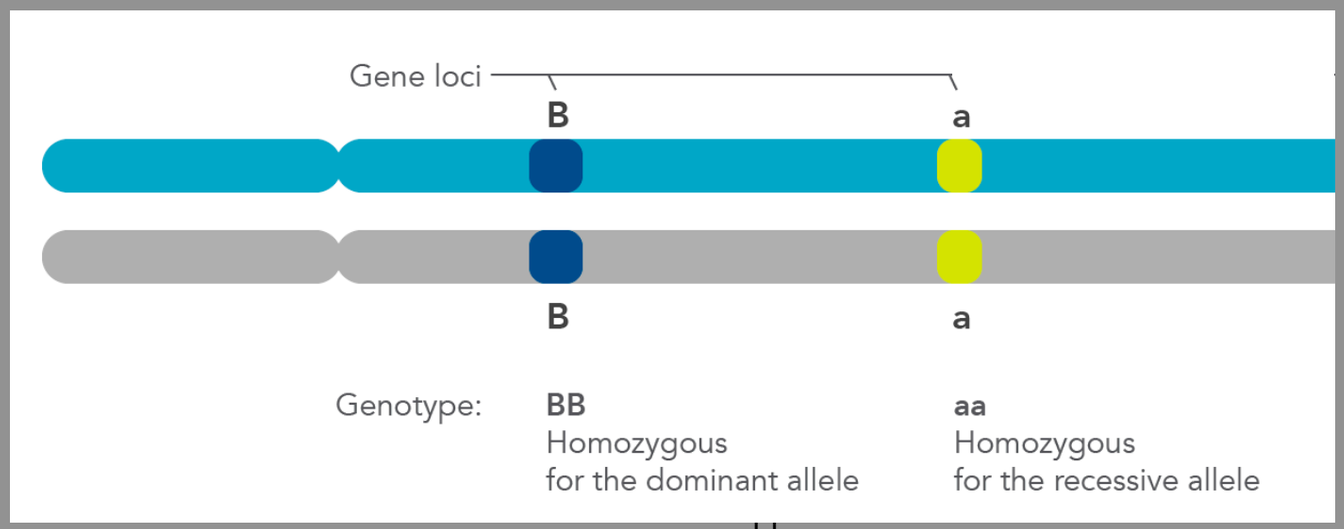
    - Natural selection
    - Rate of genetic recombination
    - Mutation rate
    - Genetic drift (the change in frequency of an existing gene variant in the population due to random chance)
    - Mating system
    - Population structure
    - …

Linkage disequilibrium

Introduction to LD

Linkage disequilibrium vs Hardy Weinberg Equilibrium

- Both concepts refer to association between alleles
- HWE refers to association between alleles at the same locus (within one marker)
- LD refers to association between alleles at different loci (between markers)



Linkage disequilibrium

Introduction to LD

Linkage disequilibrium vs Hardy Weinberg Equilibrium

- Both concepts refer to association between alleles
- HWE refers to association between alleles at the same locus (within one marker)
- LD refers to association between alleles at different loci (between markers)

	id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633	rs12534908
1	NA18939	T/G	C/T	G/A	T/G	T/G	G/A
2	NA18940	G/G	T/T	A/A	G/G	T/G	A/A
3	NA18941	G/G	T/T	A/A	G/G	T/G	A/A
4	NA18942	G/G	T/T	A/A	G/G	T/T	A/A
5	NA18943	G/G	T/T	A/A	G/G	T/T	A/A
6	NA18944	T/T	C/C	G/G	T/G	G/G	G/G
7	NA18945	G/G	T/T	A/A	G/G	G/G	A/A
8	NA18946	T/G	C/T	G/A	G/G	G/G	G/A
9	NA18947	T/G	C/T	G/A	G/G	T/G	G/A
10	NA18948	G/G	T/T	A/A	G/G	G/G	A/A
11	NA18949	T/G	C/T	G/A	T/G	T/G	G/A
12	NA18950	G/G	T/T	A/A	G/G	T/G	A/A
13	NA18951	G/G	T/T	A/A	G/G	T/G	A/A
14	NA18952	T/G	C/C	G/G	T/G	T/G	G/G

Linkage disequilibrium

Haplotype

RECALL:

- An haplotype is a combination of alleles at different chromosomal regions that are closely linked and that tend to be inherited together.
- An haplotype is a group of alleles in an organism that are inherited together from a single parent.
- In practice, a haplotype often refers to a set of SNPs on a single chromosome that are statistically associated.

Content

Linkage disequilibrium

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium

Measures of LD

- D (coefficient of linkage disequilibrium)
- Lewontin's $D' = \frac{D}{D_{max}}$
- R^2
- χ^2 statistic of a contingency table
- p – value in a chi-square test or in an exact test
- ...

All measures attempt to quantify the difference between the observed frequency of a particular combination of alleles at two loci and the frequency expected for the random association

Linkage disequilibrium

Measures of LD

Coefficient of linkage disequilibrium D

- Consider a population of n individuals
- Consider two sites (two bi-allelic markers) on the same chromosome
- One marker with alleles A and a, and one marker with alleles B and b
- Four possible haplotypes: AB, Ab, aB and ab with frequencies pAB, pAb, paB and pab
- Allele frequencies: pA, pa, pB and pb
- Expected probabilities of each haplotype **under independence**:

		SNP2		
		B	b	
SNP1	A	$p_A p_B$	$p_A p_b$	p_A
	a	$p_a p_B$	$p_a p_b$	p_a
		p_B	p_b	1

$$p_{AB} = p_A p_B$$

Linkage disequilibrium

Measures of LD

Coefficient of linkage disequilibrium D

- Consider a population of n individuals
- Consider two sites (two bi-allelic markers) on the same chromosome
- One marker with alleles A and a, and one marker with alleles B and b
- Four possible haplotypes: AB, Ab, aB and ab with frequencies pAB, pAb, paB and pab
- Allele frequencies: pA, pa, pB and pb
- Expected probabilities of each haplotype **in presence of LD**:

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

There is said to be a linkage disequilibrium between the two alleles whenever pAB differs from pApB for any reason:

$$p_{AB} = p_A p_B + D$$

The level of linkage disequilibrium between A and B can be quantified by the coefficient of linkage disequilibrium D which is defined as:

$$D = p_{AB} - p_A p_B$$

$D > 0$ known as ‘coupling’

$D < 0$ known as ‘repulsion’

Linkage disequilibrium

Measures of LD

Coefficient of linkage disequilibrium D

- Consider a population of n individuals
- Consider two sites (two bi-allelic markers) on the same chromosome
- One marker with alleles A and a, and one marker with alleles B and b
- Four possible haplotypes: AB, Ab, aB and ab with frequencies pAB, pAb, paB and pab
- Allele frequencies: pA, pa, pB and pb
- Expected probabilities of each haplotype **in presence of LD**:

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

We can obtain D from any haplotype frequency

$$D = p_{AB} - p_A p_B$$

or

$$p_{ab} = p_a p_b + D \longrightarrow D = p_{ab} - p_a p_b$$

$$p_{aB} = p_a p_B - D \longrightarrow -D = p_{aB} - p_a p_B$$

$$p_{Ab} = p_A p_b - D \longrightarrow -D = p_{Ab} - p_A p_b$$

Linkage disequilibrium

Measures of LD

Coefficient of linkage disequilibrium D

- The coefficient of linkage disequilibrium $D = p_{AB} - p_A p_B$ quantifies the deviation from independence.
- If D = 0 then we have $p_{AB} = p_A p_B$ and then the alleles A and B are said to be in linkage equilibrium.
- –0.25 ≤ D ≤ +0.25

How to compute D?

- pA and pB can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$
- pAB is unobserved and thus unknown
  - Recall $p_{AB} = p_A p_B + D$
  - We have data at the genotype level, and pAB is at the haplotype level.
  - Directly depends on the frequencies of the alleles under study pA and pB.

Linkage disequilibriu

Measures of LD

Coefficient of linkage disequilibrium D

Observed genotype data				
		SNP2		
		BB	Bb	bb
SNP1	AA	n_{AABB}	n_{AABb}	n_{AAbb}
	Aa	n_{AaBB}	n_{AaBb}	n_{Aabb}
	aa	n_{aaBB}	n_{aaBb}	n_{aabb}

- Given an observed genotype data we aim to estimate $D = p_{AB} - p_A p_B$
  - pA and pB can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$
  - pAB is unobserved and thus unknown. Recall $p_{AB} = p_A p_B + D$
- This data can be considered a sample from a multivariate normal distribution with 9 categories, where the probability of each of the 9 categories ultimately depends on the four haplotype probabilities pAB,pAb,paB and pab.

$$\boldsymbol{\theta} = (p_{AB}, p_{Ab}, p_{aB}, p_{ab}), \quad \mathbf{x} = (n_{AABB}, n_{AABb}, \ldots n_{aabb})$$

- It gets better…as the problem can be reparametrized in terms of pA,pB and pAB

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

$$p_A = p_{AB} + p_{Ab}$$
$$p_B = p_{AB} + p_{aB}$$
$$p_{AB} = 1 - (p_{Ab} + p_{aB} + p_{ab})$$

Linkage disequilibriu

Measures of LD

Coefficient of linkage disequilibrium D

Observed genotype data				
		SNP2		
		BB	Bb	bb
SNP1	AA	n_{AABB}	n_{AABb}	n_{AAbb}
	Aa	n_{AaBB}	n_{AaBb}	n_{Aabb}
	aa	n_{aaBB}	n_{aaBb}	n_{aabb}

- Given an observed genotype data we aim to estimate $D = p_{AB} - p_A p_B$
  - pA and pB can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$
  - pAB is unobserved and thus unknown. Recall $p_{AB} = p_A p_B + D$
- We will use a maximum likelihood approach

We maximize the likelihood by a Newton-Raphson algorithm
Alternatively the EM algorithm may be used

$$\boldsymbol{\theta} = (p_{AB}, p_{Ab}, p_{aB}, p_{ab}), \quad \mathbf{x} = (n_{AABB}, n_{AABb}, \ldots n_{aabb})$$

$$L(\boldsymbol{\theta}|\mathbf{x}) = \frac{n!}{n_{AABB}! \cdots n_{aabb}!} \cdot (p_{AB}^2)^{n_{AABB}} \cdots (p_{ab}^2)^{n_{aabb}}$$

Likelihood function

$$l(\boldsymbol{\theta}|\mathbf{x}) = C + 2n_{AABB} \ln(p_{AB}) + \cdots + 2n_{aabb} \ln(p_{ab})$$

Log-likelihood function

Maximum likelihood to estimate mean and variance: https://www.youtube.com/watch?v=XepXtl9YKwc
Maximum likelihood estimation in linear regression: https://www.youtube.com/watch?v=bhTIpGtWtzQ

Linkage disequilibrium

Measures of LD

- Other measures o assess linkage disequilibrium:
  - Lewontin's $D' = \frac{D}{D_{max}}$
- R^2
- χ^2 statistic of a contingency table
- p – value in a chi-square test or in an exact test
- ...

All measures attempt to quantify the difference between the observed frequency of a particular combination of alleles at two loci and the frequency expected for the random association

Linkage disequilibrium

Measures of LD

RECALL

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

Lewontin's D'

- D' is an attempt to standardize D

$$D' = \frac{D}{D_{max}}$$

- D_{max} is defined as:

$$D_{max} = \begin{cases} \min(p_A p_b, p_a p_B) & D > 0 \quad \text{(coupling)} \\ \min(p_A p_B, p_a p_b) & D < 0 \quad \text{(repulsion)} \end{cases}$$

- $-1 \leq D' \leq 1$.
- $D' \approx 0$: low LD
- $|D'|$ close to 1 : high LD.

Linkage disequilibrium

Measures of LD

R^2 and χ^2 statistic

- The genotype data can be recoded as indicator data, creating indicators for the carriers of the A and B allele.

- R^2 is the squared correlation between these indicators.

- R^2 is related to the χ^2 statistic of a 2 × 2 contingency table $R^2 = \chi^2/(2n)$

- The R^2 and the χ^2 statistic are related to D:

$$R^2 = \chi^2/(2n) = \frac{D^2}{p_A p_B p_a p_b}$$

Content

Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium

A statistical test for LD

- Hypothesis:

$$H_0 : D = 0$$

$$H_1 : D \neq 0$$

- Test statistic:

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

ML estimation

Once the haplotype frequencies $\hat{p}_{AB}$ are estimated, D can be estimated

- Asymptotically $\hat{D}$ will be normally distributed
- Theoretically:

$$E\left(\hat{D}\right) = \frac{2n-1}{2n} D \qquad V\left(\hat{D}\right) = \frac{1}{2n}\left(p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D - D^2\right)$$

- Imposing the null

$$Z = \frac{\hat{D}}{\sqrt{V\left(\hat{D}\right)}} \sim N(0,1) \quad \text{or, equivalently} \quad X^2 = \frac{2n\hat{D}^2}{\hat{p}_A(1-\hat{p}_A)\hat{p}_B(1-\hat{p}_B)} \sim \chi_1^2$$

Content

Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium

Example for LD

- Data from the FAMuSS (Functional SNPs Associated with Muscle Size and Strength) study (Foulkes, 2009)
- n = 1397 individuals and 225 SNPs Muscle performance variables
- We first consider the pair of SNPs, actn3 r577x and actn3 rs540874 and look at the data at the genotype level

```
                    actn3_rs540874
actn3_r577x           AA    GA   GG
                CC     3    24  184
                CT     4   296   15
                TT   127    42   30

chisq.test(Z)
Pearson's Chi-squared test
data:  Z
X-squared = 818.87, df = 4, p-value < 2.2e-16
```

Reject $H_0 : D = 0$
BUT
What exactly is the coefficient of linkage disequilibrium D?

ML estimation of probabilities pC, pA and pCA

It.	$l(P_{CA}, P_C, P_A \mid x)$	P_{CA}	P_C	P_A
0	-1471.8874	0.0100000	0.508276	0.434483
1	-1469.9878	0.0438867	0.503479	0.429587
2	-1460.8970	0.0375485	0.514644	0.441162
3	-1459.0183	0.0297541	0.514183	0.440727
4	-1458.2618	0.0288494	0.508727	0.435198
5	-1458.0022	0.0263196	0.509216	0.435692
6	-1457.9928	0.0257361	0.507443	0.433847
7	-1457.9840	0.0251530	0.509738	0.432716
8	-1457.9716	0.0253836	0.508019	0.434685
9	-1457.9709	0.0257321	0.507963	0.434594
10	-1457.9696	0.0256473	0.508296	0.434473
11	-1457.9696	0.0256113	0.508247	0.434500
12	-1457.9696	0.0256208	0.508278	0.434481
13	-1457.9696	0.0256212	0.508276	0.434483

After convergence

Linkage disequilibrium

Example for LD

- Data from the FAMuSS (Functional SNPs Associated with Muscle Size and Strength) study (Foulkes, 2009)
- n = 1397 individuals and 225 SNPs Muscle performance variables
- We first consider the pair of SNPs, actn3 r577x and actn3 rs540874 and look at the data at the genotype level

- After convergence
  - $p_{CA} = 0.025612$
  - $p_{CG} = p_C - p_{CA} = 0.4826544$
  - $p_{TA} = p_A - p_{CA} = 0.408862$
  - $p_{TG} = 1 - (p_{CA} + p_{GC} + p_{TA}) = 0.08286$
  - $D = p_{CA} - p_A p_c = 0.19$

ML estimation of probabilities pC, pA and pCA

It.	$l(P_{CA}, P_C, P_A \mid x)$	P_{CA}	P_C	P_A
0	-1471.8874	0.0100000	0.508276	0.434483
1	-1469.9878	0.0438867	0.503479	0.429587
2	-1460.8970	0.0375485	0.514644	0.441162
3	-1459.0183	0.0297541	0.514183	0.440727
4	-1458.2618	0.0288494	0.508727	0.435198
5	-1458.0022	0.0263196	0.509216	0.435692
6	-1457.9928	0.0257361	0.507443	0.433847
7	-1457.9840	0.0251530	0.509738	0.432716
8	-1457.9716	0.0253836	0.508019	0.434685
9	-1457.9709	0.0257321	0.507963	0.434594
10	-1457.9696	0.0256473	0.508296	0.434473
11	-1457.9696	0.0256113	0.508247	0.434500
12	-1457.9696	0.0256208	0.508278	0.434481
13	-1457.9696	0.0256212	0.508276	0.434483

After convergence

Linkage disequilibrium

Example for LD - R

- Data from the FAMuSS (Functional SNPs Associated with Muscle Size and Strength) study (Foulkes, 2009)
- n = 1397 individuals and 225 SNPs Muscle performance variables
- We first consider the pair of SNPs, actn3 r577x and actn3 rs540874

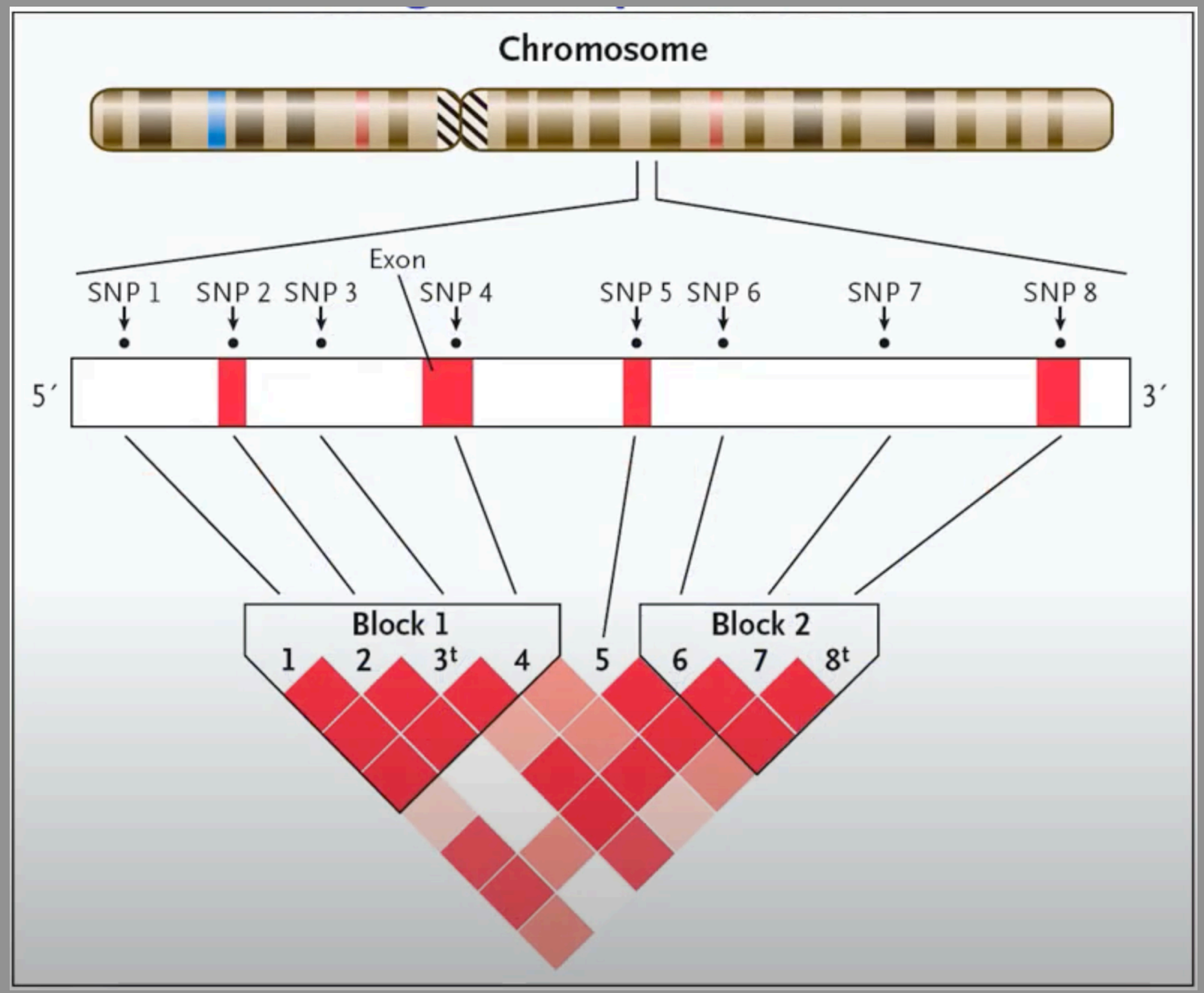
```
> actn3_r577x[1:10]
[1] "CC" "CT" "CT" "CT" "CC" "CT" "TT" "CT" "CT" "CC"
> actn3_rs540874[1:10]
[1] "GG" "GA" "GA" "GA" "GG" "GA" "AA" "GA" "GA" "GG"
> library(genetics)
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
> out <- LD(Actn3Snp1,Actn3Snp2)
> out

Pairwise LD
-----------
                  D           D'      Corr
Estimates: 0.1945726   0.8858385  0.7860811

              X^2  P-value    N
LD Test: 895.9891        0  725
```

$$D_{max} = \begin{cases} \min(p_A p_b, p_a p_B) & D > 0 \quad \text{(coupling)} \\ \min(p_A p_B, p_a p_b) & D < 0 \quad \text{(repulsion)} \end{cases}$$

Linkage disequilibrium

Example for LD - Heatmap

Heatmaps are a very common visual representation of the LD between SNPs under study. Each square shows some metric that quantifies LD between SNPs

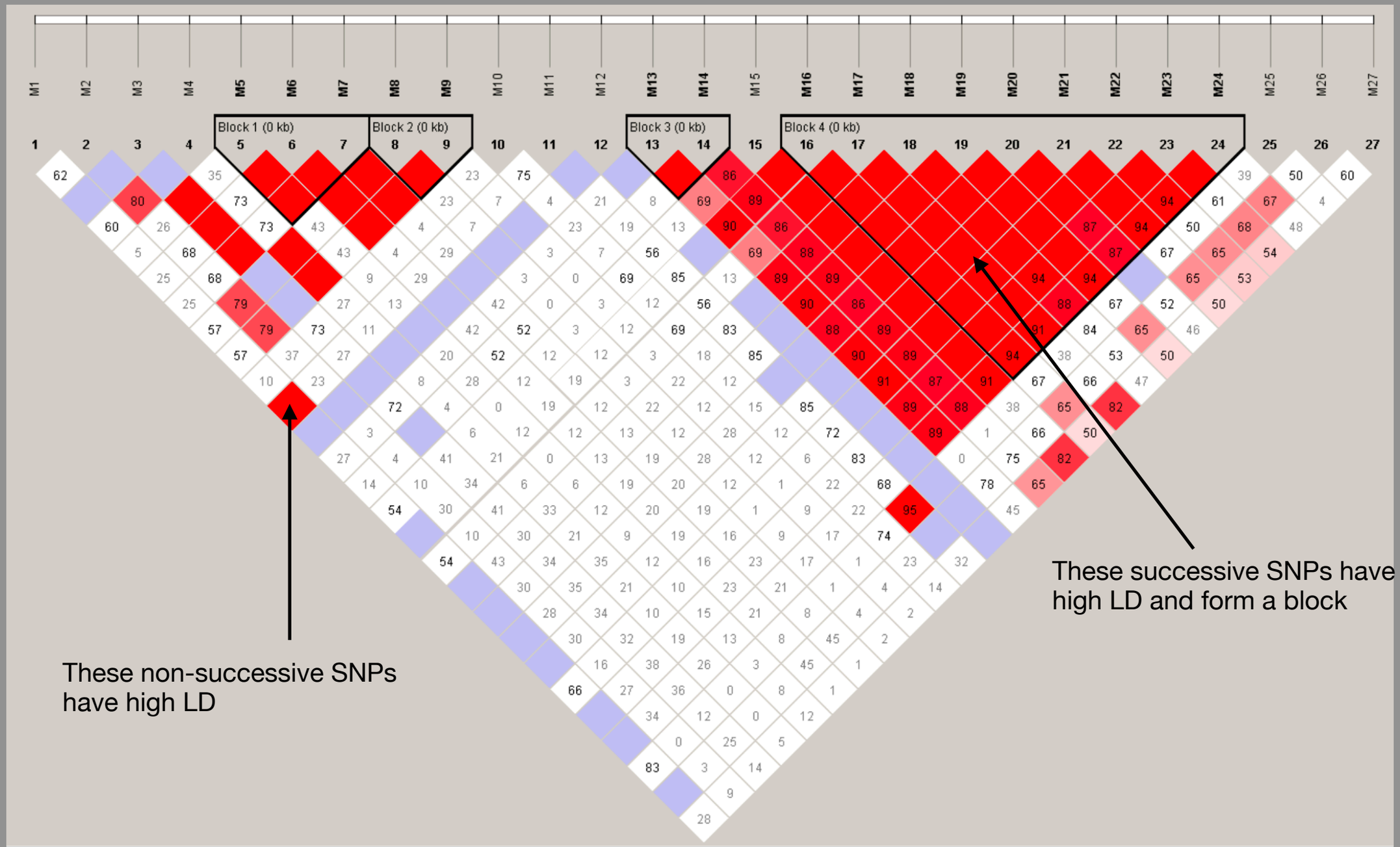


Linkage disequilibrium

HaploView is designed to simplify and expedite the process of haplotype analysis by providing a common interface to several tasks relating to such analyses

Example for LD - Heatmap from HaploView

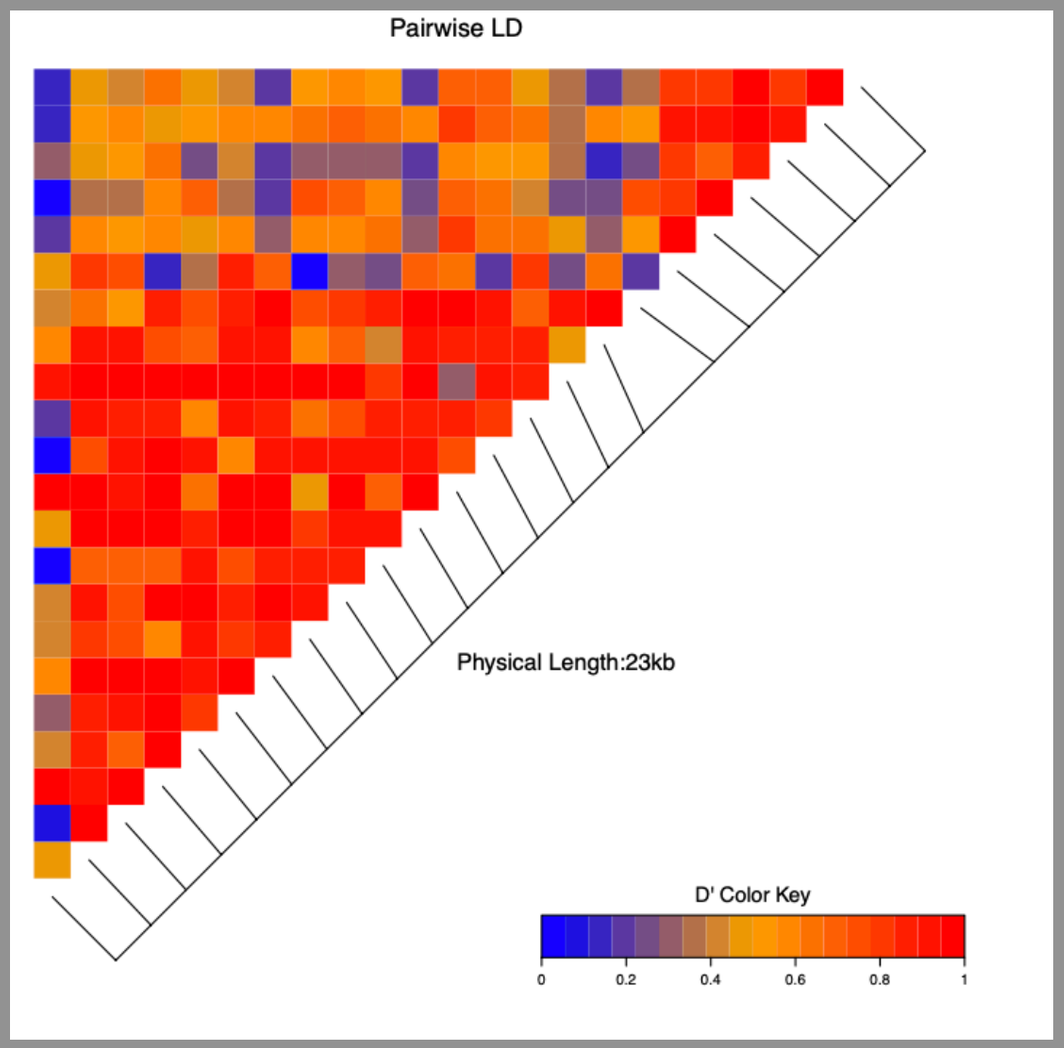
100 (successive) SNPs on chromosome 1 of a sample of 45 individuals from a Chinese population of the HapMap project (www.hapmap.org), 27 remaining after removing monomorphics



Linkage disequilibrium

Example for LD - Heatmap

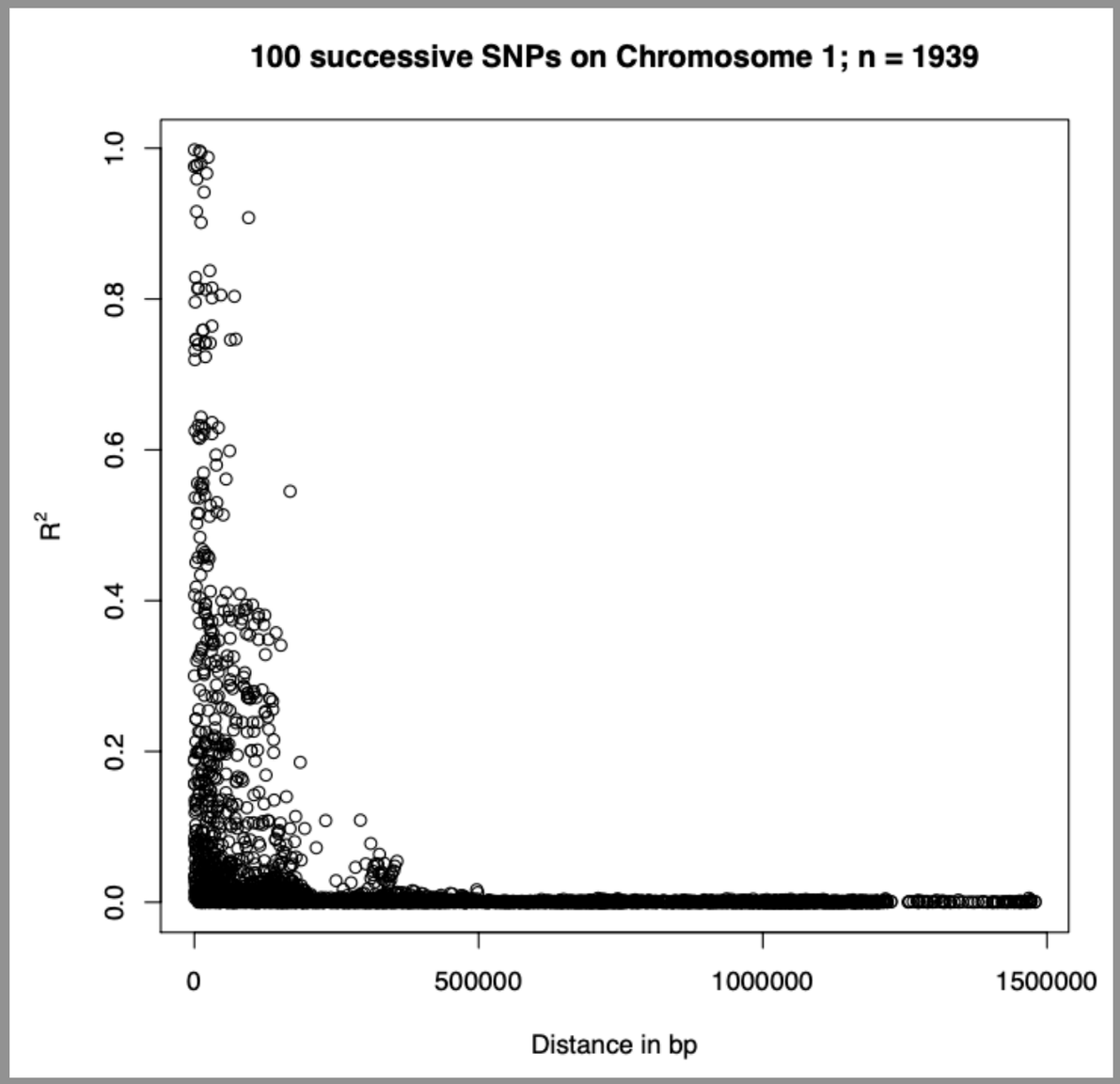
- Data from the FAMuSS (Functional SNPs Associated with Muscle Size and Strength) study (Foulkes, 2009)
- n = 1397 individuals and 225 SNPs Muscle performance variables
- Compute pairwise LD and make a heat map



Linkage disequilibrium

Example for LD - physical distance

- Visualizing LD and physical distance within the chromosome



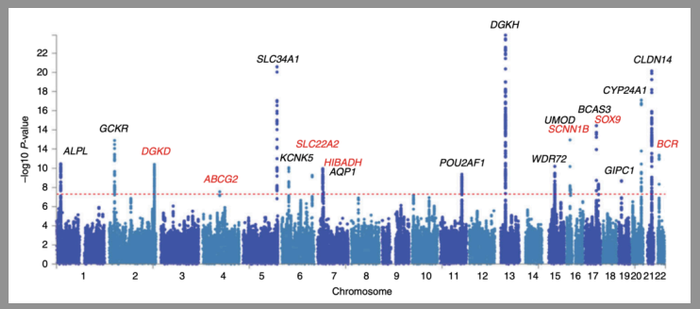
Content

Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium

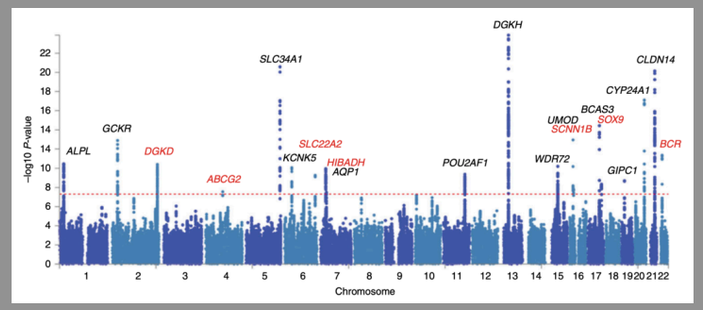
LD pruning



- It is often convenient to have independent genetic variants (recall the Manhattan plot above).
- Genetic variants that are physically close on a chromosome typically have high correlations.
- A subset of variants can be selected that is, at least approximately, independent.
- In practice… The process is known as **LD pruning**.
  - Define a window of fixed size (in kb or as a variant number).
  - Calculate R2 statistic for each pair of variants in the window.
  - Remove variants from the window until al remaining pairs of variants have R2 < t, where t is some threshold.
  - Shift the window along the chromosome, allowing for some overlap.
- Easy to do in the PLINK software.

Linkage disequilibrium

LD pruning



- It is often convenient to have independent genetic variants (recall the Manhattan plot above).
- Genetic variants that are physically close on a chromosome typically have high correlations.
- A subset of variants can be selected
- In practice… The process is known a
  - Define a window of fixed size (in
  - Calculate R2 statistic for each pa
  - Remove variants from the windo where t is some threshold.
  - Shift the window along the chrom
- Easy to do in the **PLINK** software.

A widely used program in statistical genetics is the command-line based program PLINK

- Available at https://www.cog-genomics.org/plink2/
- maintained by Shaun Purcell and Christopher Chang
- Offers many options for
  - Data manipulation, format conversion.
  - Allele frequencies, missing data,
  - Hardy-Weinberg equilibrium.
  - LD calculations, LD pruning.
  - Population substructure.
  - Kinship calculations.
  - Association analysis.
  - …

Content

Linkage disequilibrium and haplotype estimation

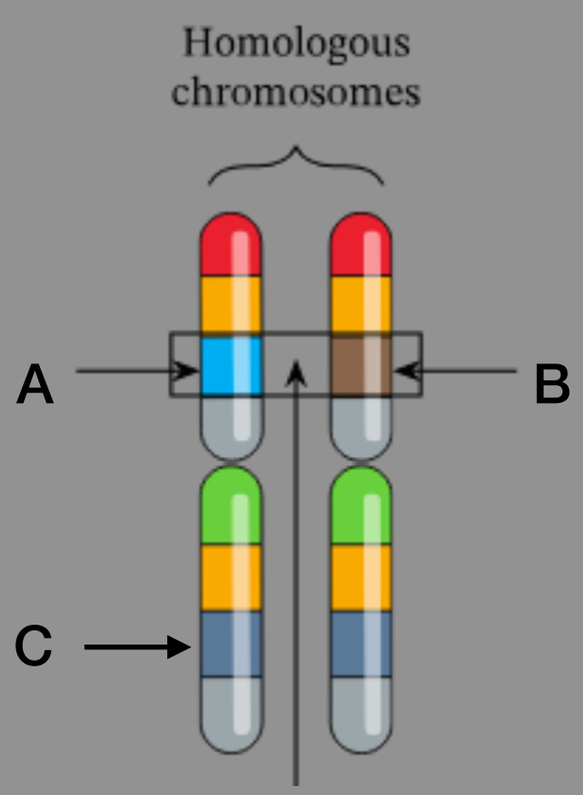
1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Haplotype estimation

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium ~~is~~ fundamental in population genetics: ~~Linkage dise~~quilibrium

Haplotype estimation



Locus of interest

All concepts refer to association between alleles but….

HWE refers to an association between alleles at the same locus

LD refers to associations between alleles at different loci

AND…

Haplotype refers to a set of alleles at different locus within one chromosome that are closely linked and that tend to be inherited together

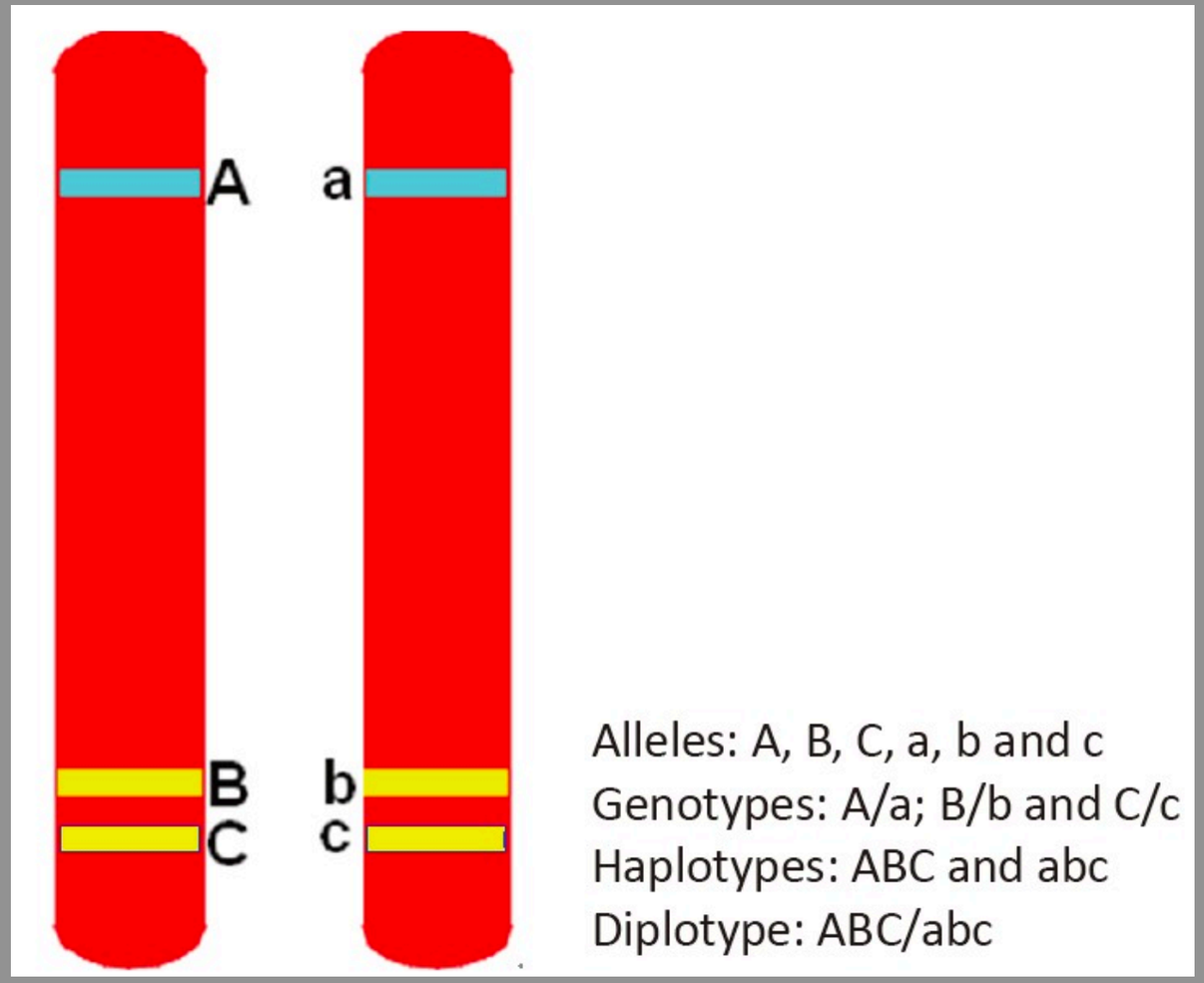
Haplotype estimation

Introduction - definition

RECALL:

- An **haplotype** is a combination of alleles at different chromosomal regions that are closely linked and that tend to be inherited together.

EXAMPLE:



- Haplotypes are clusters of genetic material that came exclusively from one parent.
- In practice, a haplotype often refers to a set of SNPs on a single chromosome that are statistically associated.
- The haplotype constitution of an individual is called its **diplotype**.

Lingjun, Z.U.O., Kesheng, W.A.N.G. and Xingguang, L.U.O., 2014. Use of diplotypes–matched haplotype pairs from homologous chromosomes–in gene-disease association studies. Shanghai archives of psychiatry, 26(3), p.165.

Haplotype estimation

Introduction

These SNPs may have statistical association (ie. LD) but we still don't have information about its chromosomal location!

BUT:

- With present-day technology it is hard to economically determine the phase of genotypes, i.e, to allocate SNP alleles to individual chromosomes.

EXAMPLE:

SNP database

Id	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	···
NA18524	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18526	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18529	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18532	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18537	CC	CC	TT	TT	AT	CC	CC	AC	CT	GG	···
NA18540	CC	CC	CT	TT	AT	CC	CG	AC	CT	GG	···
NA18542	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18545	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18547	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18550	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18552	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18555	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18558	CC	NA	CC	TT	TT	CC	CG	CC	CT	GG	···
NA18561	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18562	CC	CC	TT	TT	AT	AC	CG	AC	CT	GG	···
NA18563	CC	CC	CT	TT	AA	CC	CC	AA	CT	GG	···
NA18564	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18566	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18570	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18571	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

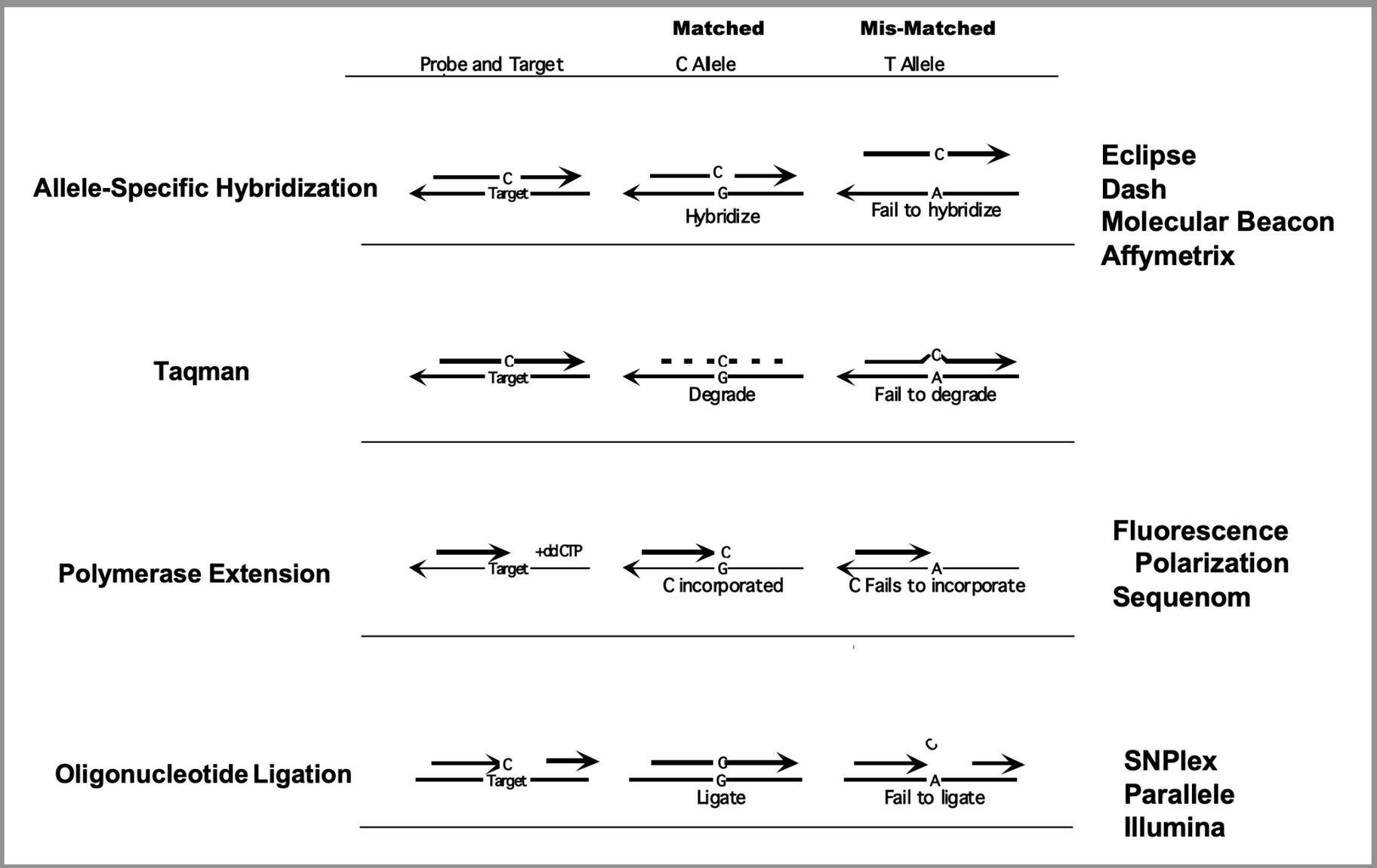
Haplotype estimation

Introduction

- Direct DNA sequencing is used to initially characterize a mutation, but is too laborious for routine screening
- Many different genotyping approaches are available, use different chemistries (image below)

BUT:

- With **present-day technology** it is hard to economically determine the phase of genotypes, i.e, to allocate SNP alleles to individual chromosomes.



Overview of SNP Genotyping
Debbie Nickerson

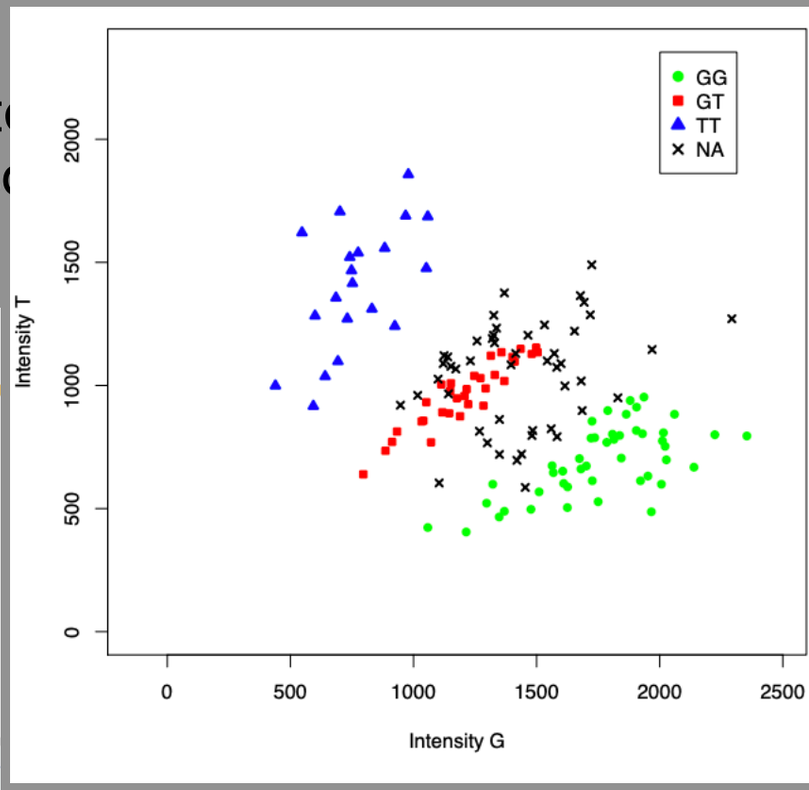
Haplotype estimation

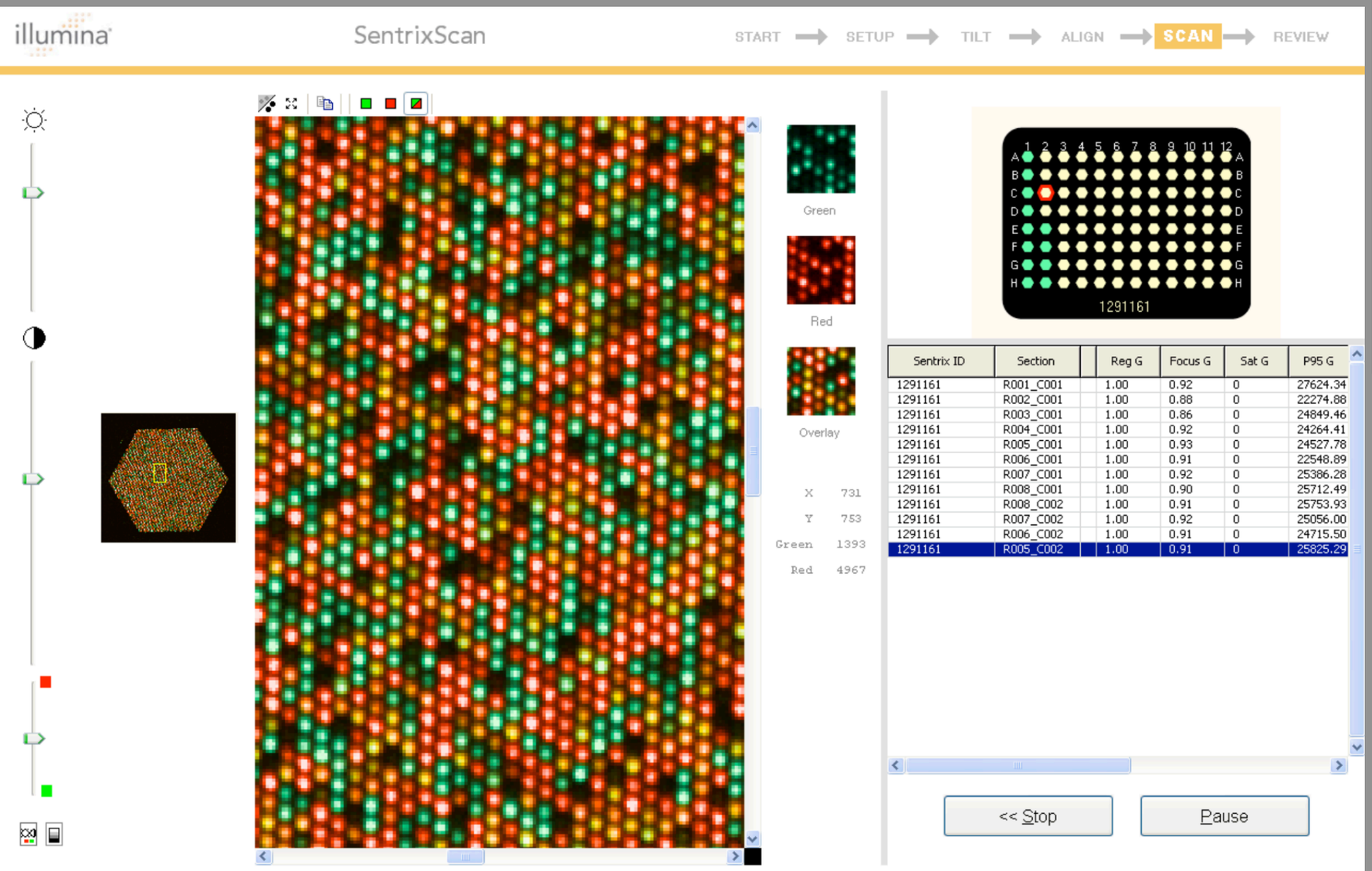
Introduction

- Direct DNA sequencing is used to initially characterize a mutation, but is too laborious for routine screening
- Many different genotyping approaches are available, use different chemistries (image below)

BUT:

- With **present-day technology** it is hard to economically dete
  genotypes, i.e, to allocate SNP alleles to individual chromos





- The underlying principle is that the signal intensity depends upon the amount of target DNA in the sample, as well as the affinity between target and probe.
- Extensive processing and analysis of these raw intensity measures yield SNP genotype inferences

Haplotype estimation

Introduction - definition

RECALL:

- It was an initiative to develop a map of common haplotypes in the human genome, known as a HapMap.
- The process implied identifying SNP loci, genotype them in individuals of various ancestries, and uncover their correlation structure in the genome.
- The HapMap project chose a sample of 269 individuals and selected several million well-defined SNPs, genotyped the individuals for these SNPs, and published the results.
- This methodology has largely been superseded by genome-wide association studies.

- The **human genome project** was launched in 1990…with a cost of $3 billion.
  - In 2003, the project announces the finished version of human genome sequence, with the 92% of sampling (exceeding 99.99% accuracy)

- Continuations:
  - 1000 Genomes Project (2008-2015) to establish the most detailed catalogue of human genetic variation.
  - International HapMap Project (2002-2009) to develop a haplotype map
  - China Kadoorie Biobank (2004-2020) largest prospective cohort study
  - …

LaFramboise, T., 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic acids research, 37(13), pp.4181-4193.

Haplotype estimation

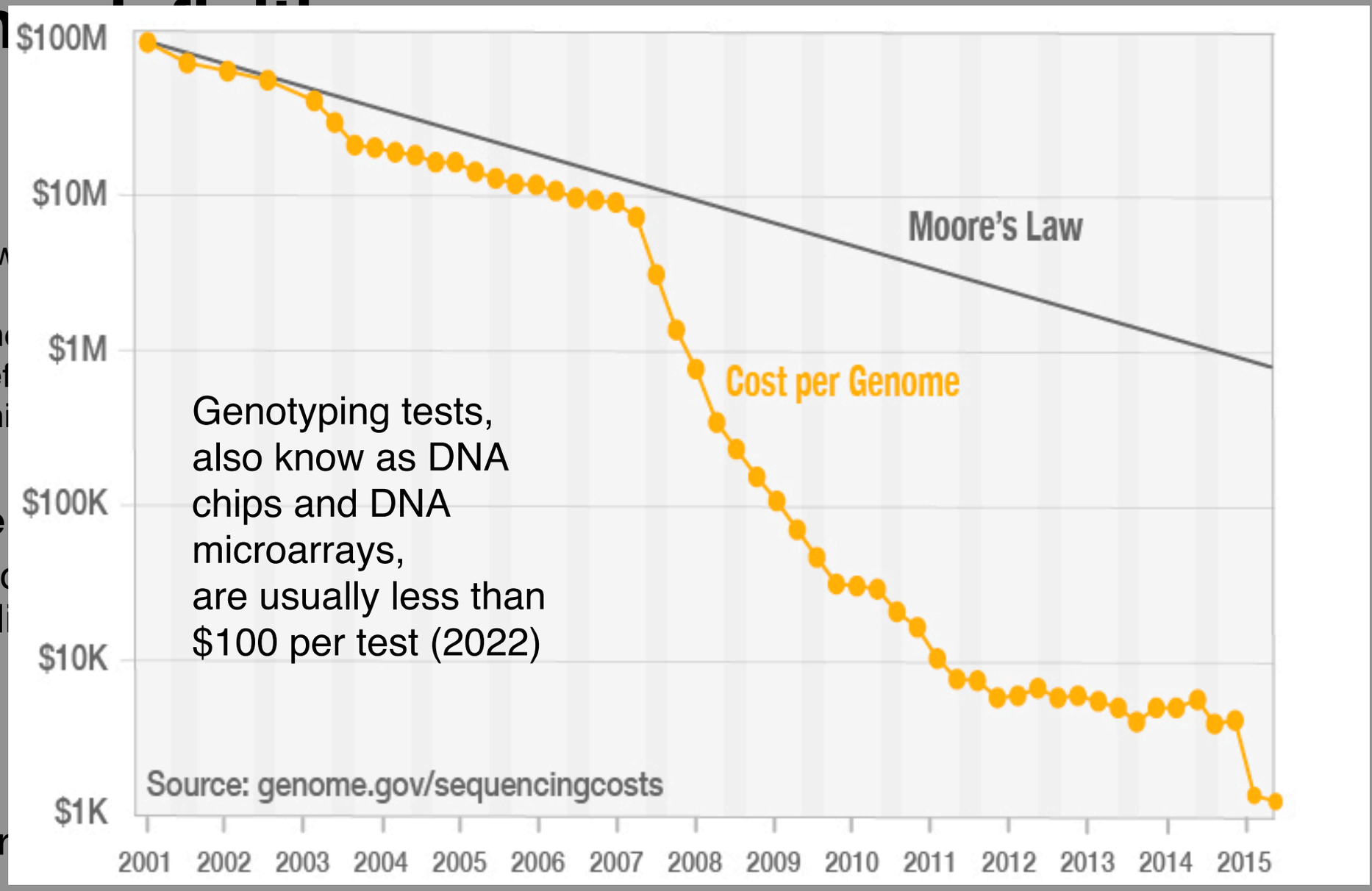
Introduction

RECALL:

- It w
  as
- The
  def
- Thi

- The **human genome**
  - In 2003, the projec
    the 92% of sampli

- Continuations:
  - 1000 Genomes Pr
    genetic variation.
  - International HapMap Project (2002-2009) to develop a haplotype map
  - China Kadoorie Biobank (2004-2020) largest prospective cohort study
  - …



Haplotype estimation

Introduction

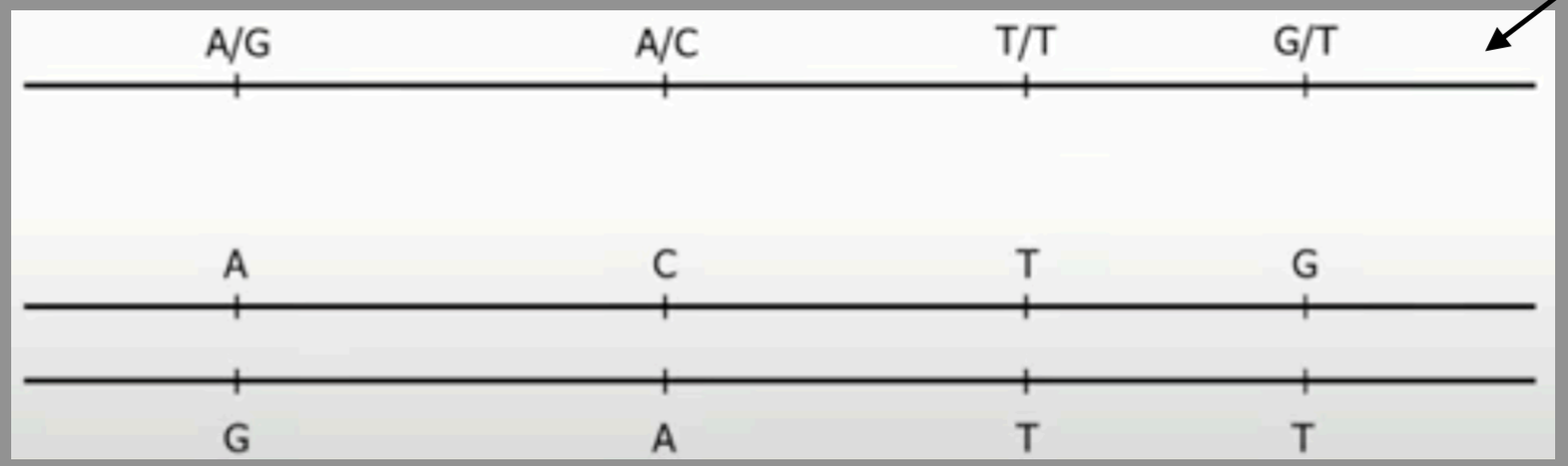
BUT:

- With present-day technology it is hard to economically determine the phase of genotypes, i.e, to allocate SNP alleles to individual chromosomes.

So…what we need (for now) is a set of methods (haplotype phasing methods) that go from genotype data to haplotype data

Example:

What is the actual set of alleles or sequence of alleles that were inherited from each parent?



Pair of alleles that come from data

One of the possibles set of haplotypes

…from one parent

…from the other parent

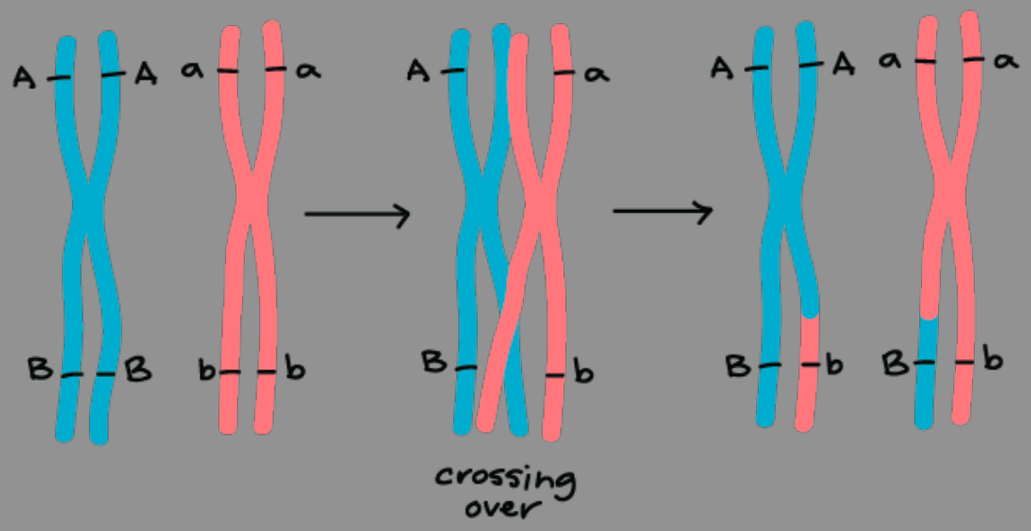
- The alleles of nearby SNPs on a single chromosome are correlated.
- A sequence of SNPs on a single chromosome that are statistically associated is called haplotype.

Haplotype estimation

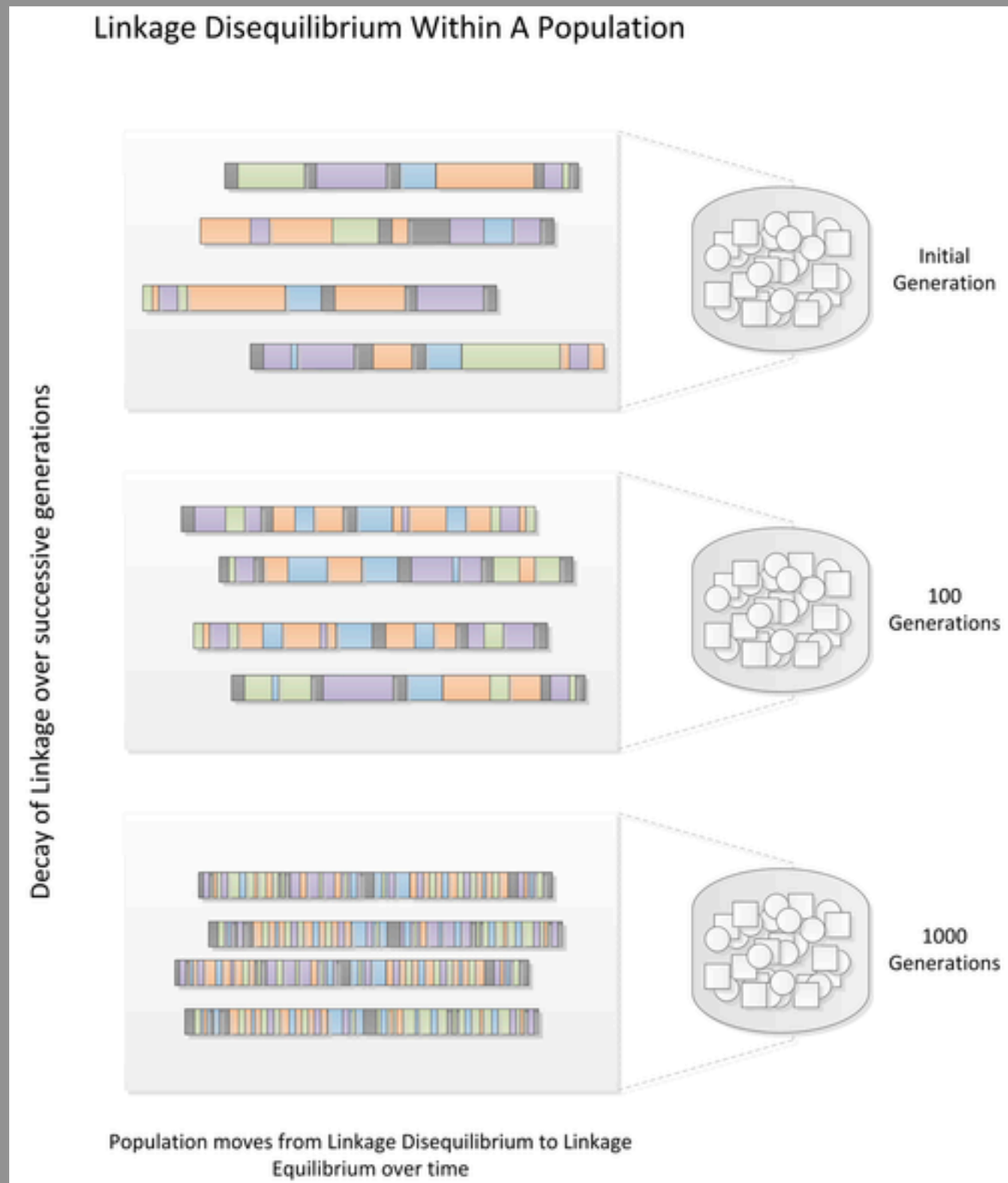
Introduction

Stronger-than-expected linkage can reflect selection for functional relationships between genes, such as combinations of alleles that are advantageous when inherited together, or shared regulatory mechanisms (positive selection)

- Recall genetic recombination during meiosis:



- Through time, over generations, the average length of the haplotype blocks will tend to get smaller
- How far back in time you have to go until you have pure chromosome types?



Haplotype estimation

Introduction

Strategies for deriving haplotypes:

- **Statistical phasing**: uses inter-marker correlation
    1. Start by randomly phasing heterozygotes
    2. Use current haplotypes to create an haplotype probability model
    3. For each individual choose the most probable pair of haplotypes that are consistent with the observed genotypes
    4. Repeat step 2-3 until convergence
- **Mendelian phasing**: uses parental genotypes

Another way of looking at the methodologies (Niu 2004):

- Population inference: Assign haplotype to an individual's genome from a database that contains population's haplotypes.
- Molecular haplotyping: Sequencing, DNA cloning or any other physical manipulation of the sample.
- Genetic analysis: Inferring haplotypes from genotype data from family relations

- Brian Browning (2022) Haplotype phase inference. European Society of Human Genetics
- Niu, T., 2004. Algorithms for inferring haplotypes. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 27(4), pp.334-347.

Haplotype estimation

Introduction - why?

- Haplotype estimation (phasing) refers o the process of inferring haplotypes from genotype data.
  - Consider a genome-wide association studies (GWAS) that collect genotypes of thousands of individuals testing from 200,000 to 5,000,000 SNPs on each individual using microarrays.
  - Consider other methods like whole exome sequencing (WES, 1-2% of the genome) or whole-genome sequencing (WGS, with 90-99-9% accuracy).
  - Genotype data is *unordered*: the specific assignment of alleles on chromosomes (haplotypes) remains unknown. No information on how these markers are transmitted from parents.
- The number of markers used in genotyping studies is often very large. If we combine markers into haplotypes, the number of variables decreases.
  - Haplotypes may be biologically more relevant than single SNPs.
  - Association studies can be carried out between haplotypes and traits and reduce the problem of multiple testing.
  - Association studies with haplotypes tend to have more power than those using single markers.

Haplotype estimation

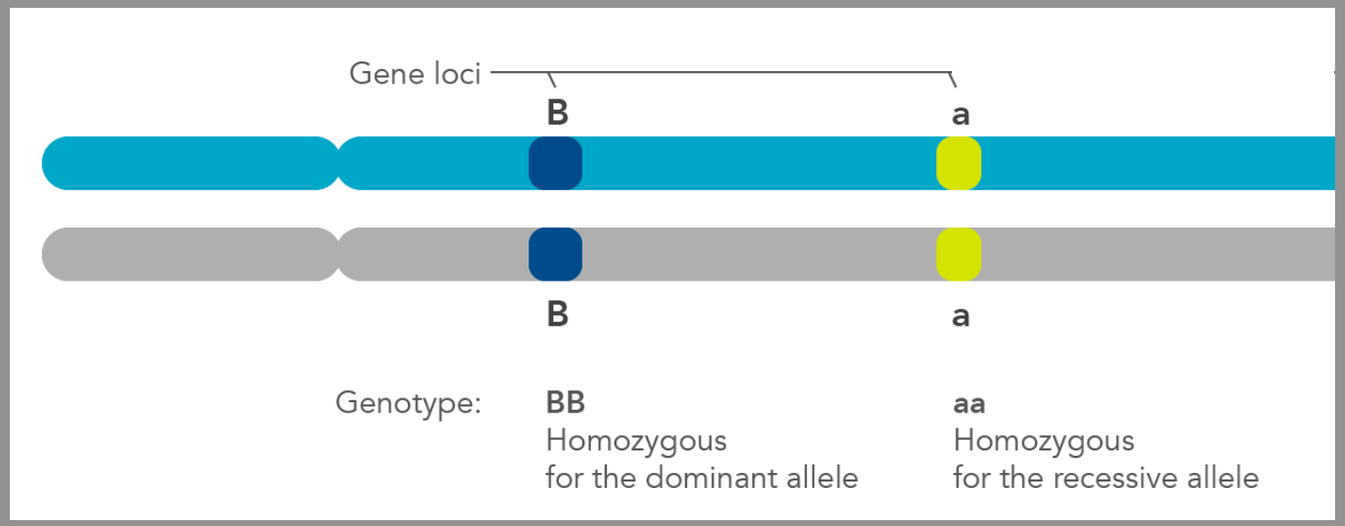
Introduction

- The haplotypic constitution of an individiual is called its **diplotype**.
  - With two loci and two alleles (A, a and B, b), for the double heterozygote AaBb there are two possible diplotypes (AB,ab) and (Ab,aB).
  - With three loci and two alleles (A,a), (B,b) and (C,c), for the triple heterozygote AaBbCc there are four possible diplotypes (ABC,abc), (ABc,abC), (AbC,aBc) and (aBC,Abc).
  - For a multilocus genotype consisting of k heterozygous SNPs there are 2^{k-1} diplotypes.

- Note that for homozygous or single-site heterozygous individuals the haplotypes (and diplotypes) are known.

Haplotype estimation

Introduction



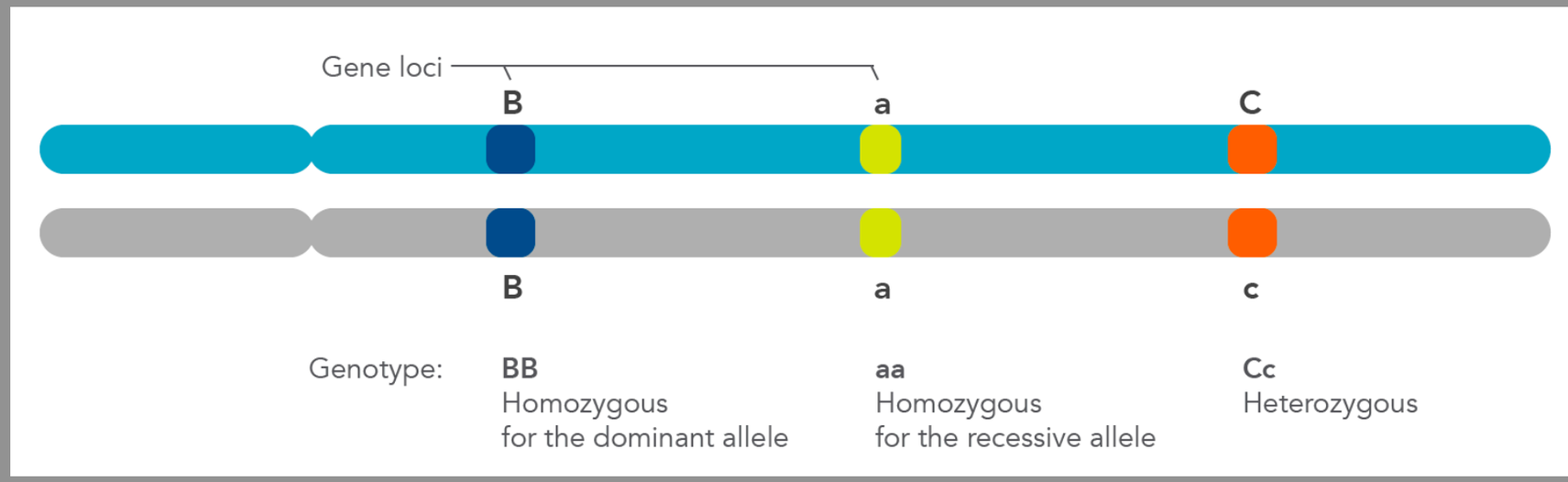
- Consider a two sites (two bi-allelic markers) on the same chromosome. The first locus has alleles A an a, second locus has alleles B and b.
- If an individual's genotype is:
  - AA/BB ⟶ corresponding haplotype: AB and AB
  - aa/BB ⟶ corresponding haplotype: aB and aB (like in the example image)
  - Aa/BB ⟶ corresponding haplotype: AB and aB
  - …
  - Aa/Bb ⟶ corresponding haplotype: (AB and ab) OR (Ab and aB)

The haplotypes for homozygotes and single-site heterozygotes are unambiguous!

- In total, there are four possible haplotypes for the fully heterozygous individual: AB, Ab, aB and ab with frequencies pAB, pAb, paB and pab
- Note that for homozygous or single-site heterozygous individuals the haplotypes (and diplotypes) are known.

Haplotype estimation

Introduction



- Consider a three sites (two bi-allelic markers) on the same chromosome. The first locus has alleles A an a, second locus has alleles B and b, and the third locus is C and c.
- If an individual's genotype is:
  - AA/BB/CC ⟶ corresponding haplotype: ABC and ABC
  - aa/BB/CC ⟶ corresponding haplotype: aBC and aBC
  - aa/BB/Cc ⟶ corresponding haplotype: aBC and aBc (like in the example image)
  - …
  - Aa/Bb/Cc ⟶ corresponding haplotype: (ABC and abc) OR (ABc and abC) OR (Abc and aBC) OR (aBc and AbC)
- In total, there are four possible diplotypes for the fully heterozygous individual. Every time you ad an heterozygous on the sequence of SNP…the number of possible phasing doubles.
- For $k = 300$ heterozygotes (about 500kb of DNA sequence), we have $2^{k-1} = 10^{90}$ diplotypes

Content

Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Haplotype estimation

Methods for haplotype estimation

- Several statistical methods for inferring haplotypic phase have been proposed.
  - **Parsimony methods**
  - Likelihood based methods
  - Bayesian methods
  - ….

- The purpose of haplotype estimation methods can be twofold:
  - Resolving the haplotype constitution of each individual in the database.
  - Estimation of population haplotype frequencies.

NOTE:

In the previous module on LD we learned how to estimate haplotype frequencies by maximum likelihood.

Haplotype estimation methods will ALSO determine frequency of haplotypes…but it goes further, also tells you which are the haplotypes given a particular genetic data

Haplotype estimation

Methods for haplotype estimation

- Several statistical methods for inferring haplotypic phase have been proposed.
  - **Parsimony methods**
  - Likelihood based methods
  - Bayesian methods
  - ….
- The purpose of haplotype estim
  - Resolving the haplotype constitut
  - Estimation of population haplotyp

Haplotype estimation is and has been a topic of intense research, and there are many programs for it:

- PHASE (Stephens et. al., 2001)
- fastPHASE (Scheet & Stephens, 2006)
- BEAGLE (Browning et. al., 2007)
- IMPUTE2 (Howie et. al., 2009)
- SHAPEIT (Delaneau et al., 2011, 2012)
- …

Many of these programs also:

- Estimate missing values
- Can infer intervening SNPs that have not been typed.

NOTE:

In the previous module on LD we learned how to estimate haplotype frequencies by maximum likelihood.

Haplotype estimation methods will ALSO determine frequency of haplotypes…but it goes further, also tells you which are the haplotypes for each individual given a particular genetic data

Haplotype estimation

Methods for haplotype estimation: Parsimony methods

- Objective: to resolve the minimum possible number of haplotypes.
- Steps

    1. Identify all unambiguous haplotypes (homozygotes and single-site heterozygotes)
    2. Check if any of the resolved haplotypes is compatible for an unresolved individual. If not, stop.
    3. Identify the complementary haplotype for this last unresolved individual and add it to the set of resolved haplotypes. Return to (2)
    4. Continue till all haplotypes are resolved, or no new haplotypes can be found.

Problems:

- There may be no unambiguous haplotypes.
- Unresolved haplotypes may remain.
- The solution is order-dependent.

Haplotype estimation

Methods for haplotype estimation: Parsimony methods

Toy example

	A/T	A/G	C/T	type
ID1	AT	AA	CT	(double heterozygote)
ID2	AT	GG	CT	(double heterozygote)
ID3	AA	AG	CC	(single-site heterozygote)
ID4	TT	GG	CT	(single-site heterozygote)
ID5	AA	AA	TT	(only homozygous)

- For ID3, ID4, and ID5 the haplotypes are unambiguous, and the resolved diplotypes are: (AAC,AGC), (TGC,TGT), (AAT,AAT)
- For ID1, AAC is compatible with the resolved haplotypes. Its complementary haplotype is TAT. Its diplotype may be inferred as (AAC,TAT). TAT is added to the pool of available haplotypes
- For ID2, AGC and TGT are both compatible with the resolved haplotypes. Its diplotype may be inferred as (AGC,TGT)
- Final haplotype set: (AAC,AGC,TGC,TGT,AAT,TAT)

Notes:

- Resolving ID2 as (AGT,TGC) would require creating a new haplotype
- If ID1 would be resolved as (AAT,TAC), the final haplotype set would be (AAC,AGC,TGC,TGT,AAT,TAC), which may be considered more plausible

Haplotype estimation

Methods for haplotype estimation

- Several statistical methods for inferring haplotypic phase have been proposed.
  - Parsimony methods
  - **Likelihood based methods**
  - Bayesian methods
  - ….

- The purpose of haplotype estimation methods can be twofold:
  - Resolving the haplotype constitution of each individual in the database.
  - Estimation of population haplotype frequencies.

NOTE:

In the previous module on LD we learned how to estimate haplotype frequencies by maximum likelihood.

Haplotype estimation methods will ALSO determine frequency of haplotypes…but it goes further, also tells you which are the haplotypes given a particular genetic data

Haplotype estimation

Methods for haplotype estimation: Likelihood-based methods

- Some of the earliest approaches:
  - A simple multinomial model is defined, in which each possible haplotype consistent with the sample was given an unknown frequency parameter and these parameters were estimated with an EM algorithm.
  - Sequential approaches were also developed (i.e. SNPHAP method by Clayton 2001)
  - Allows for the fact that many different haplotypes can be constructed from individuals genotype
  - Can handle missing data

- Clayton, D., 2001. Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. Nature Genetics, 29(2).
- Fallin, D. and Schork, N.J., 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. The American Journal of Human Genetics, 67(4), pp.947-959.

Haplotype estimation

Methods for haplotype estimation: Likelihood-based methods

- We have n individuals genotyped at m SNPs, yielding genotype data G
- There are 2^m possible haplotypes with frequencies $h = (h_1, \ldots, h_{2^m})$ that we want to estimate
- We wish to maximize likelihood function over the parameter space:

$$L(h) = f(G, h) = \prod_{i=1}^{2^m} f(G_i \mid h)$$ where f is a density function (i.e. normal distribution)

- Estimation of the possible haplotypes h can be done by the EM algorithm:
  - Initial set of haplotype frequencies h^0 defined
  - Haplotype frequencies h^t at iteration t are updated from frequencies at iteration $t-1$ using expectation and maximization steps until h^t converges.

Expectation Maximization algorithm: https://www.youtube.com/watch?v=AnbiNaVp3eQ

Haplotype estimation

Methods for haplotype estimation: Likelihood-based methods

Toy example in R

```
>  library(haplo.stats)
>  snp1 <- c("AT","AT","AA","TT","AA")
>  snp2 <- c("AA","GG","AG","GG","AA")
>  snp3 <- c("CT","CT","CC","CT","TT")
>  Geno <- cbind(substr(snp1,1,1),substr(snp1,2,2),
                 substr(snp2,1,1),substr(snp2,2,2),
                 substr(snp3,1,1),substr(snp3,2,2))

>  snpnames <- c("snp1","snp2","snp3")
>  HaploEM <- haplo.em(Geno,locus.label=snpnames)
>  HaploEM
=============================================================================
                    Haplotypes
=============================================================================
  snp1 snp2 snp3 hap.freq
1    A    A    C      0.1
2    A    A    T      0.3
3    A    G    C      0.2
4    A    G    T      0.0
5    T    A    C      0.1
6    T    A    T      0.0
7    T    G    C      0.1
8    T    G    T      0.2
=============================================================================
Details
=============================================================================
lnlike = -14.18484
lr stat for no LD =  4.49868 , df =  2 , p-val =  0.10547
```

Content

Linkage disequilibrium and haplotype estimation

1. Introduction to LD
2. Measures of LD
3. Statistical test for LD
4. Example for LD
5. LD pruning
6. Haplotype estimation
7. Methods for haplotype estimation
8. Computer exercise

Linkage disequilibrium and Haplotype estimation

References

- Weir, B.S. (1996) Genetic Data Analysis II, Chapter 3, Sinauer Associates, Massachusetts.

- Foulkes, A.S. (2009) Applied statistical genetics with R. Springer.

- Neale, B.M. (2008) Statistical genetics: gene mapping through linkage and association. Chapter 17. Taylor & Francis.

- Gonick, L. and Wheelis, M., 1991. The cartoon guide to genetics.

- Niu, T., 2004. Algorithms for inferring haplotypes. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 27(4), pp.334-347.