Announcements

- Today last lecture, afterwards tutorial for biologists
- Next Thursday Dec.16.:
  – last tutorial for comp.sci.
  – optional presentations of journal club during lecture time (interest?)
  – tutorial for biologists possibly cancelled
- End of semester deadlines
  – journal club reports Friday Dec. 17
  – HW3 Tuesday Dec. 21

Exam (comp.sci. only)

The main part is **written:**

- You need at least 50% of points
- Time 3 hours
- About 50% of points for simple questions,
  – examples will be on the course website
  – in case of interest tutorial session before exam
- The rest of the questions mostly designing/modifying an algorithm or model
- **Date?**
- Online or in person, depending on circumstances
- You can use pen, simple calculator
  and a cheat sheet up to 2 A4 two-sided sheets

Written exam, online version (comp.sci. only)

- Exam questions and submission in Moodle
- MS teams: annoucements, questions
- Write in an editor, create pdf
  or write on paper, scan/photo, convert to pdf
- Allowed aids:
  Same as in person (incl. cheat sheet)
  Text and image editors, software for digitization of handwritten pages
  MS Teams to communicate with instructors
  Moodle for getting and submitting exam
- Not allowed:
  Communication with other persons except instructors
  Other webpages
  Other software (e.g. specialized bioinformatics programs, compilers)

Oral exam

- Only for online exam
- Videocall in MS Teams
- After written exam, time slots over several days
- We will discuss your exam
- You should be able to explain your answers in detail
- Oral exam influences exam grade
- If you are unable to explain your answers, you will get Fx

"Second chance" exam: the same for as the first or oral-only
the dates arranged with those who need them

Population Genetics

Broňa Brejová

December 9, 2021

Population genetics

- Genomes of different individuals of the same species differ
- These differences cause differences in phenotype (appearance, behaviour, diseases,. . .)
- We can sequence multiple individuals and compare with reference sequence

Possible applications:

- Impact of individual genetic differences
- History and structure of populations (subpopulations, migration, historical changes in size)

SNPs (Single Nucleotide Polymorphisms)

- SNP: a single base mutation (present in $> 1\%$ individuals)
- Usually only two forms : **major** and **minor** allele
- Small change at some places in the genome can cause large phenotypic changes

Systematic mapping of SNPs:

1000 Genomes Project 2008-2015
identify 95% of SNPs with 1% minor allele frequency
using next generation genome sequencing

Trait/Disease Association Mapping

- Traits and diseases emerge by the combination of genetic and environmental influences
- Goal: Identify genetic influences.
  - Disease mechanisms?
  - What is the risk of inheritance?
  - How can we design and target new drugs (pharmacogenomics)?
    E.g. mutations of cytochrome family P450 genes
    influence metabolism of drugs in the liver,
    thus influence necessary dose

Diploid genomes

- Human has a **diploid genome**:
  each human cell contains two copies of chromosomes 1...22
  plus sex chromosomes X,X or X,Y
- From each pair, one chromosome comes from mother and one from father
- For a SNP with alleles (forms) a and A,
  an individual is **homozygote** (aa or AA),
  or **heterozygote** (aA)
- A disease caused by allele a can appear only in homozygotes aa,
  or also in heterozygotes aA, or more severe for aa than aA

Diploid genomes

- Human has a **diploid genome**:
  each human cell contains two copies of chromosomes $1 \ldots 22$
  plus sex chromosomes X,X or X,Y
- From each pair, one chromosome comes from mother and one from father
- For a SNP with alleles (forms) a and A,
  an individual is **homozygote** (aa or AA),
  or **heterozygote** (aA)
- **Haplotype:** combination of alleles of different SNPs on the same chromosome (inherited from one parent)
  Diploid individual has two haplotypes

chr1 from mother: $\ldots$ A $\ldots$ T $\ldots$ G $\ldots$ $\ldots$

chr1 from father: $\ldots$ T $\ldots$ C $\ldots$ A $\ldots$ $\ldots$

Testing a single SNP

Contingency table - the number of haplotypes

Dog size vs allele at chr15:44,228,468 [Sutter et al., 2007]

	allele A	allele a	total
small dog (< 9 kg)	14	535	549
large dog (> 31 kg)	339	38	377
total	353	573	

Test if columns and rows are **independent (null hypothesis)**

If null hypothesis **rejected**, there is association between SNP and size (not necessarily causal)

If null hypothesis **not rejected**, association not found (perhaps will be found with more data)

Testing independence in a contigency table

	allele A	allele a	total
small dog	14	535	549
large dog	339	38	377
total	353	573	926

Fisher's exact test: (Fisher's exact test) exact probability from hypergeometric distribution

χ^2 test (chí-kvadrát): popular approximate test, appropriate for large counts

In practice also more complex statistical methods / models (diploid genome, family relationships, . . .)

Testing independence in a contingency table by χ^2 test

	allele A	allele a	total
small dog	14	535	549
large dog	339	38	377
total	353	573	926

Under null hypothesis (independence of rows ans columns):

$\Pr(A) = 353/926 = 0.381$, $\Pr(a) = 0.619$

$\Pr(s) = 549/926 = 0.593$, $\Pr(l) = 0.407$

$\Pr(A, s) = \Pr(A)\Pr(s) = 0.226$

$\Pr(a, s) = \Pr(a)\Pr(s) = 0.367$

$\Pr(A, l) = \Pr(A)\Pr(l) = 0.155$

$\Pr(a, l) = \Pr(a)\Pr(l) = 0.252$

Under the null hypothesis we expect 926 haplotypes in the table divided in ratios 0.226:0.367:0.155:0.252

Testing independence in a contingency table by χ^2 test

Real table

$O_{i,j}$ (observed):

	A	a	total
small	14	535	549
large	339	38	377
total	353	573	926

Expected under null

$E_{i,j}$ (expected):

	A	a	total
small	209.3	339.8	549
large	143.5	233.4	377
total	353	573	926

Compute $\chi^2 = \sum_{i \in \{s,l\}} \sum_{j \in \{A,a\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

$\chi^2 = (14 - 209.3)^2/209.3 + (535 - 339.8)^2/339.8 + (339 - 143.5)^2/143.5 + (38 - 233.4)^2/233.4 = 724.3$

χ^2 is a measure of difference between tables O and E.

Always $\chi^2 \geq 0$, and $\chi^2 = 0$ only if tables equal.

Testing independence in a contingency table by χ^2 test

$O_{i,j}$ (observed):

	A	a	total
small	14	535	549
large	339	38	377
total	353	573	926

$E_{i,j}$ (expected):

	A	a	total
small	209.3	339.8	549
large	143.5	233.4	377
total	353	573	926

Compute $\chi^2 = \sum_{i \in \{s,l\}} \sum_{j \in \{A,a\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 724.3$

Under null hypothesis, χ^2 is approximately from $\chi^2(1)$ distribution, i.e. **chi squared with one degree of freedom**.

1 degree: if we know E and 1 number from O, the rest of O can be computed

The probability that under null we get by chance $\chi^2 \geq 724.3$ is $1.6 \cdot 10^{-159}$ (P-value)

To **reject null hypothesis** use threshold e.g. $P < 0.05$, $\chi^2 > 3.841$

Dependencies between two different SNPs

Consider SNP with alleles p/P and another with alleles q/Q.
Count haplotypes pq, PQ, pQ, Pq

Example: 2000 haplotypes (1000 individuals)

	Q	q
P	474	611
p	142	773

$\chi^2 = 184.78$, P-value $4.4 \cdot 10^{-42}$

Columns and rows not independent, dependency between the SNPs

Example 2: Similar ratios of counts, but only 30 haplotypes:

	Q	q
P	7	9
p	2	12

$\chi^2 = 3.0867$, P-value 0.07893

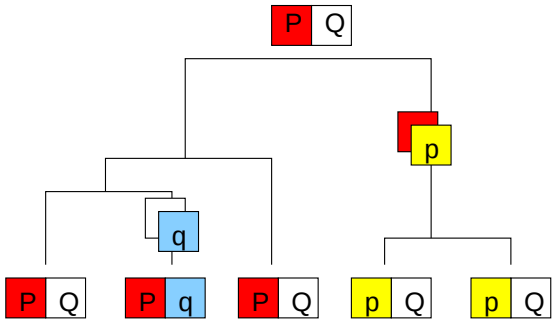
Null hypothesis not rejected for threshold P<0.05 ($\chi^2 > 3.841$)
Beware, χ^2 not appropriate for such low counts

Why are SNPs dependent?

SNPs on different chromosomes:

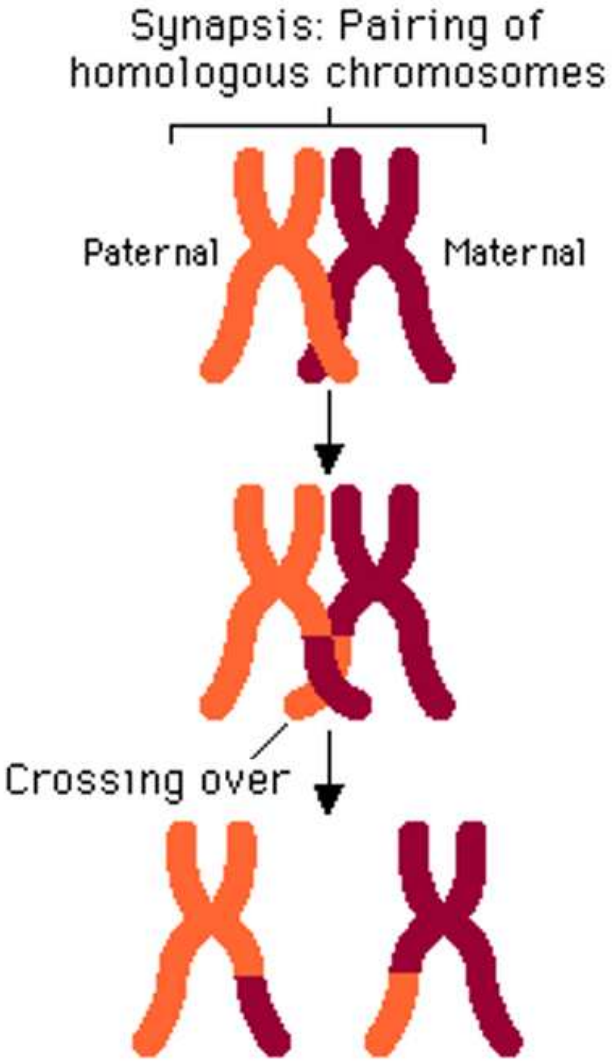
- Probabilities of individual alleles often independent
- $\Pr(pq) = \Pr(p)\Pr(q)$, $\Pr(PQ) = \Pr(P)\Pr(Q)$, etc.
- **linkage equilibrium (LE, väzbová rovnováha)**

SNPs nearby on the same chromosome:



- The same mutation happening twice is rare, recombination also relatively rare
- Allele combinations not completely random
- Correleation between SNPs
  $\Rightarrow$ **linkage disequilibrium (LD, väzbová nerovnováha)**

Recombination



Approx. 1-3 **recombinations** on 1 human chromosome during meiosis (production of sperm/eggs)

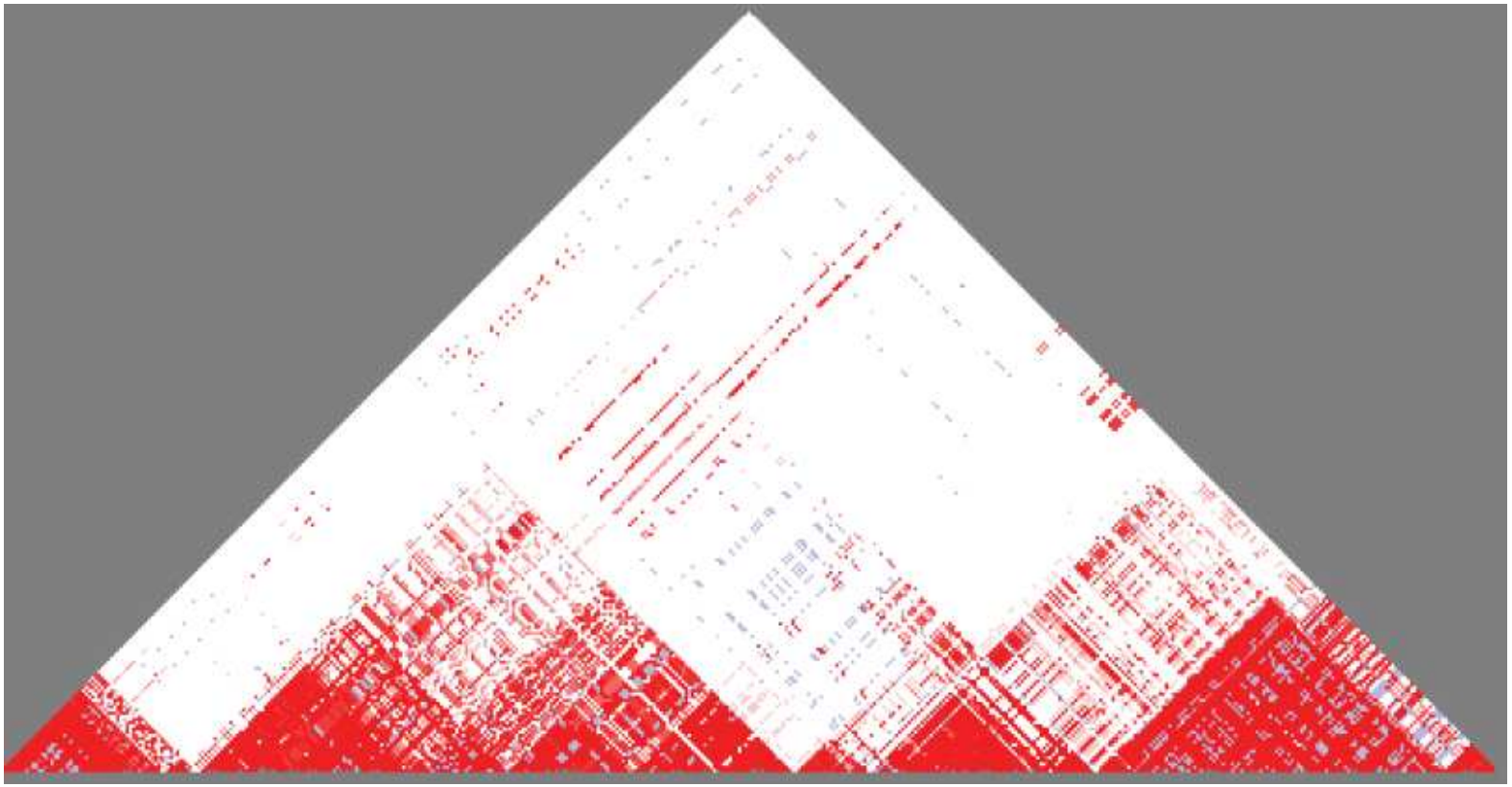
Recombination lowers LD

Assuming uniform recombination

- LD decreases with SNP distance on a chromosome
- LD decreases with SNP age
- Other factors: population structure, natural selection, recombination hotspots

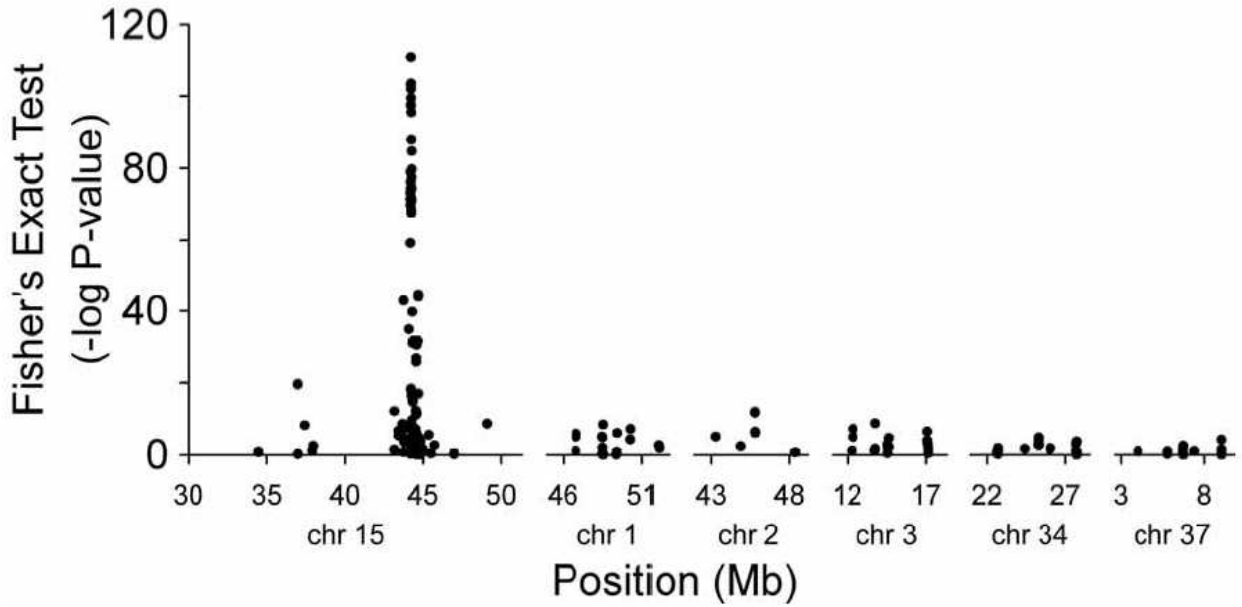
Linkage disequilibrium (LD) in the human genome

[The International HapMap Consortium, 2005]



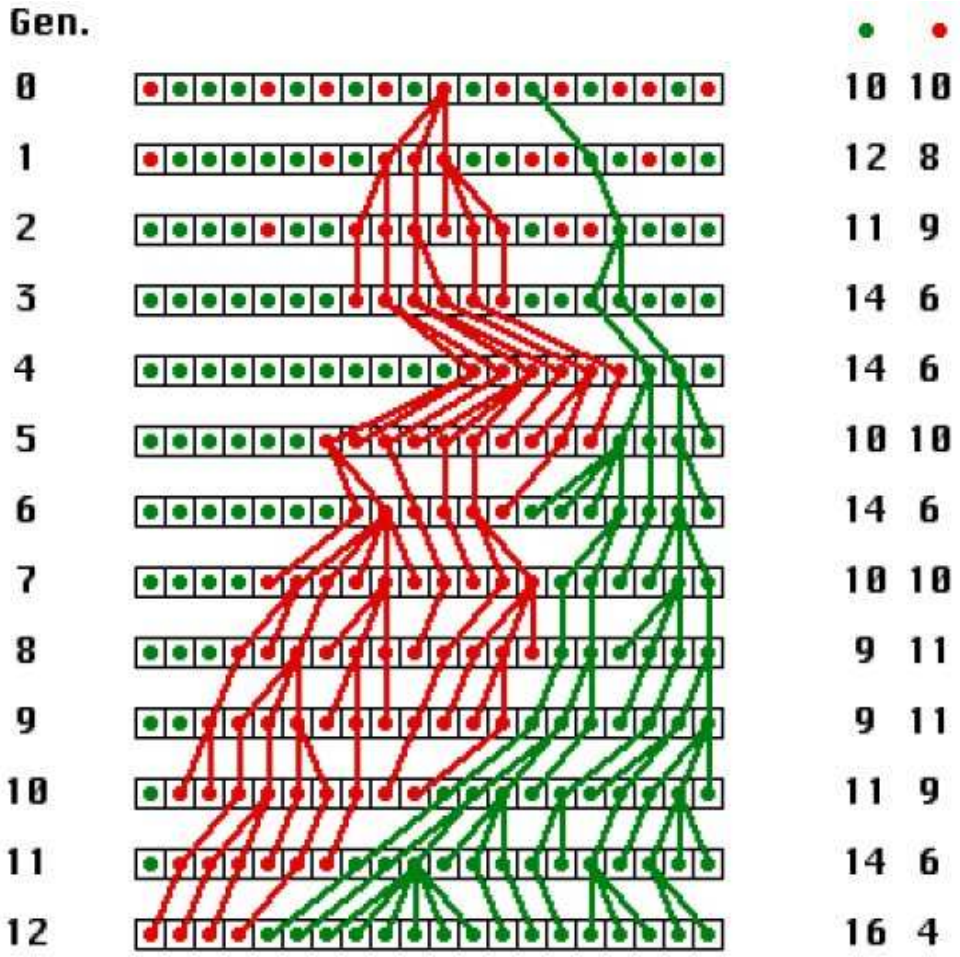
Region ENm014 (500kB, chr 7), 90 people from Utah

Back to dogs: Whole-Genome Association Scan (WGAS)



- For dog size, WGAS identified 84 kB region
- Causal SNP has to be more finely mapped by additional experiments
- **Large LD blocks** $\Rightarrow$ only can identify large regions

Basic model of population genetics: Wright-Fischer model



Lifecycle of SNPs in Wright-Fisher model

- Population of N haploid organisms
- One allele per organism (A or a)
- New generation created as a copy of a random parent (random mating), no influence of natural selection
- X_t: the count of allele a in generation t
- **Markov chain** with states $X_t \in \{0, 1, \ldots, N\}$

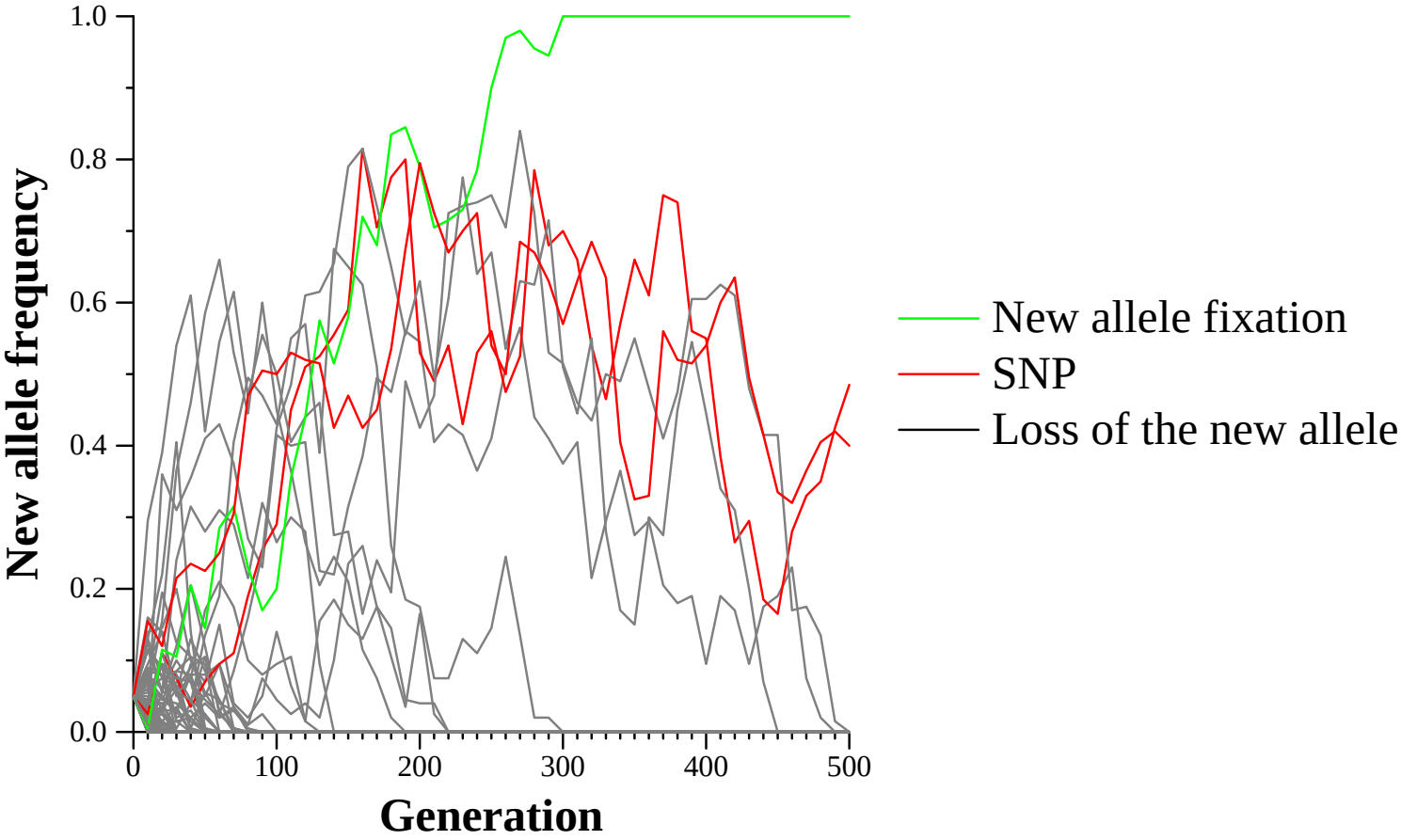
$$\Pr(X_t = j \mid X_{t-1} = i) = \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \binom{N}{j}$$

(Probability that we have j copies of a in generation t, given i copies in generation $t-1$

- States 0 and N are **absorbing**

Random genetic drift

$N = 200$, $X_0 = 10$, 500 generations



More complex models of population

- **Mutations** introduce new alleles, these get eliminated or fixed by random genetic drift
- Speed of fixation influenced by **population structure** or **natural selection**.
- $\Rightarrow$ More complex probabistic models.

Analysis of population history using probabilistic models

Typical model parameters:

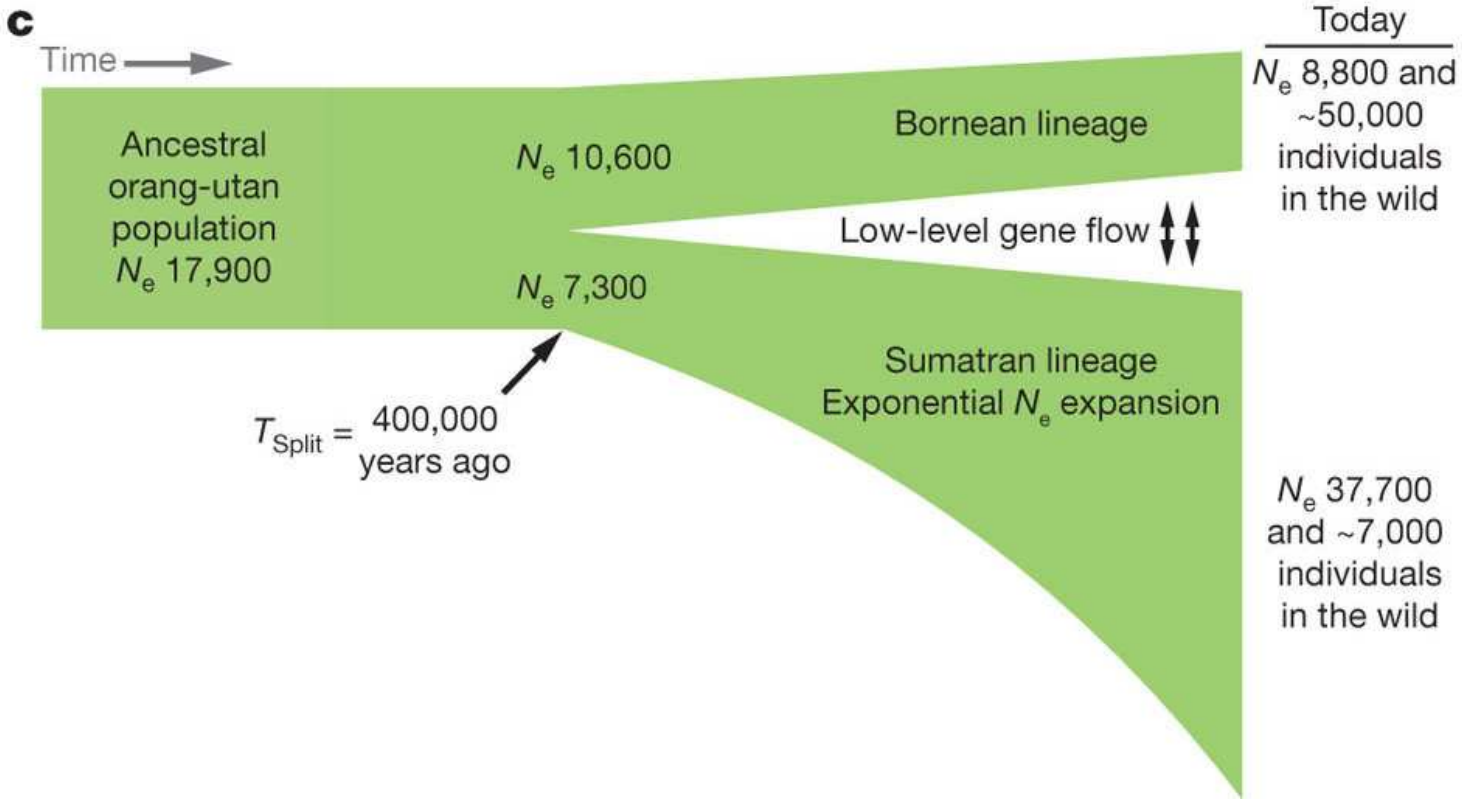
- efective population size
- frequencies of mutation and recombination

These parameters influence observed data:

- SNP frequencies (frequency of minor allele)
- Heterozygocity in diploid individuals
- The number and size of LD blocks

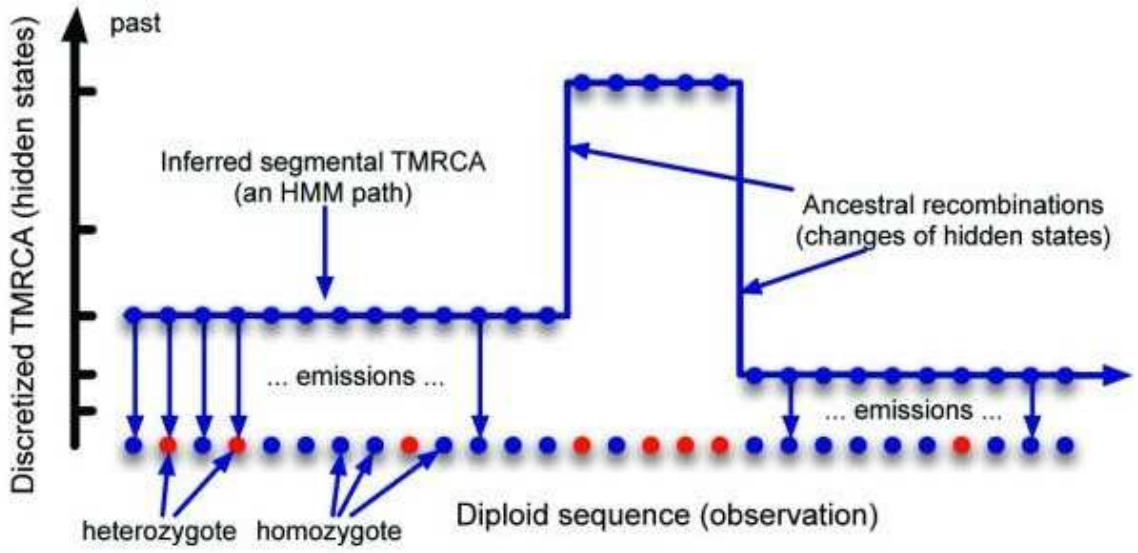
Standard approach: Find parameters of the model best explaining observed data in sequenced individuals

Example: Population history of orangutans



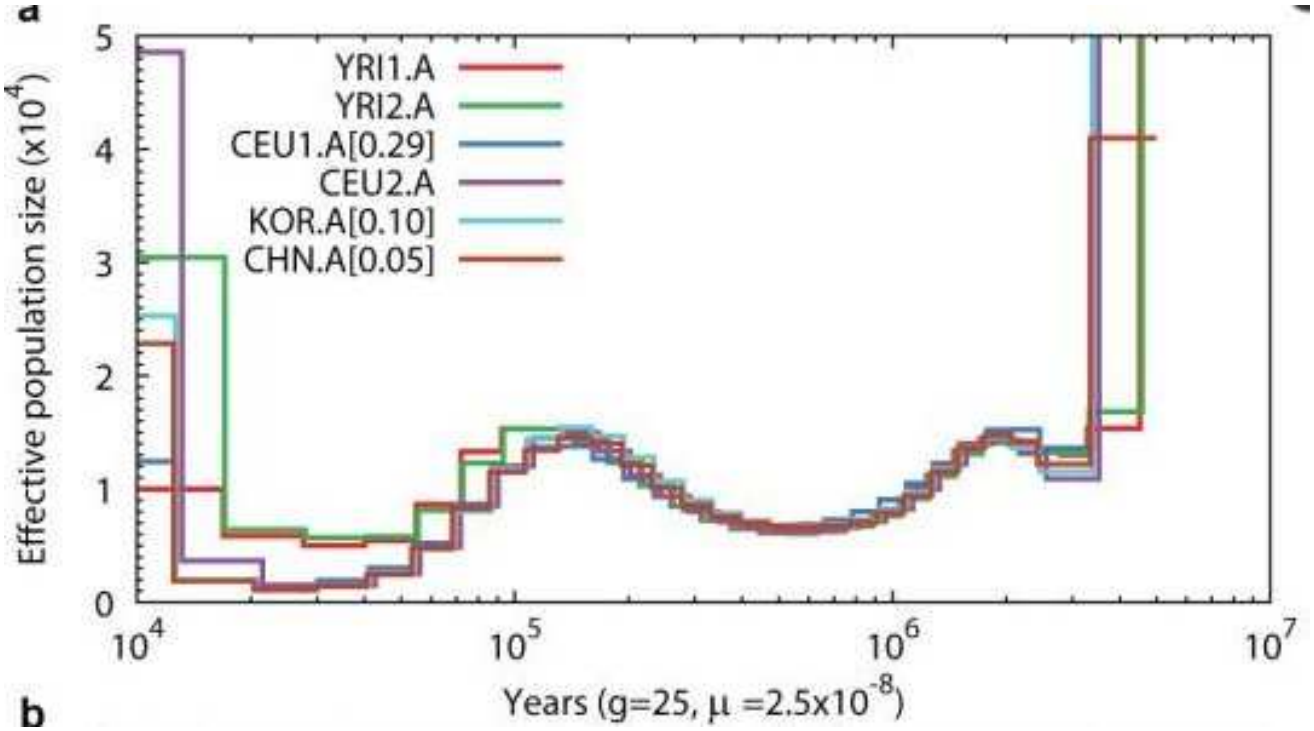
History of a human population from a single human genome (Li, Durbin 2011)

- **Model parameters:** effective human population time changing over time
- **Observed data:**
  - sizes of recombination blocks
  - distribution of time to the most recent common ancestor (TMRCA)



History of a human population from a single human genome

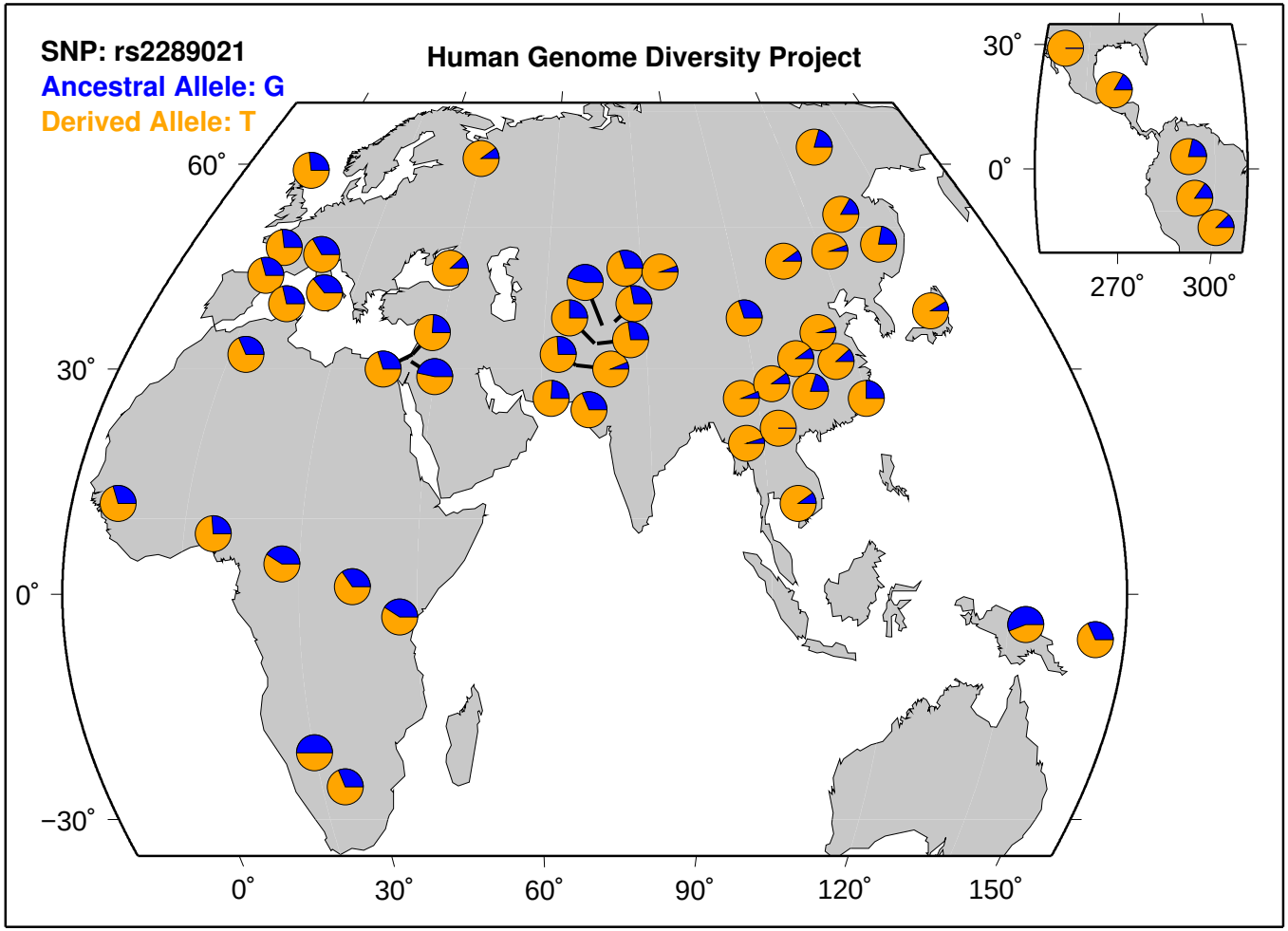
Task: Find historical population sizes best explaining observed statistics



Population structure

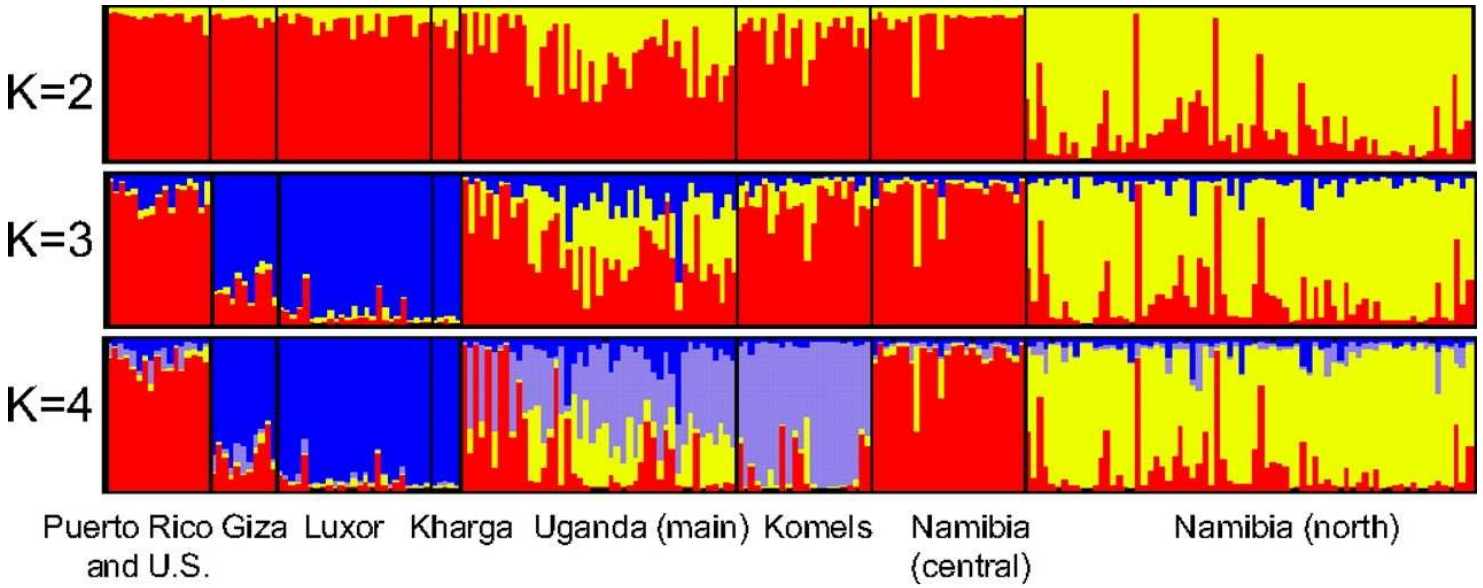
- Assumption so far: new generation produced by random mating
- Most organisms evolve in **subpopulations**,
  with limited migration between populations
- Frequencies of the same SNP in two different populations can be very different
- $\Rightarrow$ "false" long-range correlations between SNPs (e.g., even between chromosomes) if we work with a mix of subpopulations
- $\Rightarrow$ erroneous results in WGAS, LD studies, etc.

Example: allele frequencies of a particular SNP in different regions



from genome.ucsc.edu

Wild dog population structure



Boyko et al. PNAS 2009; software STRUCTURE Pritchard et al. Genetics 2000

- Program STRUCTURE splits population into K subpopulations (colors)
- Each column represents an individual from the population
- Ratio of colors represents ratio of SNPs in the mixture of the K subpopulations.

Algorithm used in STRUCTURE

- **Input:** Set of haplotypes X, which we want to separate into K subpopulations
- Define probabilistic model with the following variables:
  - $P_{i,j}$ - frequency of SNP j in subpopulation i
  - $Z_{i,j}$ - assignment of subpopulation to SNP j in haplotype i
  - Q_i - what portion of SNPs in haplotype i belong to which subpopulation
- Model defines $\Pr[X \mid P, Q, Z]$ and prior distribution for P, Q
- **Output:** $E[Q \mid X]$

Algorithm Markov Chain Monte Carlo (MCMC)

- Variables:
  - $P_{i,j}$ - frequency of SNP j in subpopulation i
  - $Z_{i,j}$ - assignment of subpopulation to SNP j in haplotype i
  - Q_i - what portion of SNPs in haplotype i belong to which subpopulation
- Start with some initial values $P^{(0)}, Z^{(0)}, Q^{(0)}$.
  In each iteration obtain a new random sample:
  - Sample $P^{(i)}, Q^{(i)}$ from $\Pr(P, Q \mid X, Z^{(i-1)})$
  - Sample $Z^{(i)}$ from $\Pr(Z \mid X, P^{(i)}, Q^{(i)})$
- For sufficently large m and c mean of sequence

$$Q^{(m)}, Q^{(m+c)}, Q^{(m+2c)}, \ldots$$

  converges to $E[Q \mid X]$

Summary

- **SNPs (single nucleotide polymorphisms)** appear and disappear in populations
- Their frequency influenced by natural selection
- Without recombination, dependency between SNPs on the same chromosome
  (**linkage disequilibrium**)
- Recombination creates LD blocks
- LD blocks influence the results of whole-genome association mapping
- Probabilistic models of LD block size, allele frequencies, heterozygocity etc. can reveal population history
- We should consider population structure, which can be estimated using computational methods

Other types of polymorphisms

- **Short indels**
- **Microsatellites a minisatellites**
  (simple short repeating sequences)
  13 locuses as a standard "fingerprint" for comparison of DNA samples in the US courts
- **Transposons** (Alu, LINE, SINE)
  Alu has approx. million copies,
  approx. 1 new copy in 20 newly born
- **Large scale copy number variations**