Oznamy

- Dnes posledná prednáška, po prednáške cvičenia pre biológov
- Budúci štvrtok 14.12.:
  – posledné cvičenia pre informatikov
  – nepovinné prezentácie journal clubu v čase prednášky (chceme?)
  – cvičenia pre biológov dohodneme
- Termíny na konci semestra
  – DÚ3 streda 13.12., správy zo journal clubu piatok 15.12.

Skúška pre informatikov (BIN, INF, DAV, AIN)

- Treba získať aspoň polovicu bodov
- Na stránke sú ukážky jednoduchých príkladov, cca 50% bodov
  – v prípade záujmu pred skúškou konzultačné hodiny
- Zvyšné príklady budú prekvapením, v minulosti sa vyskytli:
  – Krátke príklady na pochopenie základných pojmov
  – Navrhnite/modifikujte algoritmus alebo model
- Povolené pomôcky:
  – písacie potreby, ťahák 2 listy A4, jednoduchá kalkulačka
- **Termín?**
  Ak súhlasíte, dátumy opravných termínov dohodneme s tými, ktorých sa budú týkať

Polymorfizmus a populačná genetika

Broňa Brejová

7.12.2023

Populačná genetika

- Rôzne jedince toho istého druhu nemajú identický genóm
- Tieto rozdiely vplývajú na fenotyp (výzor, správanie, choroby, . . .)
- Genómy viacerých jedincov môžeme sekvenovať a porovnávať s referenčnou verziou

Možné aplikácie populačnej genetiky:

- Úloha jednotlivých genetických rozdielov
- História a charakter populácie (podpopulácie, migrácia, historická veľkosť populácie)

SNPy (Single Nucleotide Polymorphisms)

- SNP: jednobázová variabilita medzi jedincami ($> 1\%$ jedincov)
- Obvykle iba dve formy: **väčšinová** a **menšinová** alela
- Aj malá zmena v DNA môže spôsobiť veľké fenotypické zmeny

Systematické mapovanie SNPov:

Projekt 1000 ľudských genómov 2008-2015

identifikácia $> 95\%$ SNPov s aspoň 1% frekvenciou menšinovej alely pomocou NGS sekvenovania

UK Biobank:

500 000 genómov (Nov.2023) plus rozsiahle medicínske dáta o účastníkoch

Plus mnoho ďalších veľkých projektov

Mapovanie asociácií (Trait/Disease Association Mapping)

- Znaky (a choroby) vznikajú kombináciou genetických a environmentálnych vplyvov
- Cieľ: Identifikovať genetické vplyvy
  - Aký je risk choroby s dedičným faktorom u danej osoby?
  - Ako fungujú choroby (na základe génov, ktorých mutácie ich spôsobujú)?
  - Vývoj nových liekov, ich správne cielenie (farmakogenomika)

    Napr. mutácie v génoch rodiny cytochrómu P450 majú vplyv na odbúravanie liekov v pečeni, ovplyvňujú veľkosť potrebnej dávky

Diploidné genómy

- Človek má **diploidný genóm**:
  má v bunkách po dva chromozómy 1. . . 22
  plus pohlavné chromozómy X,X alebo X,Y
- Jeden chromozóm z páru od matky, jeden od otca
- Pre daný SNP s alelami (formami) a, A
  môže byť **homozygot** (aa alebo AA),
  alebo **heterozygot** (aA)
- Ak nejaká choroba zapríčinená alelou a,
  tak sa môže prejaviť iba pri homozygotoch aa,
  alebo aj pri heterozygotoch aA,
  alebo môže byť pri aa silnejšia ako pri aA

Diploidné genómy

- Človek má **diploidný genóm**:
  má v bunkách po dva chromozómy 1...22
  plus pohlavné chromozómy X,X alebo X,Y
- Jeden chromozóm z páru od matky, jeden od otca
- Pre daný SNP s alelami (formami) a, A
  môže byť **homozygot** (aa alebo AA),
  alebo **heterozygot** (aA)
- **Haplotyp:** kombinácia aliel rôznych SNPov na tom istom chromozóme
  (zdedená od jedného rodiča)
  Diploidný jedinec má teda dva haplotypy

chr1 od matky: ...A...T...G... ...

chr1 od otca: ...T...C...A... ...

Testovanie asociácie jedného SNPu

Kontingenčná tabuľka - počet haplotypov

Veľkosť psa vs. alela na pozícii chr15:44,228,468

	pôvodná alela	odvodená alela	spolu
malý pes (< 9 kg)	14	535	549
veľký pes (> 31 kg)	339	38	377
spolu	353	573	926

[Sutter a kol. 2007]

Štatisticky testujeme či sú riadky a stĺpce **nezávislé (nulová hypotéza).**

Ak **vylúčime nulovú hypotézu**, našli sme asociáciu, nemusí však ísť o príčinu

Ak ju nevylúčime, nepreukázali sme súvis SNPu s veľkosťou

(môže ale existovať, možno treba viac dát)

Testovanie nezávislosti v kontingenčnej tabuľke

	pôvodná alela	odvodená alela	spolu
malý pes	14	535	549
veľký pes	339	38	377
spolu	353	573	926

Fisherov test: (Fisher's exact test) presný výsledok z hypergeometrického rozdelenia

Chí-kvadrát (χ^2) test: obľúbený približný test, vhodný ak máme vysoké počty

Používajú sa aj zložitejšie štatistické modely

(napr. diploidný genóm, príbuzenské vzťahy, . . .)

Testovanie nezávislosti v kontingenčnej tabuľke χ^2 testom

	alela A	alela a	spolu
malý pes (m)	14	535	549
veľký pes (v)	339	38	377
spolu	353	573	926

V nulovej hypotéze (nezávislosť riadkov a stĺpcov) máme:

$$\Pr(A) = 353/926 = 0.381, \Pr(a) = 0.619$$
$$\Pr(m) = 549/926 = 0.593, \Pr(v) = 0.407$$
$$\Pr(A, m) = \Pr(A)\Pr(m) = 0.226$$
$$\Pr(a, m) = \Pr(a)\Pr(m) = 0.367$$
$$\Pr(A, v) = \Pr(A)\Pr(v) = 0.155$$
$$\Pr(a, v) = \Pr(a)\Pr(v) = 0.252$$

Podľa nulovej hypotézy by sme teda čakali, že 926 haplotypov bude v tabuľke rozdelených v pomeroch 0.226:0.367:0.155:0.252

Testovanie nezávislosti v kontingenčnej tabuľke χ^2 testom

Skutočná tabuľka

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

	A	a	spolu
malý	14	535	549
veľký	339	38	377
spolu	353	573	926

Očakávané podľa nulovej hypotézy

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

	A	a	spolu
malý	209.3	339.8	549
veľký	143.5	233.4	377
spolu	353	573	926

Spočítame veličinu $\chi^2 = \sum_{i\in\{m,v\}} \sum_{j\in\{A,a\}} \frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$

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

χ^2 je určitá miera rozdielnosti tabuliek O a E.

Platí, že $\chi^2 \geq 0$ a χ^2 je nula, iba ak sa tabuľky úplne zhodujú.

Testovanie nezávislosti v kontingenčnej tabuľke χ^2 testom

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

	A	a	spolu
malý	14	535	549
veľký	339	38	377
spolu	353	573	926

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

	A	a	spolu
malý	209.3	339.8	549
veľký	143.5	233.4	377
spolu	353	573	926

Spočítame veličinu $\chi^2 = \sum_{i \in \{m,v\}} \sum_{j \in \{A,a\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 724.3$

Ak platí nulová hypotéza, χ^2 je približne z rozdelenia $\chi^2(1)$,

t.j. **chí kvadrát s jedným stupňom voľnosti**.

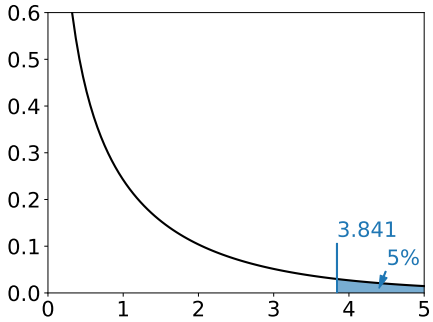
1 stupeň: ak poznáme E a 1 políčko z O, zvyšok O vieme dopočítať.

Šanca, že pri nulovej hypotéze nám náhodne vyjde $\chi^2 \geq 724.3$ je $1.6 \cdot 10^{-159}$

(P-hodnota)

Na **odmietnutie nulovej hypotézy** často používame

prah $P < 0.05$, t.j. $\chi^2 > 3.841$



Závislosť medzi dvoma rôznymi SNPmi

Uvažujme SNP s alelami p/P a ďalší SNP s alelami q/Q.

Nameriame počty haplotypov pq, PQ, pQ, Pq

Príklad: 2000 haplotypov (1000 jedincov)

	Q	q
P	474	611
p	142	773

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

Stĺpce a riadky teda nie sú nezávislé, medzi SNPmi je závislosť

Príklad 2: Podobné pomery počtov, ale iba 30 haloptypov:

	Q	q
P	7	9
p	2	12

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

Nulovú hypotézu nevylúčime pre prah P<0.05 ($\chi^2 > 3.841$)

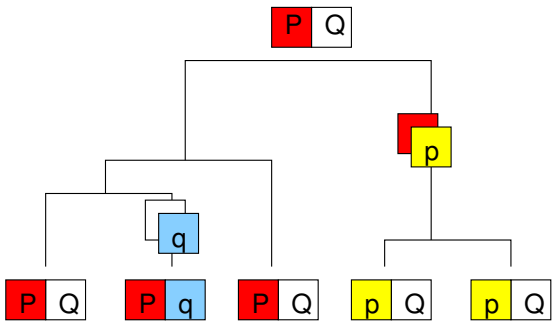
Ale pozor, pre takéto malé hodnoty χ^2 **nepresný**

Ako vzniká závislosť medzi dvoma rôznymi SNPmi

Na rozdielnych chromozómoch:

- Pravdepodobnosti výskytu jednotlivých alel sú nezávislé
- $\Pr(pq) = \Pr(p)\Pr(q), \Pr(PQ) = \Pr(P)\Pr(Q)$, atď
- **väzbová rovnováha, linkage equilibrium (LE)**

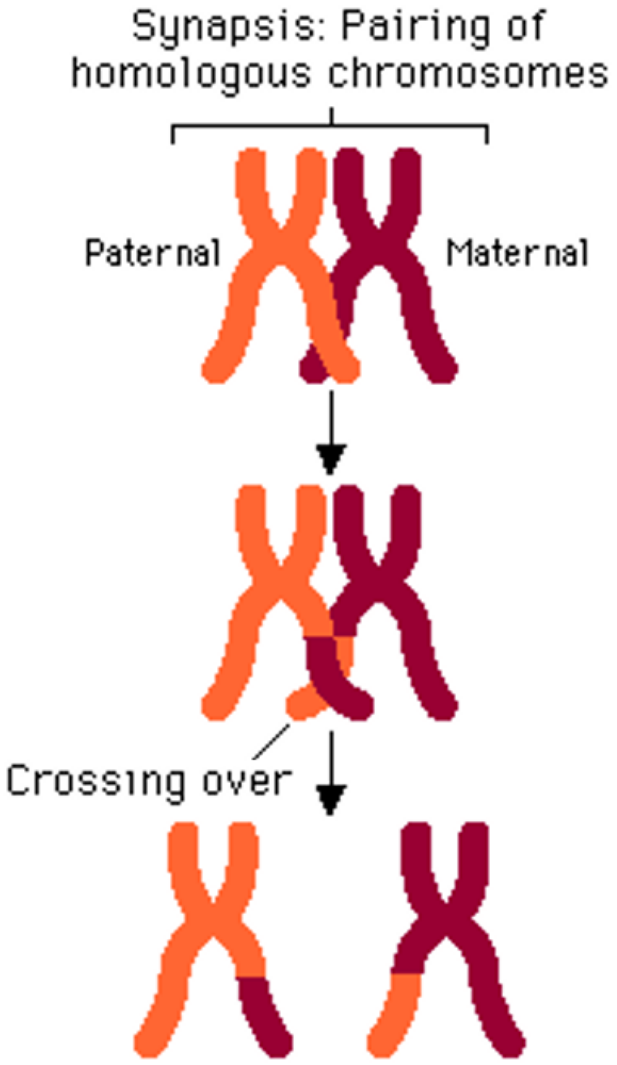
Blízko seba na tom istom chromozóme:



- Málokedy mutácia na to istom mieste 2x, zriedkavá rekombinácia
- Kombinácie nie sú úplne náhodné
- Korelácie medzi SNPmi

  $\Rightarrow$ **väzbová nerovnováha, linkage disequilibrium (LD)**

Rekombinácia



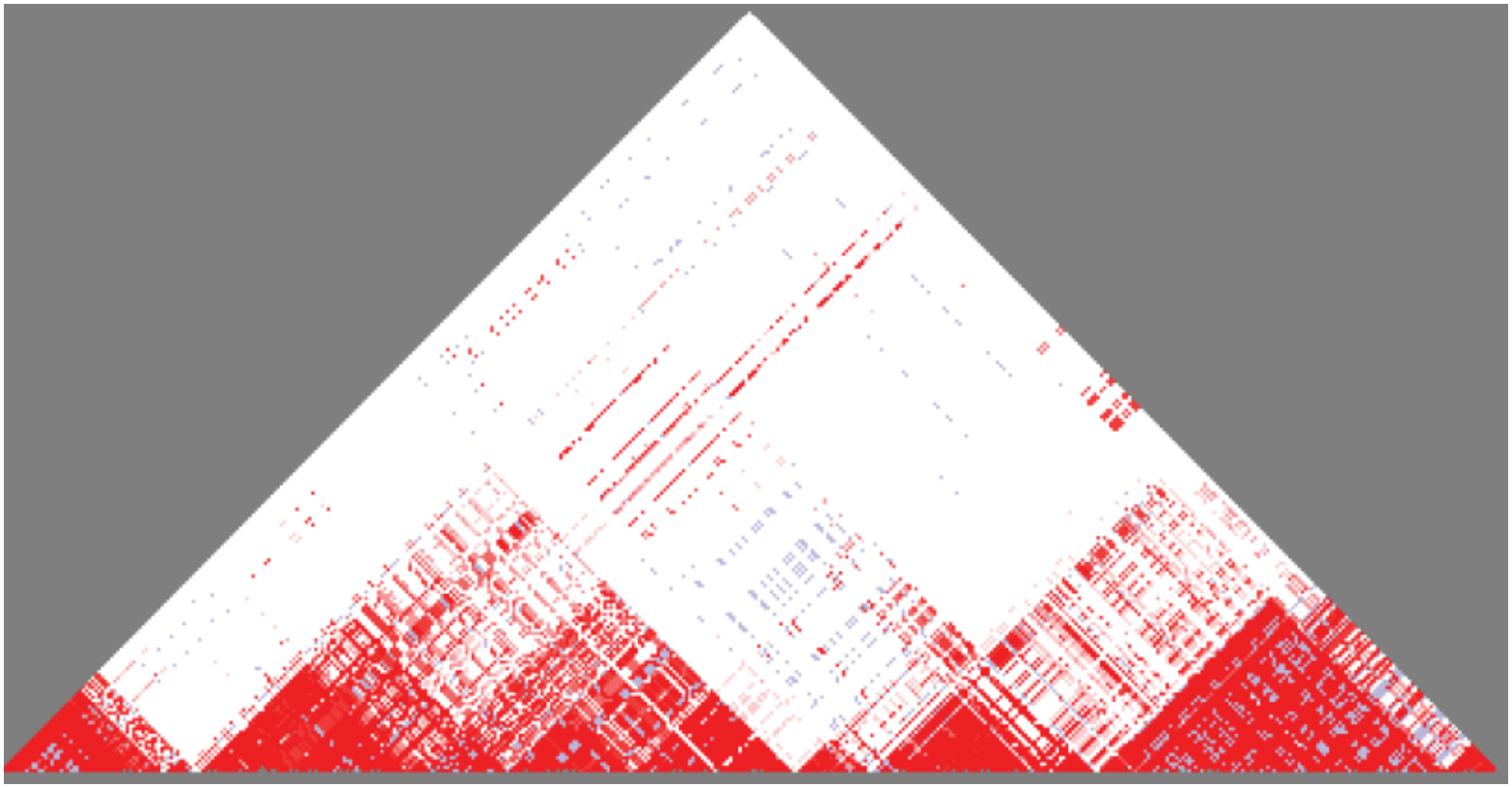
Cca 1-3 **rekombinácie** v 1 ľudskom chromozóme počas meiózy (tvorba pohlavných buniek)

Rekombinácia znižuje LD

Ak predpokladáme rovnomernú rekombináciu:

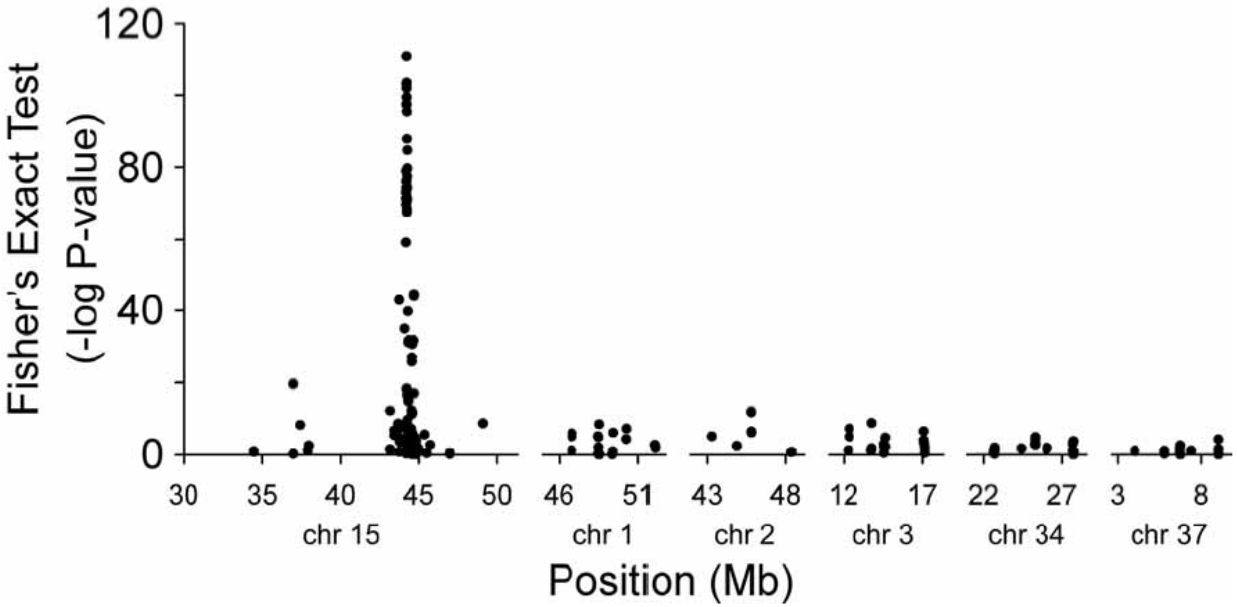
- Čím vzdialenejšie SNPy, tým nižšie LD
- Čím staršie SNPy, tým nižšie LD
- Ďalšie aspekty: štruktúra populácie, prirodzený výber, rekombinačné hotspoty

Väzbová nerovnováha (LD) v ľudskom genóme



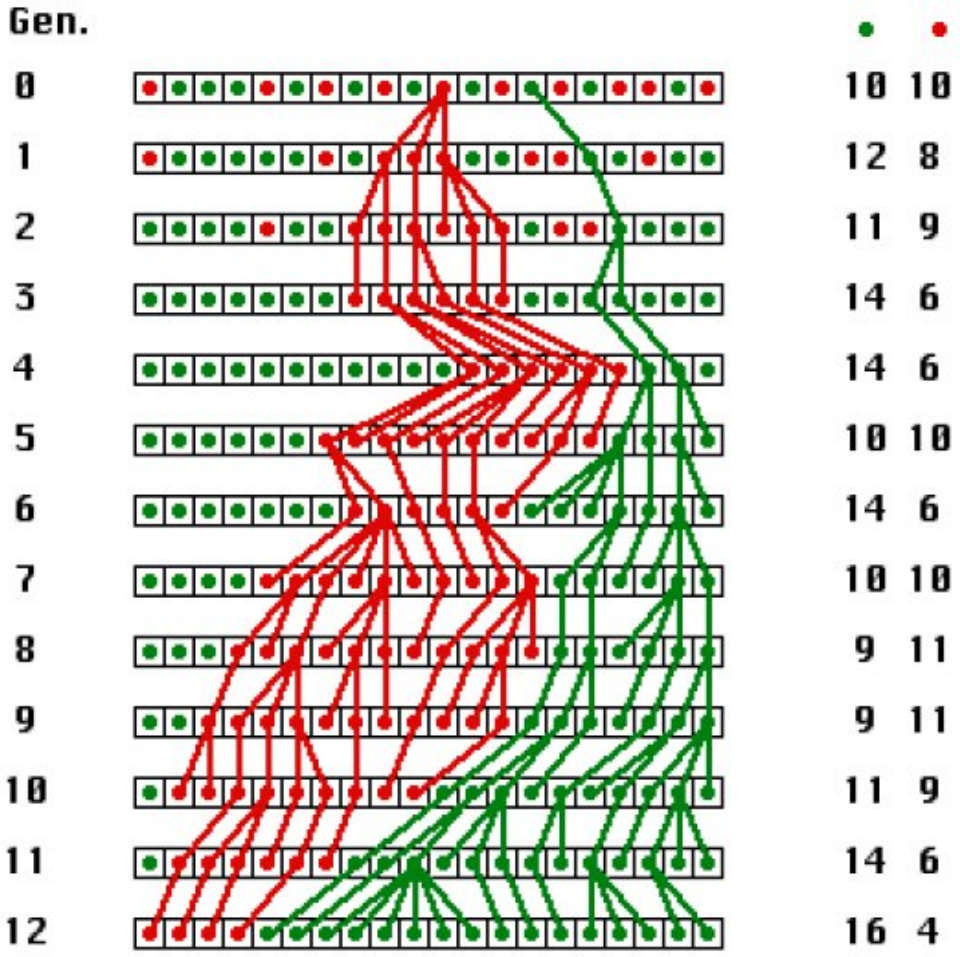
Región ENm014 (500kB, chr 7), 90 ľudí Utah, project HapMap

Späť k psom: Hľadanie asociácií v celom genóme (Genome-Wide Association Study, GWAS)



- V prípade štúdie veľkosti psov: GWAS identifikoval 84 kB región
- Pozíciu ďalej treba spresniť ďalšími experimentami
- **Veľké LD bloky** $\Rightarrow$ veľké výsledné regióny

Základný model populačnej genetiky: Wrightov-Fisherov model



Životný cyklus SNPov vo Wrightovom-Fisherovom modeli

- Populácia N jedincov (stabilná veľkosť)
- Jedinec = jedna alela (A or a)
- Nová generácia vzniká "skopírovaním" náhodného rodiča (random mating), bez vplyvu prirodzeného výberu
- X_t: počet jedincov s alelou a v generácii t
- **Markovovský reťazec** so stavmi $X_t \in \{0, 1, \ldots, N\}$

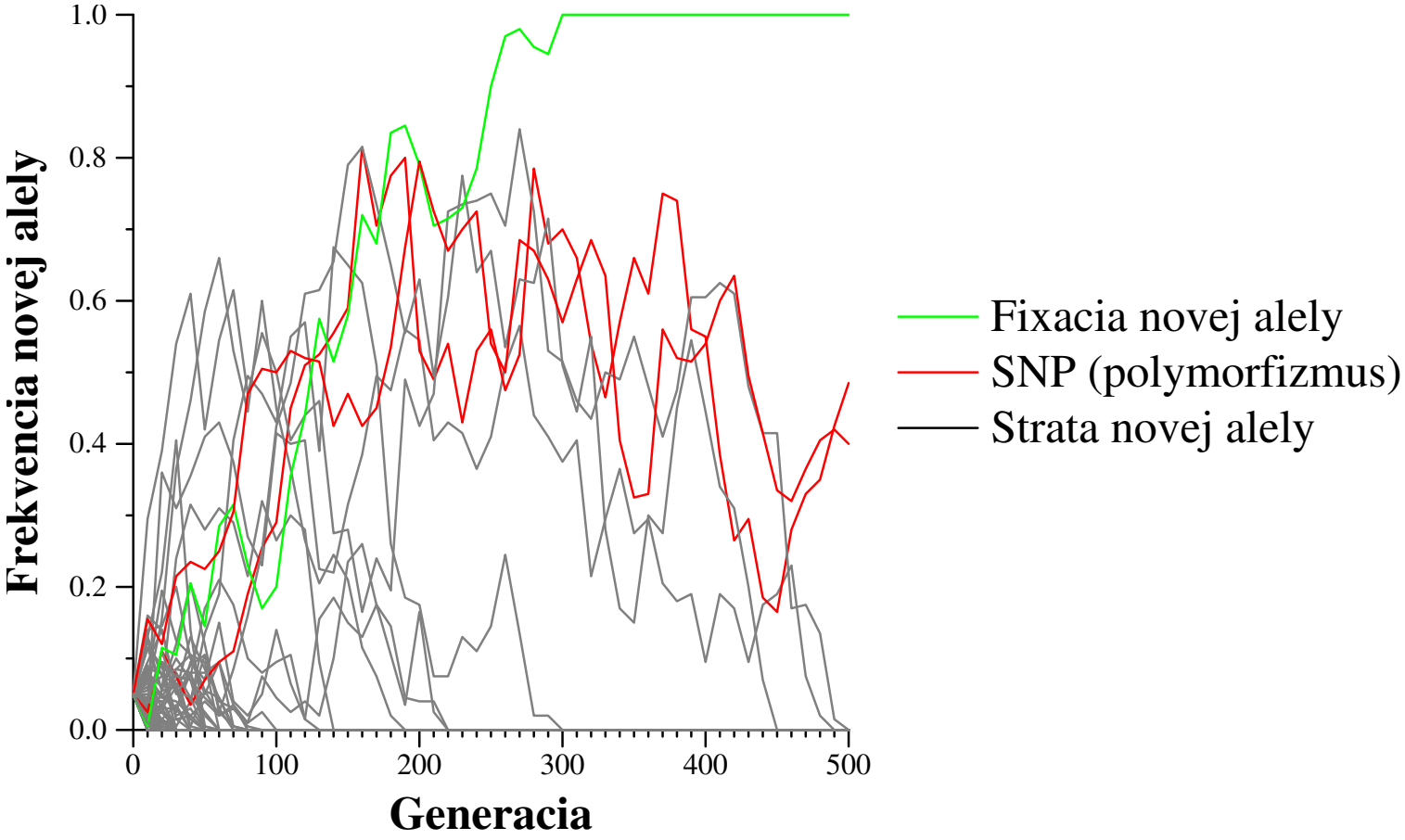
$$\Pr(X_t = j \mid X_{t-1} = i) = \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \binom{N}{j}$$

(Pravdepodobnosť, že v generácii t máme j kópií alely a, ak v generácii $t-1$ ich bolo i)

- Stavy 0 and N sú **pohlcujúce**

Náhodný genetický drift

$N = 200$, $X_0 = 10$, 500 generácií



Zložitejšie modely populácie

- **Mutácie** zavádzajú do populácie nové alely, ktoré po čase náhodným genetickým driftom zaniknú, alebo ovládnu populáciu (fixation).
- Rýchlosť procesu je ovplyvnená efektami ako **štruktúra populácie** alebo **prirodzený výber**
- $\Rightarrow$ Zložitejšie pravdepodobnostné modely

Analýza histórie populácie na základe pravdepodobnostných modelov

Typické parametre pravdepodobnostného modelu:

- efektívna veľkosť populácie
- frekvencia rekombinácie a mutácie

Parametre ovplyvňujú pozorované dáta:

- Frekvencie menšinovej alely SNPov
- Heterozygotnosť u diploidných jedincov
- Počet a veľkosť LD blokov

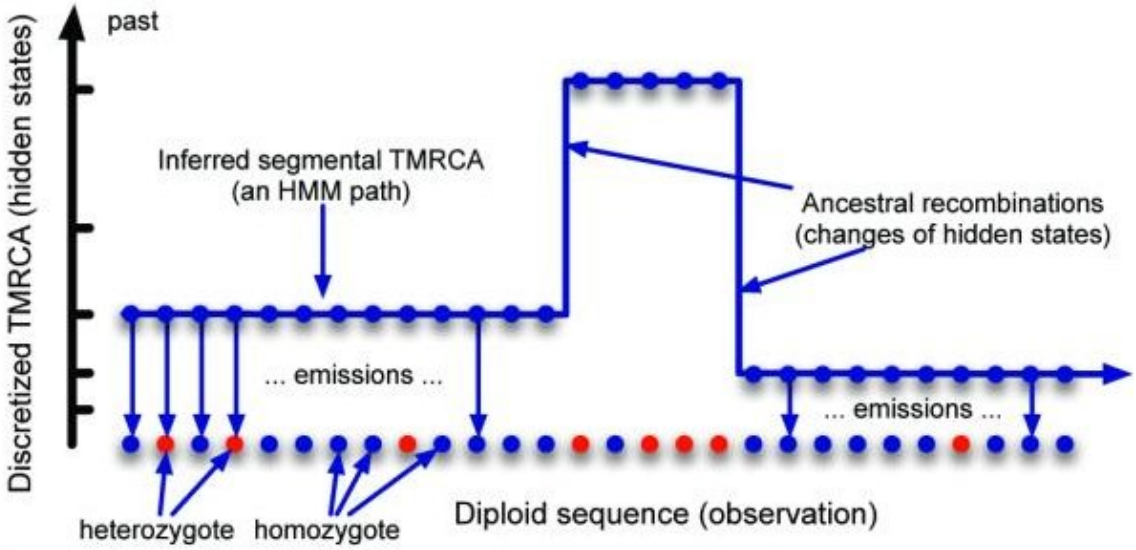
Použitie:

Snažíme sa nájsť parametre modelu, ktoré najlepšie vysvetľujú pozorované dáta u osekvenovaných jedincov.

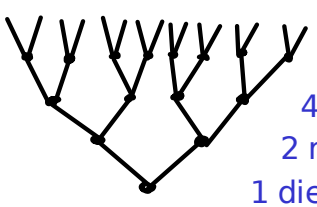
História ľudskej populácie z genómu jedinca (Li, Durbin 2011)

- **Parametre modelu:** história vývoja efektívnej veľkosti ľudskej populácie v čase
- **Použité dáta:** Pozície heterozygotných SNPov v rámci genómu

  Z ich premenlivej hustoty určí rozdelenie časov ku najbližšiemu spoločnému predkovi (TMRCA)

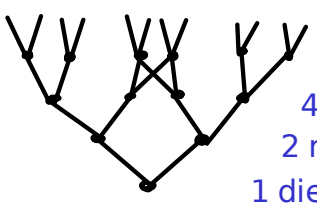


Čas k najbližšiemu spoločnému predkovi a počet mutácií



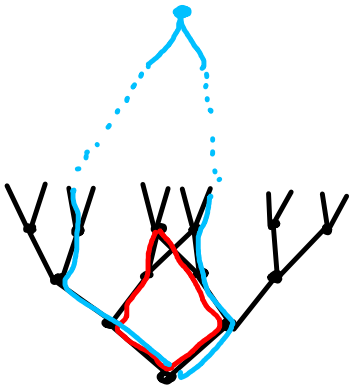
16 praprarodičov
8 prarodičov
4 starí rodičia
2 rodičia
1 dieťa

bez sobášov medzi blízkymi príbuznými



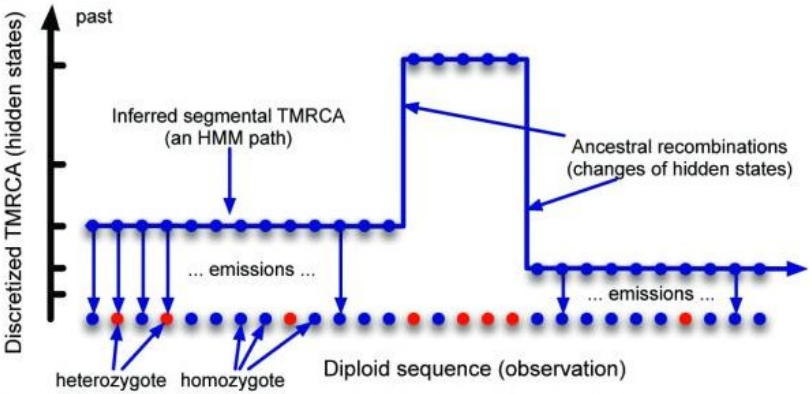
12 praprarodičov
6 prarodičov
4 starí rodičia
2 rodičia
1 dieťa

bratranec a sesternica majú dieťa



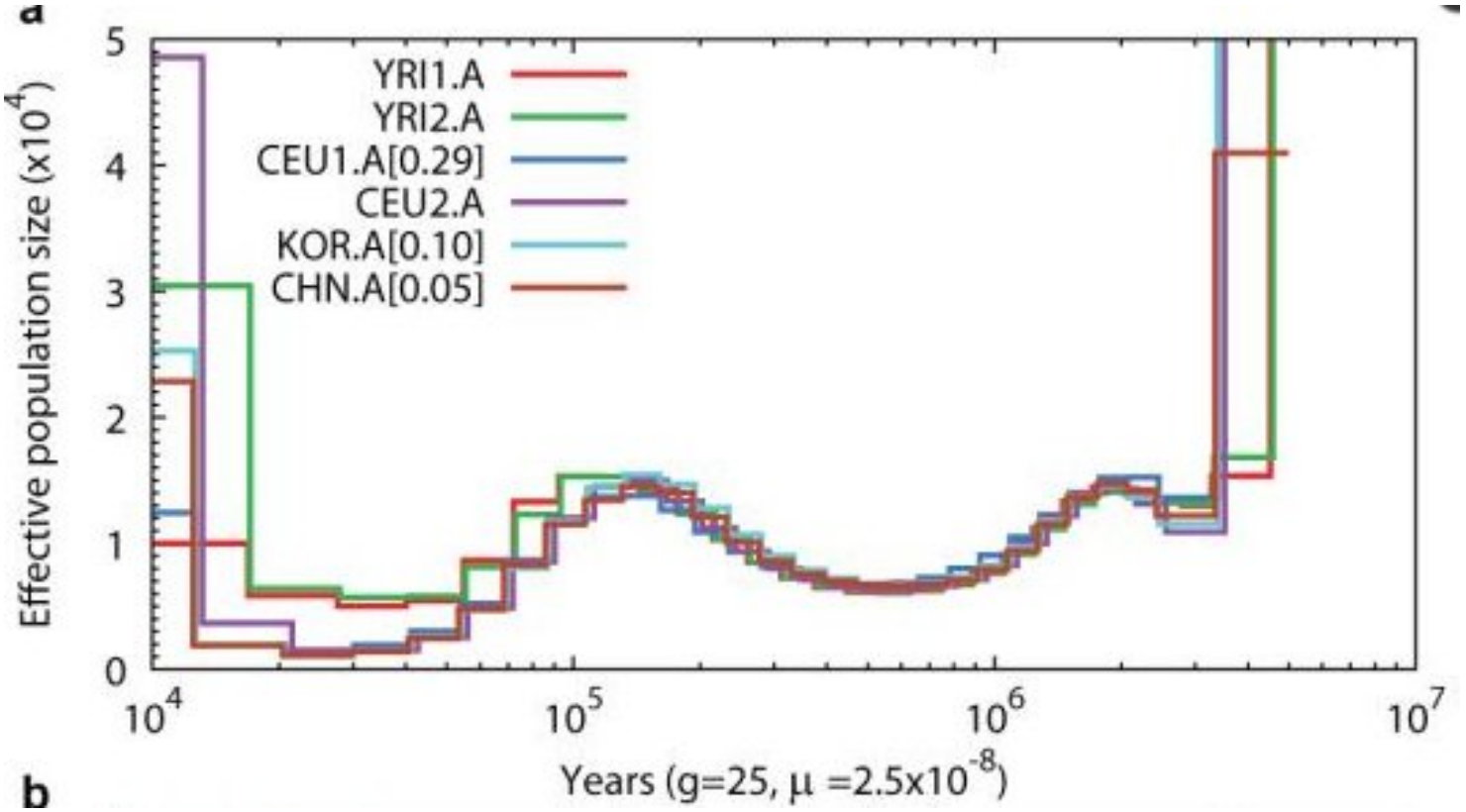
dávny spoločný predok,
veľa mutácií

nedávny spoločný predok,
málo mutácií



Príklad: História ľudskej populácie z genómu jedinca

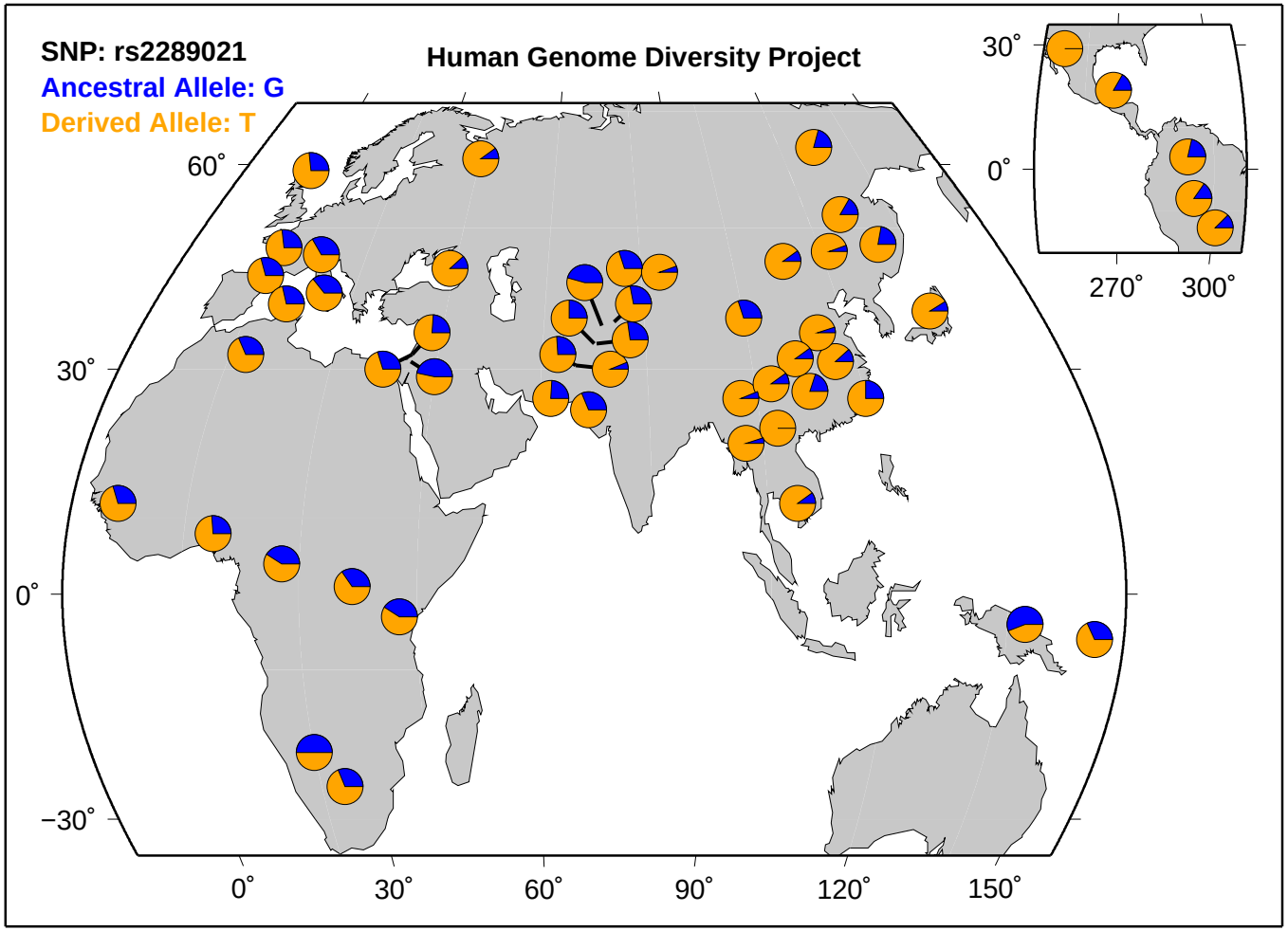
Úloha: Nájdi históriu vývoja efektívnej veľkosti ľudskej populácie, ktorá najlepšie vysvetľuje pozorované dáta



Štruktúra populácie

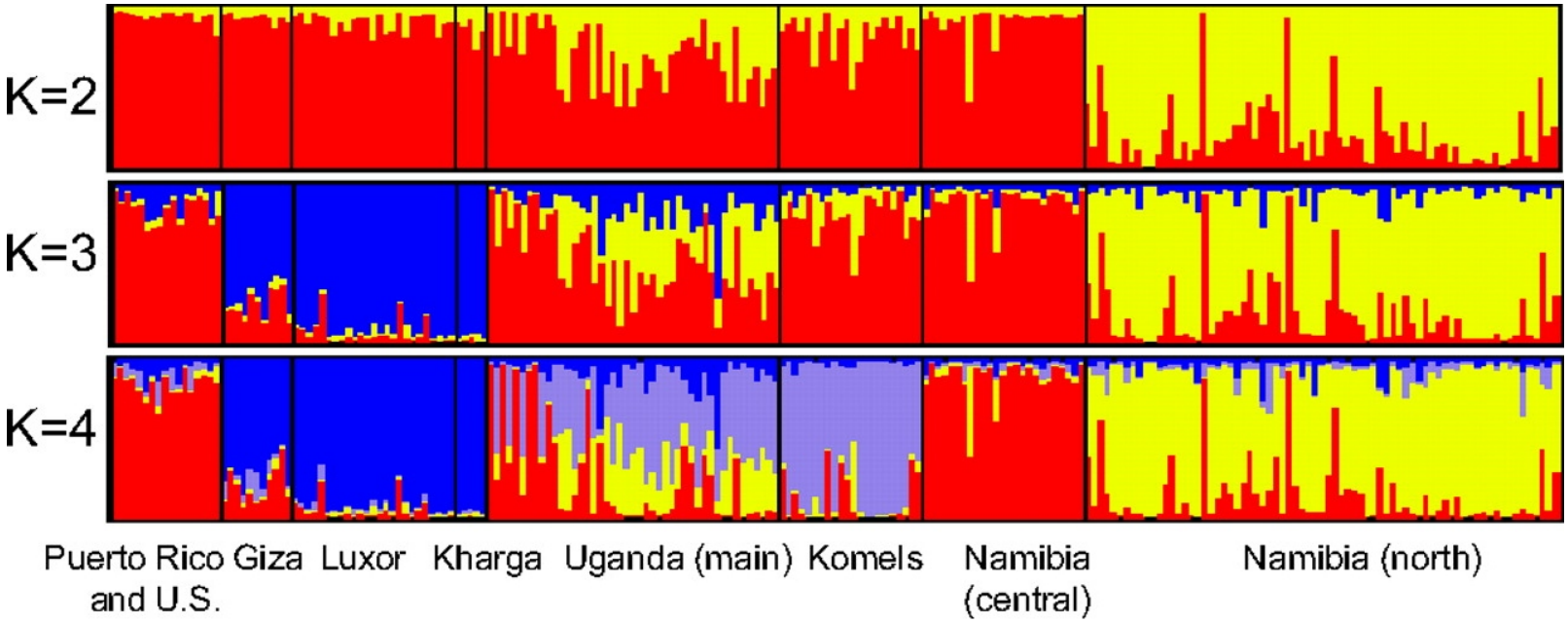
- Doteraz sme predpokladali, že nová generácia vzniká **náhodným párovaním** (random mating)
- Väčšina organizmov sa vyvíja v **subpopuláciách**, s obmedzeným prenosom genetického materiálu medzi subpopuláciami
- Frekvencie toho istého SNPu v dvoch subpopuláciách môžu byť značne odlišné
- $\Rightarrow$ "falošné" korelácie medzi SNPami (napr. aj medzi chromozómami), ak pracujeme s viacerými subpopuláciami naraz
- $\Rightarrow$ chybné výsledky pri LD a GWAS

Príklad: frekvencie alel jedného konkrétneho SNPu u ľudí v rôznych častiach sveta



zdroj: genome.ucsc.edu

Štruktúra populácie psov



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

- Program STRUCTURE rozdelí populáciu na K subpopulácií (farby)
- Každý stĺpec je jedinec z populácie
- Pomer farieb zodpovedá pomeru SNPov z každej z K populácií

Ako funguje STRUCTURE?

- **Vstup:** Vzorka haplotypov X, ktorú chceme rozdeliť do K subpopulácií
- Definujeme stochastický model s nasledujúcimi premennými:
  - $P_{i,j}$ - frekvencia SNPu j v subpopulácii i
  - Q_i - aká časť SNPov v haplotype i patrí ku ktorej subpopulácii
  - $Z_{i,j}$ - priradenie subpopulácie SNPu j v haplotype i
- Model definuje $\Pr[X \mid P, Q, Z]$ a apriórne rozdelenie pre P, Q
- **Výstup:** $E[Q \mid X]$

Algoritmus Markov Chain Monte Carlo (MCMC)

- Premenné:
  - $P_{i,j}$ - frekvencia SNPu j v populácii i
  - $Z_{i,j}$ - priradenie subpopulácie SNPu j v haplotype i
  - Q_i - aká časť SNPov v haplotype i patrí ku ktorej populácii
- Začni s hodnotami $P^{(0)}, Z^{(0)}, Q^{(0)}$. V každej ďalšej iterácii získame novú náhodnú vzorku:
  - Vyber náhodnú vzorku $P^{(i)}, Q^{(i)}$ z distribúcie $\Pr(P, Q \mid X, Z^{(i-1)})$
  - Vyber náhodnú vzorku $Z^{(i)}$ z distribúcie $\Pr(Z \mid X, P^{(i)}, Q^{(i)})$
- Pre vhodné m, c, priemer postupnosti

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

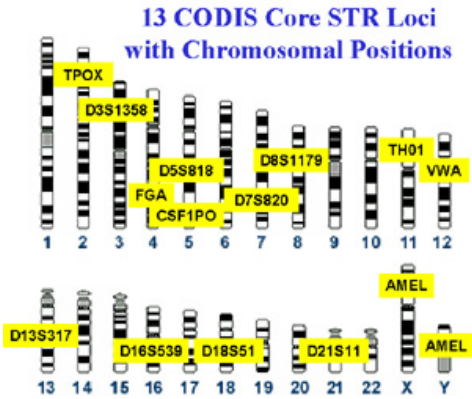
konverguje k hodnote $E[Q \mid X]$

Zhrnutie

- **SNPy (single nucleotide polymorphizms)** priebežne vznikajú a zanikajú v populáciách
- Ich frekvencia ovplyvnená navyše prirodzeným výberom
- Bez rekombinácie korelácia medzi SNPmi na tom istom chromozóme (**linkage disequilibrium**)
- Rekombinácie vytvárajú v genóme LD bloky
- Prítomnosť LD blokov vplýva na výsledky mapovania asociácií znakov (**genome-wide association mapping**)
- Pravdepodobnostné modely veľkosti LD blokov, frekvencií alel, heterozygocity a pod. nám môžu veľa prezradiť o **histórii populácie**
- Pri analýzach treba brať do úvahy **štruktúru populácie**, ktorú možno odhadnúť pomocou výpočtových metód

Ďalšie typy polymorfizmov

- **Krátke indely**
- **Mikrosatelity a minisatelity** (jednoduché krátke opakujúce sa sekvencie)
  13 lokusov ako štandarný “odtlačok” pre porovnávanie DNA vzoriek na súdoch v USA



- **Transpozóny** (Alu, LINE, SINE)
  Alu má cca milión kópií, cca 1 nová kópia na 20 novorodencov
- **Veľké úseky s variabilnou multiplicitou** (Large scale copy number variations)