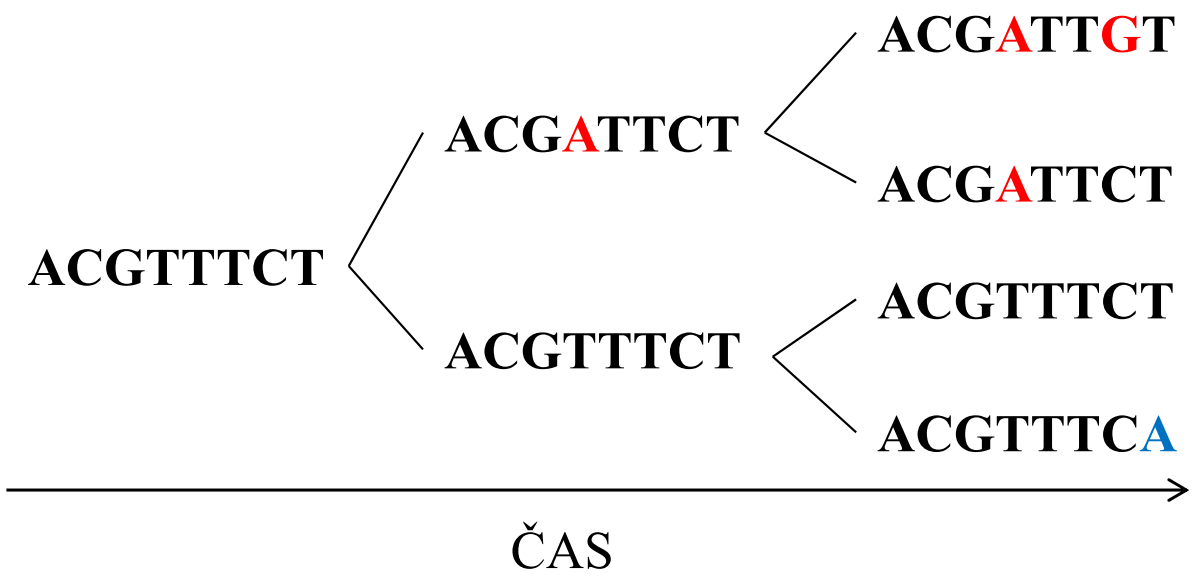
Typy populačně-genetických dat a jejich základní analýza

Fylogenetika vs. populační genetika



- mutační rychlost - datování

- fylogenetické vztahy mezi jednotlivými variantami

- selekce na úrovni kódujících sekvencí

- většinou nás nezajímají frekvence alel

- založená na frekvencích alel

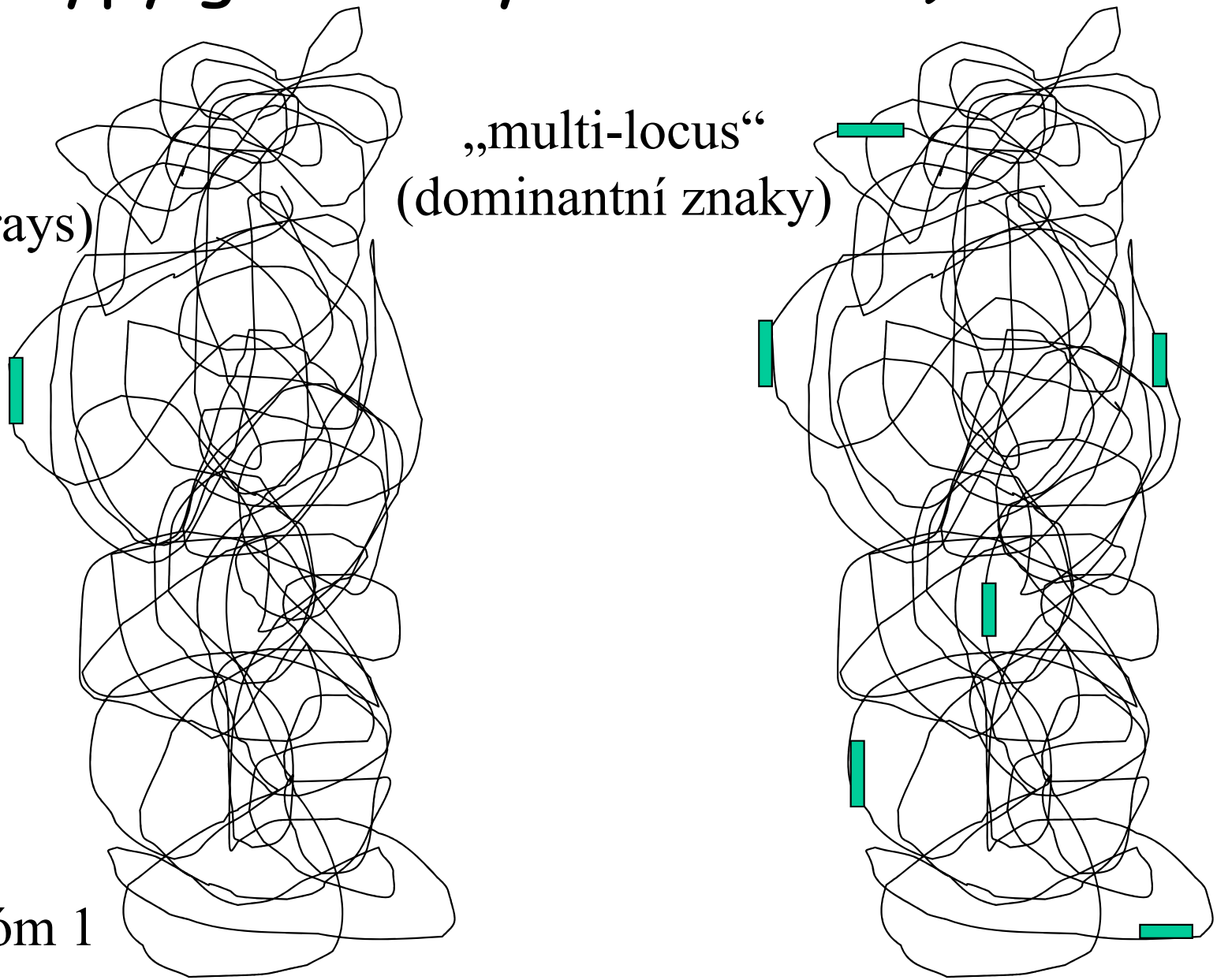
- relativně recentní procesy: genetický drift, populační struktura, tok genů, efektivní velikost populace

- selekce – srovnání populační struktury na kódujících a nekódujících znacích

Laboratorní techniky
(= typy genetických markerů)

„single-locus“
(PCR, microarrays)

„multi-locus“
(dominantní znaky)



Př.: chromozóm 1

Typy populačně-genetických dat

mikrosatelity

Jedinec	Marker 1	Marker 2
Ind_1	170/172	133/136
Ind_2	168/172	133/139
Ind_3	168/168	136/139

SNPs

Jedinec	Marker 1	Marker 2
Ind_1	A/T	C/T
Ind_2	A/T	T/T
Ind_3	T/T	C/T

SINE

Jedinec	Marker 1	Marker 2
Ind_1	+/-	-/-
Ind_2	+/+	+/+
Ind_3	-/-	+/-

AFLP

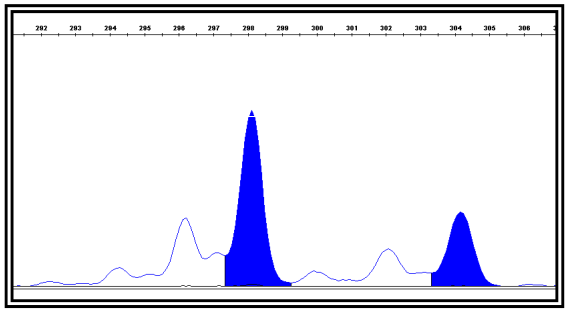
Jedinec	Marker 1	Marker 2
Ind_1	+	-
Ind_2	+	+
Ind_3	-	-

Populačně-genetická analýza

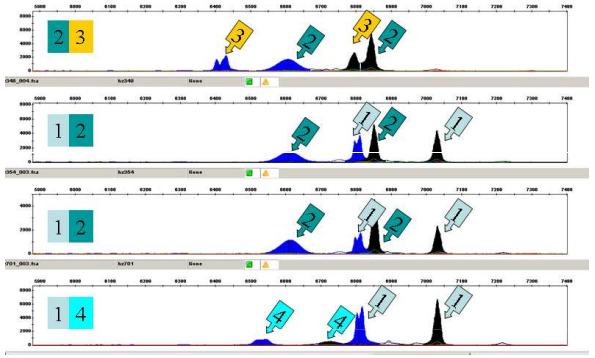
- Jak je variabilní daná populace a jaká je její efektivní velikost
- Nachází se daná populace v období demografické expanze nebo poklesu?
- Existuje mezi dvěma subpopulacemi bariéra toku genů a jak je silná? Jaká je prostorová genetická struktura?
- Vyskytují se v populacích imigranti nebo jejich potomci (hybridi)?

Populačně-genetická data

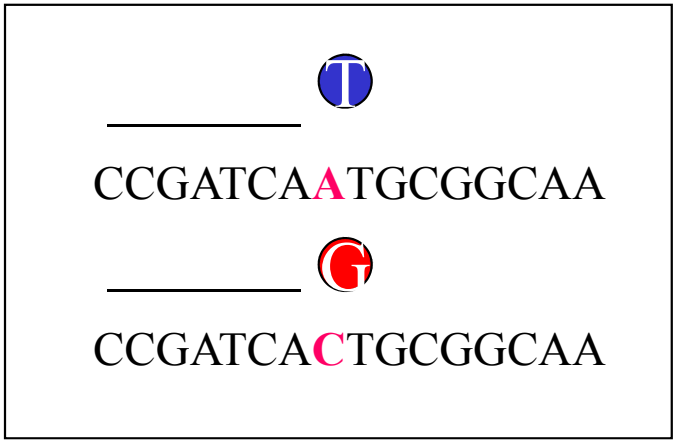
- v tomto kurzu omezena na diploidní kodominantní znaky (Mendelovská dědičnost)



mikrosatelity



jaderné sekvence (např. SSCP)



SNPs

Velké množství populačně-genetických programů

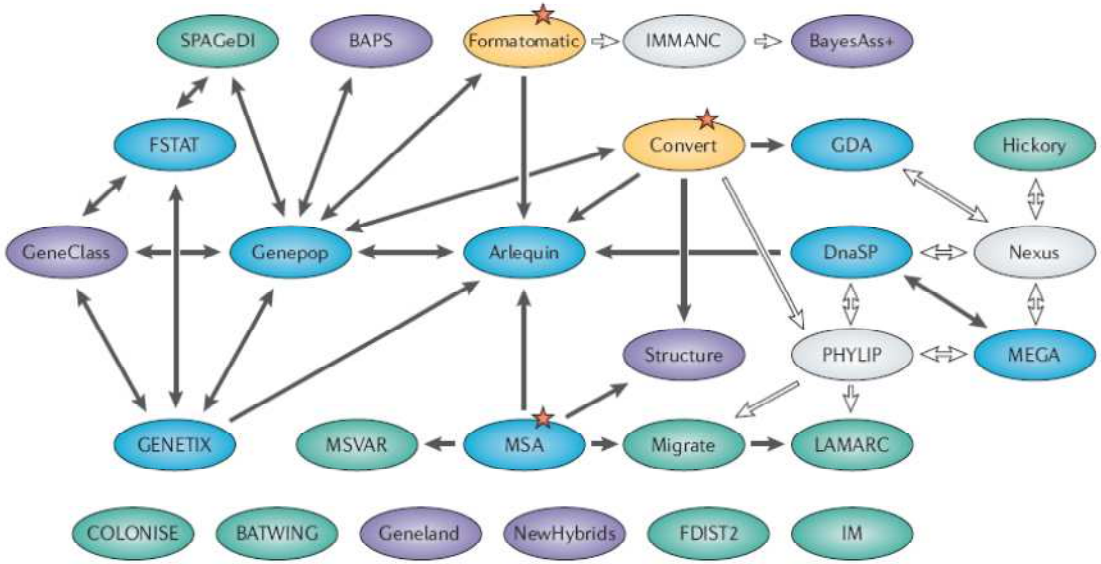
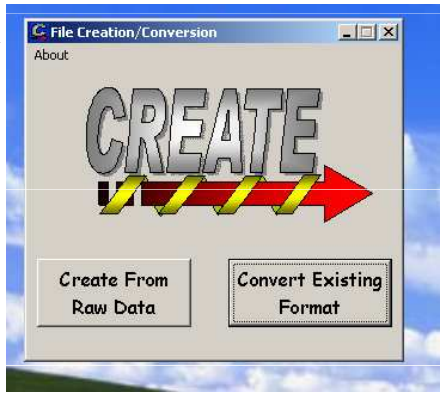
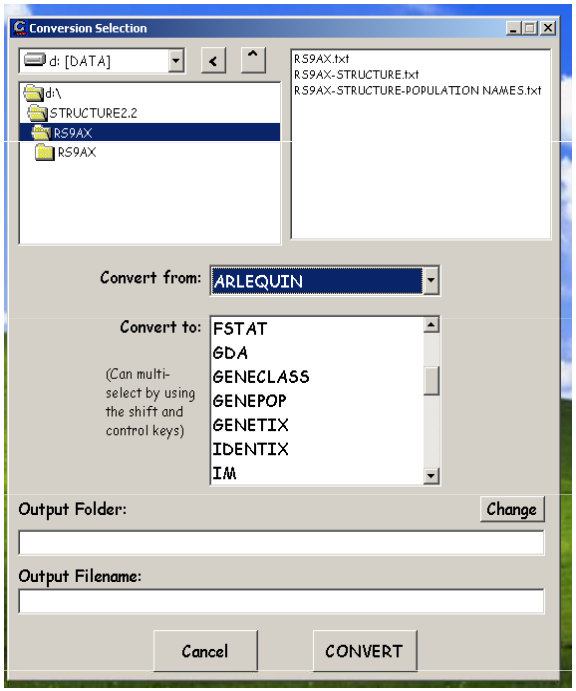


Figure 1 | **Flow chart of possible data exchange between different population genetics programs.** Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.





Computer programs for population genetics data analysis: a survival guide

Laurent Excoffier and Gerald Heckel

© 2006 Nature Publishing Group

Kodominantní znaky (např. mikrosatelity) – GenAlEx formát

počet lokusů

počet jedinců

počet populací

počet vzorků v 1. populaci

počet vzorků v 2. populaci, atd.

geografické koordináty

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	A
1	11	233	10	30	22	25	22	18	23	12	27	25	29															
2	**Allele frequency**																											
3	**ID**	**Pop**	**Ppip01**		**Ppip02**		**Ppip04**		**Ppip06**		**EF1**		**EF4**		**EF6**		**Paur05**		**NN18**		**NnP217**		**NnP219**			**X**	**Y**	
4	**TYN1**	**TYN**	174	176	128	128	213	215	120	132	166	175	236	247	174	188	249	251	287	291	205	209	140	170		48.76500	17.00750	
5	**TYN2**	**TYN**	167	167	132	132	207	207	130	136	169	175	230	245	182	188	247	249	287	287	205	218	166	166		48.76500	17.00750	
6	**TYN3**	**TYN**	176	178	128	128	207	207	130	130	169	169	0	0	184	190	249	249	283	293	209	213	166	170		48.76500	17.00750	
7	**TYN4**	**TYN**	180	180	130	132	207	207	118	124	171	173	220	243	184	190	247	249	283	285	209	214	140	166		48.76500	17.00750	
8	**TYN5**	**TYN**	174	174	128	128	207	207	128	138	179	181	241	241	182	188	247	253	0	0	205	209	158	162		48.76500	17.00750	
9	**TYN6**	**TYN**	180	182	128	130	209	209	136	136	173	175	241	243	184	186	253	253	295	325	209	213	162	170		48.76500	17.00750	
10	**TYN7**	**TYN**	174	180	128	130	215	215	130	130	173	173	241	241	190	190	249	249	285	325	209	213	158	162		48.76500	17.00750	
11	**TYN8**	**TYN**	174	175	128	134	207	221	128	138	171	177	220	240	178	186	249	249	287	293	213	213	162	166		48.76500	17.00750	
12	**TYN9**	**TYN**	156	178	126	130	209	209	128	149	177	177	220	230	184	188	249	255	287	287	217	218	158	166		48.76500	17.00750	
13	**TYN10**	**TYN**	167	178	128	130	209	209	132	134	171	175	243	245	184	186	245	247	285	287	205	218	158	174		48.76500	17.00750	
14	**TYN11**	**TYN**	170	174	130	130	217	217	130	138	173	175	0	0	182	190	251	257	289	291	213	213	166	170		48.76500	17.00750	
15	**TYN12**	**TYN**	174	176	130	132	0	0	134	134	175	177	220	220	186	186	249	251	293	297	209	209	158	162		48.76500	17.00750	
16	**TYN13**	**TYN**	166																									
17	**TYN14**	**TYN**	162																									
18	**TYN15**	**TYN**	156																									
19	**TYN16**	**TYN**	168																									
20	**TYN17**	**TYN**	178	178	126	130	207	211	132	132	171	175	220	238	176	180	249	255	289	289	178	209	162	166		48.76500	17.00750	
21	**TYN18**	**TYN**	161	174	130	130	213	213	132	132	173	177	245	245	178	178	243	249	289	289	205	209	162	182		48.76500	17.00750	
22	**TYN19**	**TYN**	174	180	128	132	213	213	134	134	169	169	220	220	174	182	249	249	285	285	209	213	131	131		48.76500	17.00750	
23	**TYN20**	**TYN**	176	176	130	130	0	0	138	151	171	171	247	247	186	188	249	255	287	287	209	209	131	166		48.76500	17.00750	
24	**TYN21**	**TYN**	178	178	128	134	213	213	132	134	171	173	245	245	180	188	247	249	285	287	209	209	144	154		48.76500	17.00750	
25	**TYN22**	**TYN**	156	175	128	132	217	217	124	132	177	177	220	220	180	182	251	253	287	297	209	217	170	170		48.76500	17.00750	
26	**TYN23**	**TYN**	168	176	128	128	213	213	120	153	173	173	220	236	178	180	245	253	285	287	209	213	154	162		48.76500	17.00750	
27	**TYN24**	**TYN**	176	176	130	130	219	219	132	149	175	175	216	245	182	184	255	255	283	297	205	209	131	174		48.76500	17.00750	
28	**TYN25**	**TYN**	177	179	128	130	209	213	126	126	156	173	241	241	188	188	249	255	289	297	198	213	158	166		48.76500	17.00750	
29	**TYN26**	**TYN**	177	179	126	130	201	201	118	124	171	179	218	245	174	184	249	255	295	295	205	205	158	166		48.76500	17.00750	
30	**TYN27**	**TYN**	176	176	126	128	207	207	120	153	169	173	220	220	182	188	249	255	291	291	205	209	136	162		48.76500	17.00750	
31	**TYN28**	**TYN**	168	172	128	132	219	219	130	130	169	175	249	249	186	188	245	253	285	289	209	213	132	170		48.76500	17.00750	
32	**TYN29**	**TYN**	170	180	128	130	219	219	130	130	171	171	243	243	184	184	249	255	285	289	209	213	166	178		48.76500	17.00750	
33	**TYN30**	**TYN**	170	177	124	130	215	215	138	138	171	177	232	236	184	186	249	253	285	291	209	209	162	174		48.76500	17.00750	
34	**NOV1**	**NOV**	172	176	128	134	209	219	120	120	169	175	253	253	182	184	251	255	289	295	178	209	162	166		48.83320	16.50610	
35	**NOV2**	**NOV**	178	178	130	130	209	209	128	132	173	173	238	243	182	184	249	251	285	287	205	209	136	170		48.83320	16.50610	
36	**NOV3**	**NOV**	163	165	126	126	207	213	122	124	173	175	220	245	178	186	247	247	285	285	205	213	133	179		48.83320	16.50610	
37	**NOV4**	**NOV**	167	178	128	134	205	205	118	124	173	177	220	220	186	188	253	255	287	289	205	213	162	166		48.83320	16.50610	
38	**NOV5**	**NOV**	176	178	128	130	201	201	130	130	173	177	243	245	174	184	247	249	287	293	201	209	162	170		48.83320	16.50610	
39	**NOV6**	**NOV**	176	178	126	130	209	215	128	134	166	173	0	0	184	184	245	245	287	287	205	213	162	166		48.83320	16.50610	

pop1

pop2

genotypy, tj. velikosti fragmentů u jednotlivých jedinců

PYG - Poznámkový blok

Soubor Úpravy Formát Zobrazení Nápověda

```
Title line:"Genotype file for Pipistrellus data"
Ppip01
Ppip02
Ppip04
Ppip06
EF1
EF4
EF6
Paur05
NN18
NnP217
NnP219
Pop
  Pop1, 1820 0505 0809 0511 0308 1321 0209 0506 0608 1214 0513
  Pop1, 1111 0707 0505 1013 0508 0920 0609 0405 0606 1219 1212
  Pop1, 2022 0505 0505 1010 0505 0000 0710 0505 0409 1416 1213
  Pop1, 2424 0607 0505 0407 0607 0419 0710 0405 0405 1417 0512
  Pop1, 1818 0505 0505 0914 1011 1717 0609 0407 0000 1214 1011
  Pop1, 2426 0506 0606 1313 0708 1719 0708 0707 1014 1416 1113
  Pop1, 1824 0506 0909 1010 0707 1717 1010 0505 0514 1416 1011
  Pop1, 1819 0508 0512 0914 0609 0416 0408 0505 0609 1616 1112
  Pop1, 0222 0406 0606 0919 0909 0409 0709 0508 0606 1819 1012
  Pop1, 1122 0506 0606 1112 0608 1920 0708 0304 0506 1219 1014
  Pop1, 1418 0606 1010 1014 0708 0000 0610 0609 0708 1616 1213
  Pop1, 1820 0607 0000 1212 0809 0404 0808 0506 0911 1414 1011
  Pop1, 1020 0407 0910 0621 0505 0419 0408 0506 0606 1416 1113
  Pop1, 0722 0505 1111 0919 0709 0404 0406 0707 0505 0212 1212
  Pop1, 0220 0507 0606 0409 0910 0404 0406 0506 0514 1616 1214
  Pop1, 1222 0505 0910 1112 0708 1919 0404 0405 0707 1214 1113
  Pop1, 2222 0406 0507 1111 0608 0414 0305 0508 0707 0214 1112
  Pop1, 0618 0606 0808 1111 0709 2020 0404 0205 0707 1214 1118
  Pop1, 1824 0507 0808 1212 0505 0404 0206 0505 0505 1416 0101
  Pop1, 2020 0606 0000 1420 0606 2121 0809 0508 0606 1414 0112
  Pop1, 2222 0508 0808 1112 0607 2020 0509 0405 0506 1414 0609
  Pop1, 0219 0507 1010 0711 0909 0404 0506 0607 0611 1418 1313
  Pop1, 1220 0505 0808 0521 0707 0413 0405 0307 0506 1416 0911
  Pop1, 2020 0606 1111 1119 0808 0220 0607 0808 0411 1214 0114
  Pop1, 2123 0506 0608 0808 0207 1717 0909 0508 0711 0916 1012
  Pop1, 2123 0406 0202 0407 0610 0320 0207 0508 1010 1212 1012
  Pop1, 2020 0405 0505 0521 0507 0404 0609 0508 0808 1214 0411
  Pop1, 1216 0507 1111 1010 0508 2222 0809 0307 0507 1416 0213
  Pop1, 1424 0506 1111 1010 0606 1919 0707 0508 0507 1416 1216
  Pop1, 1421 0306 0909 1414 0609 1113 0708 0507 0508 1414 1114
Pop
  Pop2, 1620 0508 0611 0505 0508 2323 0607 0608 0710 0214 1112
  Pop2, 2222 0606 0606 0911 0707 1419 0607 0506 0506 1214 0413
  Pop2, 0809 0404 0508 0607 0708 0420 0408 0404 0505 1216 0317
  Pop2, 1122 0508 0404 0407 0709 0404 0809 0708 0607 1216 1112
  Pop2, 2022 0506 0202 1010 0709 1920 0207 0405 0609 1014 1113
  Pop2, 2022 0406 0609 0912 0307 0000 0707 0303 0606 1216 1112
  Pop2, 1414 0506 0000 1111 0709 2020 0206 0307 0606 1012 1212
  Pop2, 1818 0505 0710 1111 0708 2020 0709 0305 0606 1516 1213
  Pop2, 2020 0405 0208 1021 0606 0412 0607 0305 0606 1212 1012
  Pop2, 1820 0505 0808 0809 0709 0319 0710 0407 0000 1212 0212
  Pop2, 0220 0406 0707 0909 0909 0304 0808 0208 1617 1112 0213
  Pop2, 2020 0206 0505 0509 0307 0517 0308 0105 0914 1014 0213
  Pop2, 2024 0607 1010 1212 0307 1920 0809 0407 0505 1414 1213
  Pop2, 2024 0509 0202 1515 0206 0420 0307 0508 0707 1414 1112
  Pop2, 1624 0507 0205 1115 0208 0420 0709 0205 0707 1214 1215
  Pop2, 1621 0607 1010 0921 0609 2222 0708 0707 1415 1416 1113
  Pop2, 1616 0506 0508 1010 0607 0517 0306 0509 0608 1117 1215
  Pop2, 1420 0507 0909 1010 0708 0416 0710 0506 0506 1214 0912
  Pop2, 2024 0405 1111 0910 0507 1921 0709 0105 0607 1014 0912
  Pop2, 1620 0508 0611 0406 0511 0319 0610 0506 0610 0212 0211
  Pop2, 1220 0505 0609 0909 0809 1717 0208 0708 0606 1416 0412
  Pop2, 2222 0506 0808 1010 0608 1919 0708 0307 0506 1114 0210
Pop
  Pop3, 2125 0407 0000 0606 0606 0000 0409 0708 0406 1616 0513
  Pop3, 1620 0405 0609 1414 0708 1418 0203 0405 0505 1015 1011
  Pop3, 1818 0205 0910 1010 0307 0505 0609 0411 0410 1115 1112
  Pop3, 1823 0406 1010 0306 0810 1420 0208 0303 0809 1212 0712
```

Start Doručená p... GMZ_lectures GenepopPYG

Genepop file format – jednoduchý ASCI kód (.txt)

- jednotlivé alely pro daný lokus jsou seřazeny podle velikosti a očíslovány
- tj. např. 128/130 je převedeno na 10/11

Vnitropopulační variabilita

Polymorfismus

- podíl polymorfních lokusů (znaků) – např. 0,8 = 4 z pěti zkoumaných mikrosatelitů mají v populaci alespoň 2 alely, z nichž ta vzácnější dosahuje frekvence alespoň 1% nebo 5%

Počet alel (number of alleles)

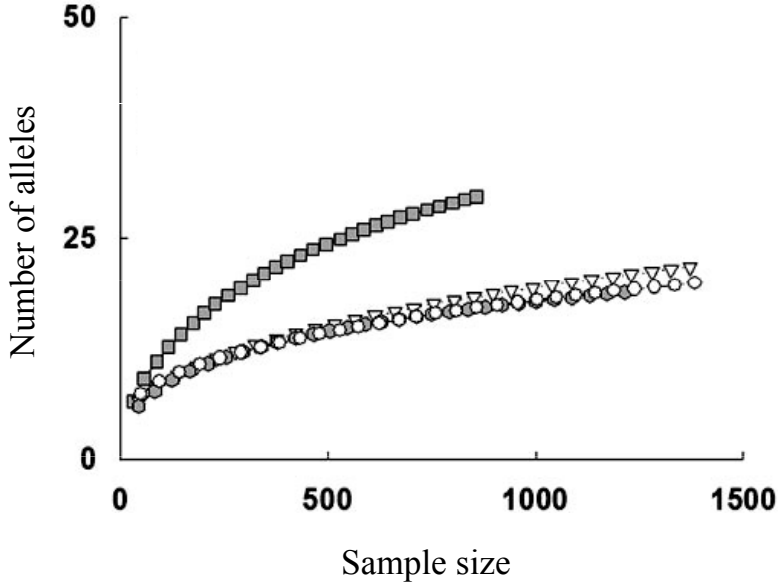
- počet alel na lokus

Alelická bohatost (allelic richness)

- počet alel na lokus vztažený k velikosti vzorku (metodou „rarefaction“) - **FSTAT**

Pozorovaná heterozygotnost (observed heterozygosity)

- průměrná četnost heterozygotů v jednotlivých lokusech



Hardy-Weinbergova rovnováha (HWE)

= rozmístění alel do genotypů v rovnovážné populaci

Alela	Četnost alely
170	p
172	q

Př. Jeden lokus se 2 alelami

$p + q = 1$

p, q - zjistíme analýzou svých vzorků

Genotyp	Očekávaná četnost genotypu
170/170	p^2
170/172	2pq
172/172	q^2

= Hardy-Weinbergova rovnováha

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

- četnosti genotypů zjistíme analýzou svých vzorků
- odchylky od očekávaných četností

$\Rightarrow$ **test HWE** (např. χ^2 test nebo exact probability test v Genepop)

Očekávaná heterozygotnost (expected heterozygosity, H_e) při HWE

$H_e=1-(p^2+q^2)$ ..... pro 1 lokus se 2 alelami s četností *p* a *q*

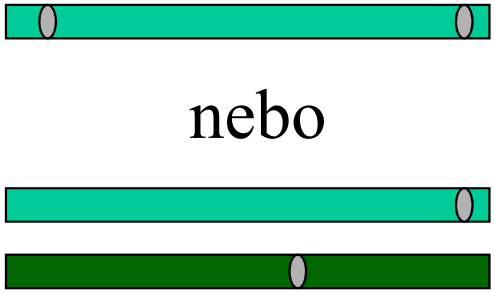
Předpoklady HWE

- náhodné párování (panmixia)
- zanedbatelný efekt mutací a migrací („closed populations“)
- nekonečně velká populace
- Mendelovská dědičnost použitých markerů
- neutrální znaky – žádná selekce

- znaky nejsou ve vazbě – kontrola na „linkage disequilibrium“ (vazebná nerovnováha; speciální test např. v **Genepop**)

2 lokusy ve fyzické blízkosti
(snížená pravděpodobnost rekombinace
linkage disequilibrium)

vs.



2 lokusy fyzicky vzdálené
(pravděpodobnost rekombinace není ovlivněna
linkage equilibrium)

Odchylky od HW rovnováhy

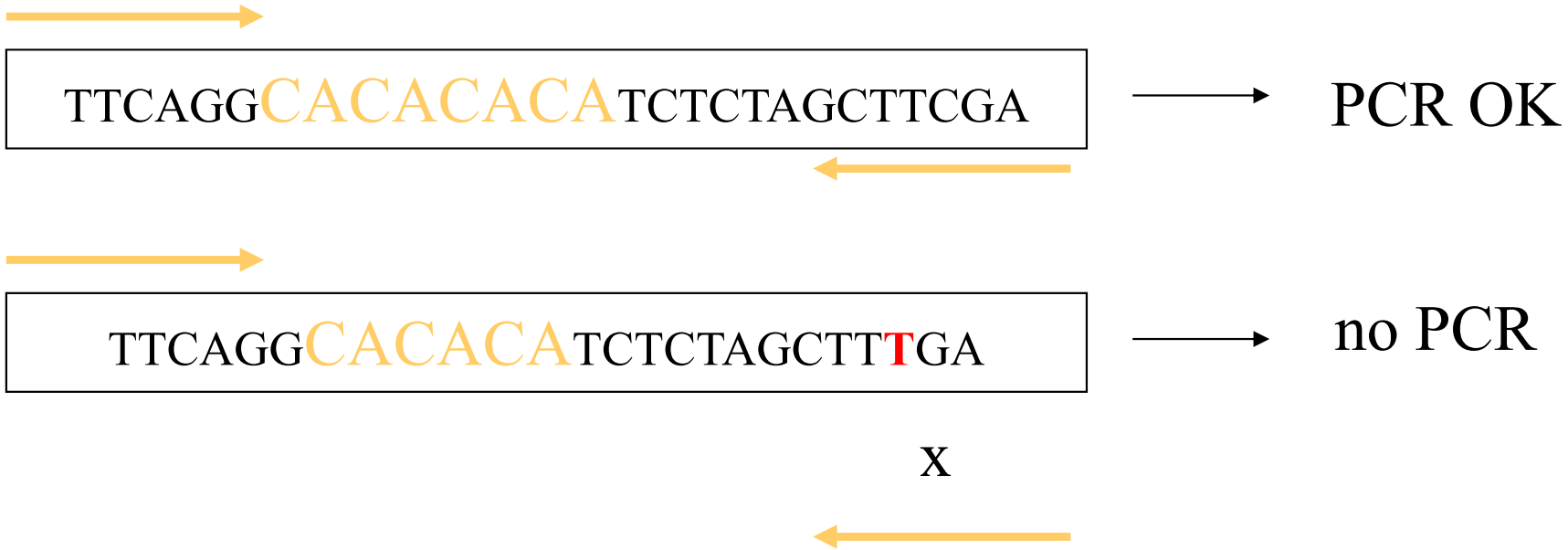
- Test HWE – např. **Genepop** („exact probability tests“) – pokud jsou odchylky, tak některý předpoklad HWE nebyl splněn

- nadbytek heterozygotů = negativní asortativní páření (tj. cílené rozmnožování nepodobných jedinců) – použité lokusy mohou být výhodné v heterozygotním stavu (např. geny MHC)

- nedostatek heterozygotů
- **inbreeding** (postihuje všechny lokusy stejně)
- **vnitřní struktura** („Wahlundův efekt“)
- **nulové alely** (jen na některých lokusech bude deficit heterozygotů) - FreeNA

Null alleles

- nulové alely (mutace v primerových sekvencích) → vyšší proporce „homozygotů“



Příklad – stanovení variability populace

Jedinec	Locus 1	Locus 2	Locus 3	Locus 4	Průměr
Ind 1	170/170	223/227	116/116	316/316	
Ind 2	170/172	223/225	112/112	316/316	
Ind 3	172/172	223/225	112/112	316/316	
Ind 4	170/172	223/227	112/112	316/316	
Počet alel	2	3	2	1	**2**
Ho	0,5	1,00	0	0	**0,375**
p	0,5	p = 0,5	0,75	1,00	
q	0,5	q = 0,25 r = 0,25	0,25	0	
He	0,5	0,625	0,375	0	**0,375**

$H_e=1-(p^2+q^2)$

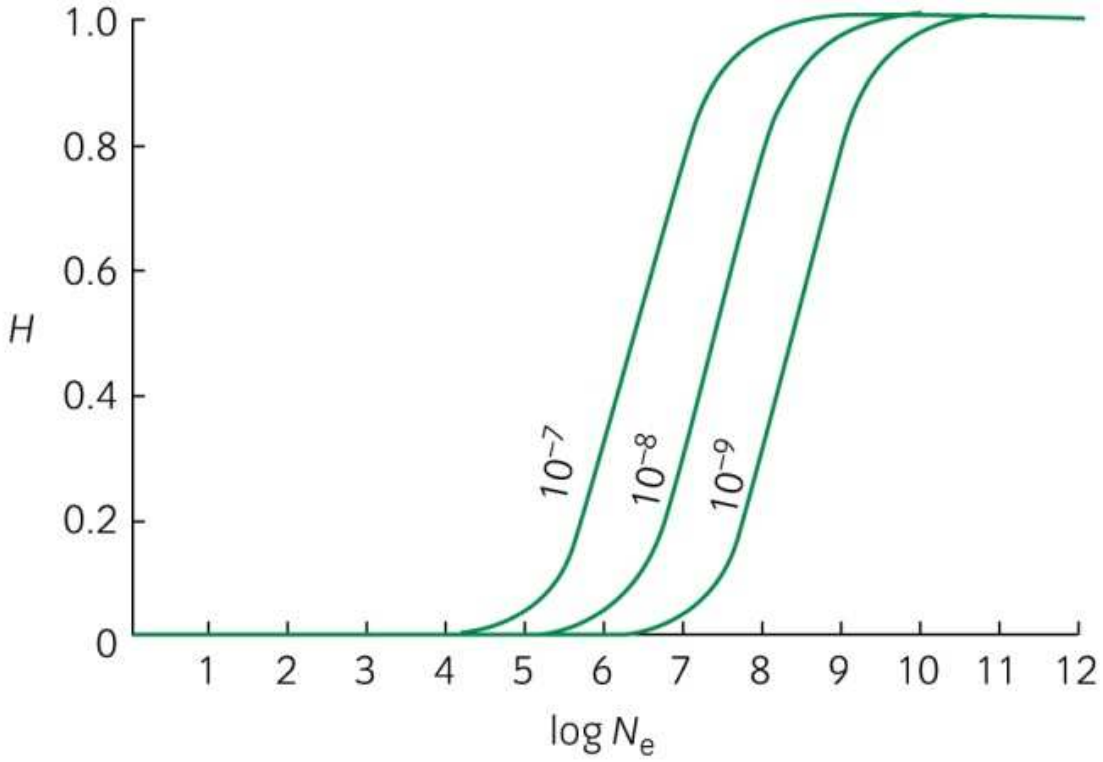
$H_e=1-(p^2+q^2+r^2)$

Polymorfismus = 0,75

Genetická variabilita (H_e) jako ukazatel efektivní velikosti populace (N_e)

- neutrální genetická teorie:
  $H_e=4N_e\mu/[4N_e\mu+1]$

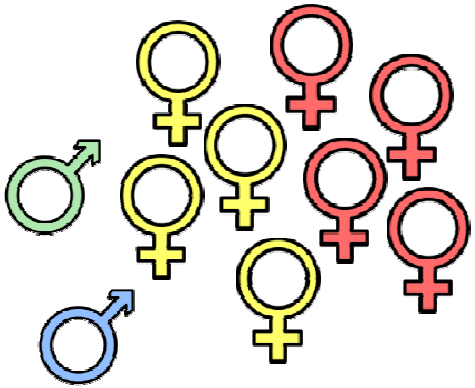
- mutation-drift equilibrium

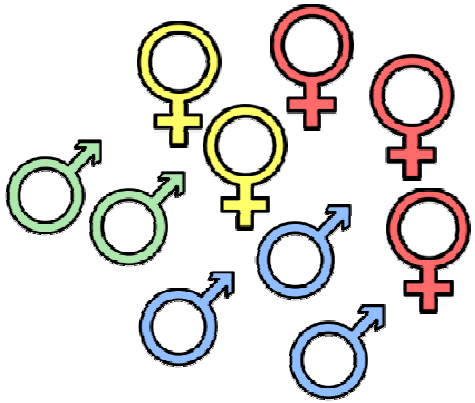


Čím je populace větší, tím má vyšší genetickou variabilitu

Efektivní velikost populace (N_e)

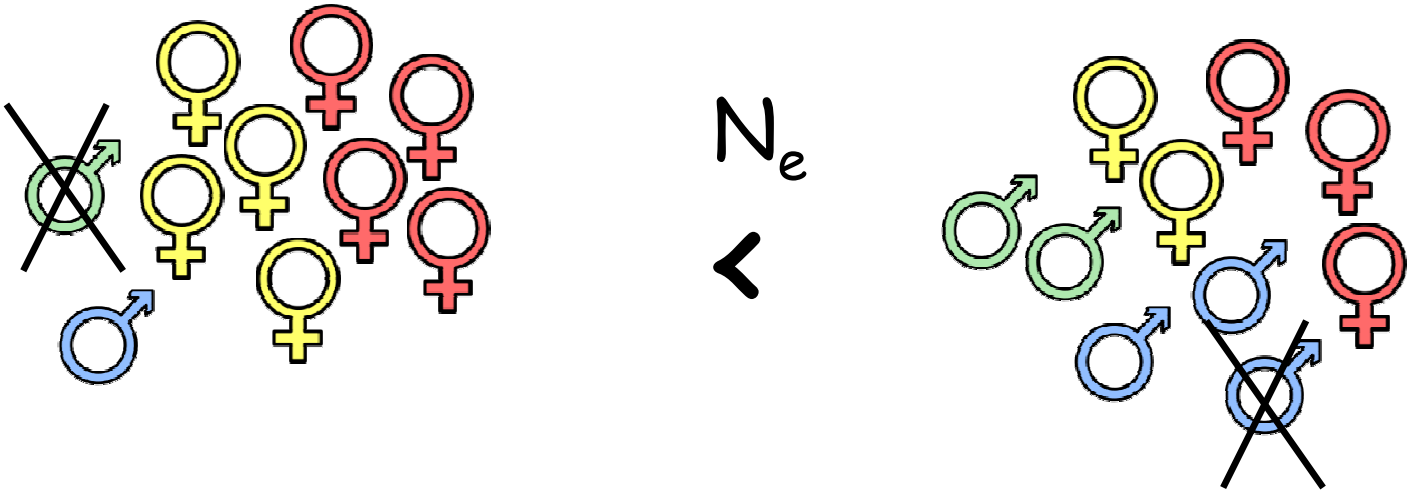
- N_e = velikost ideální populace (náhodné páření, rovnoměrný poměr pohlaví), která ztrácí genetickou diverzitu stejnou rychlostí jako aktuální populace (vlivem náhody)
- ovlivněna genetickou a věkovou strukturou, poměrem pohlaví, intenzitou příbuzenského křížení atd.





Efektivní velikost populace (N_e)

- N_e = velikost ideální populace (náhodné páření, rovnoměrný poměr pohlaví), která ztrácí genetickou diverzitu stejnou rychlostí jako aktuální populace (vlivem náhody)
- ovlivněna genetickou a věkovou strukturou, poměrem pohlaví, intenzitou příbuzenského křížení atd.



- **vývoj genetické variability v malých populacích závisí na N_e více než na N**

Důsledky poklesu N_e

Pokles pozorované variability (H_e – odhadnuta z tzv. genetických znaků) = pokles N_e

$$H_e = 4N_e\mu / [4N_e\mu + 1]$$

Pokles adaptivní variability

„Budoucnost je v genech …“

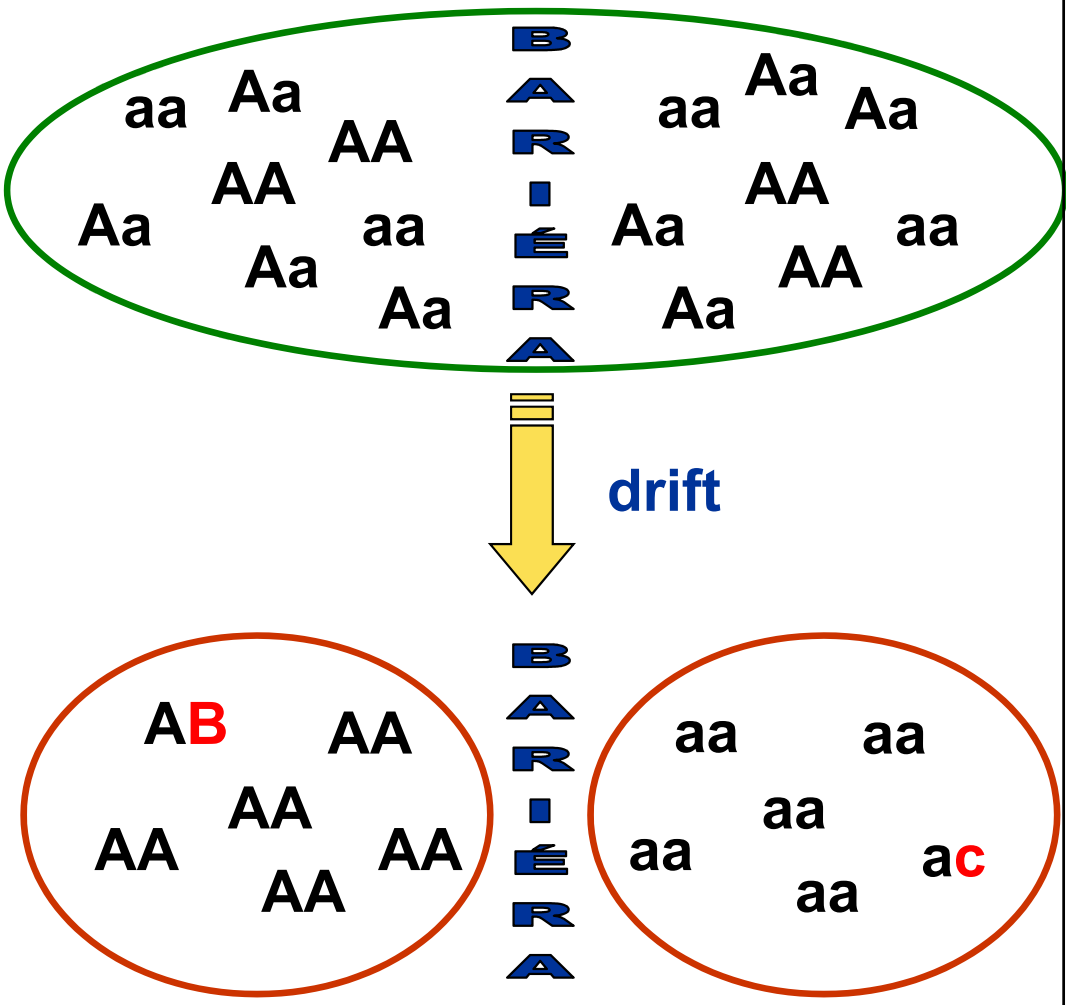
- Změny prostředí (např. klima, patogeny, aj.) → přizpůsob se nebo zmiz!
- Vysoká genetická variabilita znamená, že je zde více genetických variant, z nichž některé mohou být vhodné pro nové podmínky
- Populace s vyšší genetickou variabilitou mají „více losů v loterii“ (= **adaptivní potenciál**)

Dryas octopetala v Alpách, Německo. *Photo: Steve Smith*

Genetická struktura populací
drift, mutace

- **Drift**
  → diferenciace subpopulací
  díky změnám frekvencí (až fixaci)
  alternativních alel

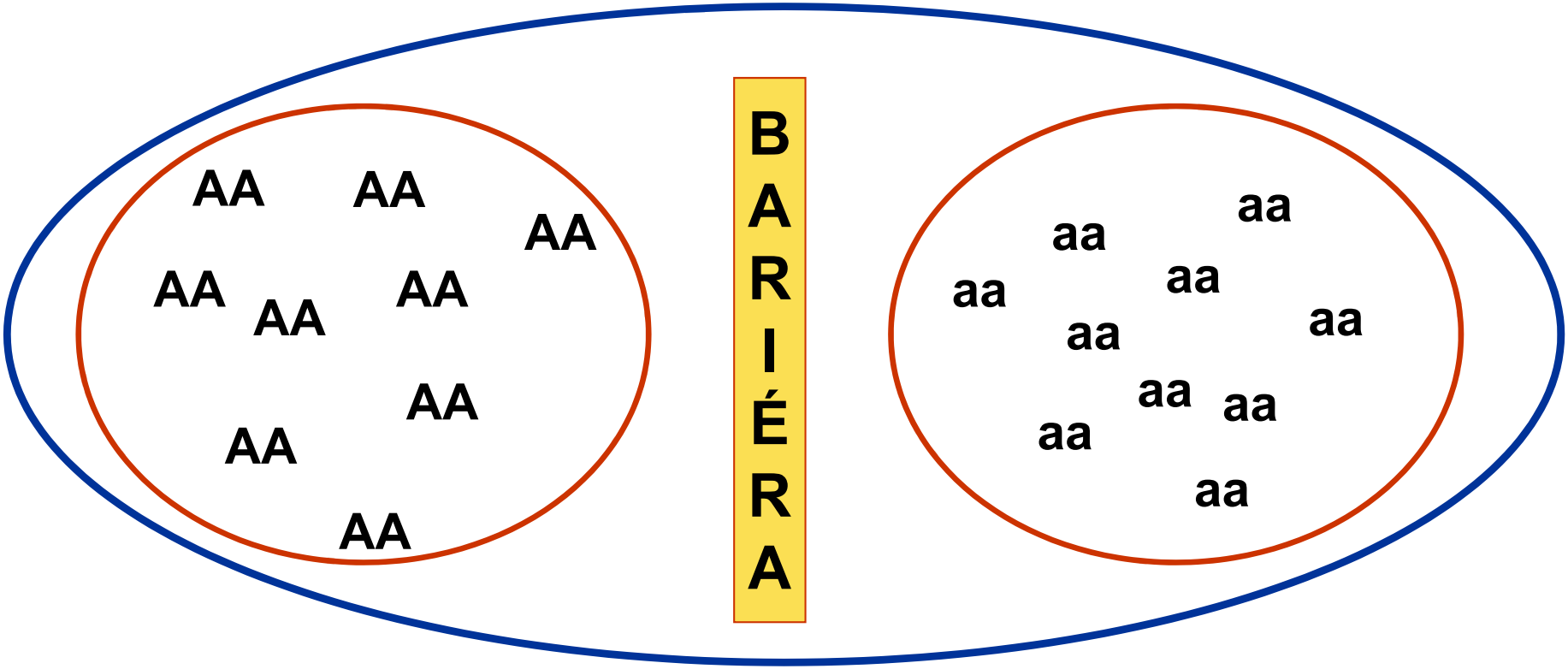
- **Mutace**
  mohou zvýšit diferenciaci
  (nebo ne – homoplázií)



Migrace (genový tok)
- působí proti diferenciaci subpopulací

Vliv populační struktury na heterozygotnost

- Wahlundův princip
- Dvě izolované subpopulace s fixovanými alelami
- Subpopulace v HW, celkově v populaci však nedostatek heterozygotů



Wahlundův princip - příklad

- Jezero Bunnersjöarna (severní Švédsko) – „brown trout“
- 2 alely na jednom znaku

	170/170	170/172 (= Ho)	172/172	Total	p	2pq (=He)
Přítok	50	0 (**0**)	0	50	1.000	**0.000**
Odtok	1	13 (**0.26**)	36	50	0.150	**0.255**
Celé jezero (expected)	51 (33.1)	13 **(0.13)** (48.9)	36 (18.1)	100	0.575	**0.489**

$p^2 = 0.575^2$ $q^2 = 0.425^2$

Ryman et al. 1979

Výpočet F_{ST} – snížení heterozygotnosti v důsledku bariéry mezi subpopulacemi

	170/170	170/172 (=Ho)	172/172	Total	p	2pq (=He)
Přítok	50	0 (**0**)	0	50	1.000	**0.000**
Odtok	1	13 (**0.26**)	36	50	0.150	**0.255**
Celé jezero	51	13 **(0.13)**	36	100	0.575	**0.489**
(expected)	(33.1)	(48.9)	(18.1)			

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} = \frac{0.489 - 0.128}{0.489} = 0.728$$

V důsledku bariéry toku genů je heterozygotnost o 72.8% nižší než by byla v panmiktické populaci

Ryman et al. 1979

Factorial correspondence analysis

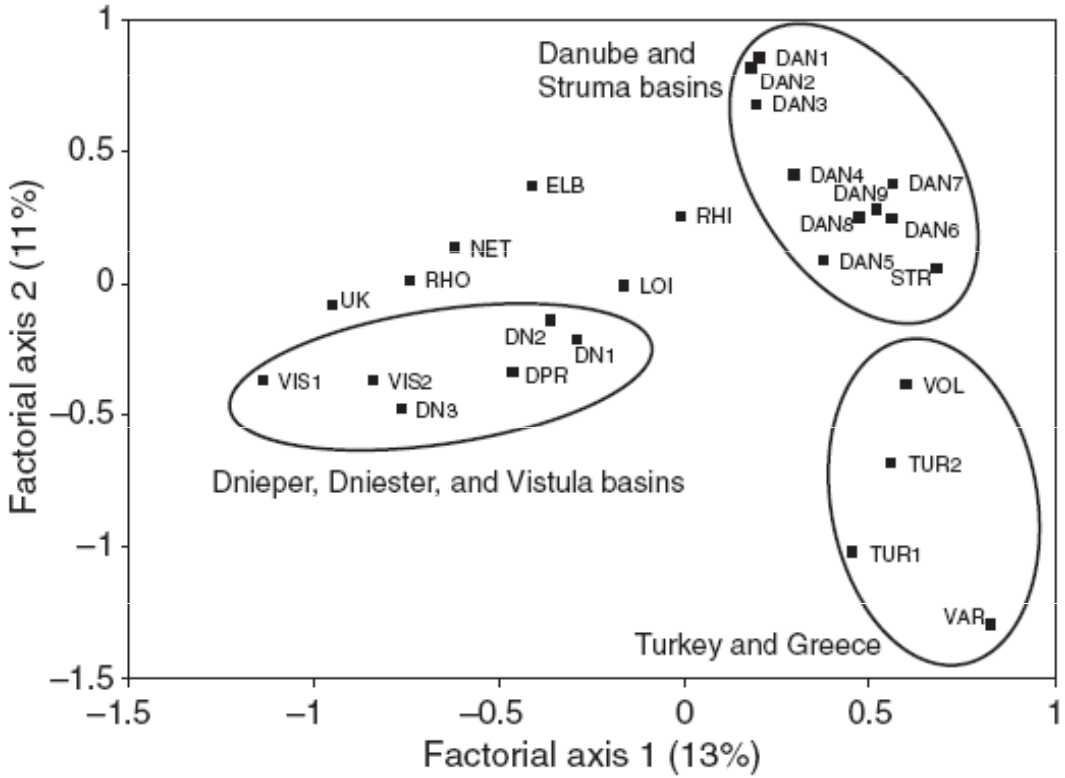


Fig. 2 A two-dimensional plot of the factorial correspondence analysis performed using GENETIX based on 12 microsatellite loci. Three geographical groups are bounded by grey lines.

- each locus as one variable, reduction of number of variables
- **Genetix** – orientační zjištění strukturovanosti populace
- individuals vs. populations

Bayesian clustering approach

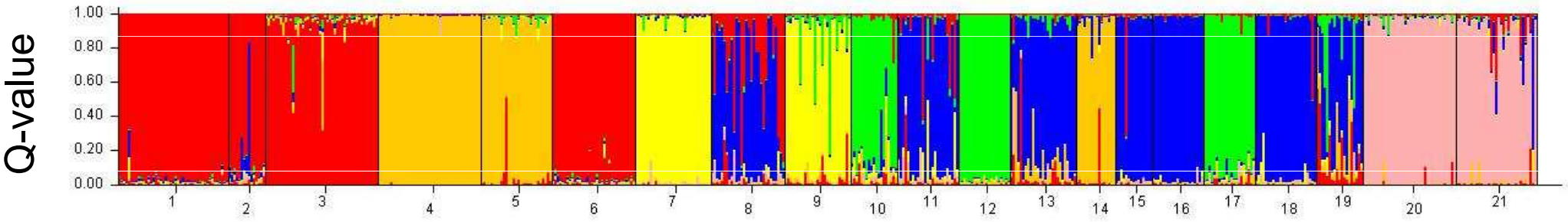
STRUCTURE - *Pritchard et al. 2000*

... Když předem strukturu neznáme

- Neznámý počet populací charakterizovaných různými frekvencemi alel → počet populací a frekvence zjišťuji
- Současně přiřazuji individua do populací
- Lokusy, které nejsou ve vazbě, HW uvnitř subpopulací (např. mikrosatelity, SNPs)
- Model se snaží vyložit HW nebo vazebnou nerovnováhu zavedením populační struktury

STRUCTURE – když předem strukturu neznáme

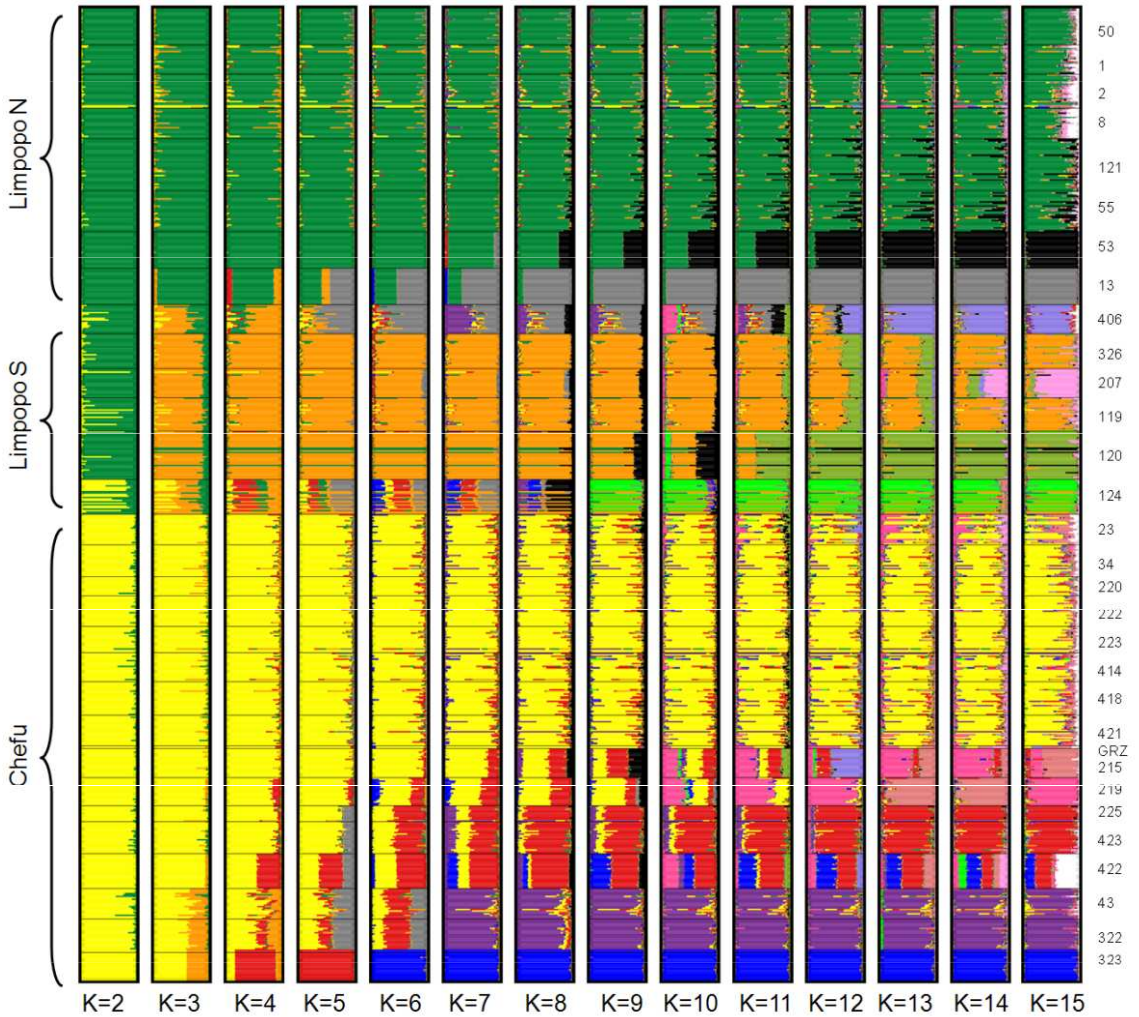
„Barplot“ pro K = 7



„Proporce genomu“ každého jedince náležející určitému „clusteru“

Existuje i jednodušší „non-admixture model“

Alternativní vizualizace výsledků ze STRUCTURE

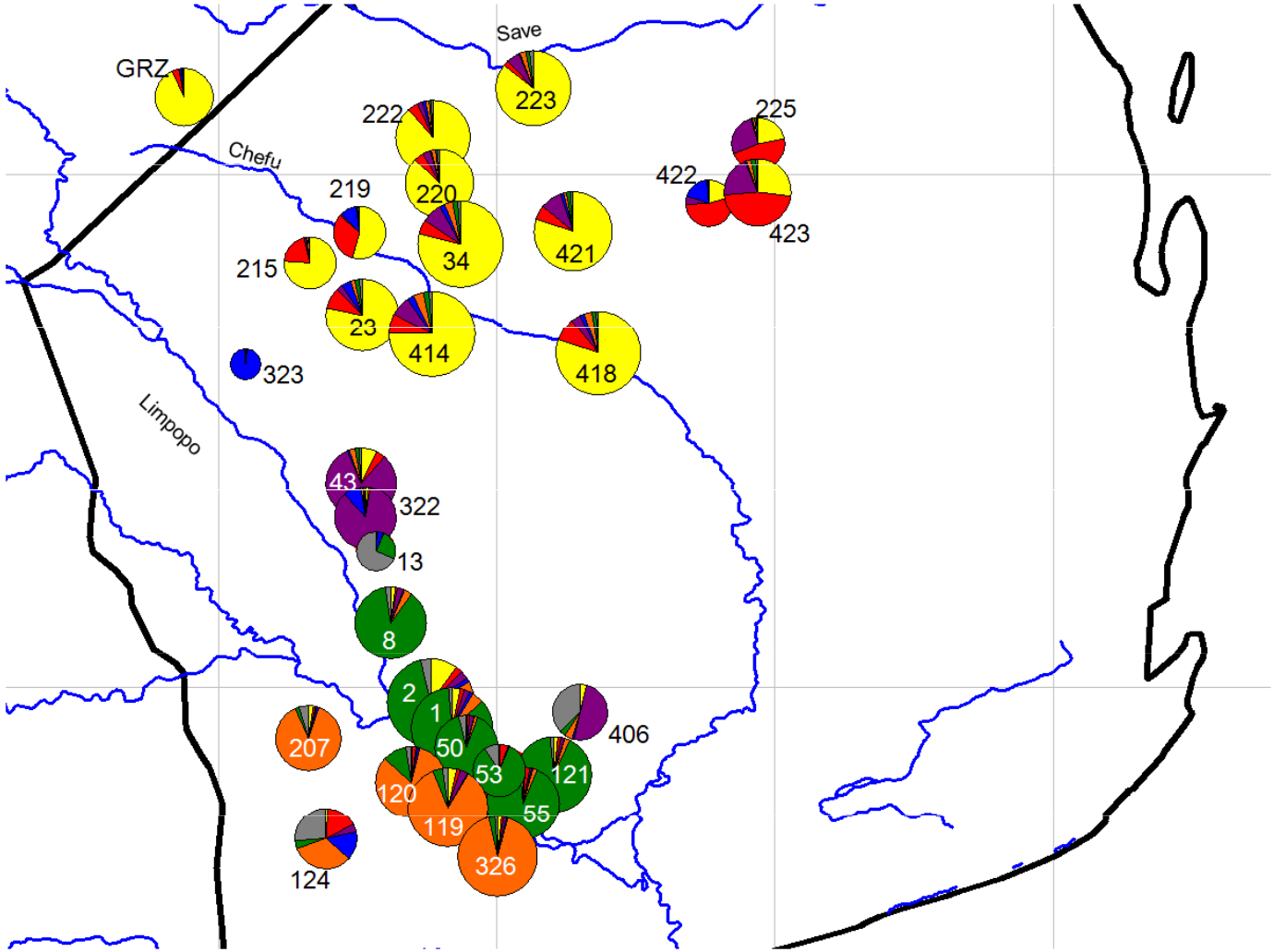


„forced clustering“

Zobrazení hierarchické struktury mezi populacemi

Bartáková et al. 2013

- Q-values pro celé populace (ne jedince)



Bartáková et al. 2013