Identity by descent across the genomes

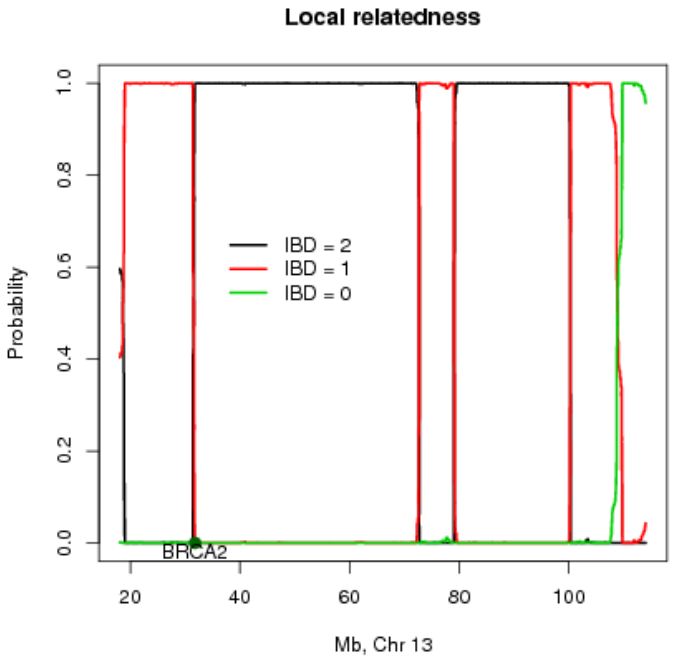
Anders Albrechtsen

March 6, 2019

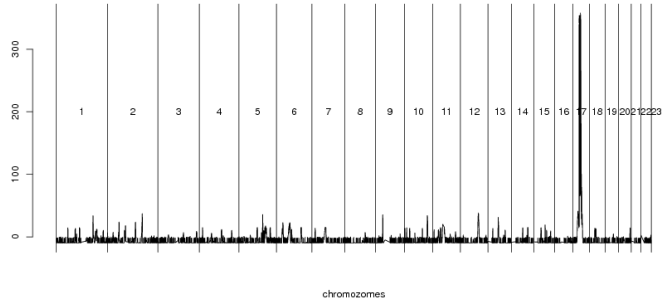
IBD estimation

$k_0 = 0.25, k_1 = 0.5$ and $k_2 = 0.25$

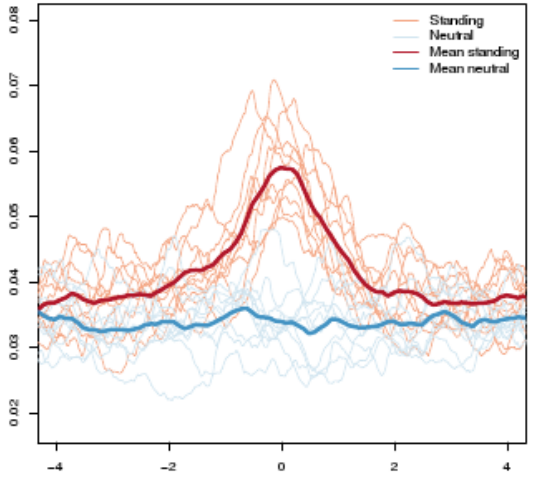
IBD tracts



IBD mapping



IBD and selection



This talk

background

- What is IBD sharing
- Definitions of IBD
- Inbreeding and relatedness
- estimate local IBD tracts (regions of relatedness or inbreeding) across the genome

learning objects

- Understand that IBD is not just a number but a pattern across the genome
- Use the patterns to infer relationships/pedigrees
- Use the patters to identify regions with a disease causing genetic variation
- Use the patterns to infer regions under adaptive selection

What is inbreeding

Inbreeding occurs when a genotype carries two copies of a gene that are **identical by descent.**

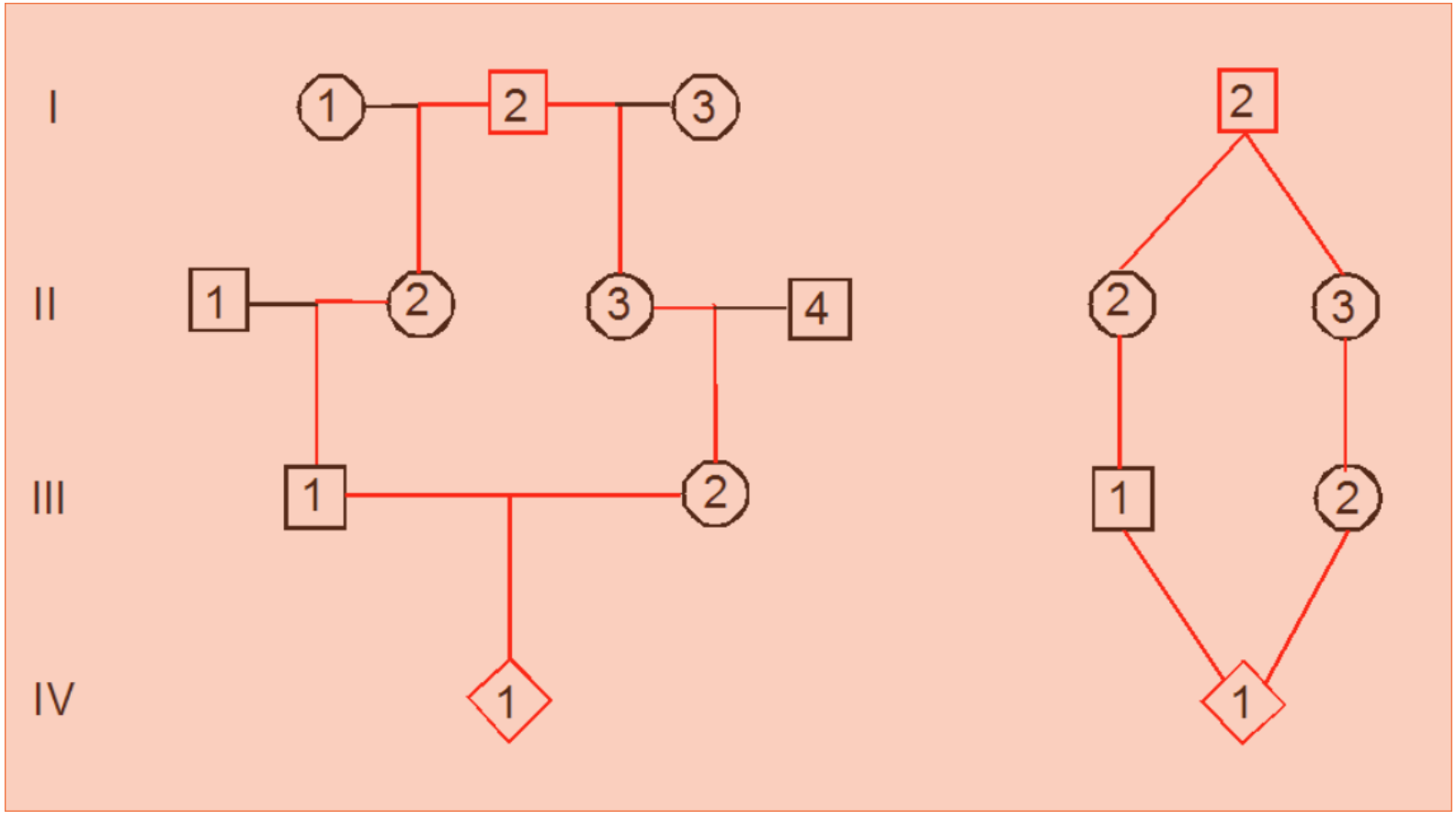
We must all be related somehow!

- How many parents do you have?
- How many grandparents do you have?
- How many ancestors did you have 10 generations ago?
- How many ancestors did you have 40 generations ago?

Inbreeding occurs when the number of ancestors is less than maximum possible ones

Expectations of Inbreeding from Pedigrees

Two versions of the same pedigree



Expectations of Inbreeding from Pedigrees

Calculation from pedigress

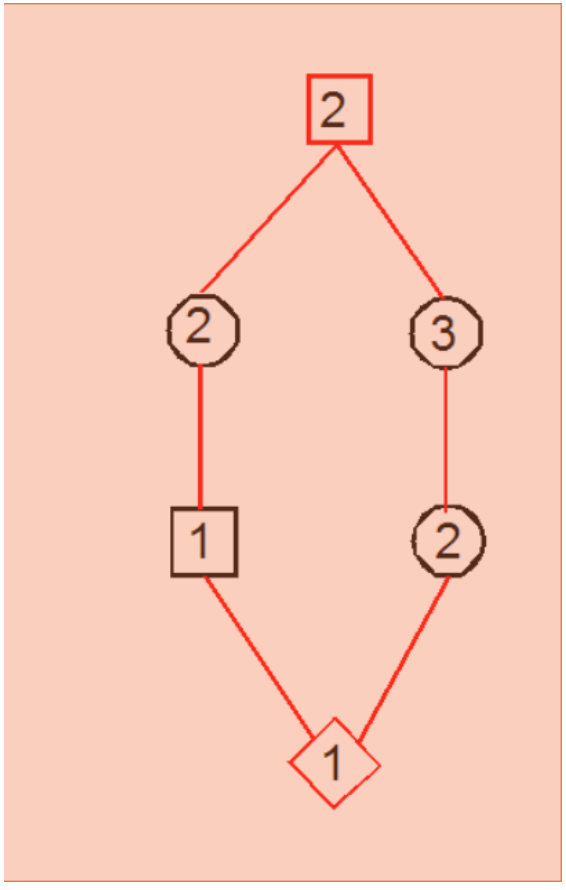
$$E(F) = \left(\frac{1}{2}\right)^{N_A}$$

N_A is the number of ancestors that form a ring

If there are multiple rings you add them together

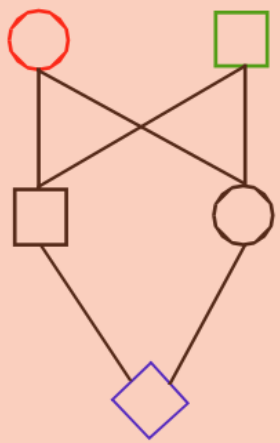
$$E(F) = \left(\frac{1}{2}\right)^{5} = 0.03125$$

Two versions of the same pedigree

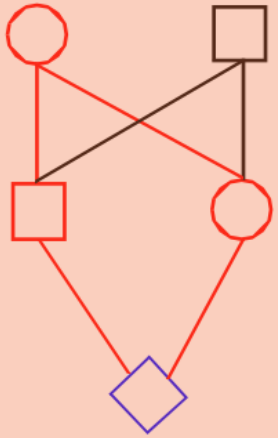


Multiple rings

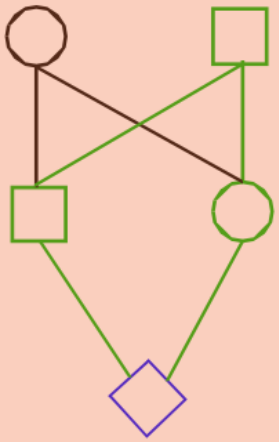
With several coancestors: add their contributions to F



Brother-sister



$$F = \left(\frac{1}{2}\right)^3$$

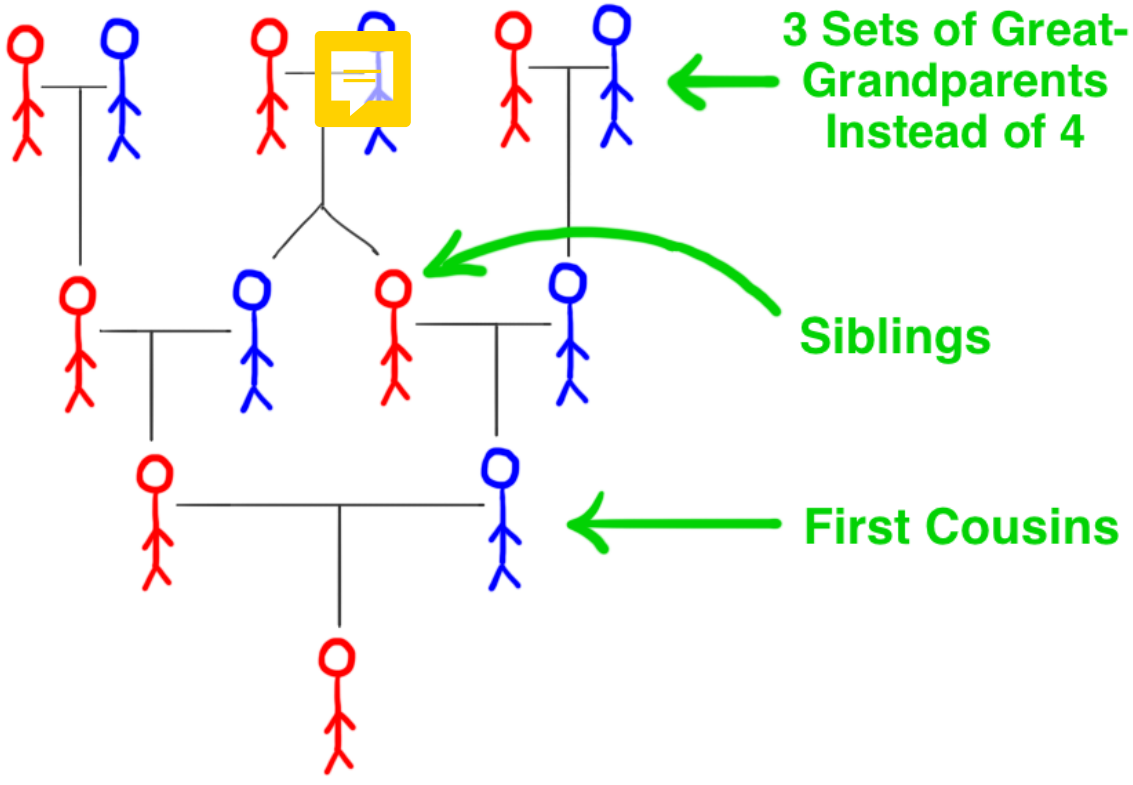


$$F = \left(\frac{1}{2}\right)^3$$

For 2 rings

$$E(F) = \sum_{r \in rings} \left(\frac{1}{2}\right)^{N_A^r} = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = 0.25$$

Exercise - calculate the expected inbreeding



What is Identical dy descent?

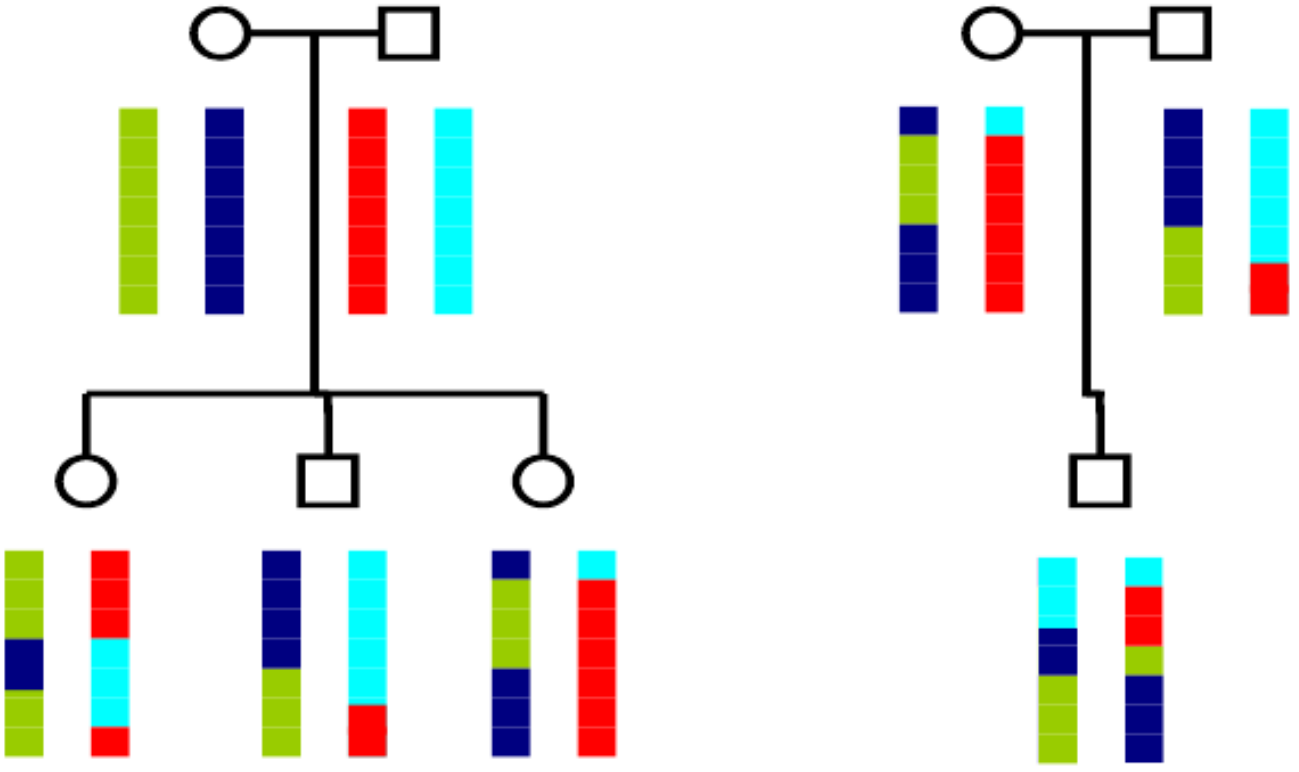
Should be easy right?

there is no single meaningful definition

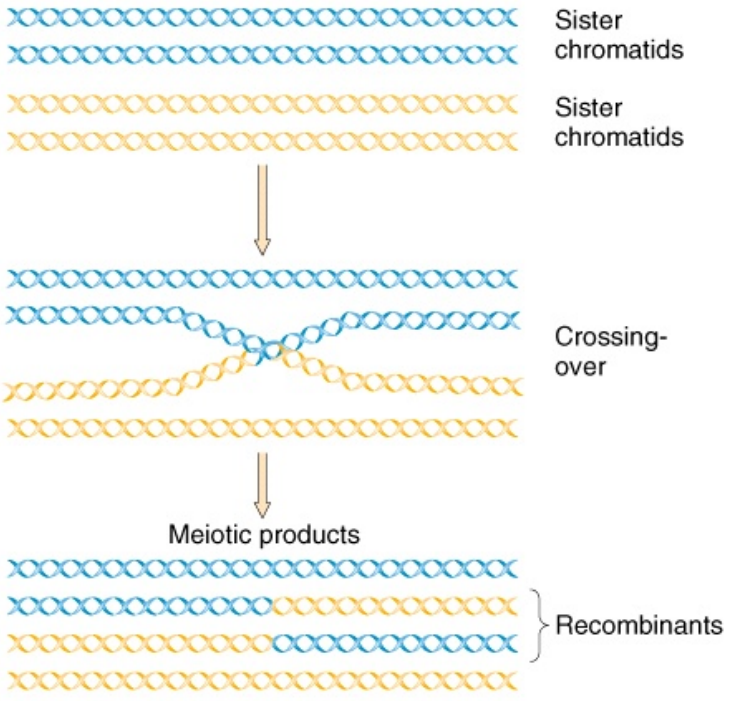
Relatedness/inbreeding and IBD

A definition

Two allele are *Identical by descent* (IBD) if they are direct copies of a single ancestral alleles



meiosis and recombination

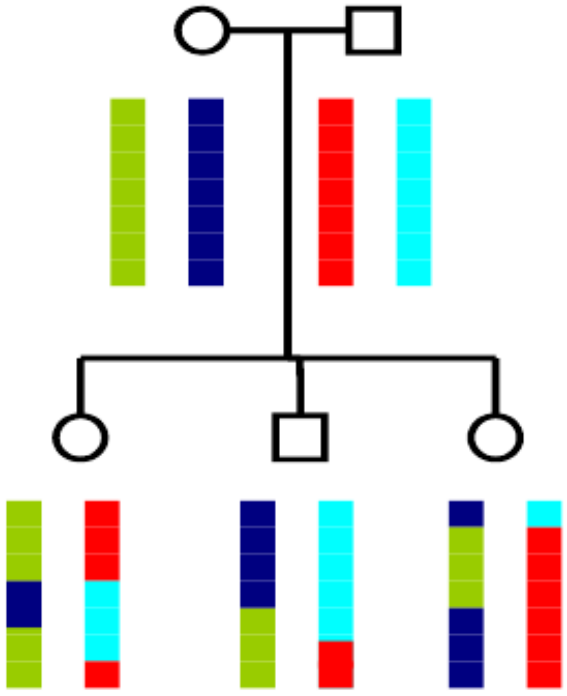


Because of a finite number of recombination events we expect the IBD state of a loci to be dependent of the state of adjacent loci

Relatedness/inbreeding and IBD

Problem

The IBD patterns are not directly observable



SNP: Single nucleotide polymorphism

Homologous pairs of chromosomes

Paternal allele

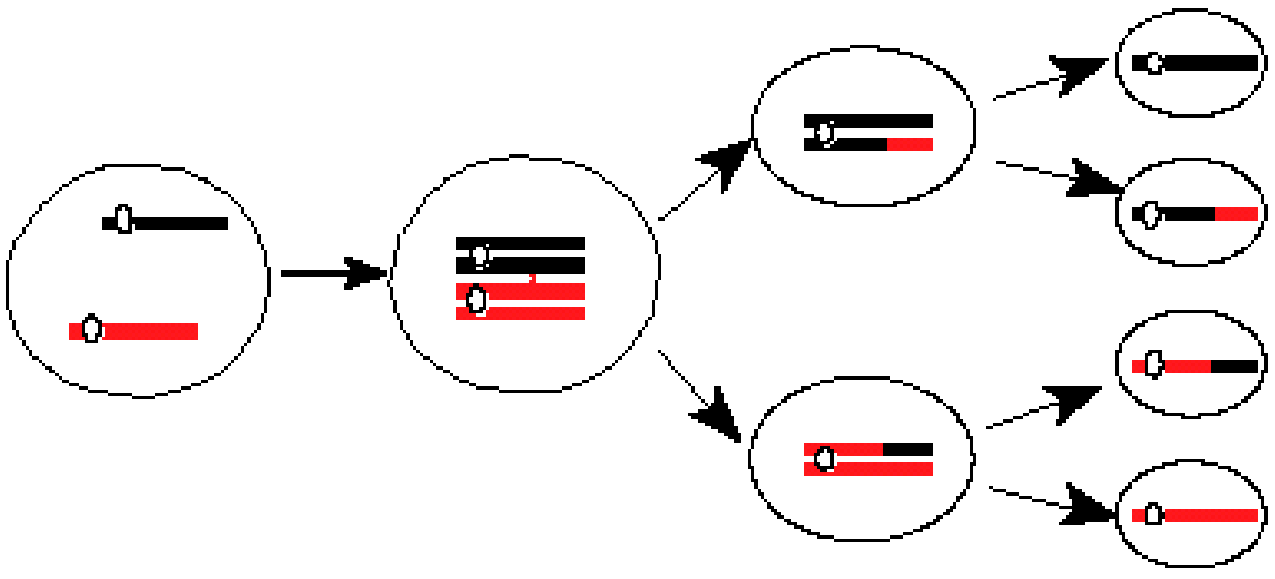
Maternal allele

Paternal allele ACGAACAGCT
TGCTTGTCGA

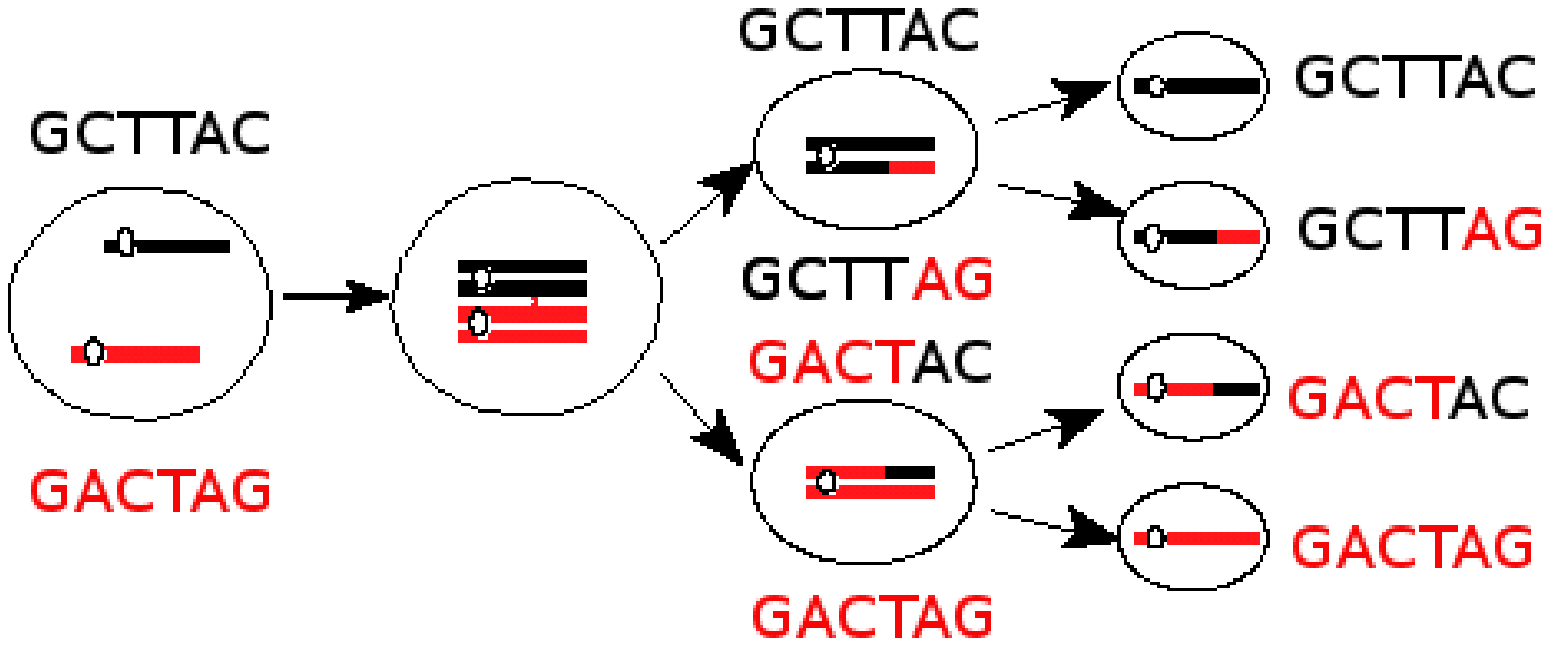
Maternal allele ACGAGCAGCT
TGCTCGTCGA

SNP A/G

meiosis and recombination



meiosis and recombination



If two alleles are IBD they are also *identical by state* (IBS)

Inbreeding as a statistical construct

A definition

The IBD is the variability that cannot be explained by the population frequency

probability of genotype assuming no Inbreeding (HWE)

$$p(AA) = f_A^2$$
$$p(Aa) = 2 f_A f_a$$
$$p(aa) = f_a^2$$

Inbreeding as a statistical construct

A definition

The IBD is the variability that cannot be explained by the population frequency

probability of genotype with Inbreeding – F is the probability of inbreeding

$$
\begin{aligned}
p(AA) &= f_A^2(1-F) + f_A F \\
p(Aa) &= 2 f_A f_a (1-F) \\
p(aa) &= f_a^2(1-F) + f_a F
\end{aligned}
$$

Genotype data

5 individuals, genotypes

SNP1	AG	AG	AG	AA	AA
SNP2	TT	TA	AA	AT	AA
SNP3	AA	AC	AC	CC	AC
SNP4	GG	GG	GC	CC	CC
SNP5	TT	TC	TC	CC	CC
SNP6	AA	AA	AC	AC	AC
SNP7	TT	TT	TC	TC	CC

5 individuals, allele counts

SNP1	1	1	1	0	0
SNP2	0	1	2	1	2
SNP3	2	1	1	0	1
SNP4	0	0	1	2	2
SNP5	2	1	1	0	0
SNP6	0	0	1	1	1
SNP7	2	2	1	1	0

How to estimate F from data - Genotype data

Genotypes (G) for 5 individuals, allele counts

SNP1	1	1	1	0	0
SNP2	0	1	2	1	2
SNP3	2	1	1	0	1
SNP4	0	0	1	2	2
SNP5	2	1	1	0	0
SNP6	0	0	1	1	1
SNP7	2	2	1	1	0

- M number of sites
- N number of individuals
- G_i^j genotypes for site i in individual j
- f_i allele frequency for site i. $f_i = \frac{1}{2} \sum_{j=1}^{N} G_i^j$

How to estimate F from data - Method of moment estimator

Observed number of heterozygous

$$O_{HE} = \sum_{i=1}^{M} I_{\{1\}}(G_i)$$

Expected number of heterozygous*

$$E_{HE} = \sum_{i=1}^{M} [2(1 - f_i)f_i]$$

Coefficient from observed and expected heterozygous

$$O_{HE} = (1 - F)E_{HE} \Rightarrow F = \frac{E_{HE} - O_{HE}}{E_{HE}}$$

*assumes the allele frequencies are known and not estimated from the data. Use $E = \sum_{i=0}^{M} \left[(1 - f_i)^2 + f_i^2\right] \frac{2N}{2N-1}$

Small exercise - estimate F

Your data - 6 SNPs

G_i (number of A)	0	0	1	0	2	0
f_i (freq of A)	0.2	0.7	0.6	0.1	0.4	0.1

How to estimate F from data - maximum likelihood

Probability of a site being inbreed

$$P(Z = 1) = F$$

probability of genotype

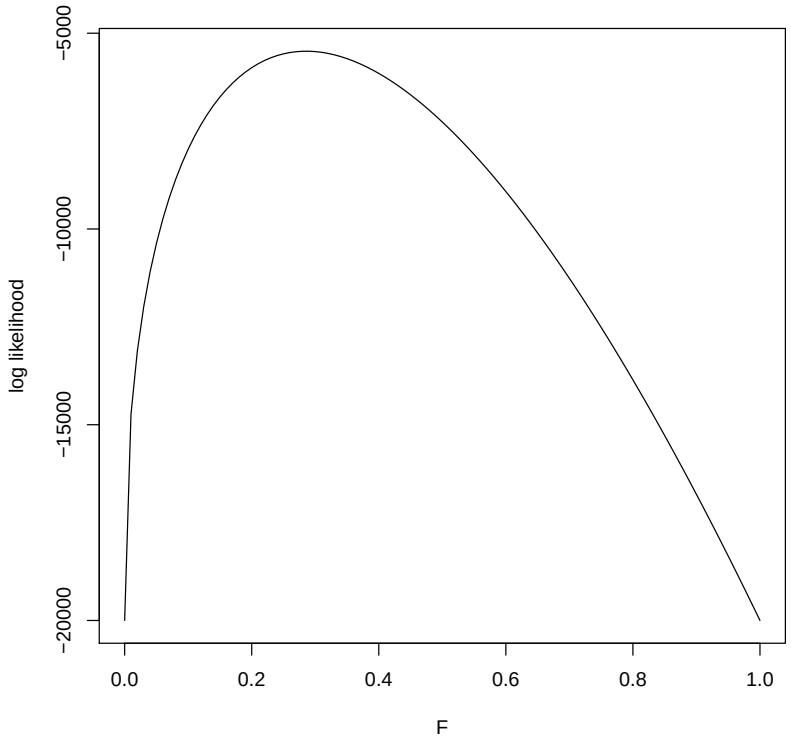
$$p(G_i|Z = z) = \begin{cases} Z = 0 & Z = 1 & \\ f^2 & f & G_i = 0 \\ 2f(-f) & 0 & G_i = 1 \\ (1-f)^2 & (1-f) & G_i = 2 \end{cases}$$

Coefficient from observed and expected homozygoes

$$p(G|F) = \sum_{i=1}^{M} [p(G_i|Z = 0)p(Z = 0) + p(G_i|Z = 1)p(Z = 1)]$$

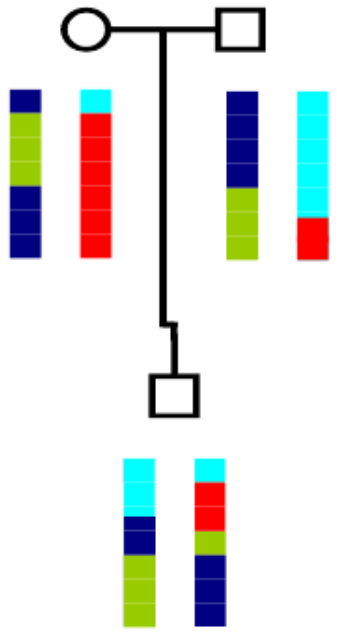
Finding F using maximum likelihood

Log likelihood surface

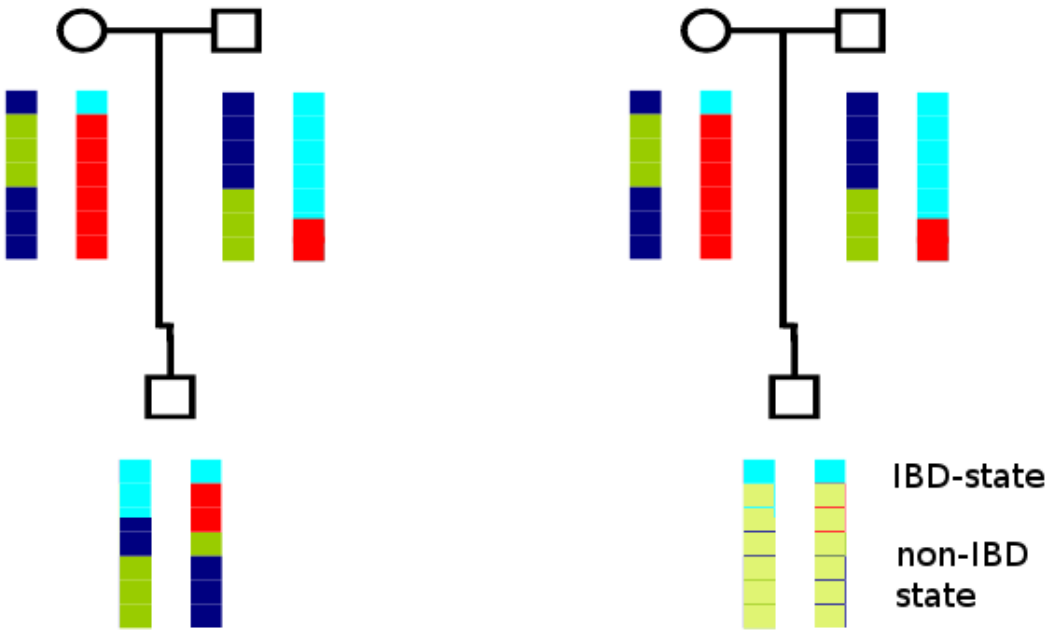


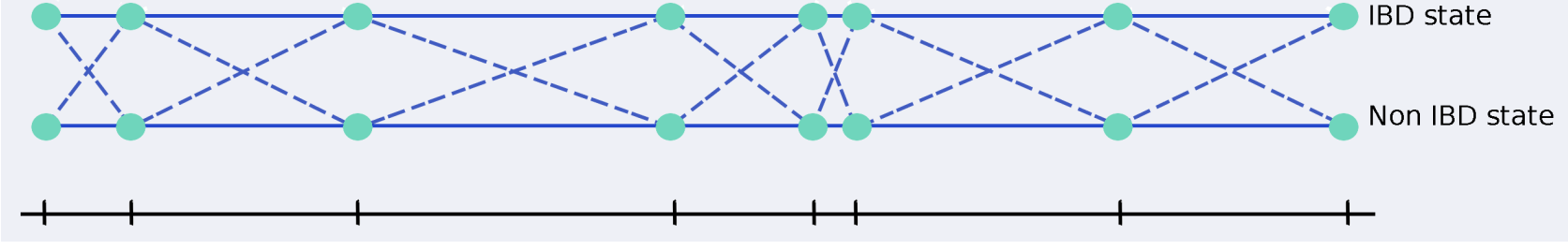
Maximum likelihood is usually obtained using numeric optimization

Use genomic information

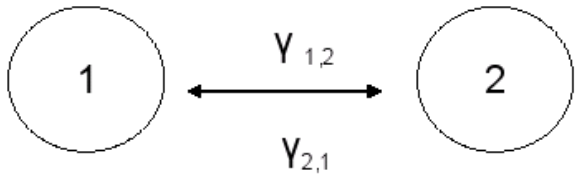


Two state hidden Markov model





IBD estimation using CTMC



Instantaneous rate matrix

$$Q = \begin{pmatrix} -\alpha F & \alpha F \\ \alpha(1-F) & -\alpha(1-F) \end{pmatrix}$$

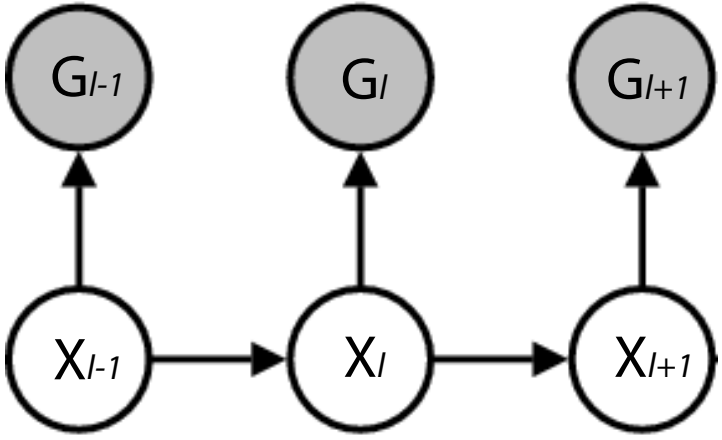
The stationary distribution is $(F, 1-F)$ the fraction shared IBD

The transition probabilities

$$P(X_i = x_i | X_{i-1} = x_{i-1}) = P(X(t+c) = x_i | X(c) = x_{i-1}),$$

where c is the genomic location of the $i-1$ marker and t is the distance between marker i and $i-1$

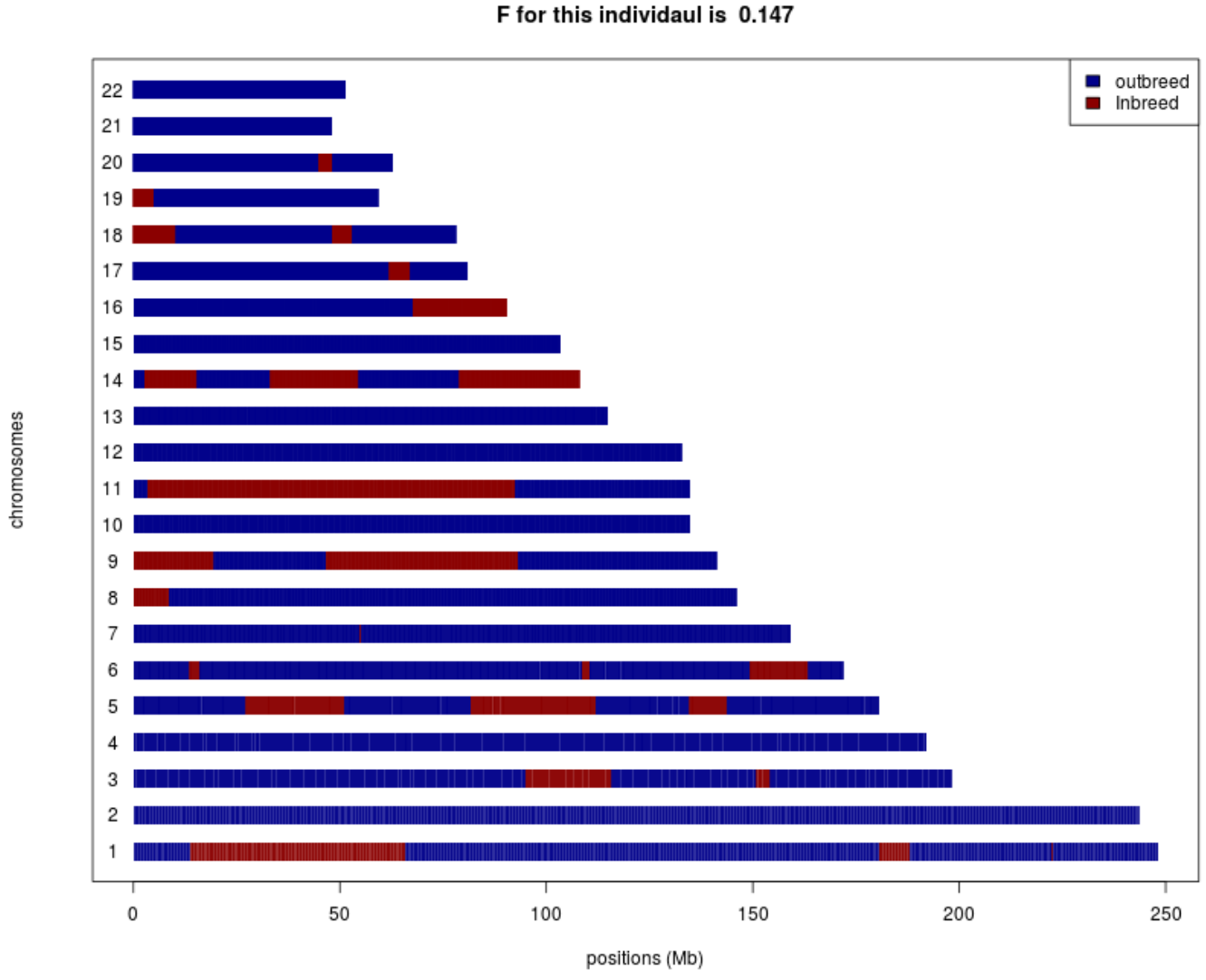
The hidden Markov model



- The model has two variables for each genotyped locus, l
  1. observed G_l (vector with a genotype for each individual)
  2. hidden X_l (vector with an IBD state for each chromosome)
- Each IBD state in X_l indicates IBD or non-IBD
- Likelihood of the SNP data $G = (G_1, G_2, ..., G_L)$

$$L(G|\alpha, F) = \sum_X \left(\prod_{l=1}^{L} P(G_l|X_l) \right) \left(P(X_1) \prod_{l=2}^{L} P(X_l|X_{l-1}) \right)$$

What we can get with good data

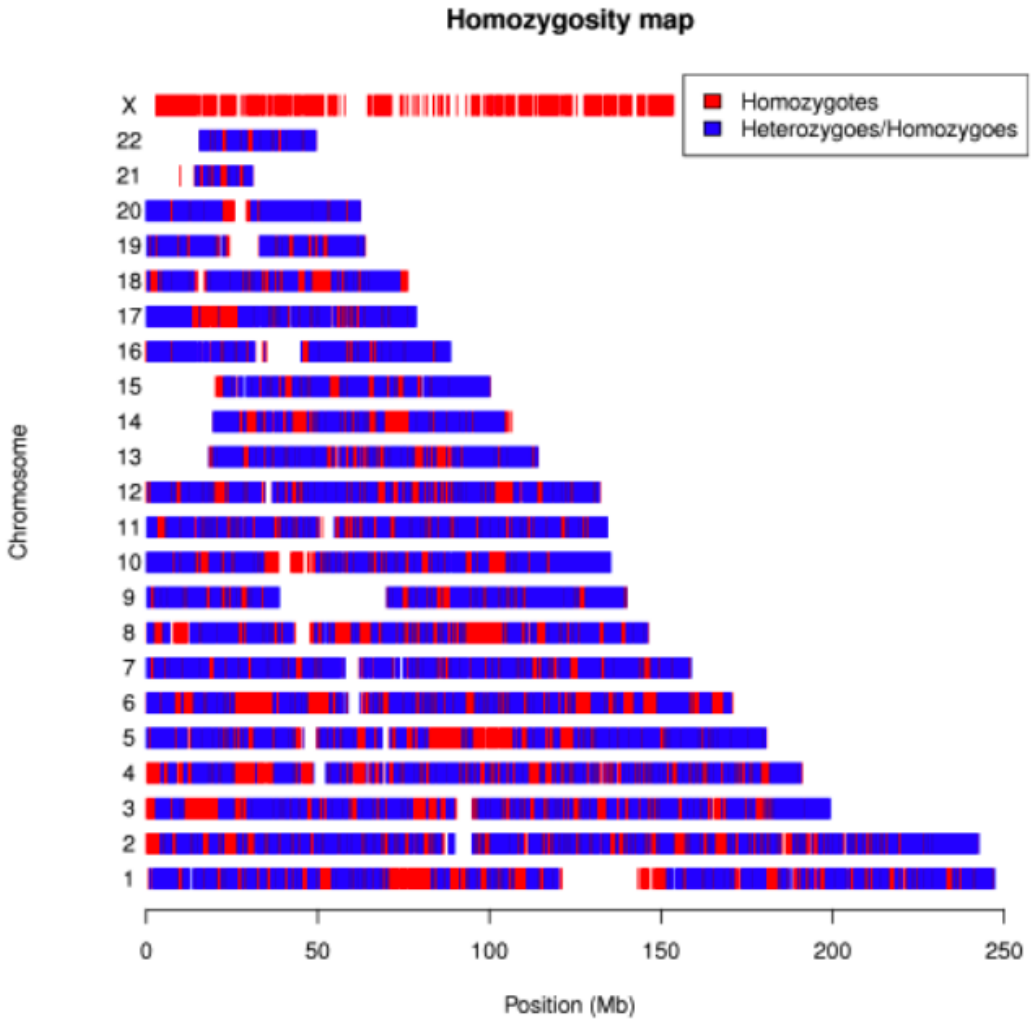


Studying a past human population

- A study where the first ancient human genome was sequenced
- 4000 year old male from the first known culture to settle in Greenland
- DNA extracted from hair found in the permafrost on the west coast

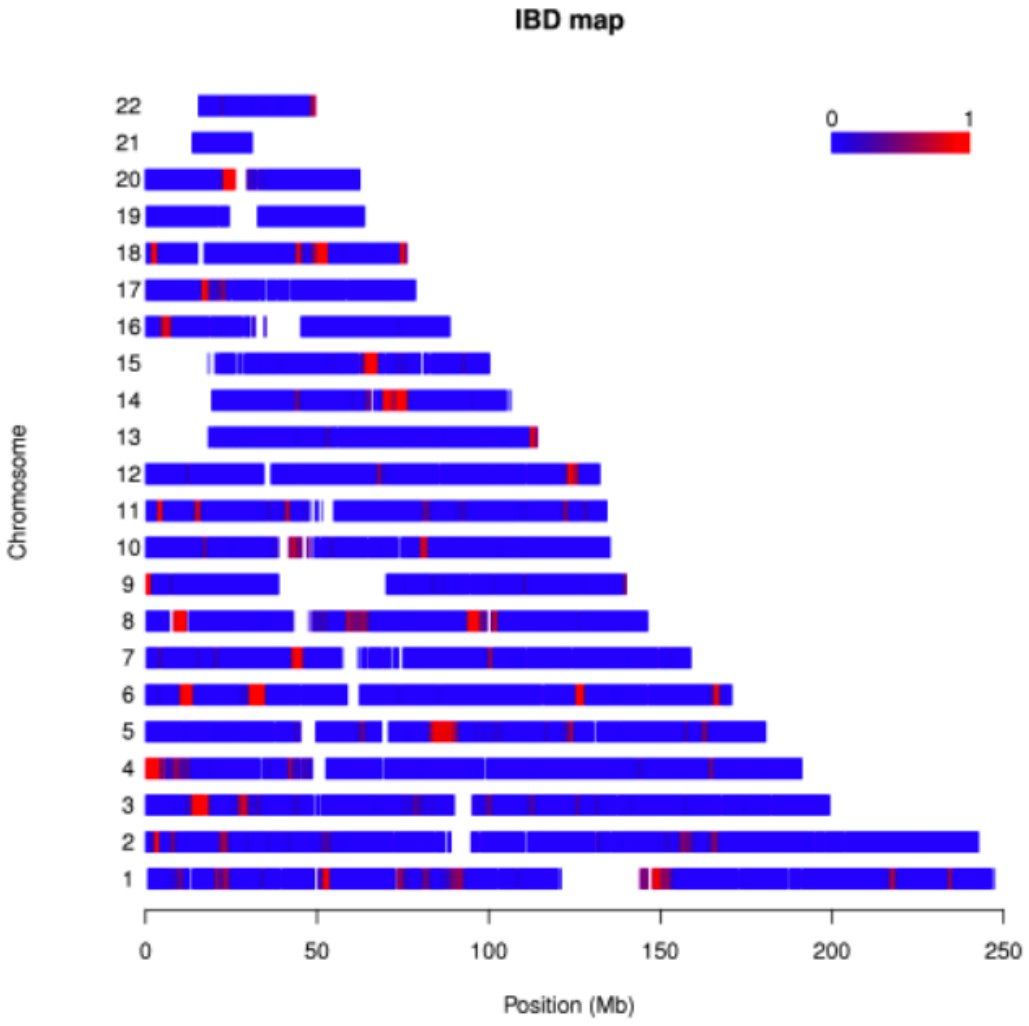
Studying a past human population (cont')

- Runs of homozygosity



Studying a past human population (cont')

- IBD analysis: to **infer IBD within the individual**



Studying a past human population (cont')

- The estimated amount of IBD is **not more than expected of an offspring of two first cousins**
- Possible reasons: very small population, tradition, an outlier.

Usher syndrome

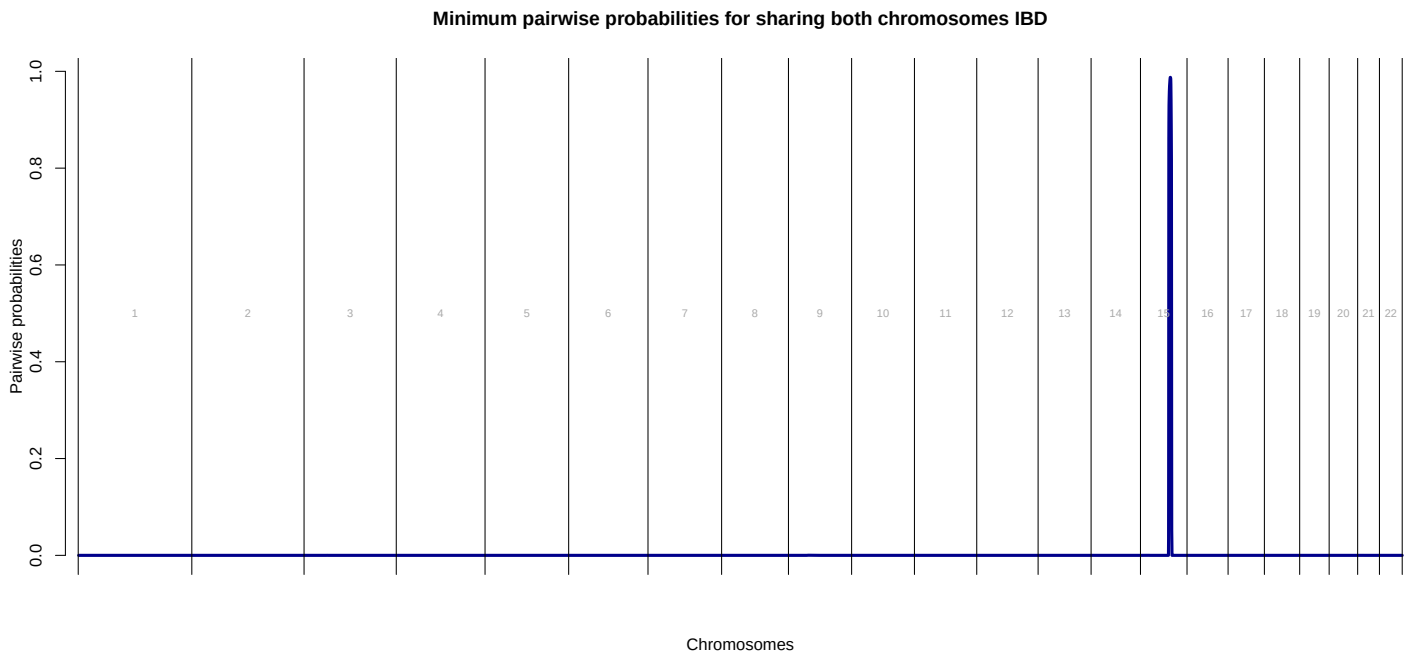
disease

- Usher syndrome (USH) is the most common genetic disease that causes both deafness and blindness
- Is a rare recessive disease
- more common in the Netherlands

data

- Five affected and five unaffected individuals
- 500,000 SNPs genotyped

Inbreeding estimation across the genome



Inbreeding estimation across the genome

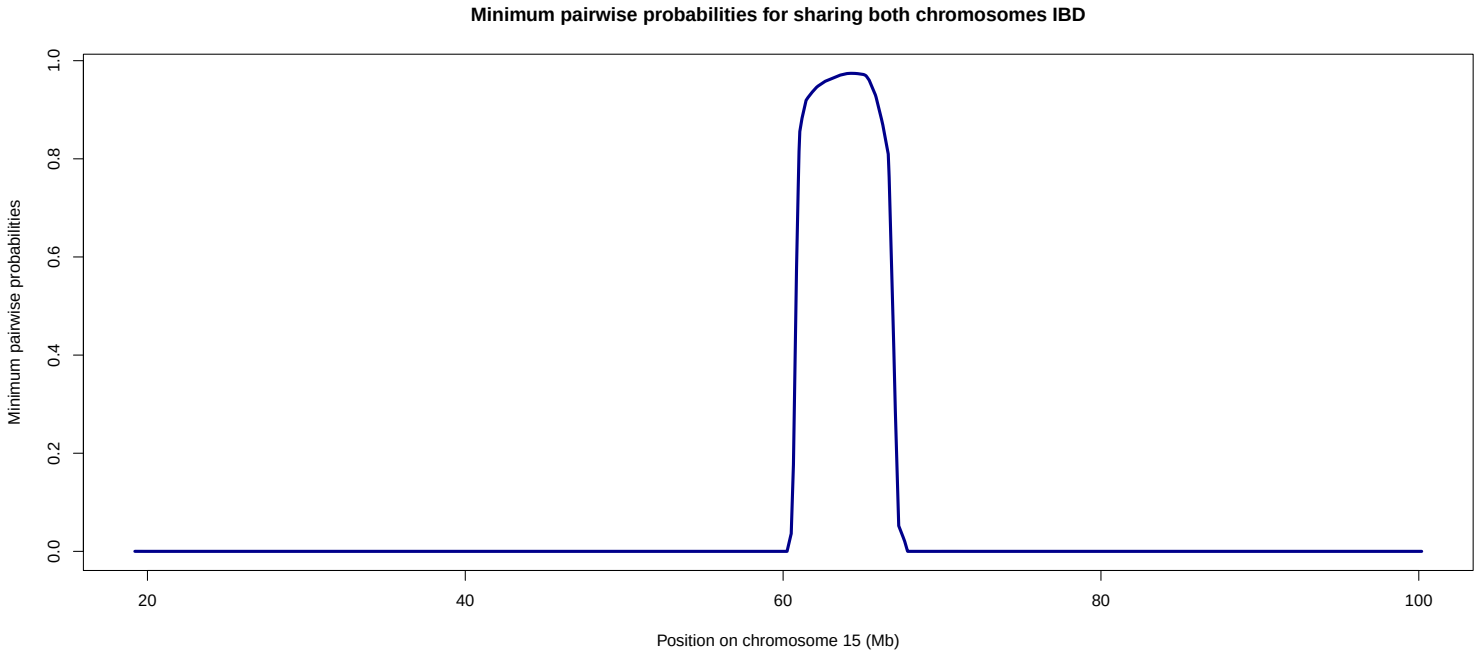


Figure: Chr 15

Conclusion

What we found

- Identified the region of the genome where the mutation is located
- Found out they are all related

How did this happen

It turns out that this was a

- a consanguineous Danish family
- there was a Dutch ancestor

Time for exercises

exercises

go to Absalon

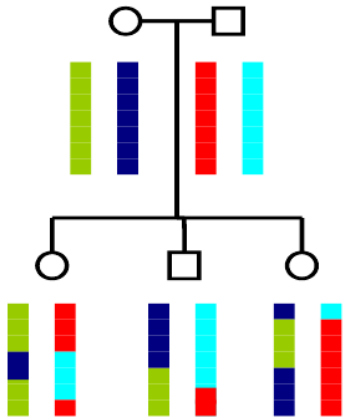
Relatedness

Relatedness definition

IBD sharing between two individuals

Another definition

The IBD is the variability that cannot be explained by the population frequency



Different IBD states

Unobservable IBD patterns

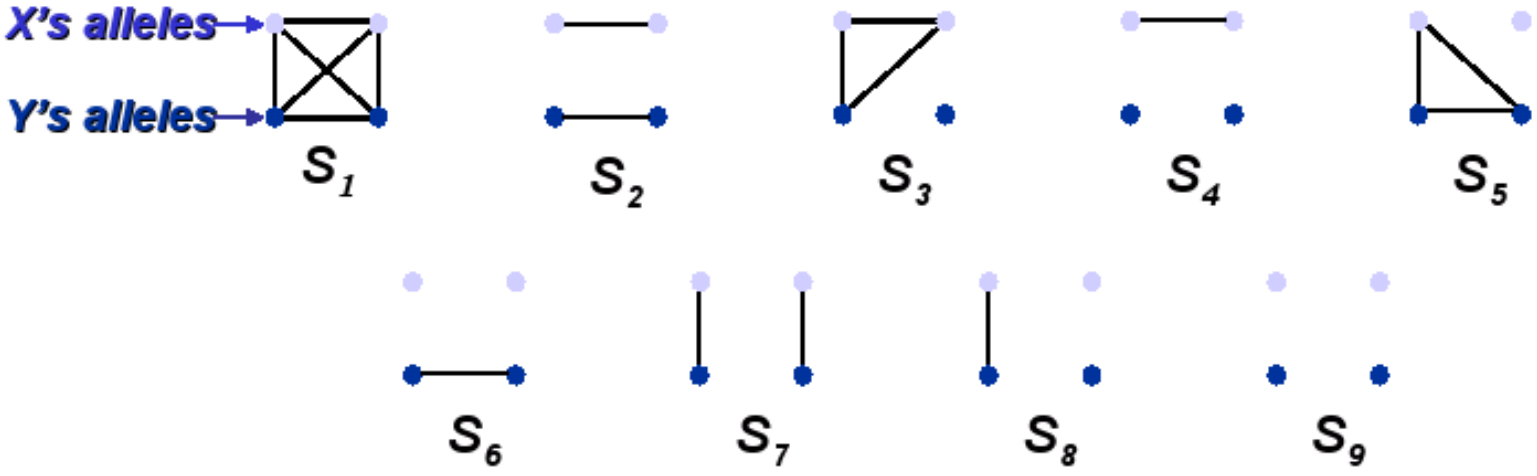


Figure: Assuming symmetry the possible IBD states can be reduced to nine states

Reducing Dimensionality

- If the both parents of X and Y are unrelated their offspring will not be inbreed

Unobservable non-inbreed IBD patterns

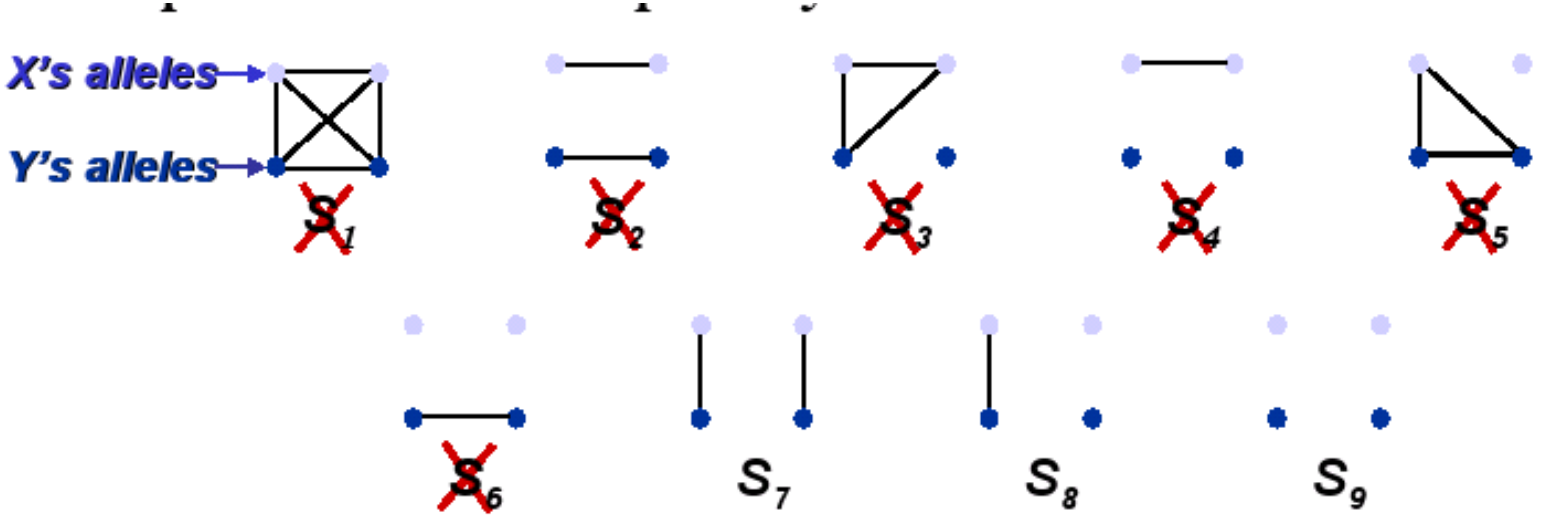


Figure: The probability k_0, k_1 and k_2 of the three posible IBD states are the parameters of interest

IBD and relatedness coefficients (R)

How much a pair shares IBD is quantified in several ways :

1. $\mathbf{R} = (\mathbf{k_0}, \mathbf{k_1}, \mathbf{k_2})$: fractions of genome with 0, 1 and 2 alleles IBD
2. **r**: fraction of alleles they share IBD ($r = 0.5k_1 + k_2$)
3. **kinship**, θ: probability that 2 random alleles are IBD ($\theta = 0.5r$)

Focus: R and how we can estimate it using only genetic information.

Inference of relationships

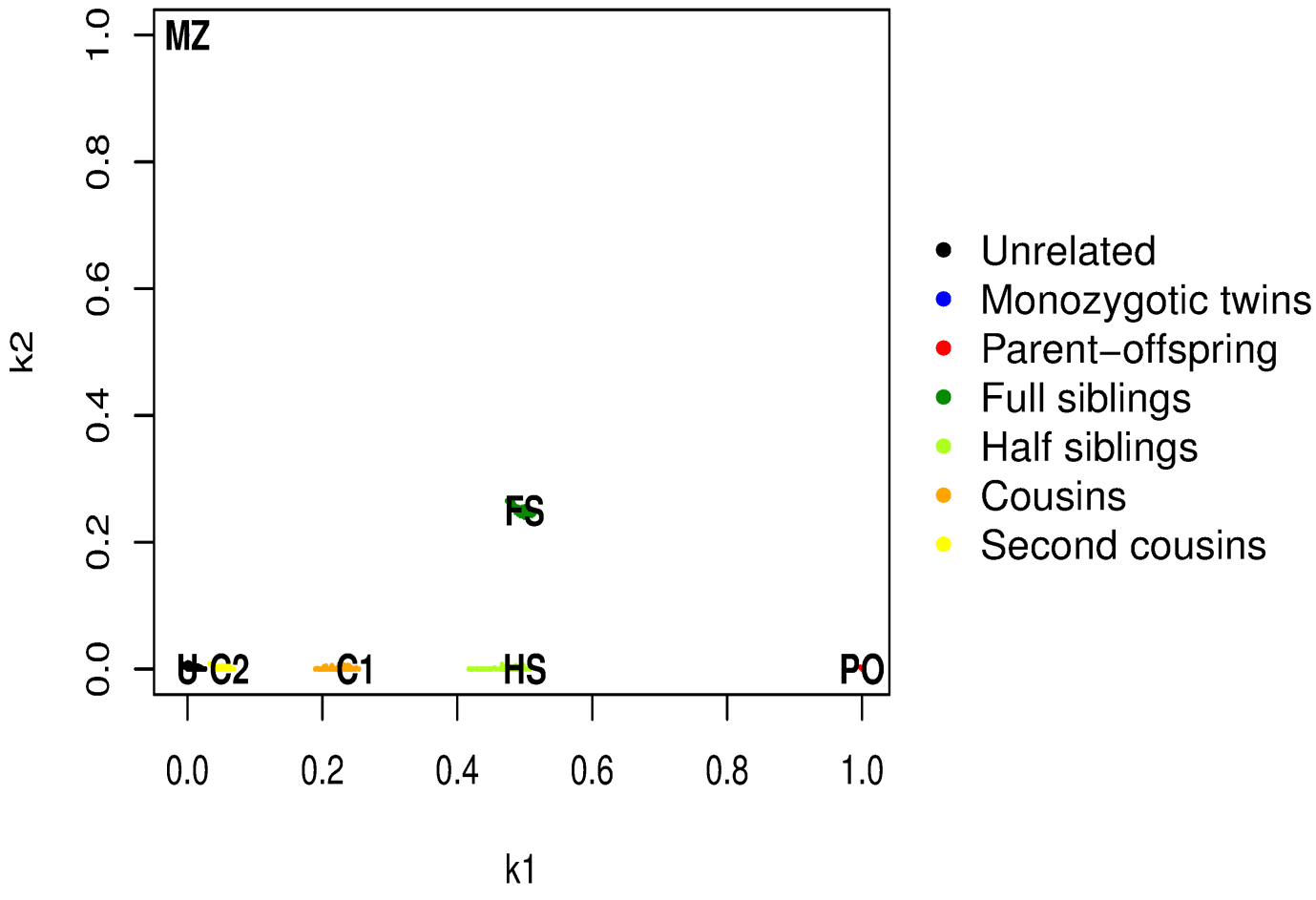
- We expect different relationships to have different values of R, e.g.:

Relationship	k_0	k_1	k_2
Monozogotic twins	0	0	1
Parent-offspring	0	1	0
Siblings	0.25	0.5	0.25
Halfsiblings/Uncle-nephew/grandparent-child	0.5	0.5	0
First cousins	0.75	0.25	0
Second cousins	0.9375	0.0625	0
Unrelated	1	0	0

- Hence we can (often) use R to infer how two individuals are related

Inference of relationships

This can e.g. be done by plotting k_1 against k_2:



Model

- For a pair of non-inbred individuals genotyped in M loci we let
  - G_j be their genotypes in locus j, e.g. $G_j = (AA, aa)$
  - Z_j indicate how many alleles they share IBD in locus j

- For a single locus, j, first we can write:

$$\begin{aligned} P(G_j|R) = P(Z_j = 0|R)P(G_j|Z_j = 0) + \\ P(Z_j = 1|R)P(G_j|Z_j = 1) + \\ P(Z_j = 2|R)P(G_j|Z_j = 2) \end{aligned}$$

- Note that $P(Z_j = i|R)$ is simply k_i for all $i \in \{0, 1, 2\}$, so we get

$$P(G_j|R) = k_0 P(G_j|Z_j = 0) + k_1 P(G_j|Z_j = 1) + k_2 P(G_j|Z_j = 2)$$

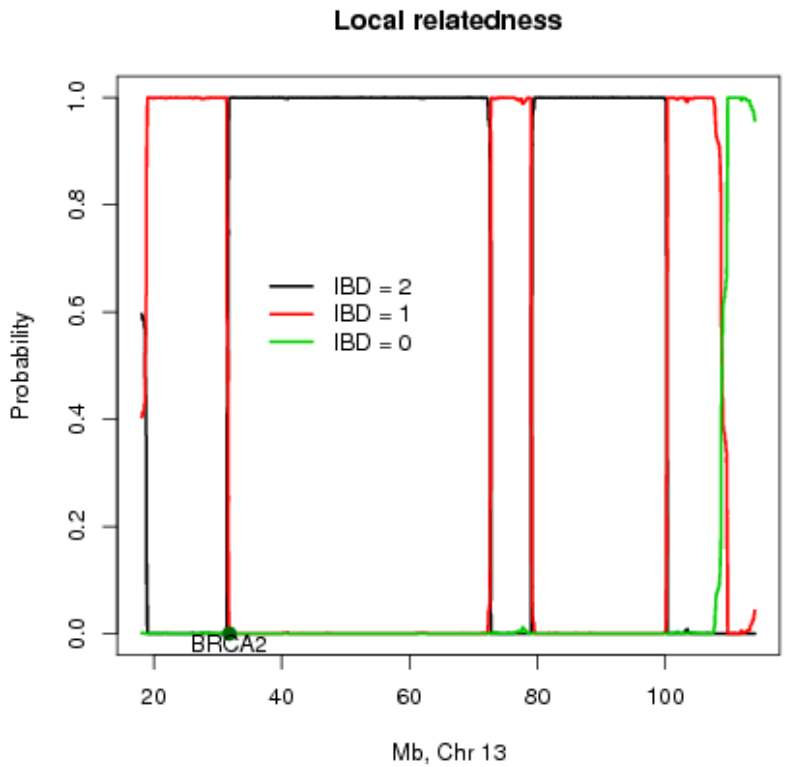
ML inference based on the model

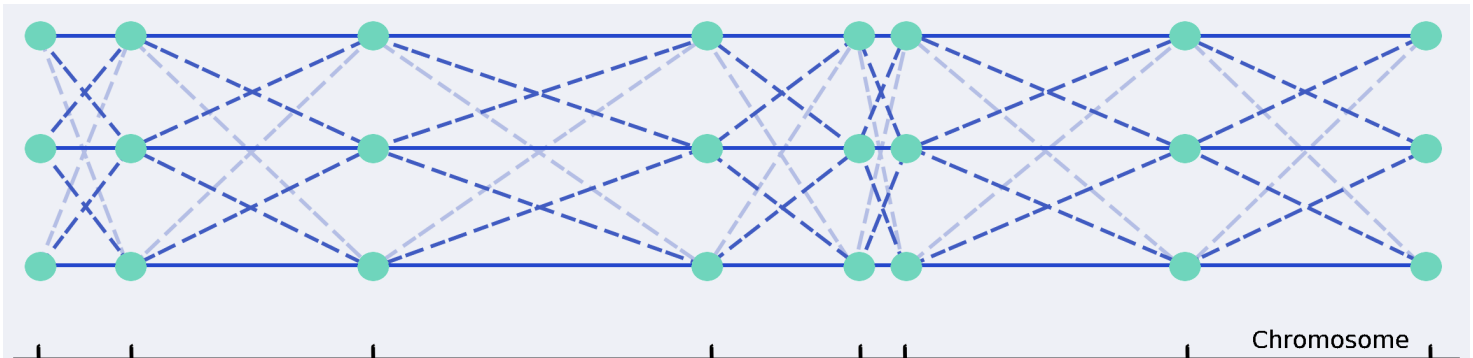
- Assuming loci are independent the model gives the likelihood:

  $P(G_1, G_2, ..., G_M|R) = \prod_{j=1}^{M} P(G_j|R)$

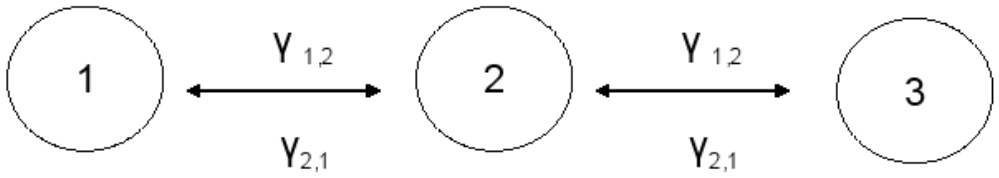
- This function is optimized for R and we get MLE of R
- Most often done using an EM algorithm

Inferred ancestral states using a Markov model





IBD estimation using CTMC



Instantaneous rate matrix

$$Q = \begin{pmatrix} -\alpha k_1 & \alpha k_1 & 0 \\ \alpha k_0 & -\alpha(k_0 + k_2) & \alpha k_2 \\ 0 & \alpha k_1 & -\alpha k_1 \end{pmatrix}$$

The stationary distribution is $K = (k_0, k_1, k_2)$ the fraction shared with $0, 1$ and 2 alleles

The transition probabilities

$$P(X_i = x_i | X_{i-1} = x_{i-1}) = P(X(t + c) = x_i | X(c) = x_{i-1}),$$

where c is the genomic location of the $i - 1$ marker and t is the distance between marker i and $i - 1$

Example of two distantly related individuals

new Definition of IBD

- The IBD is the variability that cannot be explained by the haplotype frequency

IBD is a region of the genome that have not undergone recombination

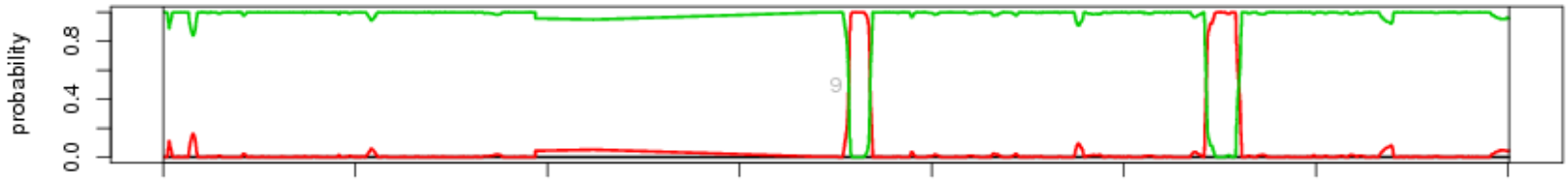
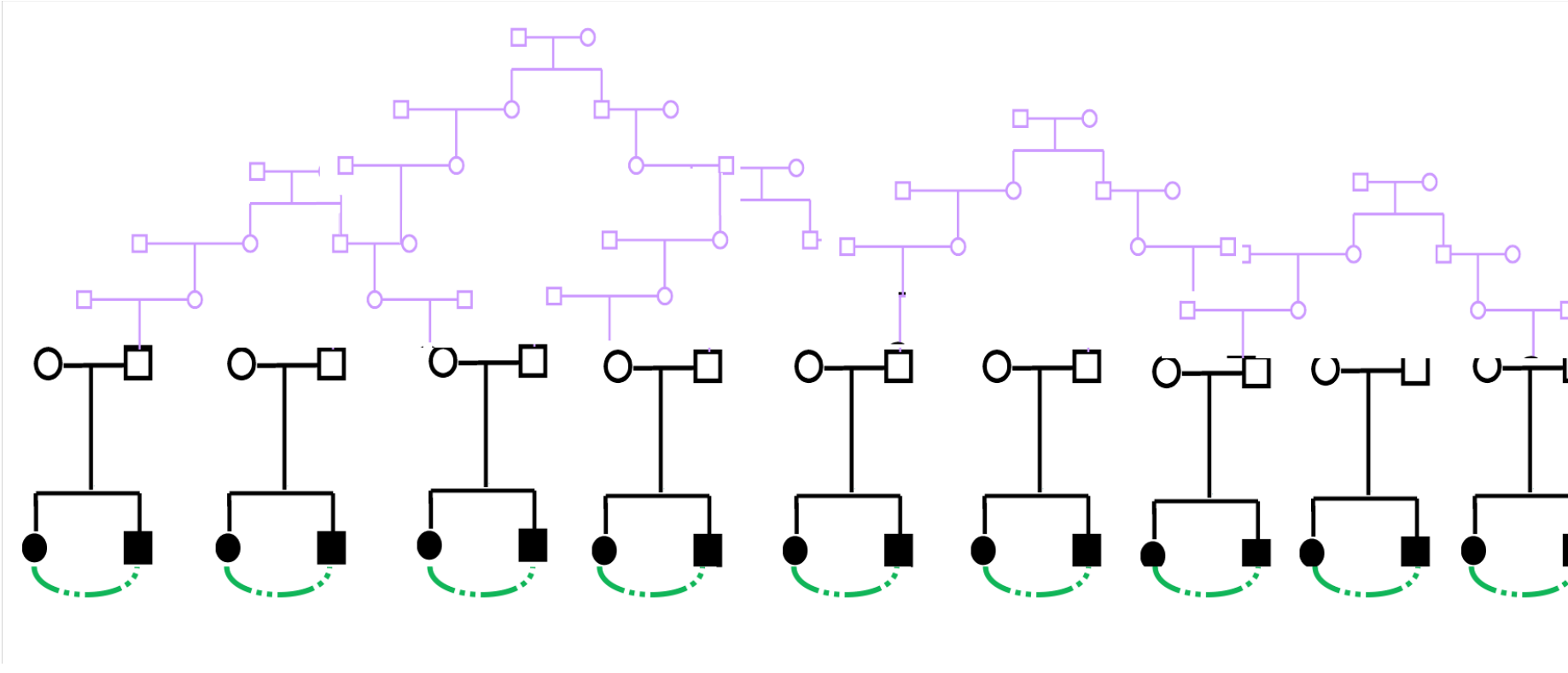


Figure: Chromosome 9 of two distantly related individuals

Relatedness mapping



Deletions in the BRCA1 gene and breast cancer

The data

- 7 Danish unrelated individuals with Breast cancer
- All individuals genotyped for $230,000$ SNPs across the genome

The Questions

- Can we identify the region/gene using only the SNPs
- Are these deletions founder deletions i.e. was the deletion introduced in the population only ones?

Pairwise relatedness

Two individuals

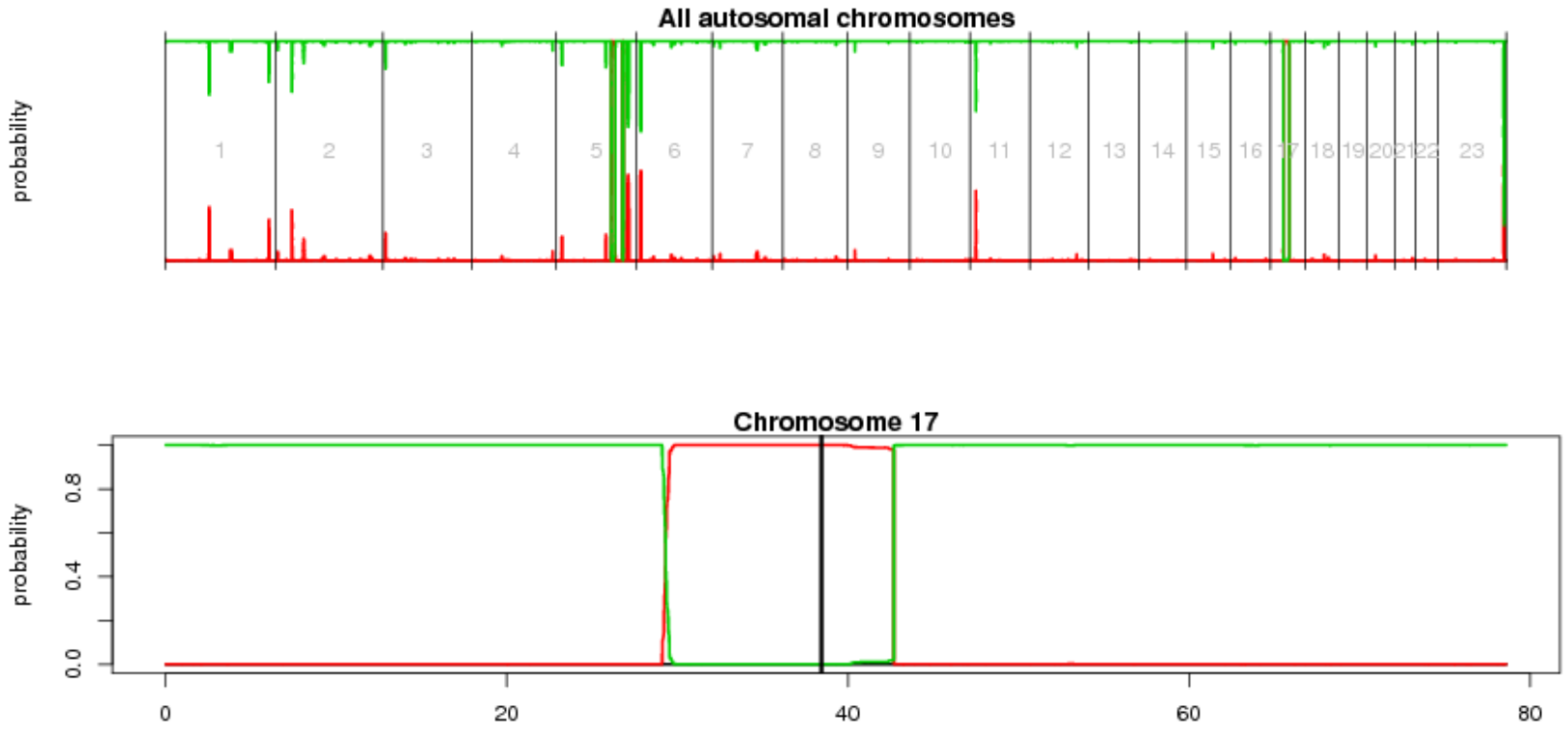


Figure: All chromosomes. Posterior probability for sharing 0 (green), 1 (red) or 2 (black) alleles IBD

Identifying the disease mutation using the only IBD patterns

Combining IBD patters between all pairs of individuals

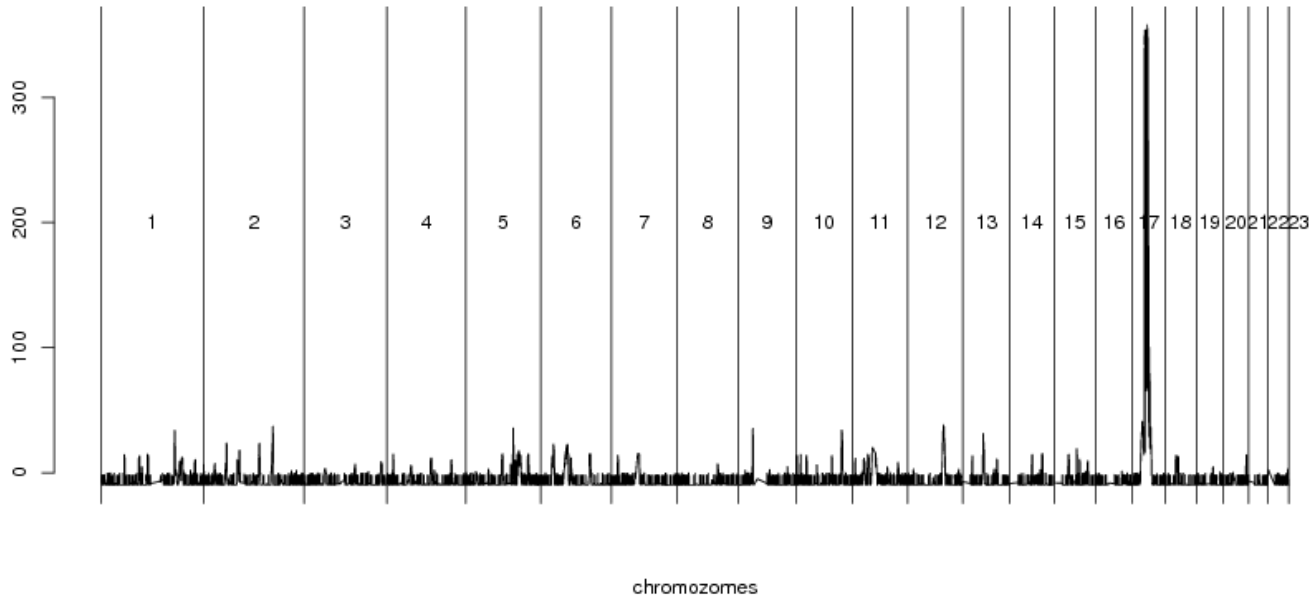


Figure: All chromosome shown

Permutation test

- permute 1000 disease vectors
- calculate your statistic for each vector
- estimate a p-value from the empirical null distribution

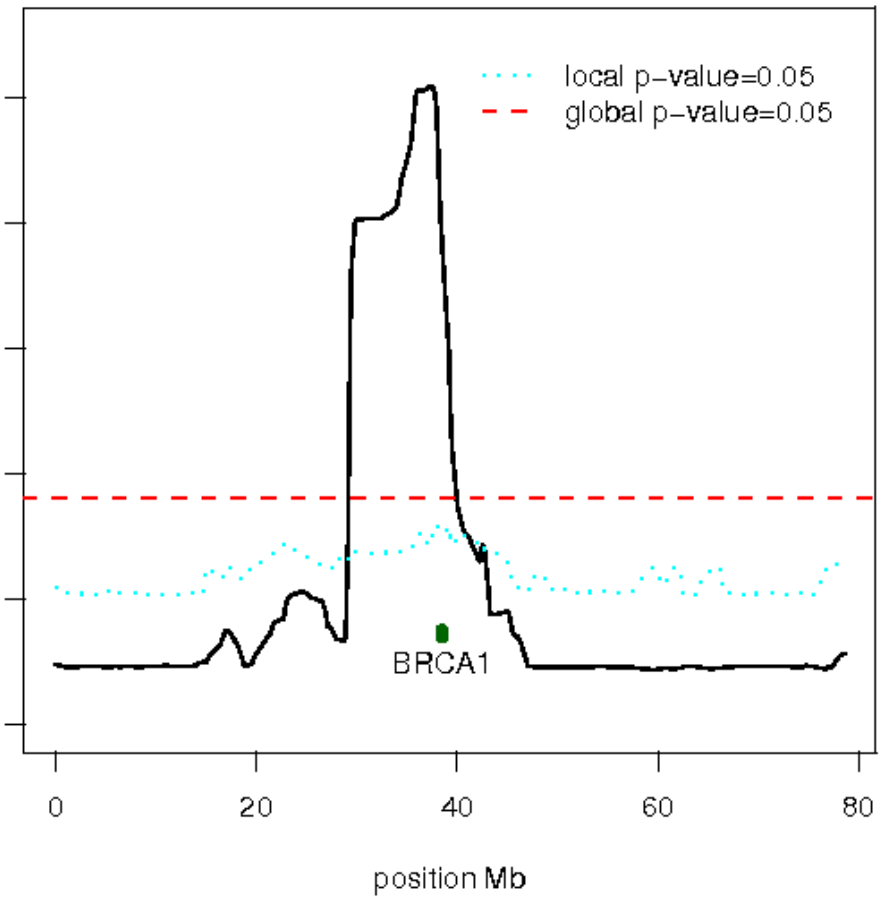


Figure: Statistic for chromosome 17

Is the mutation a founder mutation?

Across the genome

- The IBD sharing is less than 2% for all pairs of individuals

In the BRCA1 gene

- The IBD sharing is more than 95% for all pairs of individuals

conclusion

- The mutation is a founder mutation shared IBD between all individuals

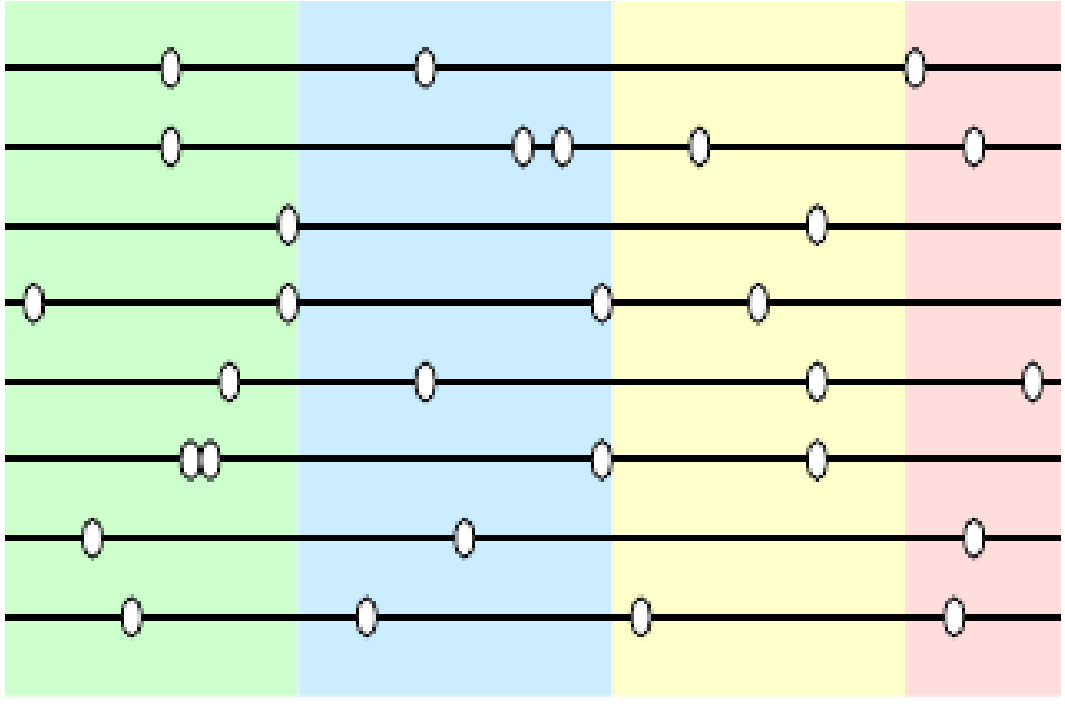
Exercise

exercise

go to absalon

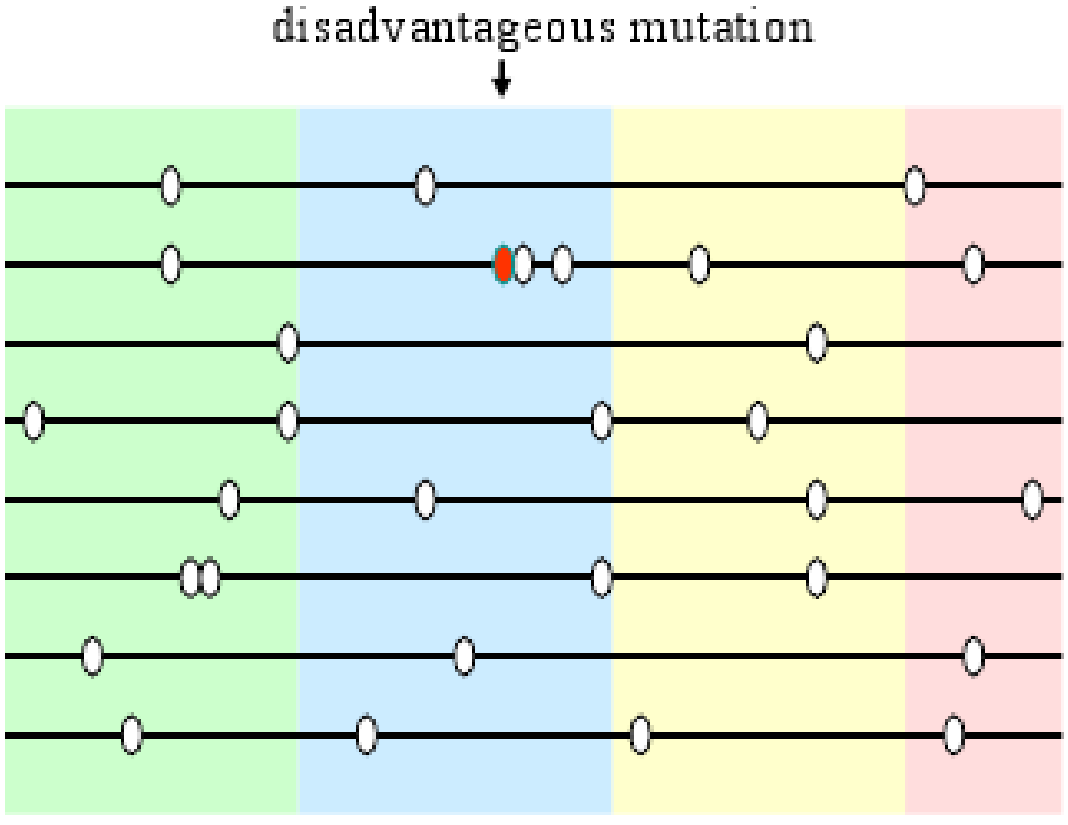
Signature of selection

- Neutral locus
- Lots of variability



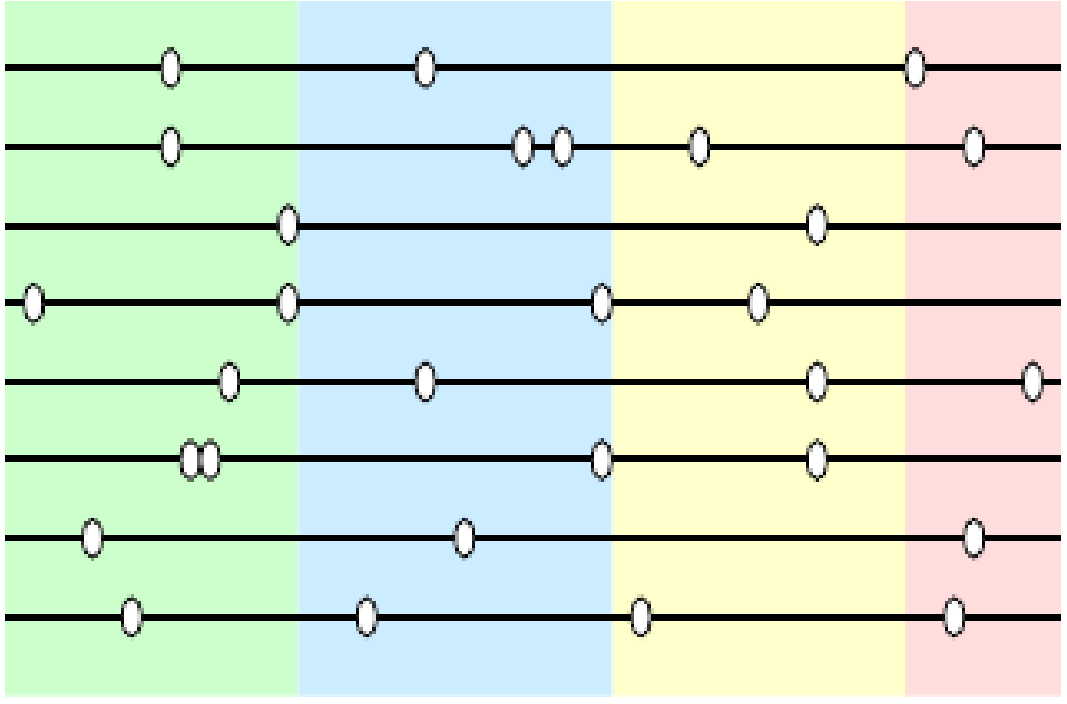
Signature of selection

- Mutation enters the population



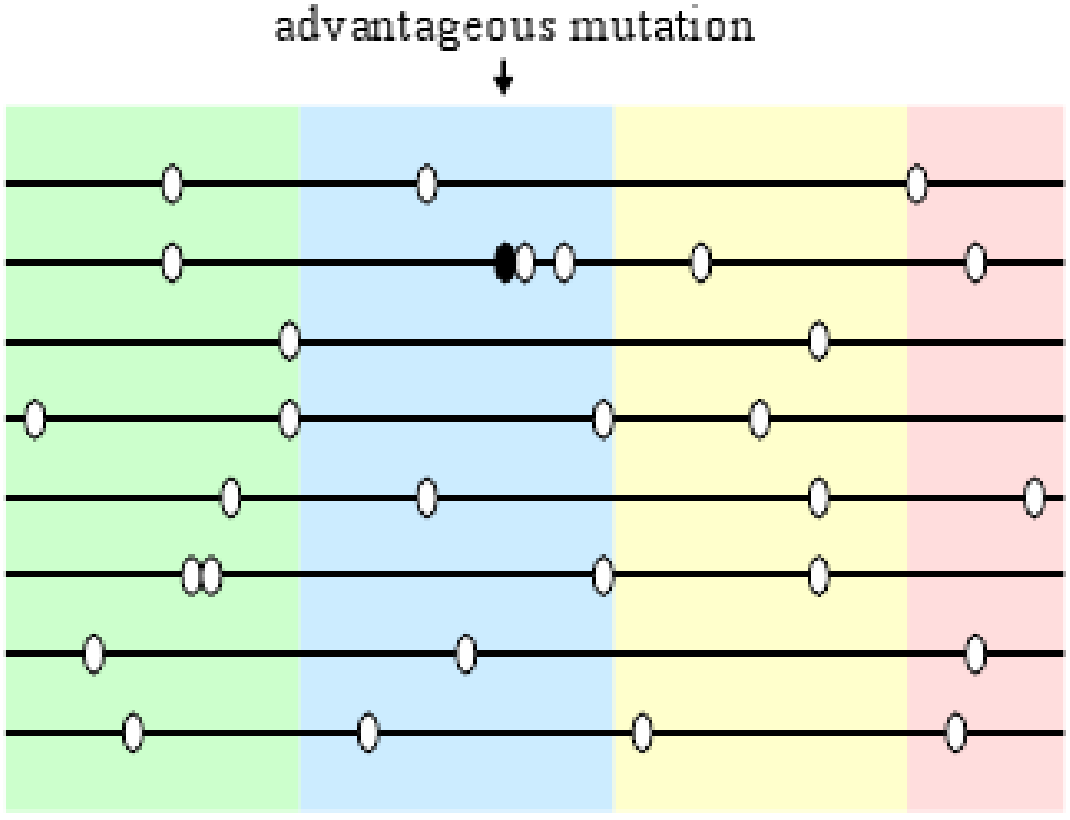
Signature of selection

- Negative selection removed the allele



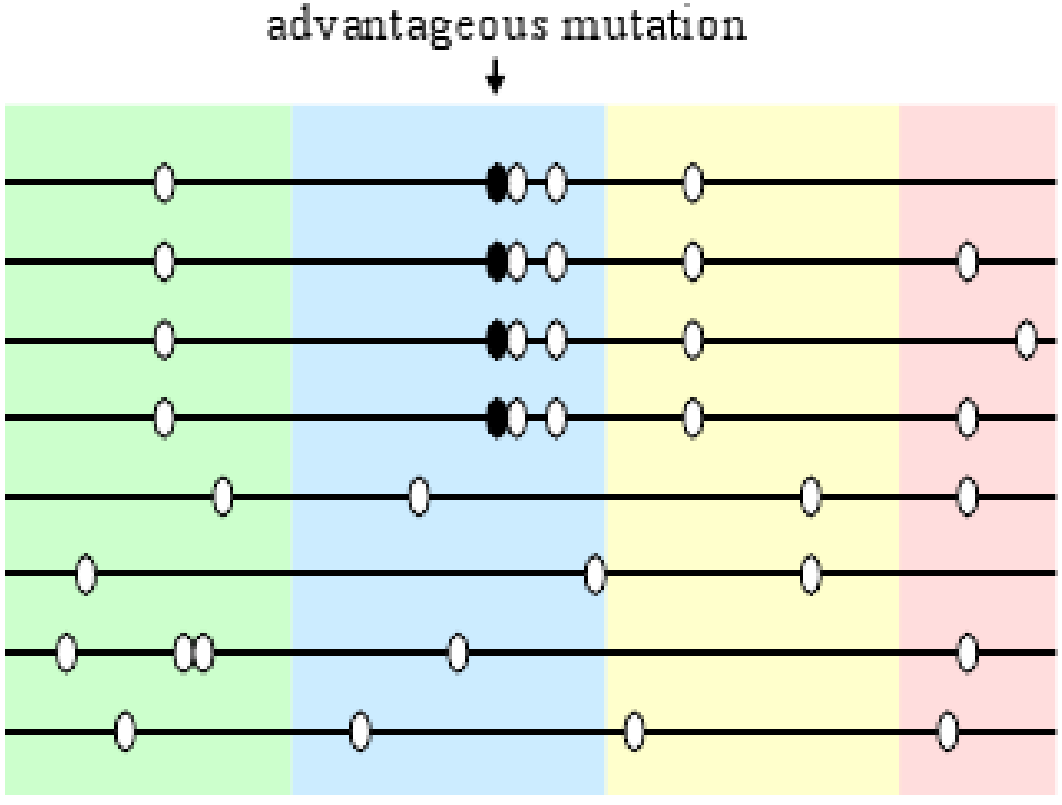
Signature of selection

- Mutation enters the population



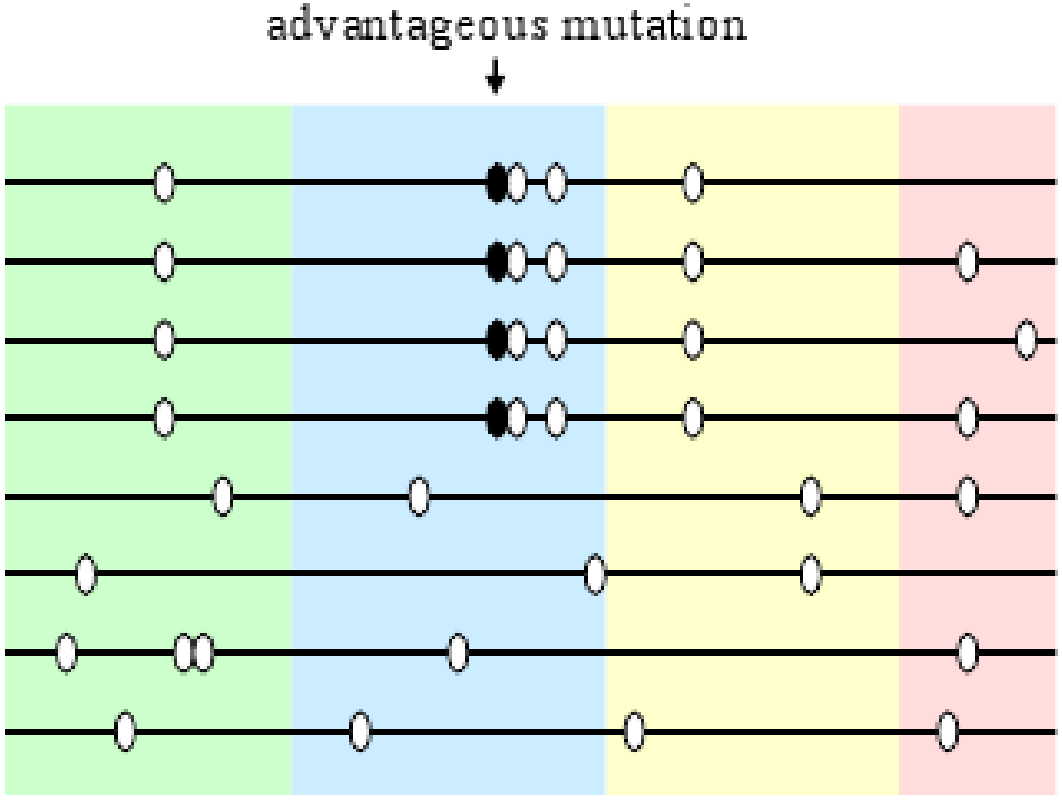
Signature of selection

- Mutation enters the population
- Mutation increases in frequency due to positive selection

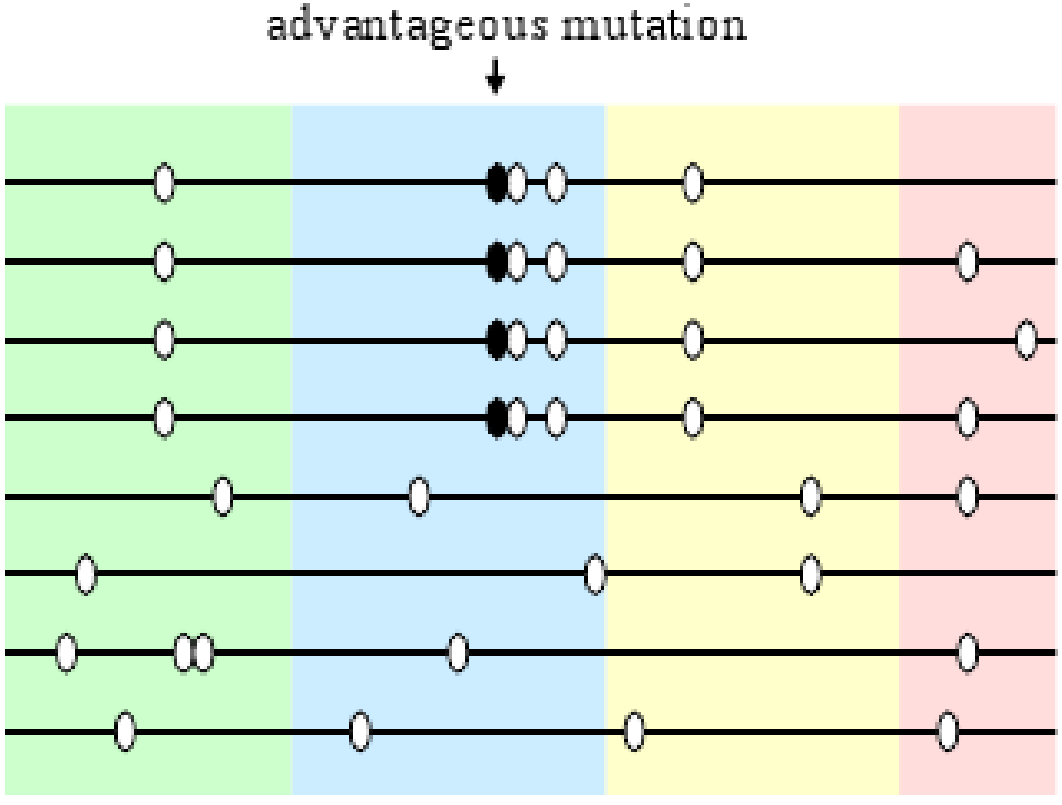


Signature of selection

- Increases LD
- Affects the variability



Signature of selection



- Increases IBD!

Deterministic approximation in continuous time

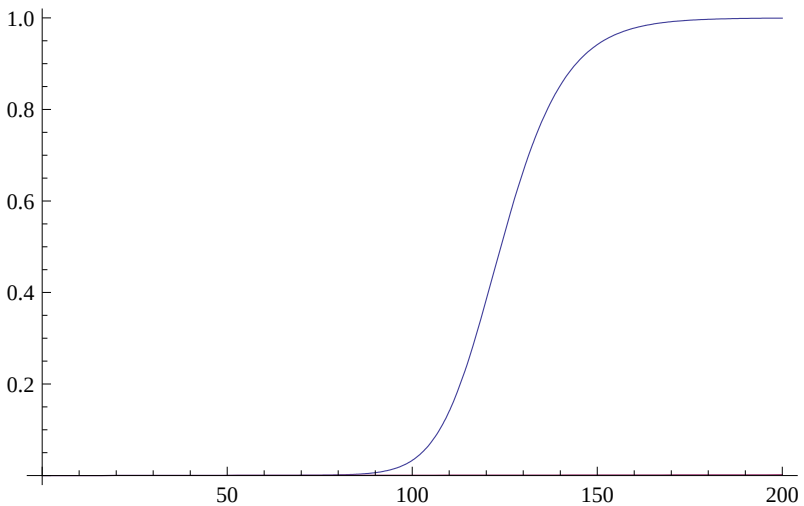


Figure: Positive selection on a new allele, 2N=100,000, s=0.1

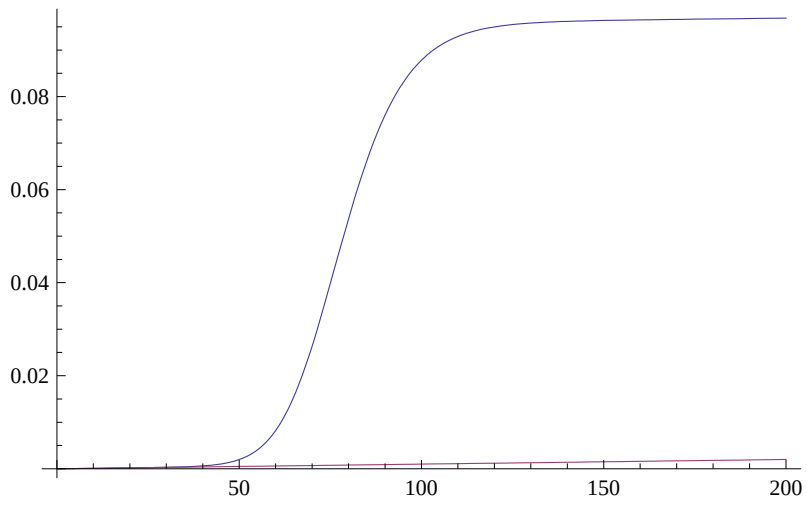
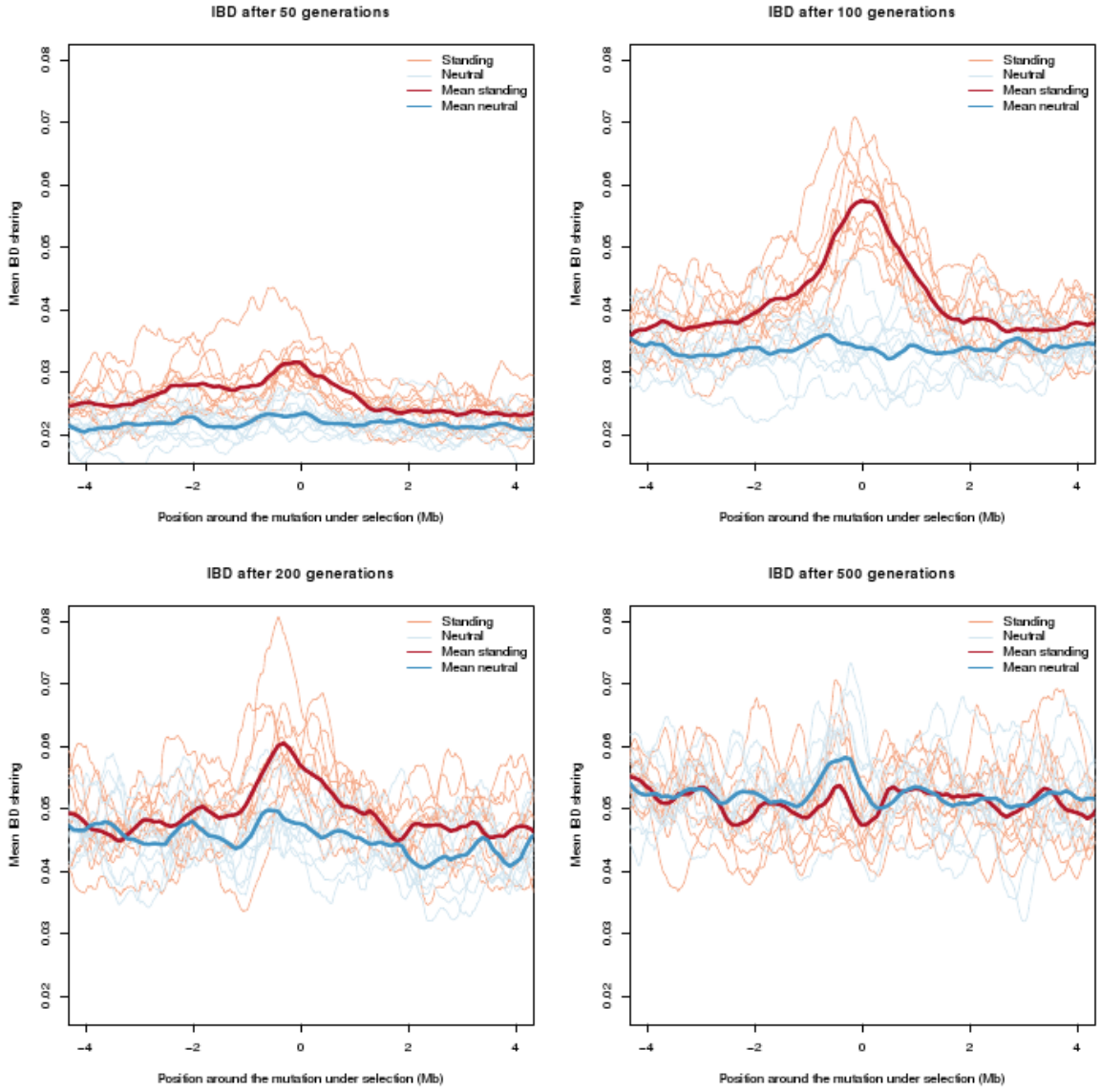


Figure: Selection on an allele of frequency 1%, 2N=100,000, s=0.1

Simulations of selection on standing variation

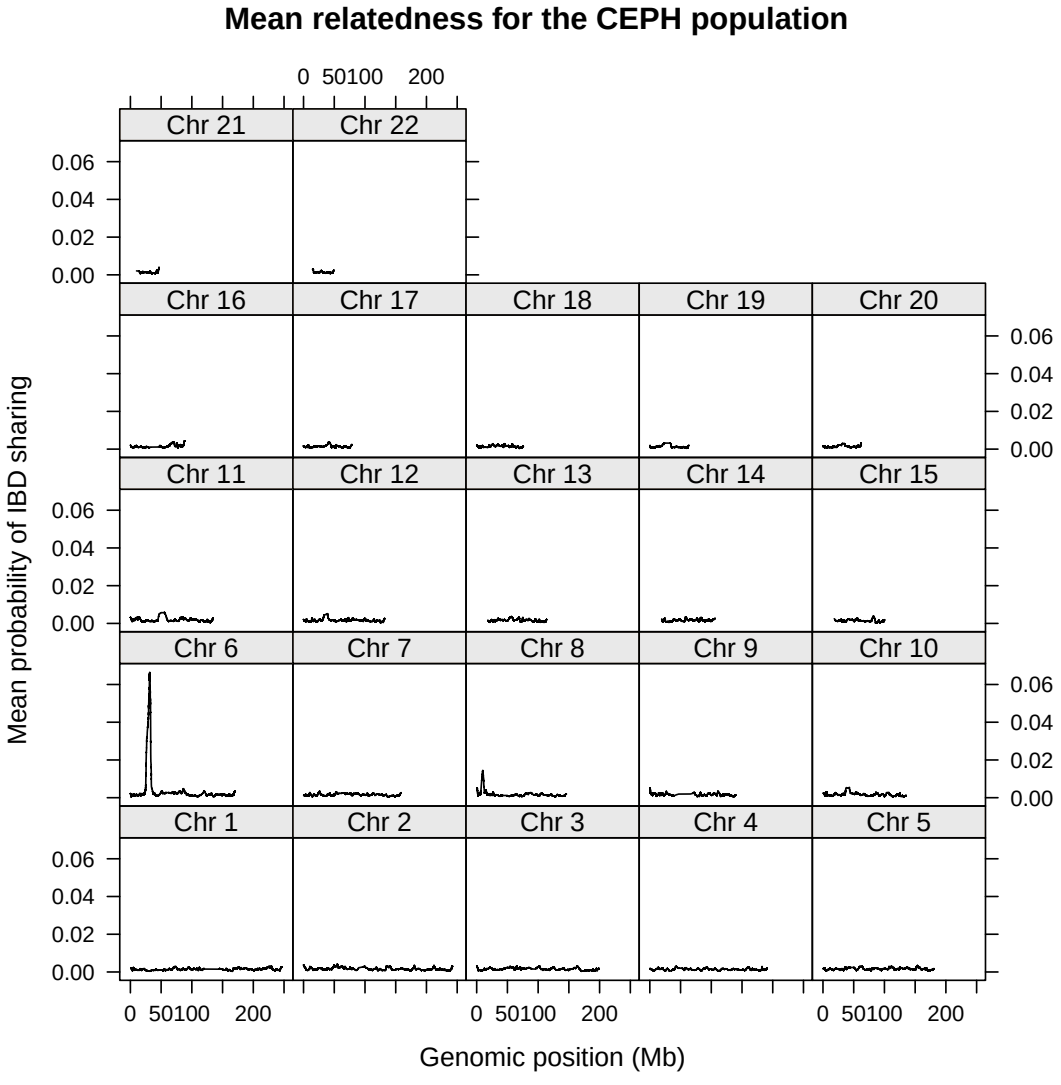


Detecting recent and strong selection on HapMap3 data

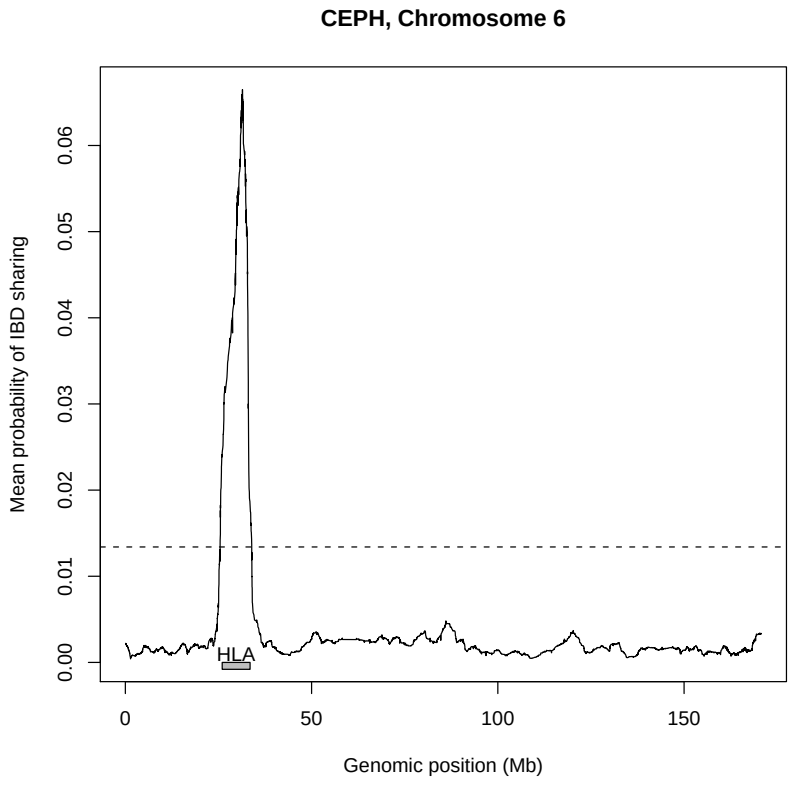
	Population	# individuals	# Unrelated	#SNP*
ASW	African ancestry in Southwest USA	90	42	120030
CEU	North European ancestry (CEPH)	180	109	105291
CHB	Han Chinese in Beijing, China	90	82	88533
CHD	Chinese in Denver	100	70	85901
GIH	Gujarati Indians in Houston	100	83	106768
JPT	Japanese in Tokyo, Japan	91	82	84557
LWK	Luhya in Webuye, Kenya	100	83	193280
MEX	Mexican ancestry in Los Angeles	90	45	94474
MKK	Maasai in Kinyawa, Kenya	180	143	196540
TSI	Toscans in Italy	100	77	104540
YRI	Yoruba in Ibadan, Nigeria	180	108	193373

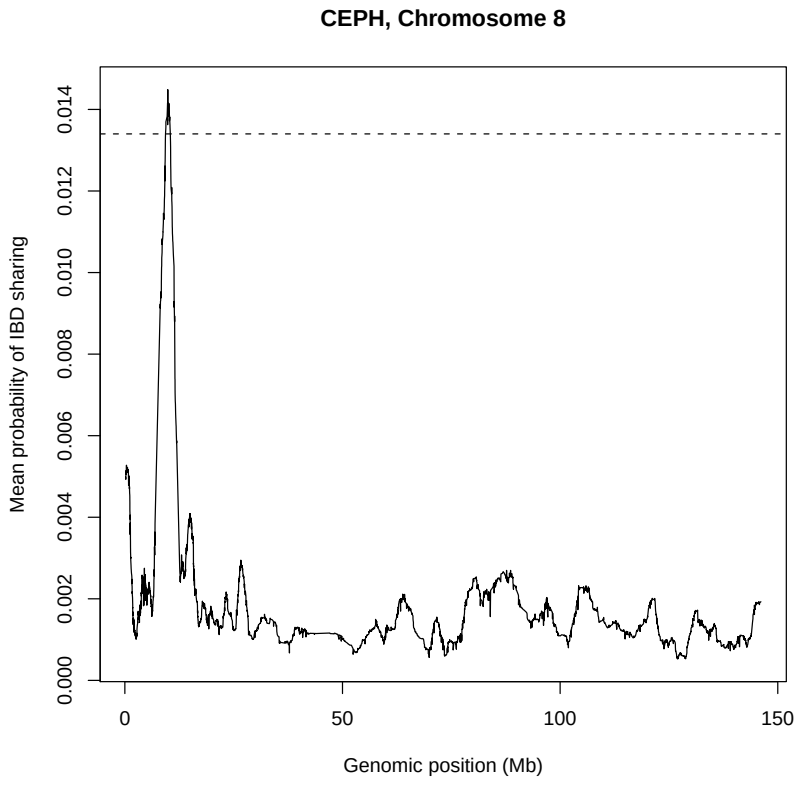
Table: Table of the HapMap phase 3 individuals used in this study. The #SNP* column contains the number of SNPs left after data cleaning and LD removal.

Results in Europeans

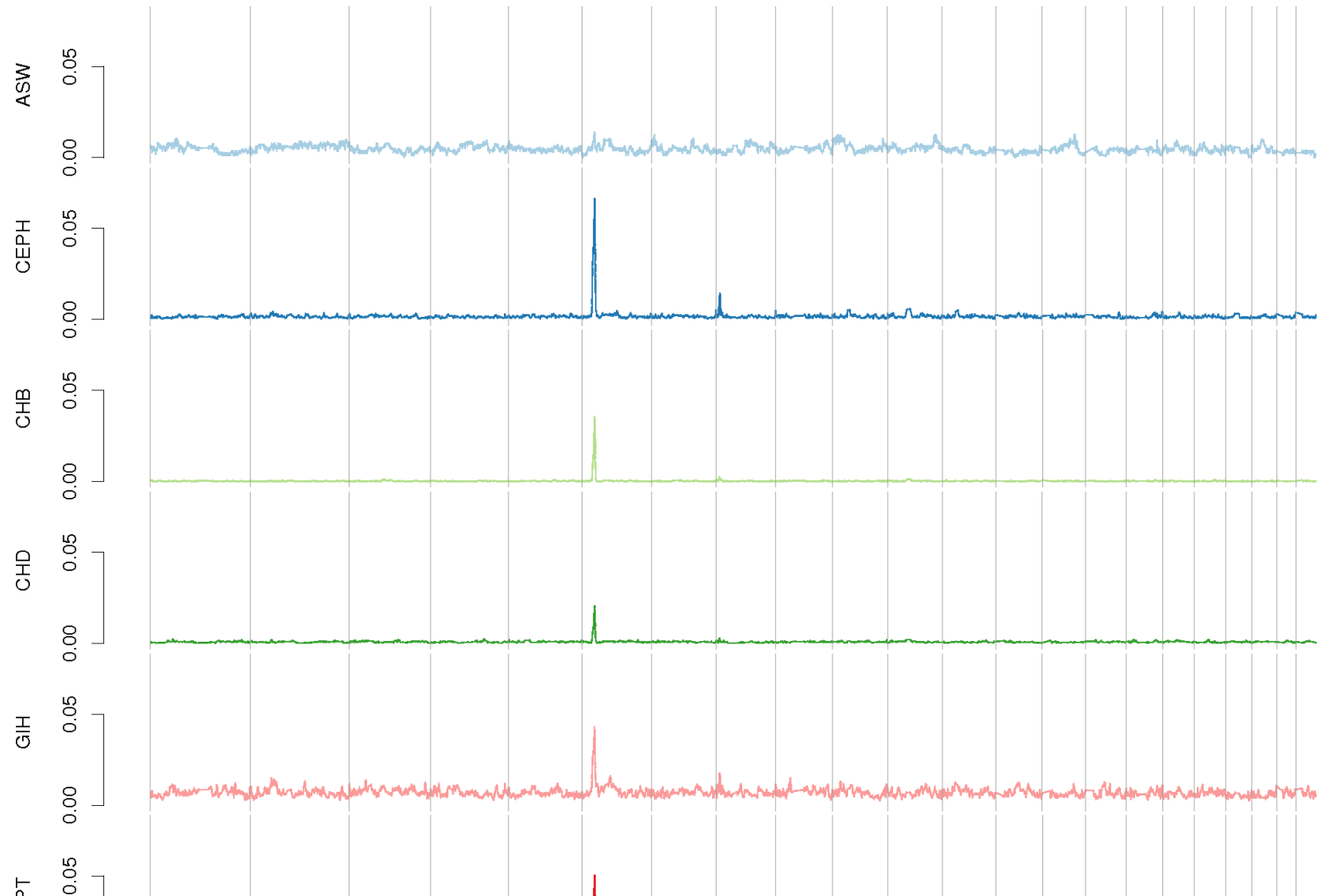


Results in Europeans

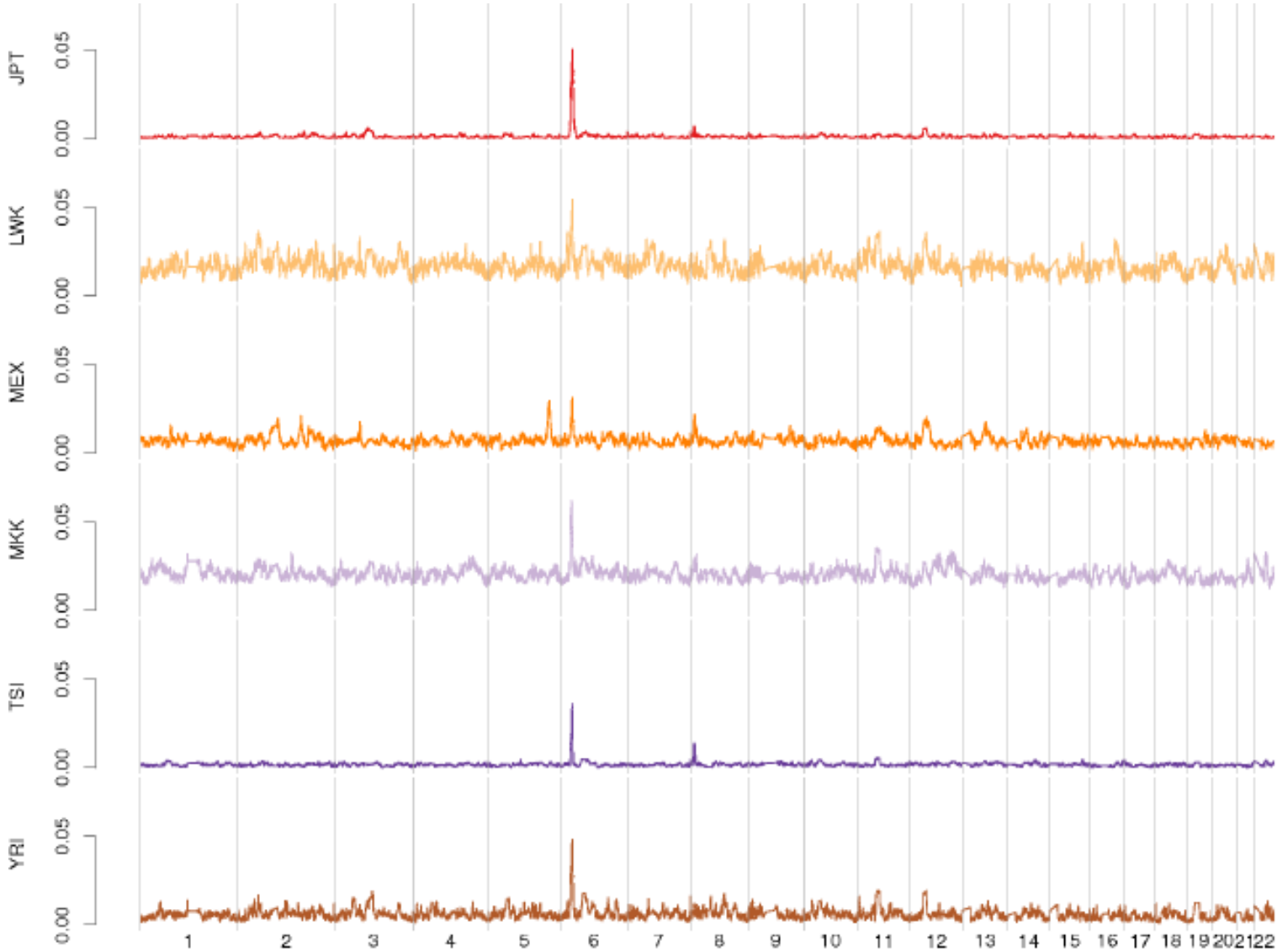




Results in all populations



Results in all populations



Conclusion

- IBD have different definitions
- IBD definitions can conveniently be extended of local IBD estimations
- IBD regions can be inferred using variable sites
- The local IBD tracts can be used for disease mapping
- By inferring the IBD tracts in a population we detect regions known to be under selection
  - The HLA region shows the highest amount of IBD in all populations