BIO306: Bioinformatics

Lecture 4

Haplotype and Linkage disequilibrium

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Linkage disequilibrium have significant implication for gene mapping

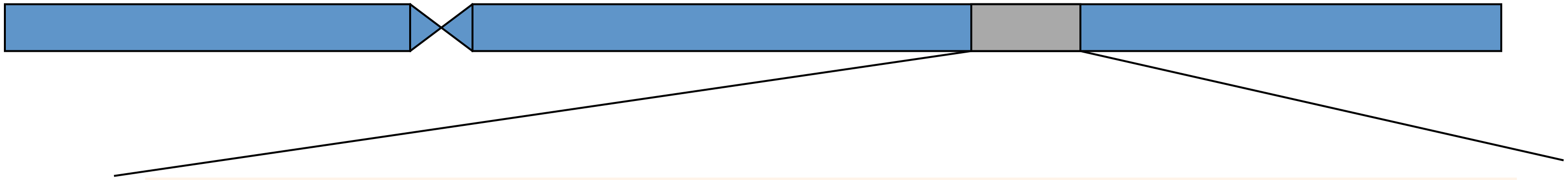
Single Nucleotide Polymorphisms

- Main form of variation between individual genomes: **single nucleotide polymorphisms (SNPs)**

… ataggtcc**C**tatttcgcgc**C**gtatacacggg**A**ctata …
… ataggtcc**G**tatttcgcgc**T**gtatacacggg**T**ctata …
… ataggtcc**C**tatttcgcgc**C**gtatacacggg**T**ctata …
… ataggtcc**C**tatttcgcgc**T**gtatacacggg**T**ctata …

- High density in the human genome: ≈ $1\text{x}10^7$ out of 3×10^9 base pairs
- Vast majority bi-allelic ➔ 0/1 encoding

Haplotypes



… ataggtcc**C**tatttcgcgc**C**gtatacacggg**A**ctata …
… ataggtcc**G**tatttcgcgc**T**gtatacacggg**T**ctata …

… ataggtcc**C**tatttcgcgc**C**gtatacacggg**T**ctata …
… ataggtcc**C**tatttcgcgc**T**gtatacacggg**T**ctata …

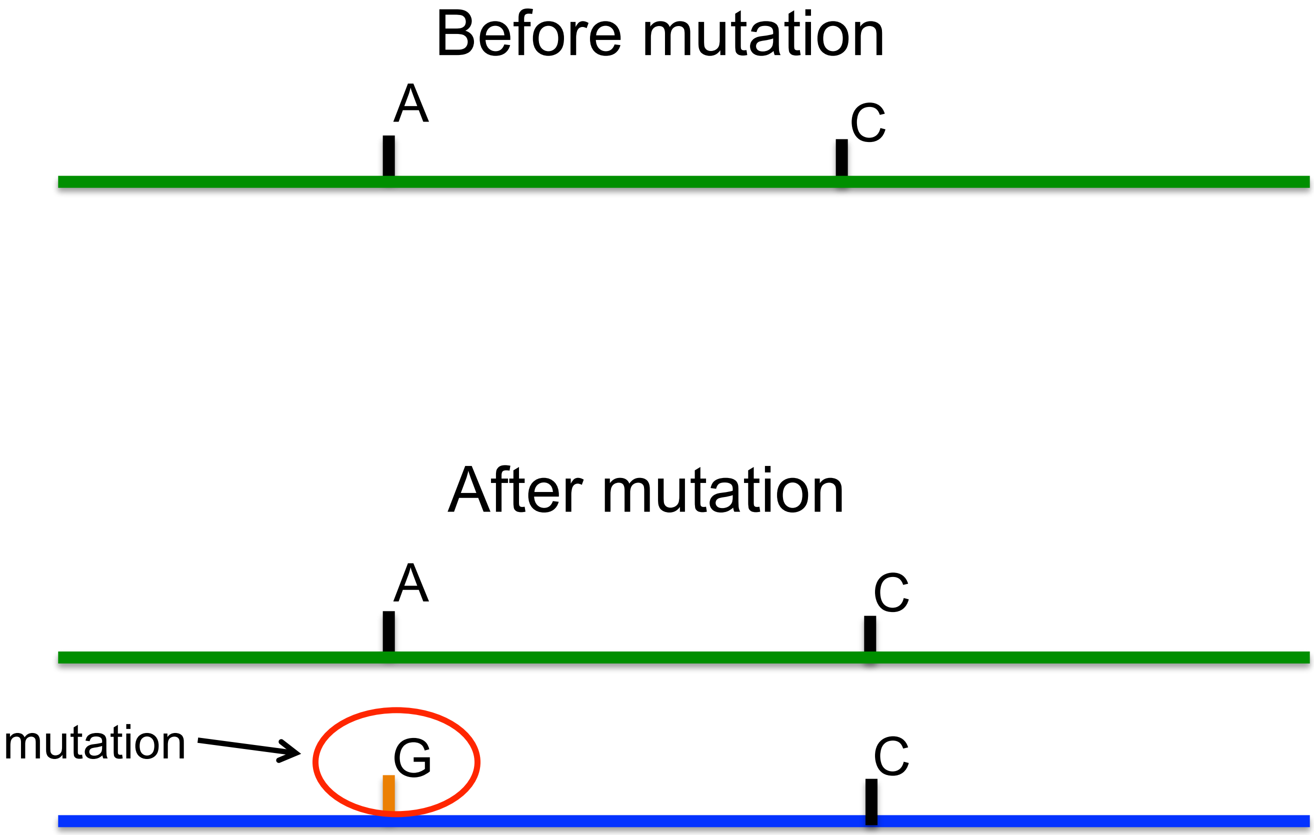
Haplotype: The combination of alleles occurring on the same chromosome

N SNPs - How many Haplotypes are possible ?

2^N (ie very large diversity possible)

Let’s consider the history of two neighboring alleles

Alleles that exist today arose through ancient mutation events...



Before mutation

A

C

After mutation

A

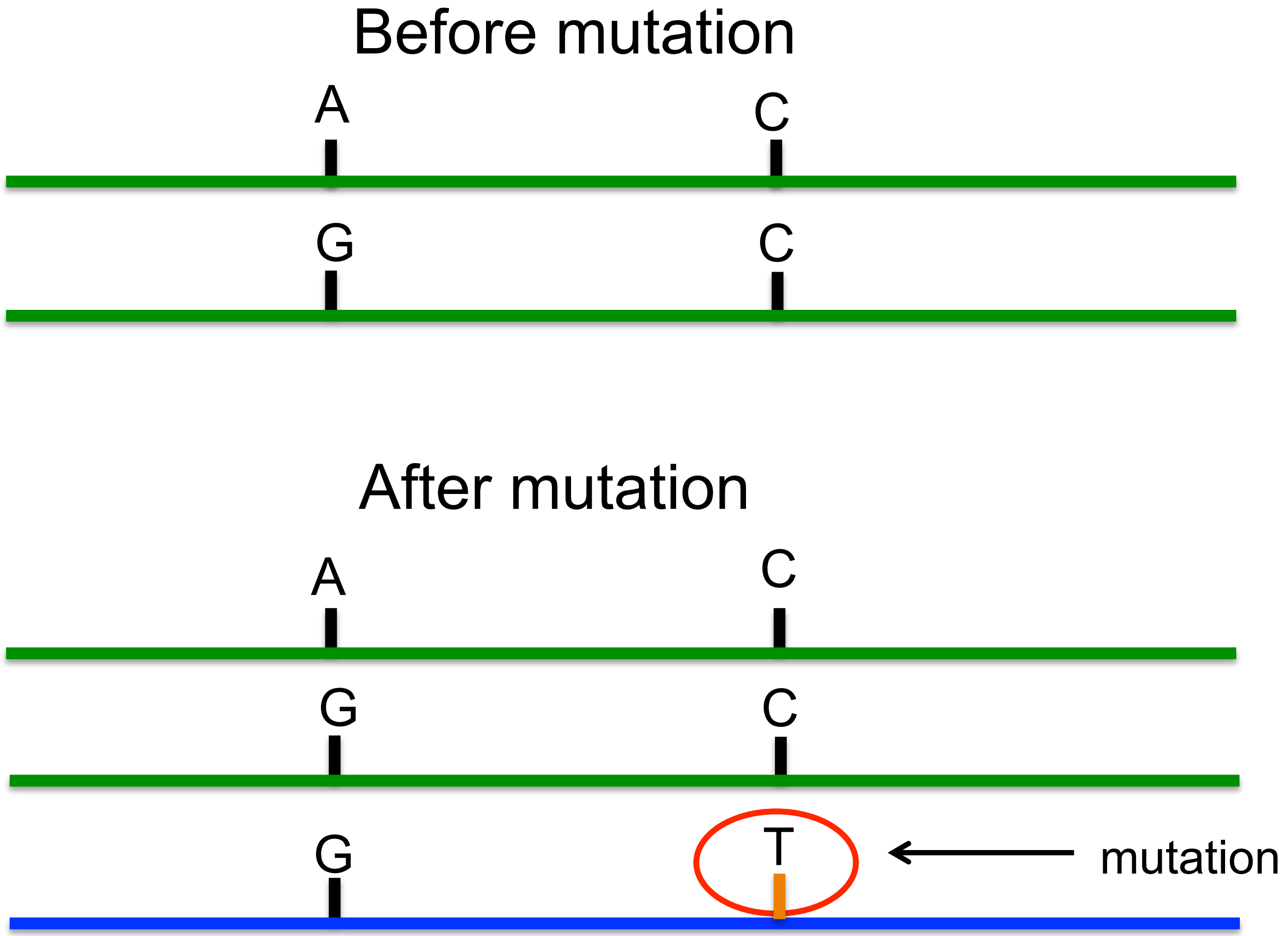
C

mutation

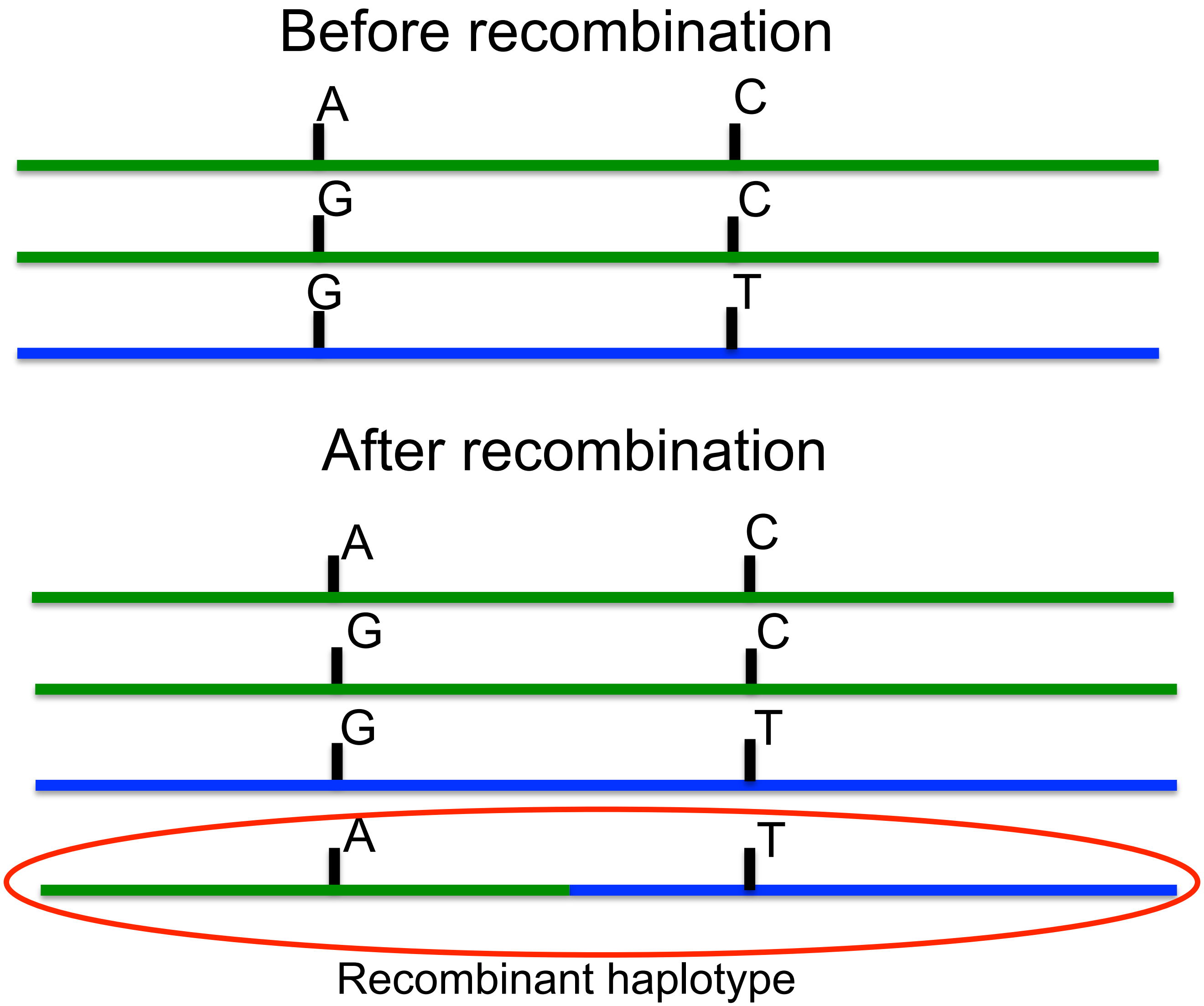
G

C

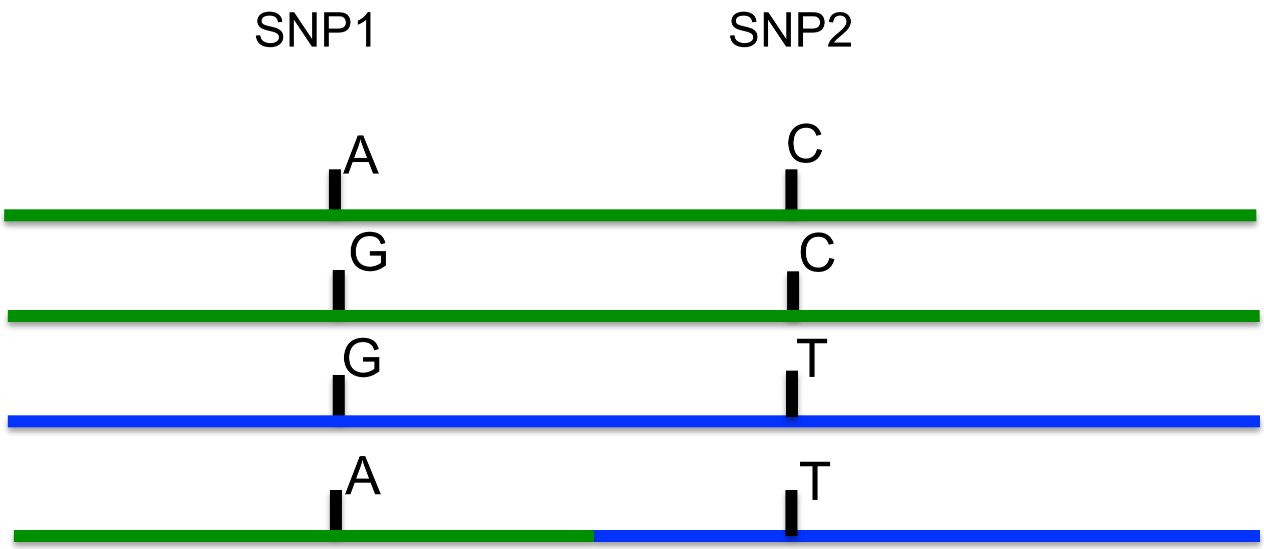
One allele arose first, and then the other...



Recombination generates new arrangements for ancestral alleles



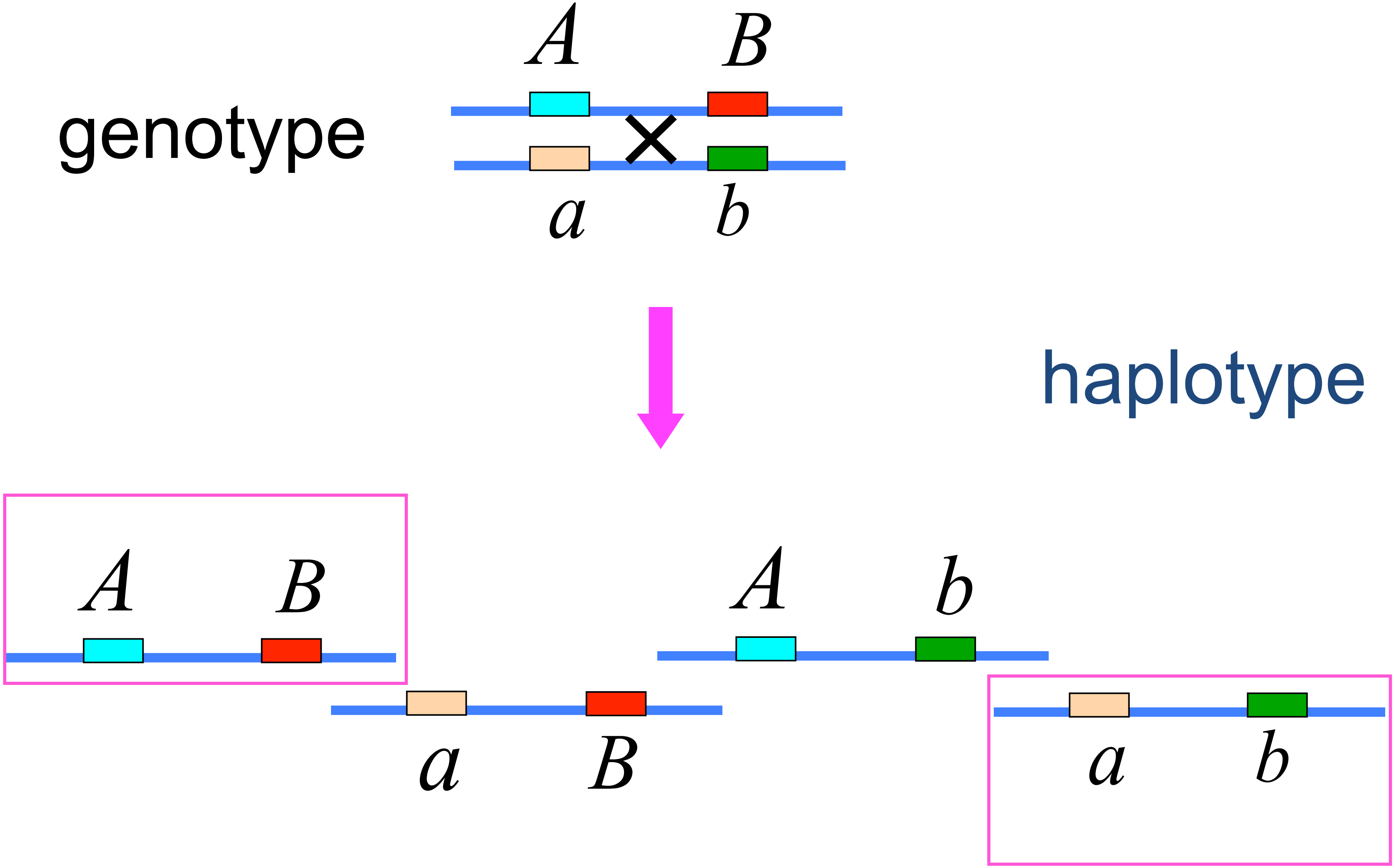
Mutations and recombination generated the haplotyes



4 haplotypes

AC
GC
GT
AT

genotype



haplotype

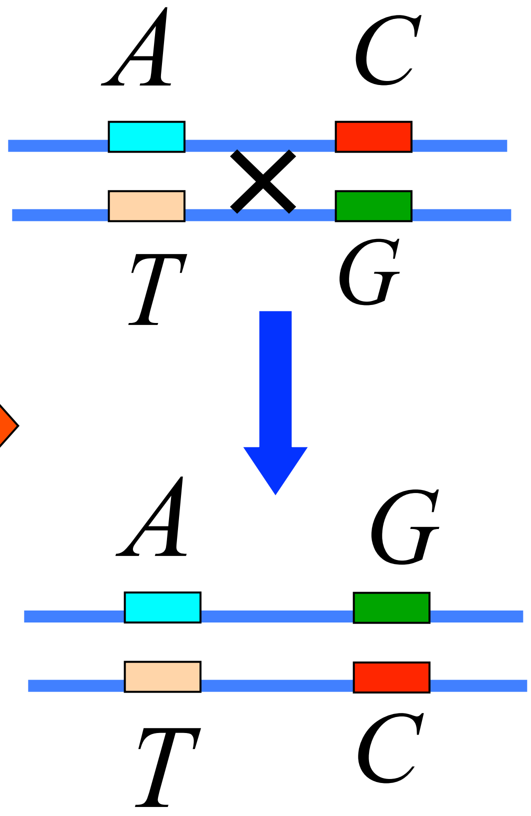
From genotype to haplotype

genotype

sample	SNP1	SNP2
1	**AT**	**CG**
2	AT	CC
3	TT	CG
4	AT	CC
5	AA	CG
6	AT	GG

unphased data

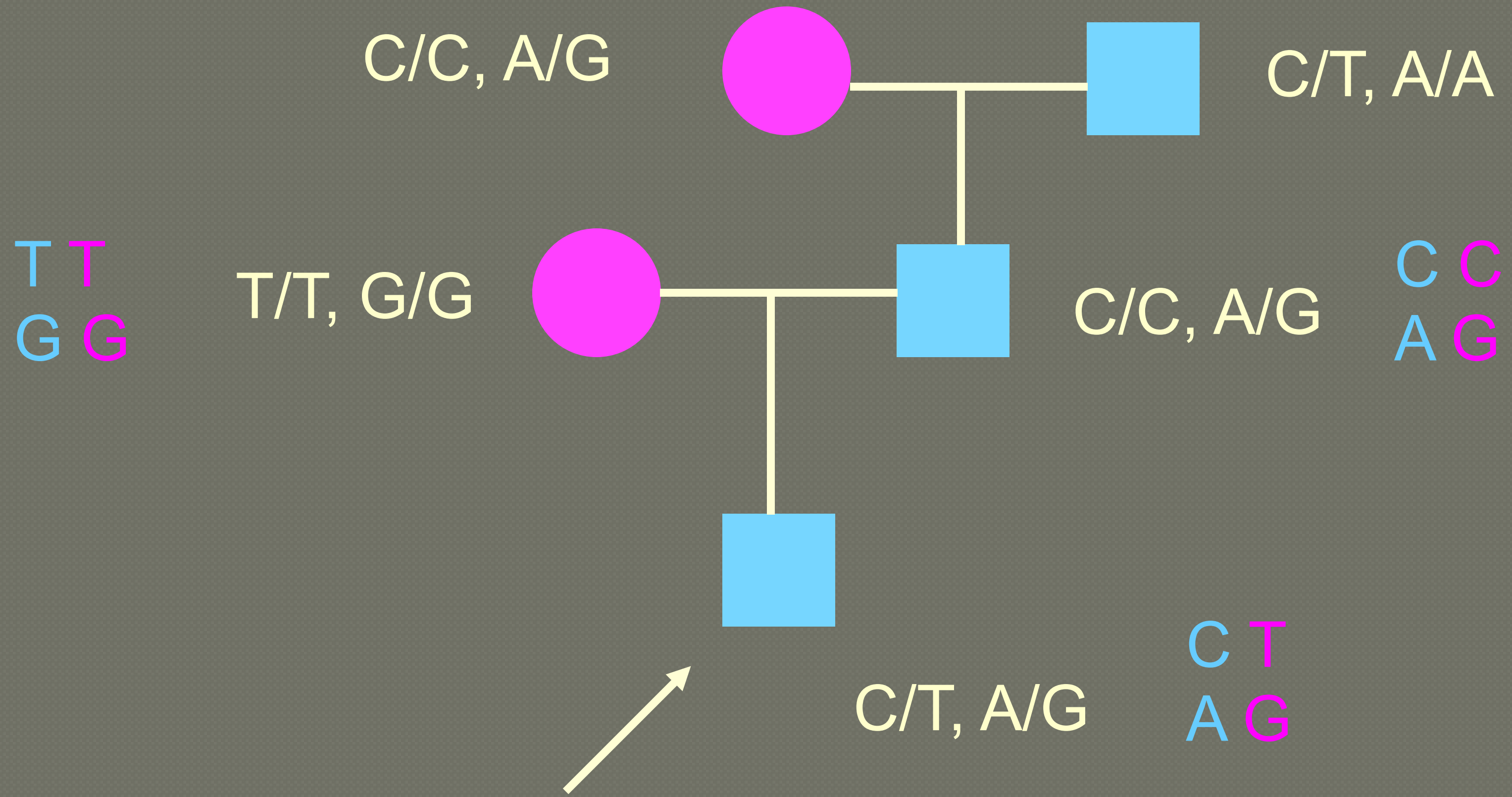
haplotype



phased data

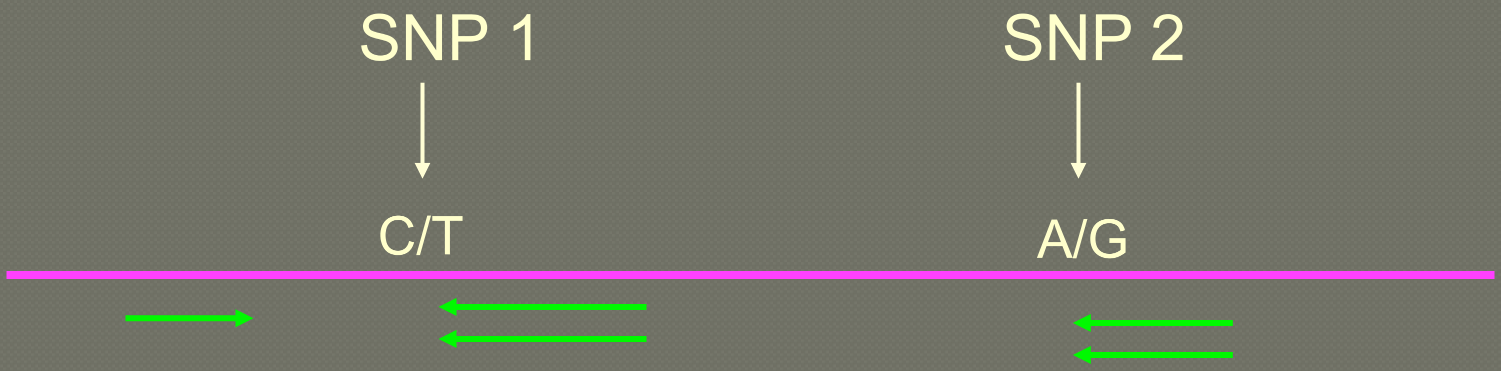
How Do You Construct Haplotypes?

1. Collect extended family members



How Do You Construct Haplotypes?

2. Allele-specific PCR



Reconstruct haplotype from genotype

- CLARK’S algorithm
  - Parsimony-based method
- E-M algorithm
  - Likelihood-based method
- PHASE algorithm
  - Bayesian method

Haplotype reconstruction:

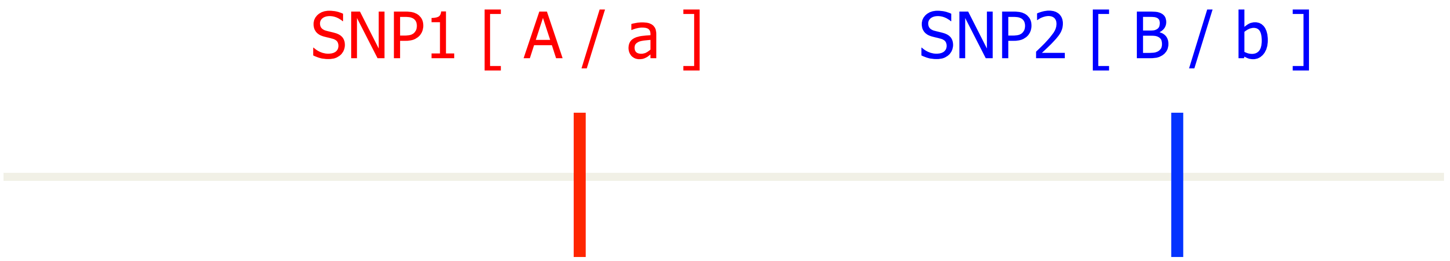
Clark's algorithm (1990)

- Choose individuals that are homozygous at every locus (e.g. TT//AA//CC)
  - Haplotype: TAC
- Choose individuals that are heterozygous at just one locus (e.g. TT//AA//CG)
  - Haplotypes: TAC or TAG
- Tally the resulting known haplotypes.
- For each known haplotype, look at all remaining unresolved cases: is there a combination to make this haplotype?
  - Known haplotype: TAC
    - Unresolved pattern: AT//AA//CG
    - Inferred haplotype: TAC/AAG. Add to list.
  - Known haplotype: TAC and TAG
    - Unresolved pattern: AT//AA//CG
    - Inferred haplotypes: TAC and TAG. Add both to list.
- Continue until all haplotypes have been recovered or no new haplotypes can be found this way.

PHASE

coalescence-based Bayesian Haplotype inference: Stephens et al (2001)

- Bayesian model to approximate the posterior distribution of haplotype configurations for each phase-unknown genotype.
- G = (G_1 , ... , G_n) observed multilocus genotype frequencies
- H = (H_1, ..., H_n) corresponding unknown haplotype pairs
- F = (F_1, ..., F_M) M unkown population haplotype frequencies
- EM algorithm: Find F that maximizes P(G|F). Choose H that maximizes P(H|F^{EM}, G).



Major Allele Freq:	p(A)	p(B)
Minor Allele Freq:	p(a)	p(b)

Independently Segregating SNPs:

Haplotype Frequency $p(ab) = p(a) \times p(b)$

Linkage Equilibrium

(How many haplotypes in total ?)

Linkage Disequilibrium

Haplotype Frequency $p(ab) \neq p(a) \times p(b)$

Linkage Equilibrium

- $p(AB)=p(A)p(B)$

- $p(Ab)=p(A)p(b)=p(A)(1-p(B))$

- $p(aB)=p(a)p(B)=(1-p(A))p(B)$

- $p(ab)=p(a)p(b)=(1-p(A))(1-p(B))$

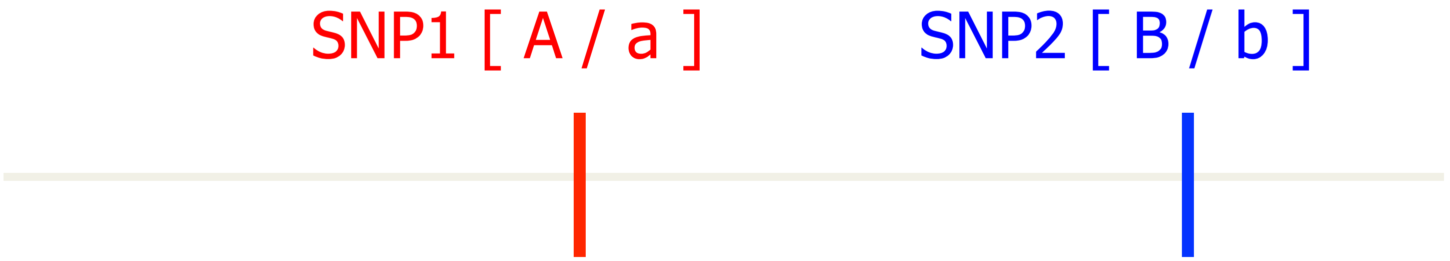
* LINKAGE EQUILIBRIUM *

Not a Punnett Square!

SNP1 Allele	SNP2 Allele		
	B	b	
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)	

Example:

$$p(A)p(B)+p(a)p(B)=p(B)\{\ p(A)+p(a)\}$$
$$=\ p(B)$$



SNP1 [A / a]

SNP2 [B / b]

Major Allele Freq:	p(A)	p(B)
Minor Allele Freq:	p(a)	p(b)

Linkage Disequilibrium

Haplotype Frequency p(ab) = p(a) p(b) + D

D=p(ab)-p(a)p(b)

(sign of D is generally arbitrary, unless comparing D values between populations or studies)

D: Lewontin's LD Parameter (Lewontin 1960)

Linkage Disequilibrium

- $D=p(AB)-p(A)p(B)$
- $p(AB) =p(A)p(B)+D$
- $p(Ab)=p(A)p(b)-D$
- $p(aB)=p(a)p(B)-D$
- $p(ab)=p(a)p(b)+D$

* LINKAGE DISEQUILIBRIUM *

SNP1 Allele	SNP2 Allele		
	B	b	
A	$p(A)p(B)+D$	$p(A)p(b)-D$	$p(A)$
a	$p(a)p(B)-D$	$p(a)p(b)+D$	$p(a)$
	$p(B)$	$p(b)$	

↓

$$p(A)p(B)+D \; + \; p(a)p(B)-D \; = p(B)$$

$$\{\, p(A)+p(a)\} \; = \; p(B)$$

	b	B	
a	0.16	0.04	p(a)=0.20
A	0.14	0.66	P(A)=0.80
	p(b)=0.30	p(B)=0.70	

What is the LD ?

$\neq 0$

$p(ab) \neq p(a)\,p(b)$

$p(ab) = p(a)\,p(b) + D$

0.16 = 0.2 x 0.3 + D

D = 0.1

Since $p(ab) = p(a)p(b) + D$

+D was used and D is +ve here, but arbitrary

eg can relabel alleles A,B as minor

Range of D values (-ve to +ve)

D has a minimum and maximum value that depends on the allele frequencies of the markers

Since haplotype frequencies cannot be -ve

$p(aB) = p(a)p(B) - D \geq 0$ $\qquad D \leq p(a)p(B)$

$p(Ab) = p(A)p(b) - D \geq 0$ $\qquad D \leq p(A)p(b)$

These cannot both be true, so $D \leq \min(p(a)p(B), p(A)p(b))$

$p(ab) = p(a)p(b) + D \geq 0$ $\qquad D \geq -p(a)p(b)$

$p(AB) = p(A)p(B) + D \geq 0$ $\qquad D \geq -p(A)p(B)$

These cannot both be true, so $D \geq \max(-p(a)p(b), -p(A)p(B))$

* Similar equations if we had defined $p(ab) = p(a)p(b) - D$

D is hard to interpret

- Sign is arbitrary ...
  - A common convention is to set A, B to be the common allele and a, b to be the rare allele
- Range depends on allele frequencies
  - Hard to compare between markers

D’ – A scaled version of D

(Lewontin, 1964)

Standardize D by rescaling to a proportion of its maximal value for the given allele frequencies (D')

$$D' = D/D_{max}$$

D’

D’ = D / D_{max}

D_{max} = max (-p(A)p(B), -p(a)p(b))   D < 0

D_{max} = min (p(A)p(b), p(a)p(B))   D > 0

Again, sign of D’ depends on definition

D’ = 1 or -1 if one of p(AB), p(Ab), p(aB), p(ab) = 0

= *Complete LD* (ie only 3 haplotypes seen)

D’=1 or -1 suggests that no recombination has taken place between markers

Beware rare markers - may not have enough power/sample size to detect 4th haplotype

D’ Interpretation

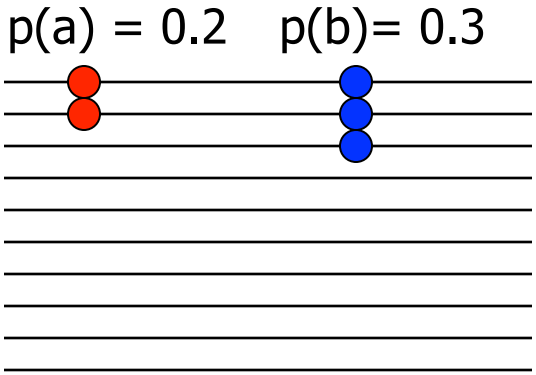
	b	B	
a	0.06	0.14	p(a)=0.20
A	0.24	0.56	p(A)=0.80
	p(b)=0.30	p(B)=0.70	

	b	B	
a	0.2	0	p(a)=0.20
A	0.1	0.7	P(A)=0.80
	p(b)=0.30	p(B)=0.70	

D=0 ; D_{max} undefined

D=D_{max} =0.14 ; D’ = +1

D’=1 (perfect LD using D’ measure
- No recombination between marker
- Only 3 haplotypes are seen



More on D’

- Pluses:
  - If allele frequencies are similar, high D’ means the markers are good surrogates for each other
- Minuses:
  - D’ estimates inflated in small samples
  - D’ estimates inflated when one allele is rare

Δ^2 (also called r^2)

$$r^2=D^2 / p(A)(1-p(A))p(B)(1-p(B))$$

- Ranges between 0 and 1
  - – 1 when the two markers provide identical information
  - – 0 when they are in perfect equilibrium
- Expected value is 1/2n

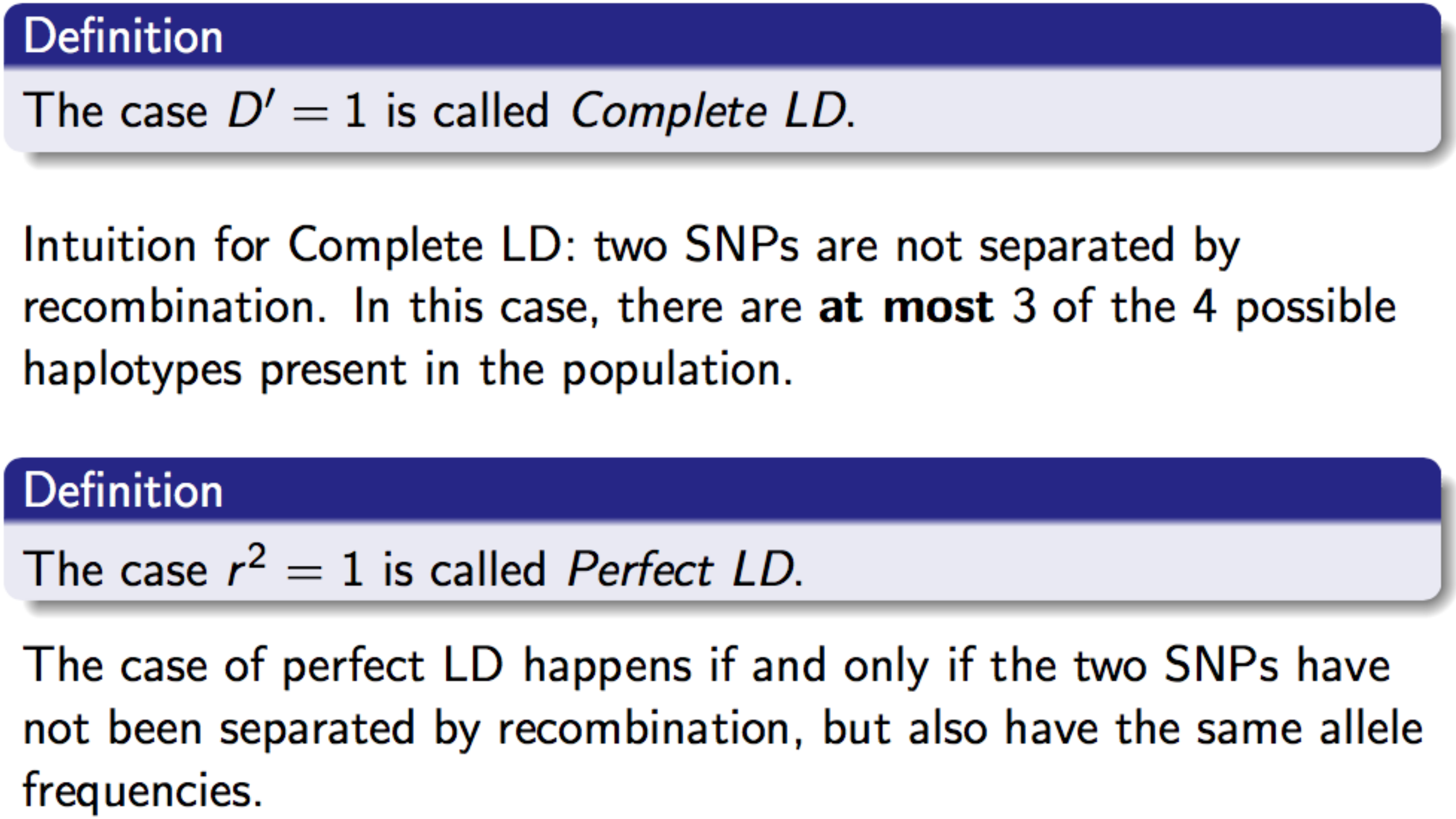
More on r^2

- $r^2 = 1$ implies the markers provide exactly the same information

- The measure preferred by population geneticists

- Measures loss in efficiency when marker A is replaced with marker B in an association study
  - With some simplifying assumptions (e.g. see Pritchard and Przeworski, 2001)

D’=1 and r^2=1



Definition

The case $D' = 1$ is called *Complete LD*.

Intuition for Complete LD: two SNPs are not separated by recombination. In this case, there are **at most** 3 of the 4 possible haplotypes present in the population.

Definition

The case $r^2 = 1$ is called *Perfect LD*.

The case of perfect LD happens if and only if the two SNPs have not been separated by recombination, but also have the same allele frequencies.

When does linkage equilibrium hold?

Equilibrium or Disequilibrium?

- We will present simple argument for why linkage equilibrium holds for most loci
- Balance of factors
  - Genetic drift (a function of population size)
  - Random mating
  - Distance between markers
  - ...

Why Equilibrium is Reached...

- Eventually, random mating and recombination should ensure that mutations spread from original haplotype to all haplotypes in the population...
- Simple argument:
  - Assume fixed allele frequencies over time

Independence test（p-value)

	B1	B2	
A1	a	b	a+b
A2	c	d	c+d
	a+c	b+d	n

Fisher exact test

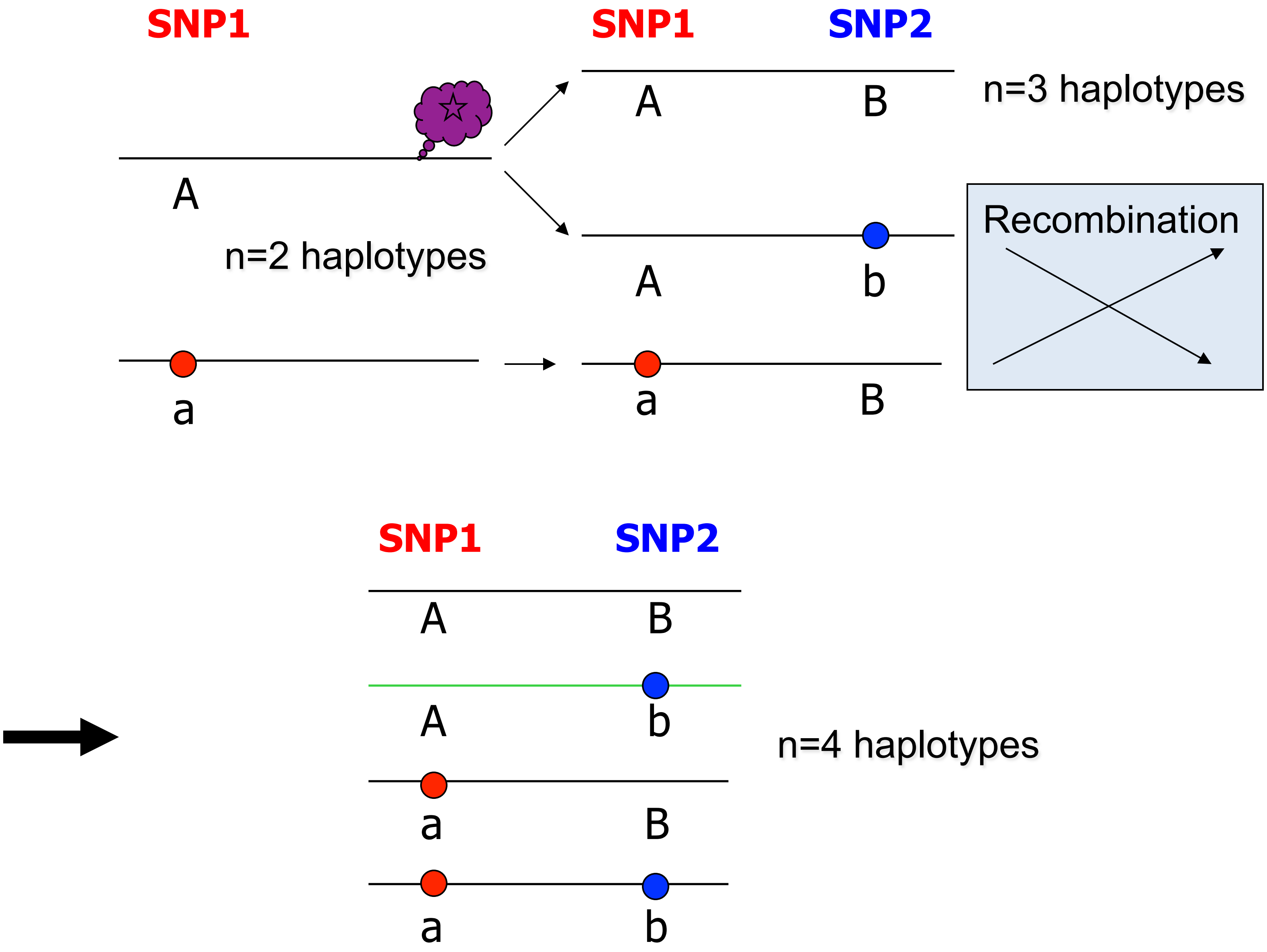
$$Pr(a,b,c,d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

2x2 table test

$$X^2 = \frac{\left(\frac{(a+b)(a+c)}{n} - a\right)^2}{\frac{(a+b)(a+c)}{n}} + \frac{\left(\frac{(a+b)(b+d)}{n} - b\right)^2}{\frac{(a+b)(b+d)}{n}} \frac{\left(\frac{(c+d)(a+c)}{n} - c\right)^2}{\frac{(c+d)(a+c)}{n}} \frac{\left(\frac{(b+d)(c+d)}{n} - d\right)^2}{\frac{(b+d)(c+d)}{n}}$$

Creation of LD

- Easiest to understand when markers are physically linked
- Creation of LD
  - Mutation
  - Founder effect
  - Admixture
  - Inbreeding / non-random mating
  - Selection
  - Population bottleneck or stratification
  - Epistatic interaction
- LD can occur between *unlinked* markers
- *Gametic phase disequilibrium* is a more general term



SNP1

A

n=2 haplotypes

a

SNP1 SNP2

A B

n=3 haplotypes

A b

a B

Recombination

SNP1 SNP2

A B

A b

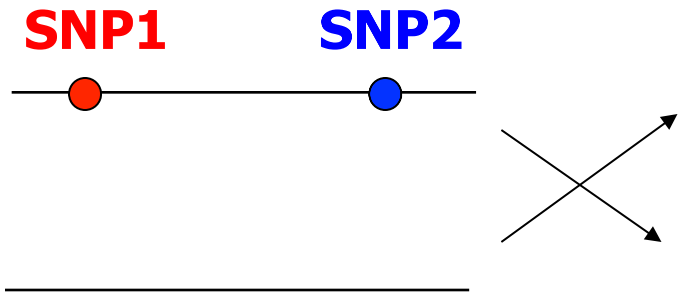
a B

a b

n=4 haplotypes

Destruction of LD

- Main force is recombination
- Gene conversion may also act at short distances (~ 100-1,000 bases)
- LD decays over time (generations of interbreeding)



Probability Recombination occurs = θ

Probability Recombination *<u>does not</u>* occur = $1-\theta$

Initial LD between SNP1 - SNP2: D_0

After 1 generation

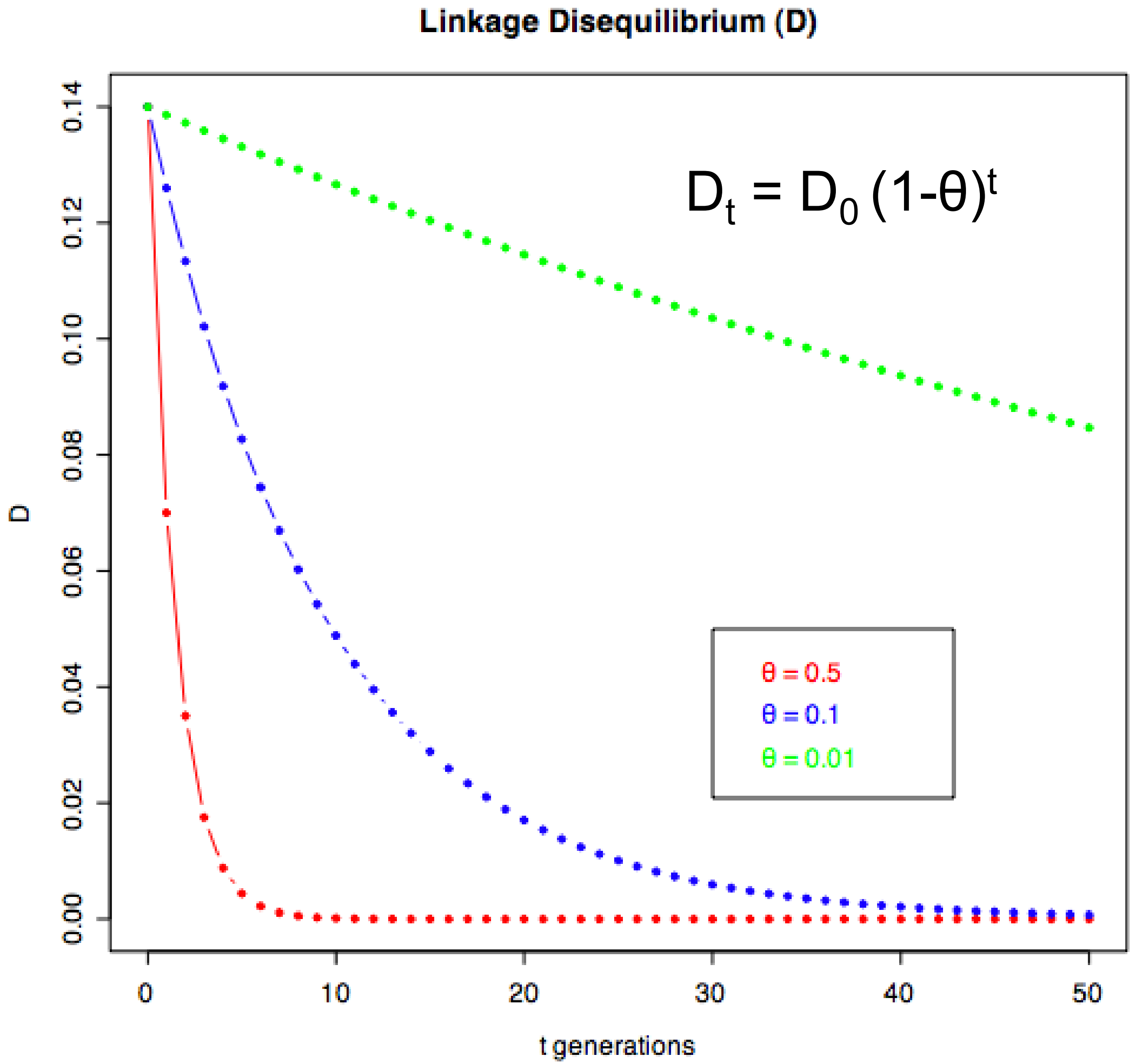
Preservation of LD:

$$D_1 = D_0(1-\theta)$$

After t generations:

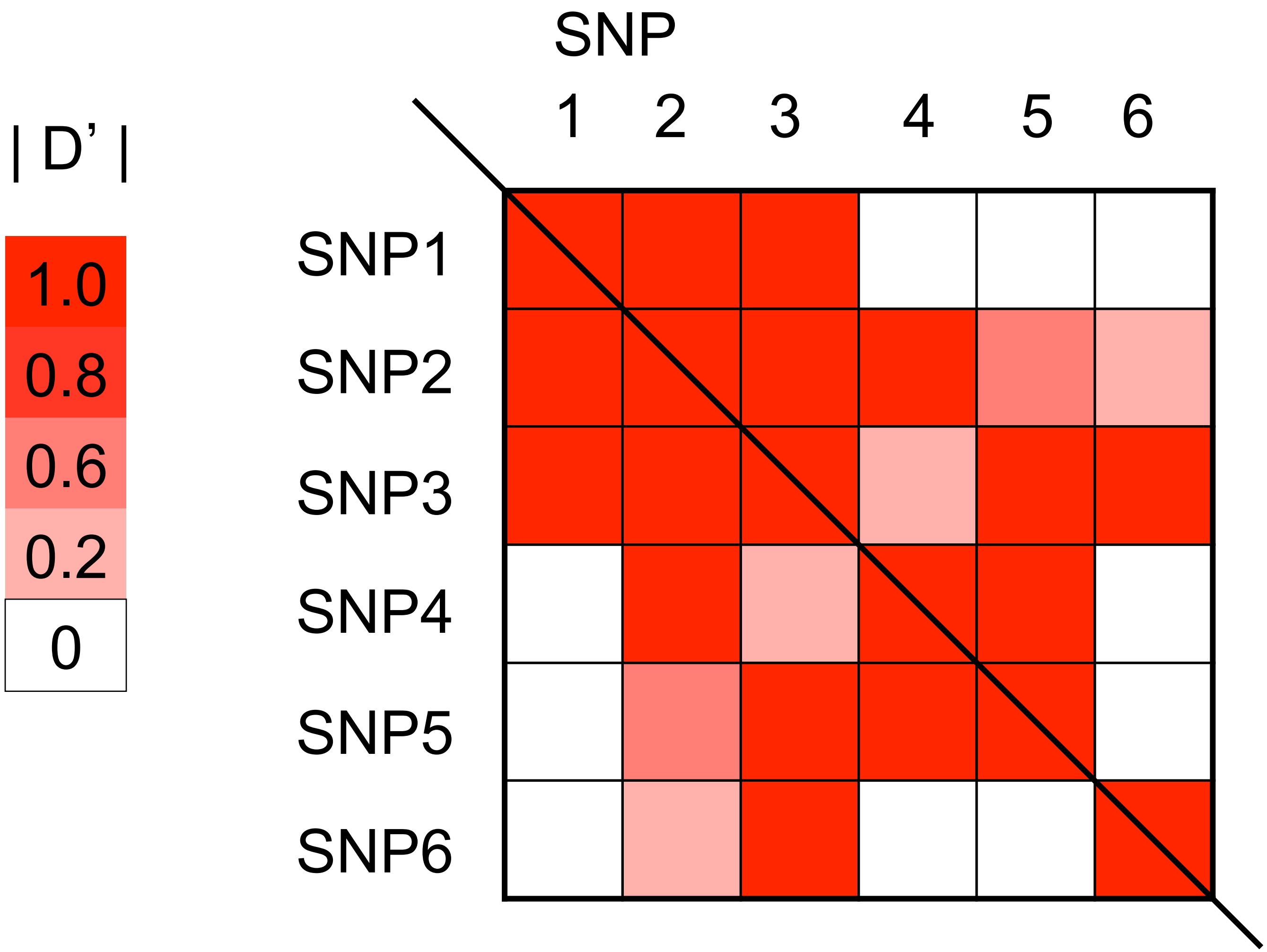
$$D_t = D_0 (1- \theta)^t$$

NB: Overly simple model - does not account for allele frequency drift over time

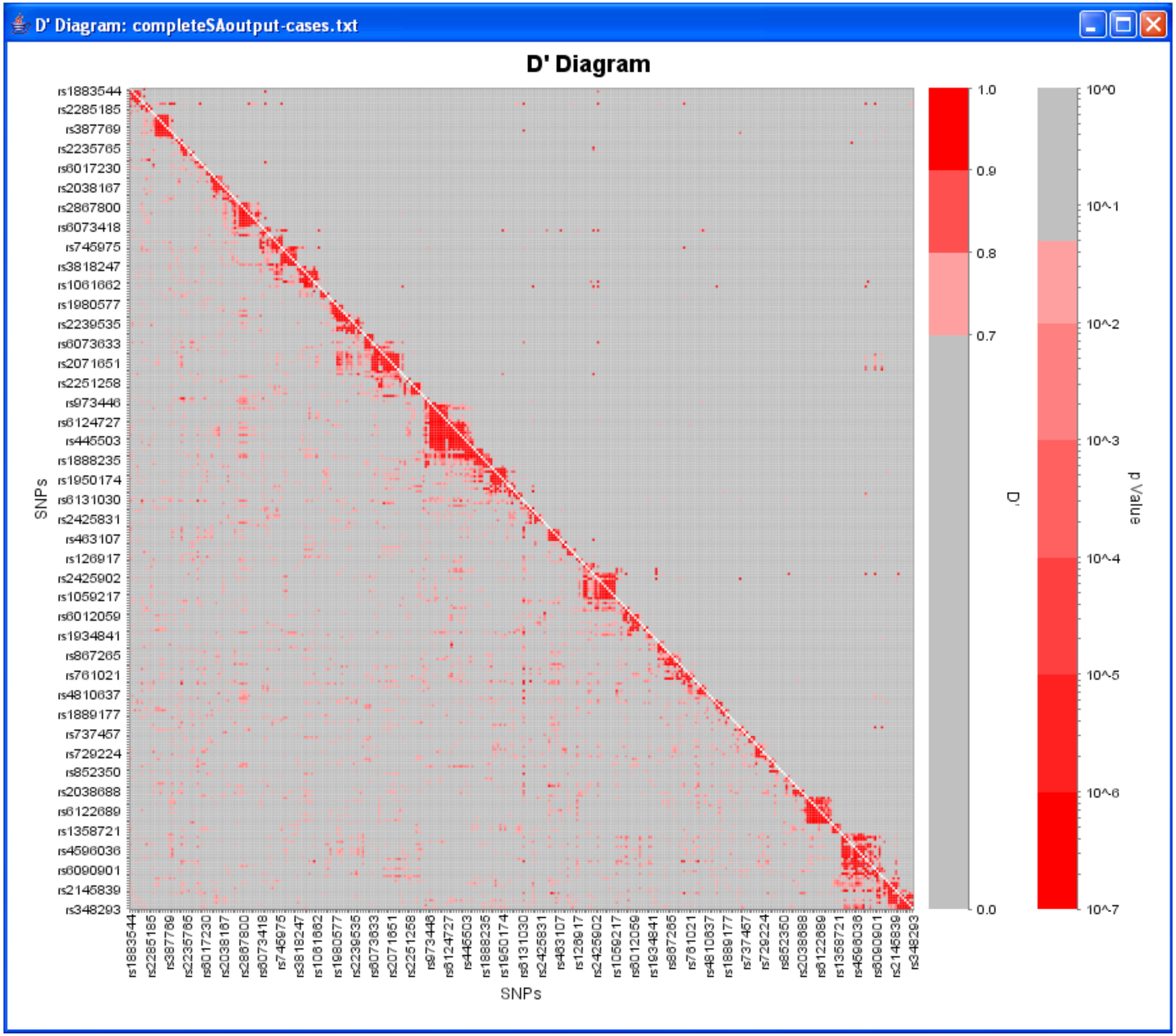


$$D_t = D_0\,(1\text{-}\theta)^t$$

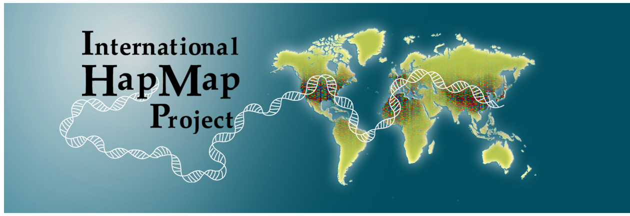
Visualizing LD metrics



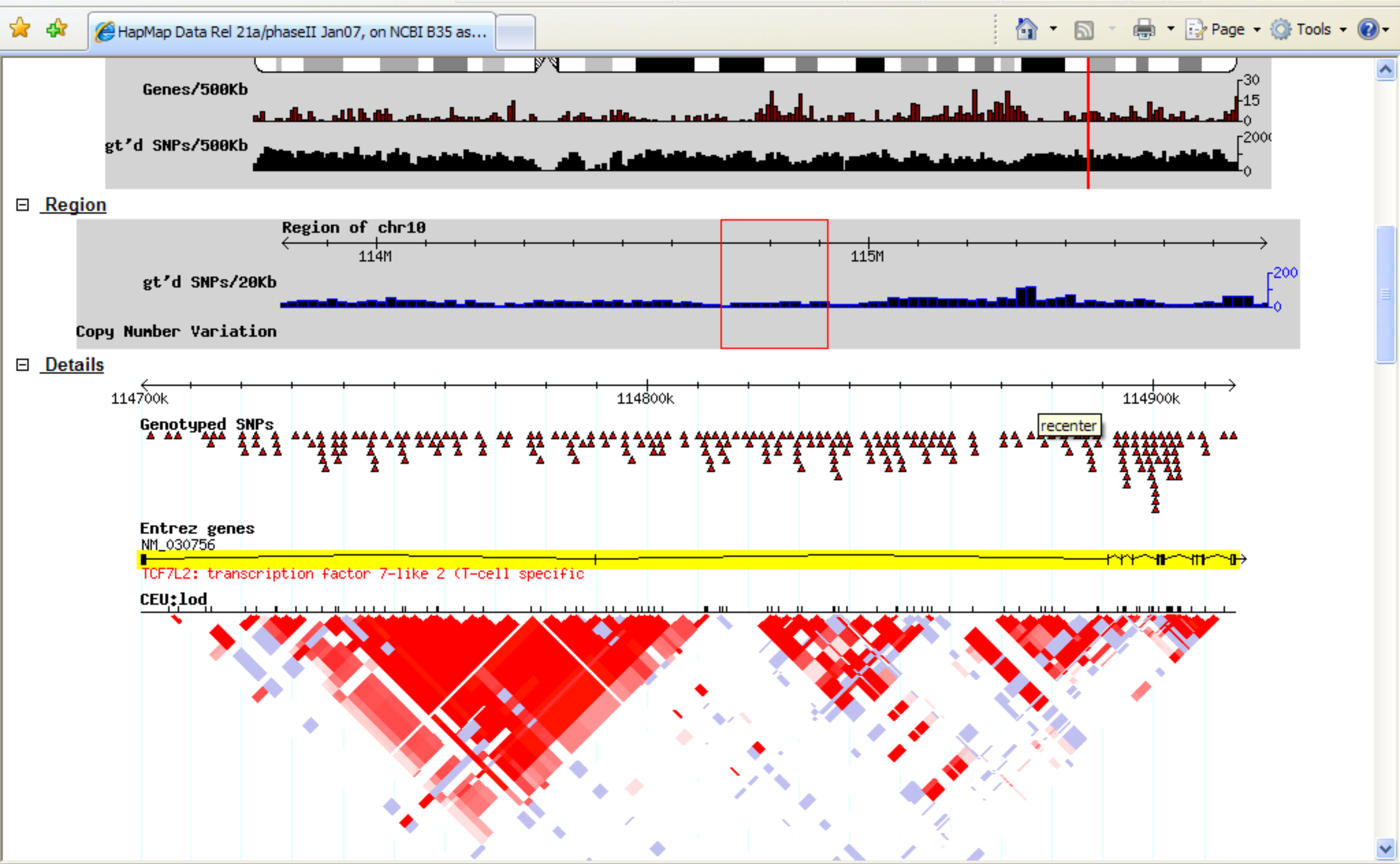
Not usually worried about sign of D’


D' Diagram: completeSAoutput-cases.txt
D' Diagram
SNPs
rs1883544
rs2285185
rs387769
rs2235765
rs6017230
rs2038167
rs2867800
rs6073418
rs745975
rs3818247
rs1061662
rs1980577
rs2239535
rs6073633
rs2071651
rs2251258
rs973446
rs6124727
rs445503
rs1888235
rs1950174
rs6131030
rs2425831
rs463107
rs126917
rs2425902
rs1059217
rs6012059
rs1934841
rs867265
rs761021
rs4810637
rs1889177
rs737457
rs729224
rs852350
rs2038688
rs6122689
rs1358721
rs4596036
rs6090901
rs2145839
rs348293
SNPs
1.0
0.9
0.8
0.7
0.0
D'
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
p Value

International HapMap Project



- Initiated Oct 2002
- Collaboration of scientists worldwide
- Goal: describe common patterns of human DNA sequence variation
- Identify LD and haplotype distributions
- Populations of different ancestry (European, African, Asian)
  - Identify common haplotypes and population-specific differences
- Has had major impact on:
  - Understanding of human popualtion history as reflected in genetic diversity and similarity
  - Design and analysis of genetic association studies


HapMap Data Rel 21a/phaseII Jan07, on NCBI B35 as...
Page
Tools
Genes/500Kb
30
15
0
gt'd SNPs/500Kb
200
0
Region
Region of chr10
114M
115M
gt'd SNPs/20Kb
200
0
Copy Number Variation
Details
114700k
114800k
114900k
Genotyped SNPs
recenter
Entrez genes
NM_030756
TCF7L2: transcription factor 7-like 2 (T-cell specific
CEU:lod

LD in Human Populations

Haplotype Blocks

N SNPs = 2^N Haplotypes possible, ie very large diversity possible

But: we do not see the full extent of haplotype diversity in human populations

Extensive LD especially at short distances eg ~20kbases.

Haplotypes are broken into blocks of markers with high *mutual* LD separated by recombination hotspots

Non-uniform LD across genome

Haplotype Blocks

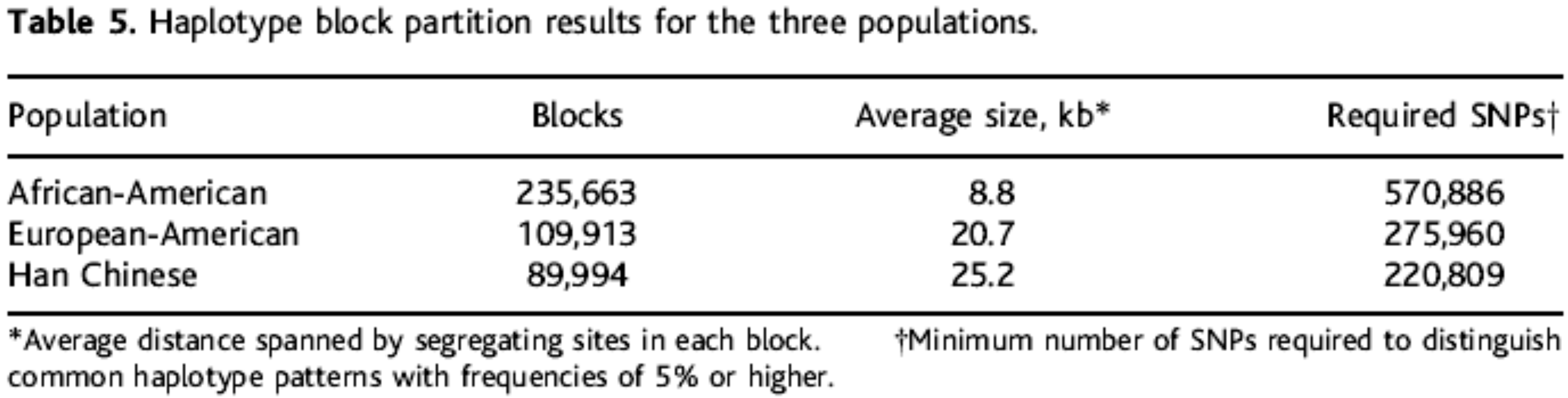


Table 5. Haplotype block partition results for the three populations.

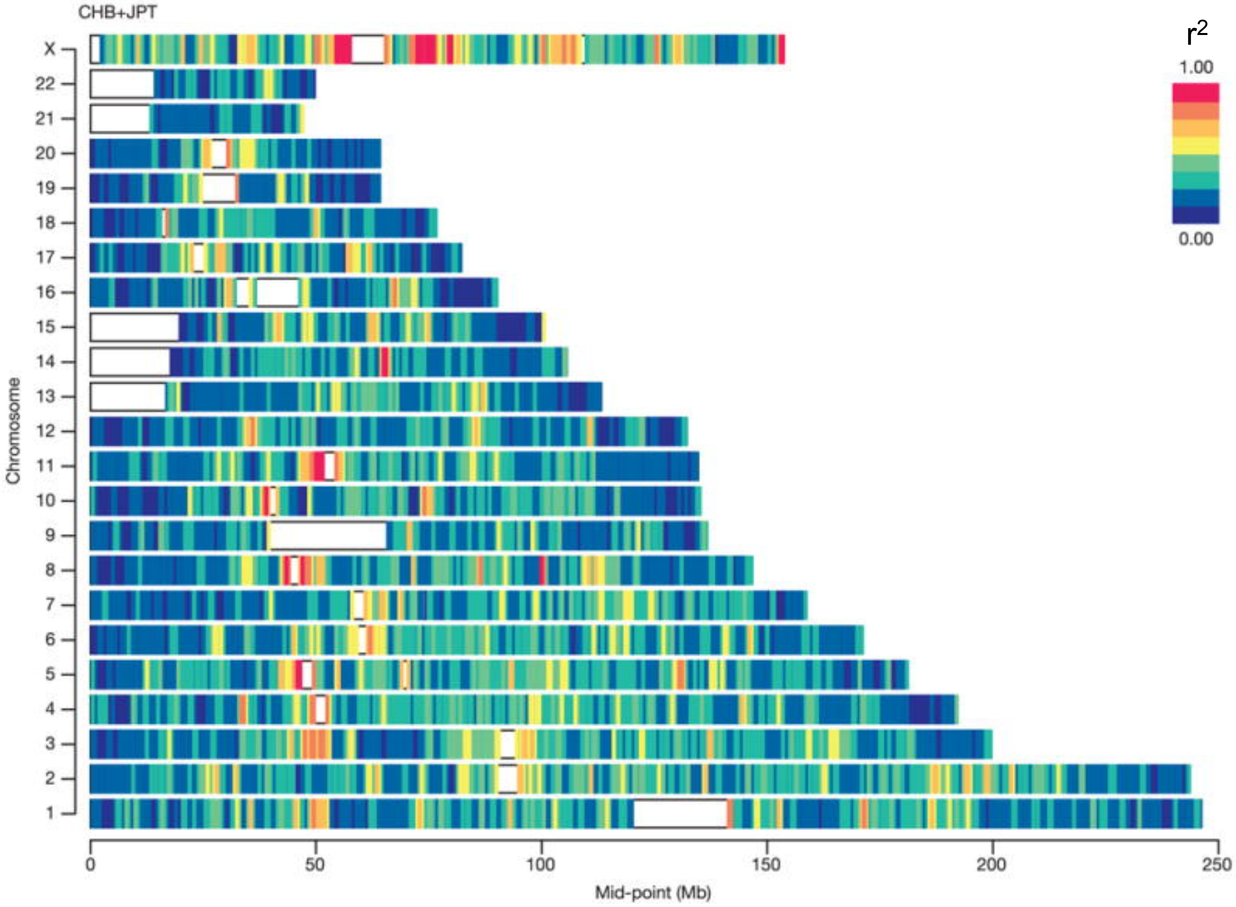
Population	Blocks	Average size, kb*	Required SNPs†
African-American	235,663	8.8	570,886
European-American	109,913	20.7	275,960
Han Chinese	89,994	25.2	220,809

*Average distance spanned by segregating sites in each block. †Minimum number of SNPs required to distinguish common haplotype patterns with frequencies of 5% or higher.

Haplotype blocks: at least 80% of observed haplotypes with frequency >= 5% could be grouped into common patterns

Whole Genome Patterns of Common DNA Variation
in Three Human Populations, Science 2005, Hinds et al.

Length of LD spans



Example: Large block of LD on chromosome 17

Cluster of common (frequent SNPs In high LD)

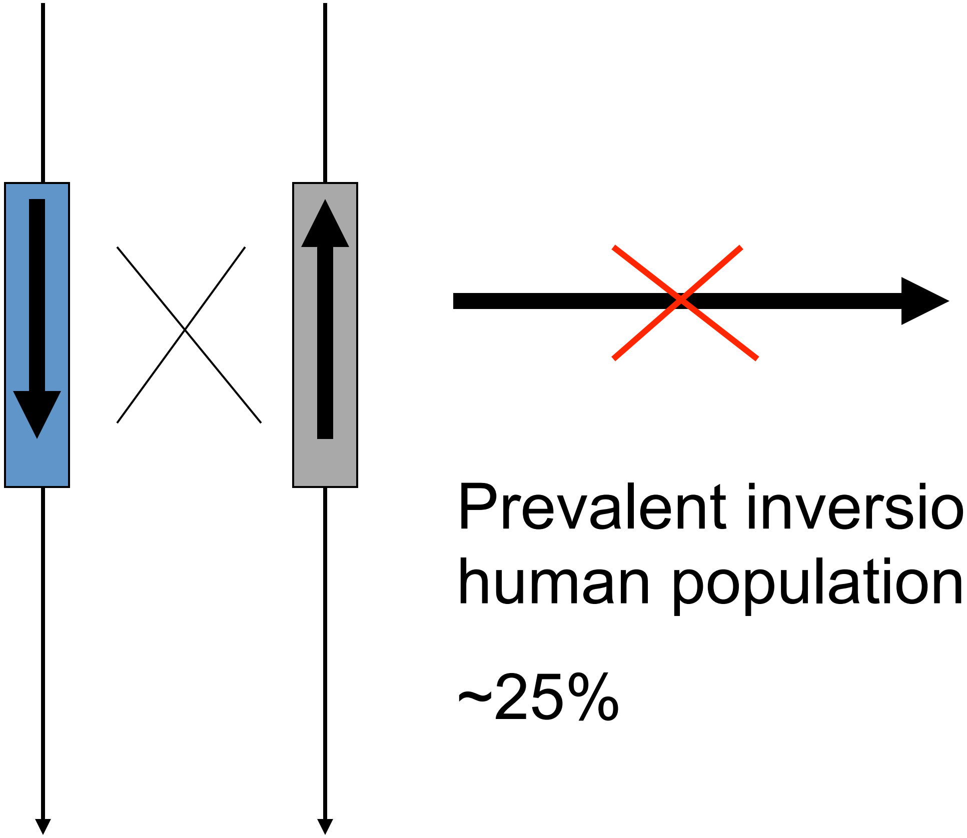
518 SNPs, spanning 800 kb

25% in EUR, 9% in AFR, missing in CHN

Genes:

- Microtubule-associated protein tau
- Mutations associated with a variety of neurodegeneartive disorders
- Gene coding for a protease similar to presenilins
- Mutations result in Alzheimer's disease
- Gene for corticotropin-releasing hormone receptor
    - Immune, endocrine, autonomic, behavioral response to stress

Chromosome 17 LD Region



Prevalent inversion in EUR human population

~25%

Thank you for your attention!