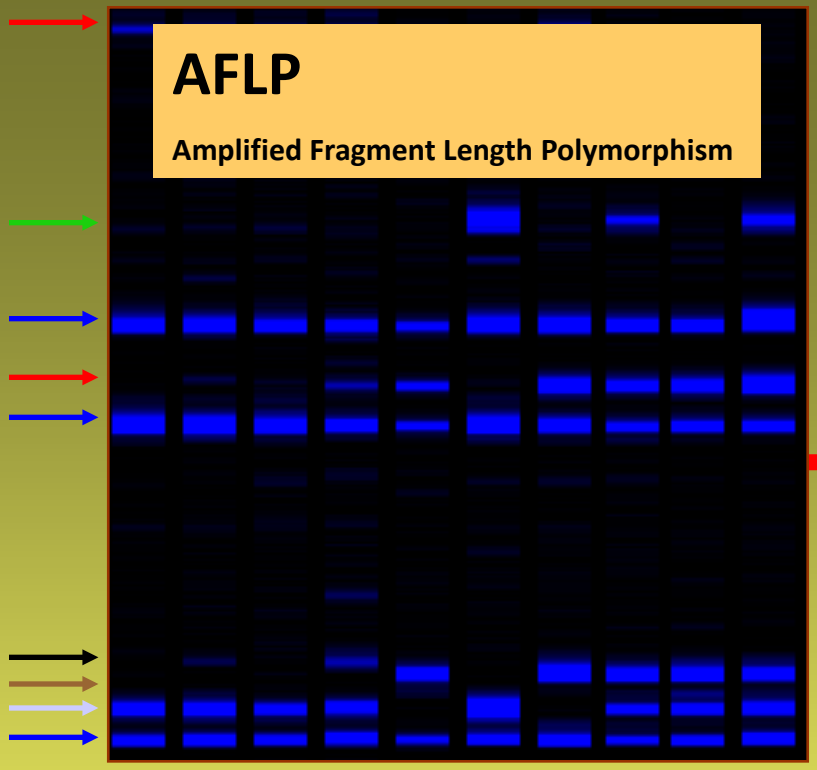
Hodnocení multilokusových molekulárních dat

Tomáš Fér & Filip Kolář
Katedra Botaniky PřF UK, Praha

Multilokusová data

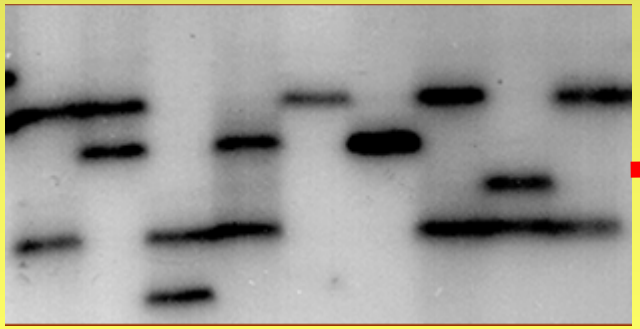
- **dominantní** – nejsme schopni odlišit heterozygoty od dominantních homozygotů
- **binární** – bialelická povaha lokusu (fragmentu)
  - přítomnost (dominantní alela/heterozygot)
  - nepřítomnost (recesivní alela)
  - tj. skorování 0-1
- **anonymní** – nevíme z jaké části genomu pocházejí
- **multilokusová** – často zároveň analyzujeme stovky lokusů, tj. analýza pokrývá „celý genom“
- RAPD, AFLP, ISSR…

- **kodominantní** – odlišíme homo- a heterozygoty, tj. detekujeme všechny alely
- **alelická** – známe frekvence alel v lokusech, v populacích…
- **anonymní** – nevíme z jaké části genomu pocházejí
- **multilokusová** – nejčastěji analýza malého množství lokusů (5-20)
- mikrosatelity (SSRs), isozymy



1	1	1	1	0	0	1	0	0	0
0	0	0	0	0	1	0	1	0	1
1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	0	1	1	1	1
1	1	1	1	1	1	1	1	1	1
0	1	0	1	0	0	0	0	0	0
0	0	0	0	1	0	1	1	1	1
1	1	1	1	0	1	0	1	0	1
1	1	1	1	1	1	1	1	1	1

2
3
4
5
6



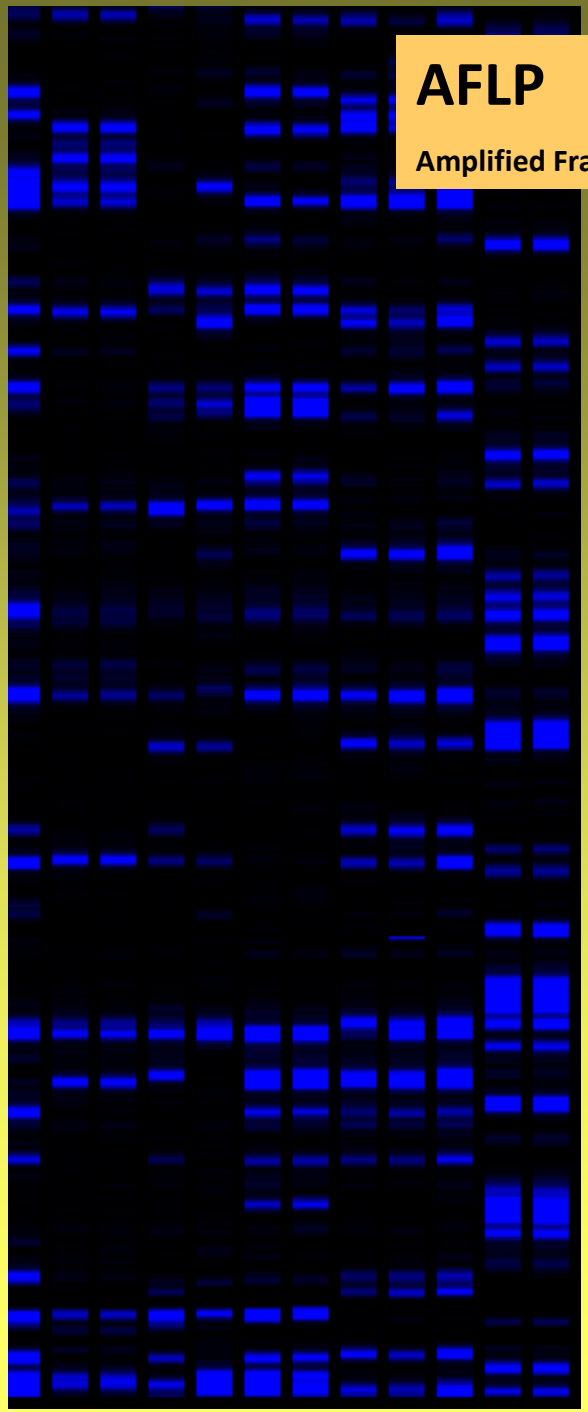
SSRs

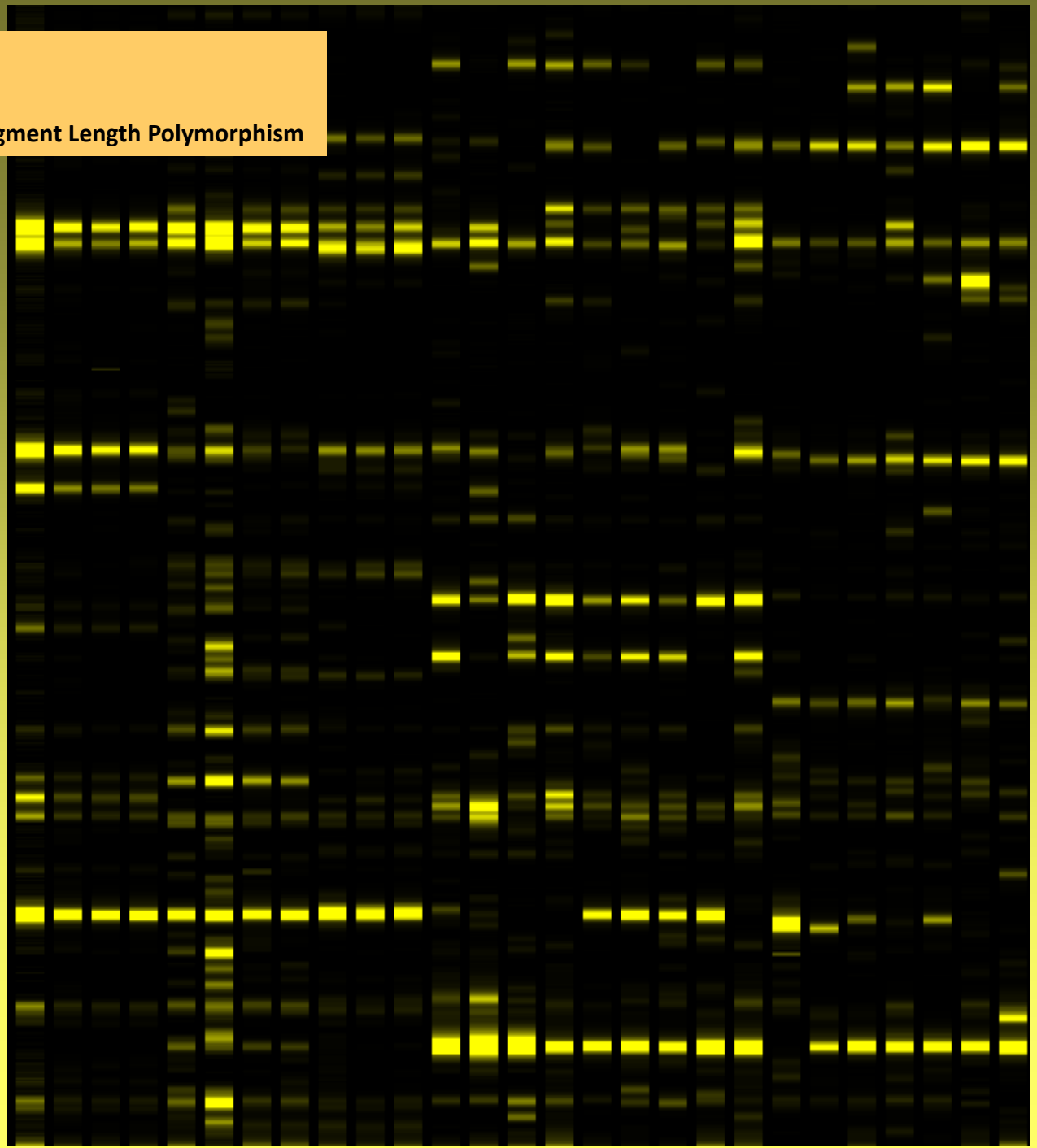
Simple sequence repeats

Alela 1	Alela 2
2	5
2	3
5	6
3	5
2	2
3	3
2	5
4	5
2	5

AFLP

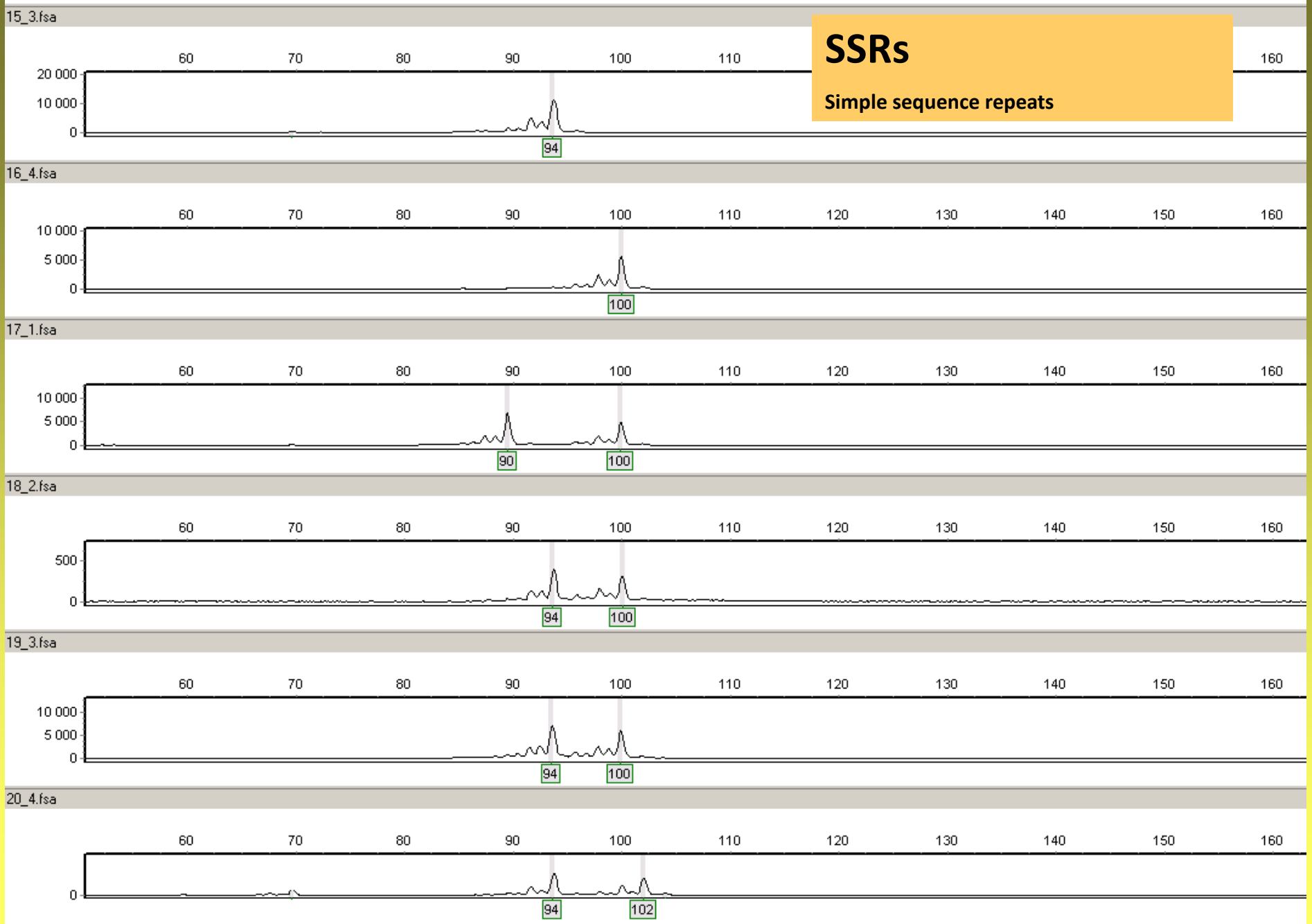
Amplified Fragment Length Polymorphism





SSRs

Simple sequence repeats



AFLP

Výhody

- vysoká variabilita – mnoho lokusů
- mnoho nezávislých lokusů (*multilocus method*)
- pokrytí „celého“ genomu
- statistický aparát k analýze dat

Nevýhody

- anonymní marker
- asymetrie v pravděpodobnosti získání a ztráty fragmentu – ano/ne?
- dominantní – nemožnost odlišit homozygoty od heterozygotů
- subjektivita při hodnocení
- neznámá rychlost akumulace mutací (nemožnost použít molekulární hodiny)
- problematické (nemožné) přidávání dalších vzorků

mikrosatelity

Výhody

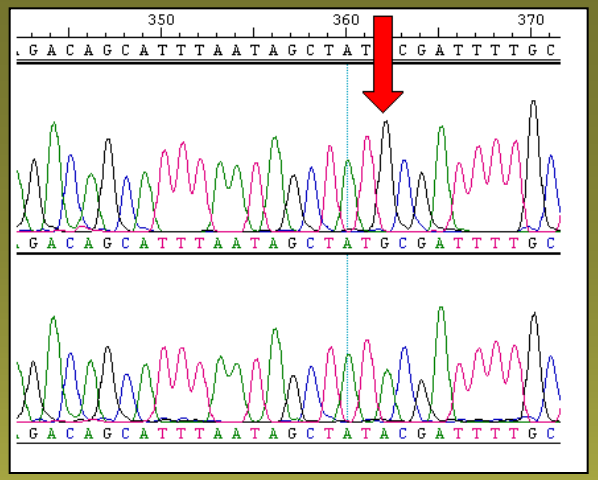
- obvykle vysoká variabilita – mnoho alel
- kodominantní – odlišení homozygotů od heterozygotů, frekvence alel
- modely evoluce alel – „známé“ vztahy mezi alelami
- objektivnější hodnocení
- statistický aparát k analýze dat
- možnosti přidávání dalších vzorků

Nevýhody

- druhově specifický marker
- současná analýza omezeného počtu lokusů
- omezenější reprezentace „celého“ genomu

SNPs

= single nucleotide polymorphisms

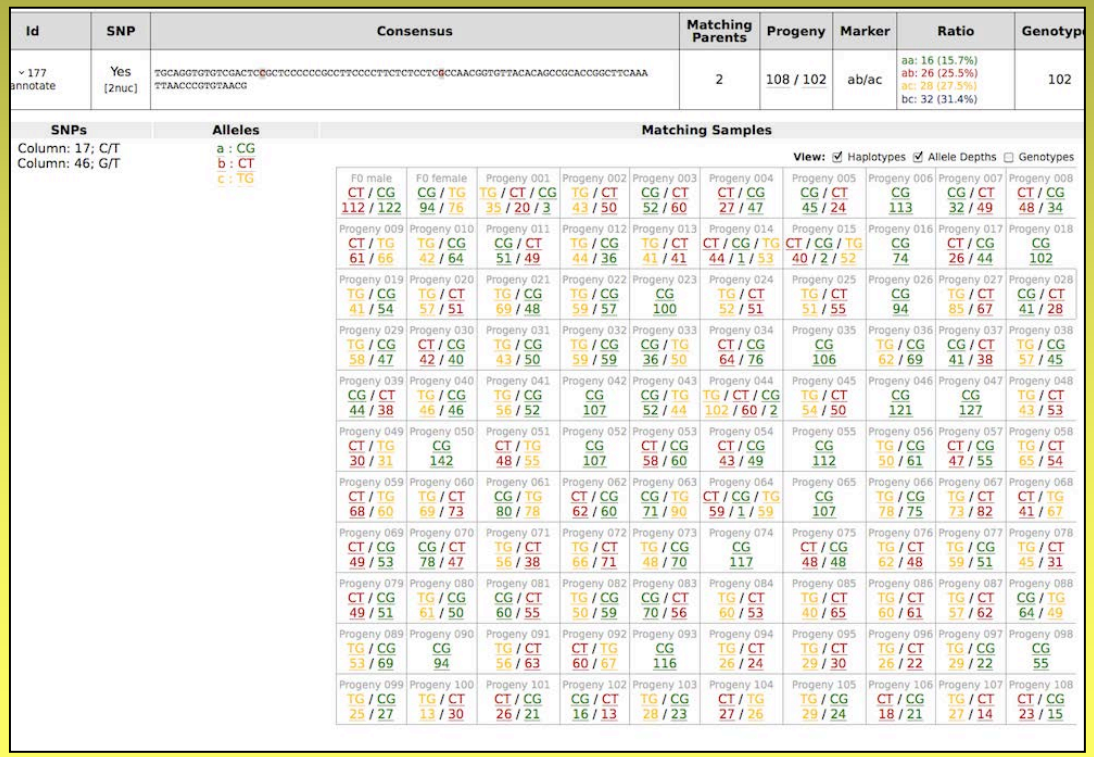


Výhody

- kombinuje výhody AFLP a SSR
- kodominantní – převážně bialelické
- sekvenční (NGS – např. RADseq)
- neanonymní
- multilocus
- substituční změny => modely evoluce
- až desítky tisíc lokusů
- studium neutrální i adaptivní variability (selekce)

Nevýhody

- neustálený statistický aparát
- neustálené laboratorní techniky
- RADseq: nulové alely a coverage bias



Id	SNP	Consensus	Matching Parents	Progeny	Marker	Ratio	Genotype
~177 annotate	Yes [2nuc]	TGCAGGTGTGTCGACTCGCTCCCCCCGCCTTCCCCTTCTCTCCTCGCCAACGGTGTTACACAGCCGCACCGGCTTCAAA TTAACCCGTGTAACG	2	108 / 102	ab/ac	aa: 16 (15.7%) ab: 26 (25.5%) ac: 28 (27.5%) bc: 32 (31.4%)	102

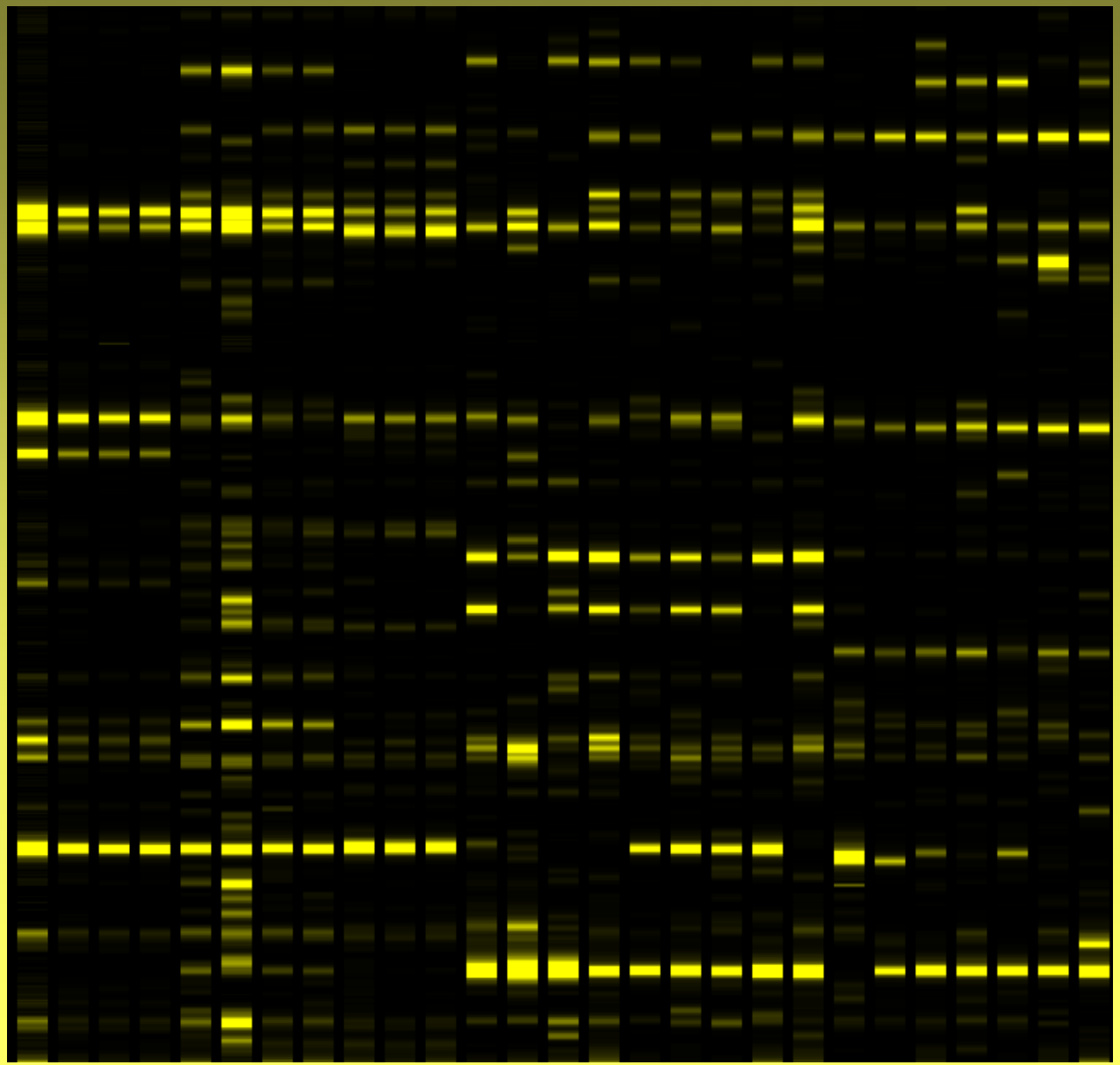
SNPs	Alleles
Column: 17; C/T	a : CG
Column: 46; G/T	b : CT
	c : TG

Matching Samples

View: ☑ Haplotypes ☑ Allele Depths ☐ Genotypes

F0 male CT / CG 112 / 122	F0 female CG / TG 94 / 76	Progeny 001 TG / CT / CG 35 / 20 / 3	Progeny 002 TG / CT 43 / 50	Progeny 003 CG / CT 52 / 60	Progeny 004 CT / CG 27 / 47	Progeny 005 CG / CT 45 / 24	Progeny 006 CG 113	Progeny 007 CG / CT 32 / 49	Progeny 008 CT / CG 48 / 34
Progeny 009 CT / TG 61 / 66	Progeny 010 TG / CG 42 / 64	Progeny 011 CG / CT 51 / 49	Progeny 012 TG / CG 44 / 36	Progeny 013 TG / CT 41 / 41	Progeny 014 CT / CG / TG 44 / 1 / 53	Progeny 015 CT / CG / TG 40 / 2 / 52	Progeny 016 CG 74	Progeny 017 CT / CG 26 / 44	Progeny 018 CG 102
Progeny 019 TG / CG 41 / 54	Progeny 020 TG / CT 57 / 51	Progeny 021 TG / CG 69 / 48	Progeny 022 TG / CG 59 / 57	Progeny 023 CG 100	Progeny 024 TG / CT 52 / 51	Progeny 025 TG / CT 51 / 55	Progeny 026 CG 94	Progeny 027 TG / CT 85 / 67	Progeny 028 CG / CT 41 / 28
Progeny 029 TG / CG 58 / 47	Progeny 030 CT / CG 42 / 40	Progeny 031 TG / CG 43 / 50	Progeny 032 TG / CG 59 / 59	Progeny 033 CG / TG 36 / 50	Progeny 034 CT / CG 64 / 76	Progeny 035 CG 106	Progeny 036 TG / CG 62 / 69	Progeny 037 CG / CT 41 / 38	Progeny 038 TG / CG 57 / 45
Progeny 039 CG / CT 44 / 38	Progeny 040 TG / CG 46 / 46	Progeny 041 TG / CG 56 / 52	Progeny 042 CG 107	Progeny 043 CG / TG 52 / 44	Progeny 044 TG / CT / CG 102 / 60 / 2	Progeny 045 TG / CT 54 / 50	Progeny 046 CG 121	Progeny 047 CG 127	Progeny 048 TG / CT 43 / 53
Progeny 049 CT / TG 30 / 31	Progeny 050 CG 142	Progeny 051 CT / TG 48 / 55	Progeny 052 CG 107	Progeny 053 CT / CG 58 / 60	Progeny 054 CT / CG 43 / 49	Progeny 055 CG 112	Progeny 056 TG / CG 50 / 61	Progeny 057 CT / CG 47 / 55	Progeny 058 TG / CT 65 / 54
Progeny 059 CT / TG 68 / 60	Progeny 060 TG / CT 69 / 73	Progeny 061 CG / TG 80 / 78	Progeny 062 CT / CG 62 / 60	Progeny 063 CG / TG 71 / 90	Progeny 064 CT / CG / TG 59 / 1 / 59	Progeny 065 CG 107	Progeny 066 TG / CG 78 / 75	Progeny 067 TG / CT 73 / 82	Progeny 068 CT / TG 41 / 67
Progeny 069 CT / CG 49 / 53	Progeny 070 CG / CT 78 / 47	Progeny 071 TG / CT 56 / 38	Progeny 072 TG / CT 66 / 71	Progeny 073 TG / CG 48 / 70	Progeny 074 CG 117	Progeny 075 CT / CG 48 / 48	Progeny 076 TG / CT 62 / 48	Progeny 077 TG / CG 59 / 51	Progeny 078 TG / CT 45 / 31
Progeny 079 CT / CG 49 / 51	Progeny 080 TG / CG 61 / 50	Progeny 081 CG / CT 60 / 55	Progeny 082 TG / CG 50 / 59	Progeny 083 CG / CT 70 / 56	Progeny 084 TG / CT 60 / 53	Progeny 085 TG / CT 40 / 65	Progeny 086 TG / CT 60 / 61	Progeny 087 TG / CT 57 / 62	Progeny 088 CG / TG 64 / 49
Progeny 089 TG / CG 53 / 69	Progeny 090 CG 94	Progeny 091 TG / CT 56 / 63	Progeny 092 CT / TG 60 / 67	Progeny 093 CG 116	Progeny 094 TG / CT 26 / 24	Progeny 095 TG / CT 29 / 30	Progeny 096 TG / CT 26 / 22	Progeny 097 TG / CG 29 / 22	Progeny 098 CG 55
Progeny 099 TG / CG 25 / 27	Progeny 100 TG / CT 13 / 30	Progeny 101 CT / CG 26 / 21	Progeny 102 CG / CT 16 / 13	Progeny 103 TG / CG 28 / 23	Progeny 104 CT / TG 27 / 26	Progeny 105 TG / CG 29 / 24	Progeny 106 CT / CG 18 / 21	Progeny 107 TG / CT 27 / 14	Progeny 108 CT / CG 23 / 15

Hodnocení binárních dat

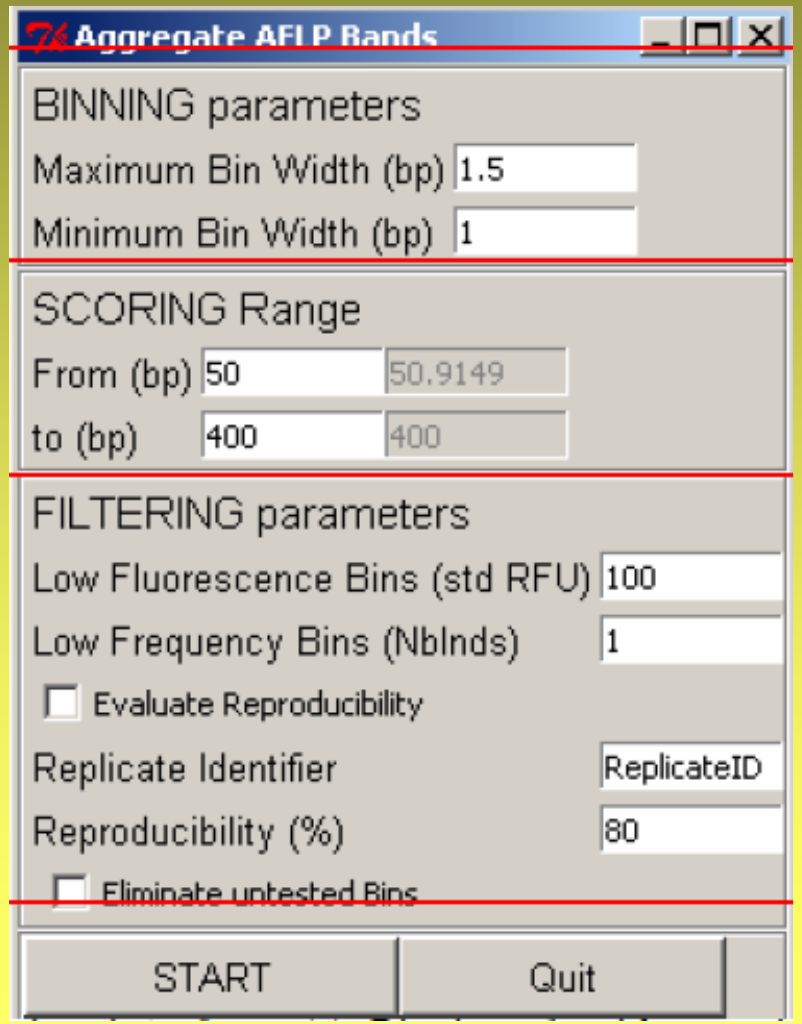


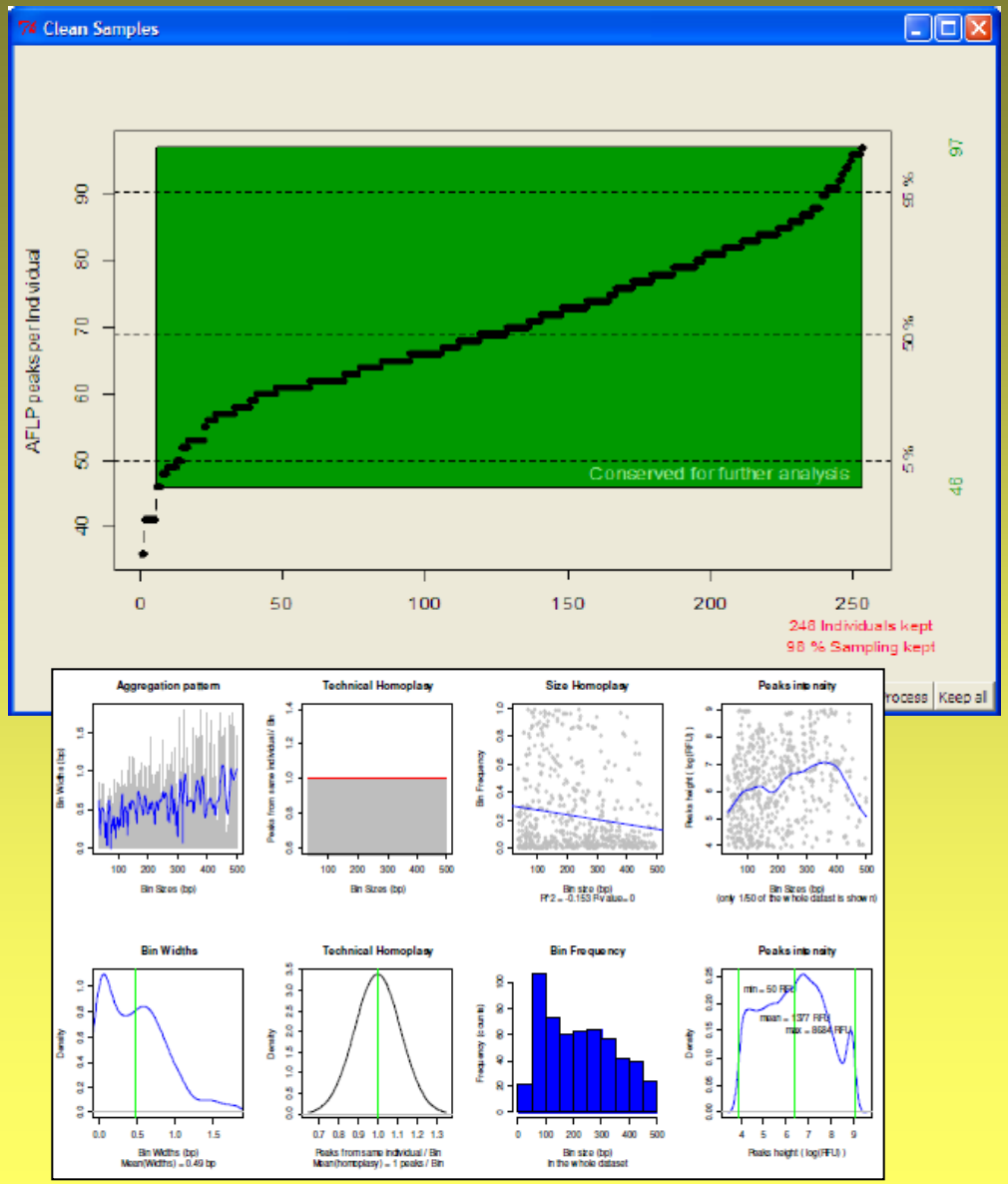
Interpretace – předpoklady

- komigrující fragmenty jsou homologní
  - nemusí být splněno u více diverzifikované skupiny
  - drobné rozdíly v mobilitě fragmentů – artefakty?
- fragmenty jsou nezávislé, tj. variabilita je dána mutací v restrikčním místě (vznikem, zánikem)
  - pokud je variabilita dána délkovou mutací → nerozpoznatelná kodominance (1 rozdíl je kodován jako dva znaky)

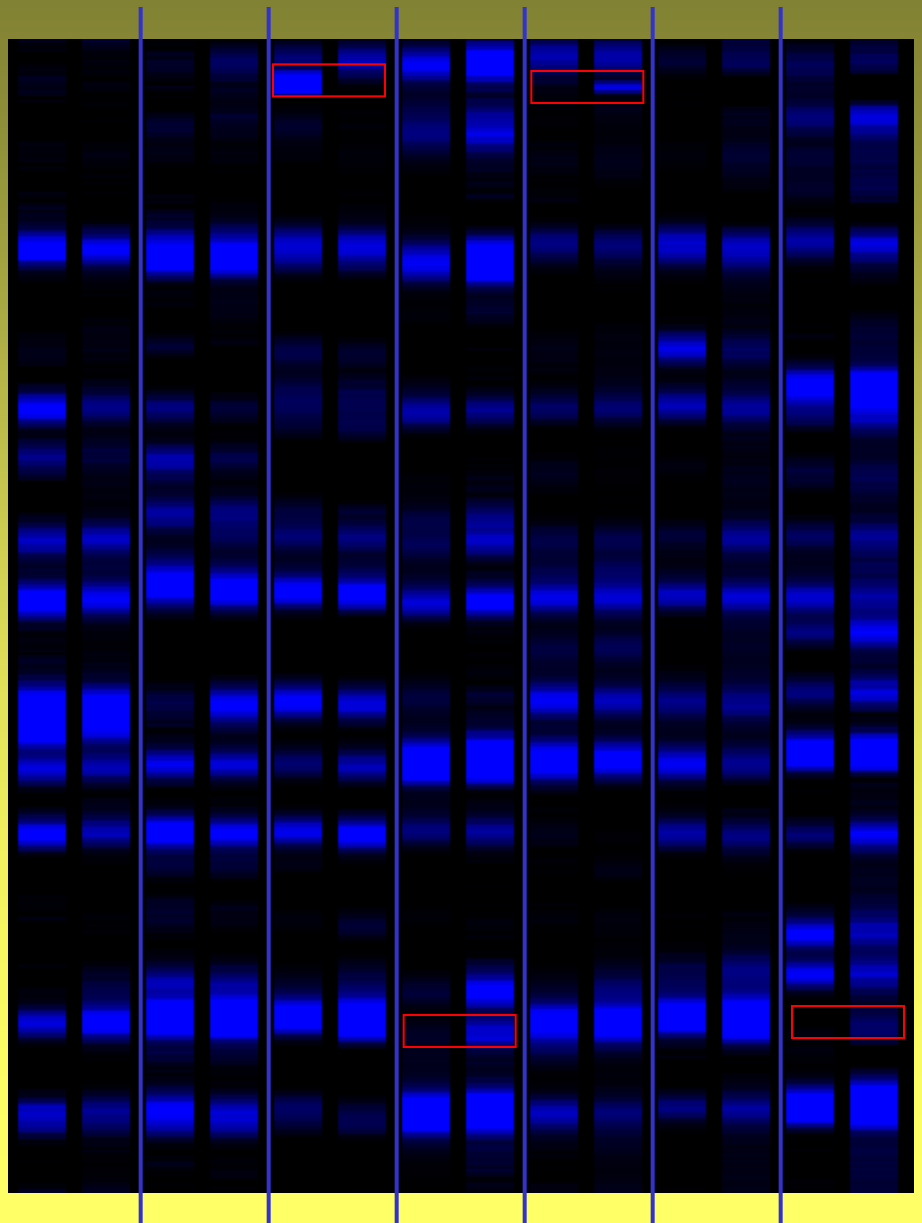
Automatické skórování

- Raw Geno (R)





Error rate



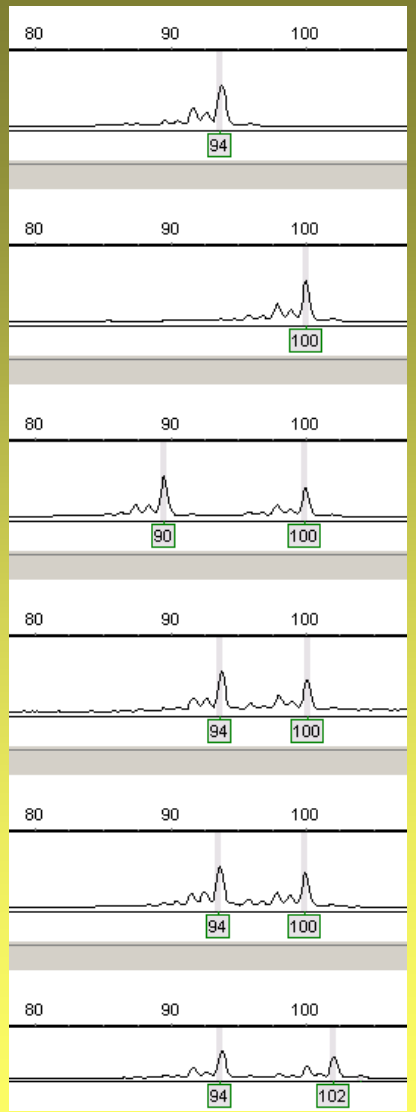
$$\frac{\text{počet rozdílů}}{\text{počet porovnávaných lokusů}}$$

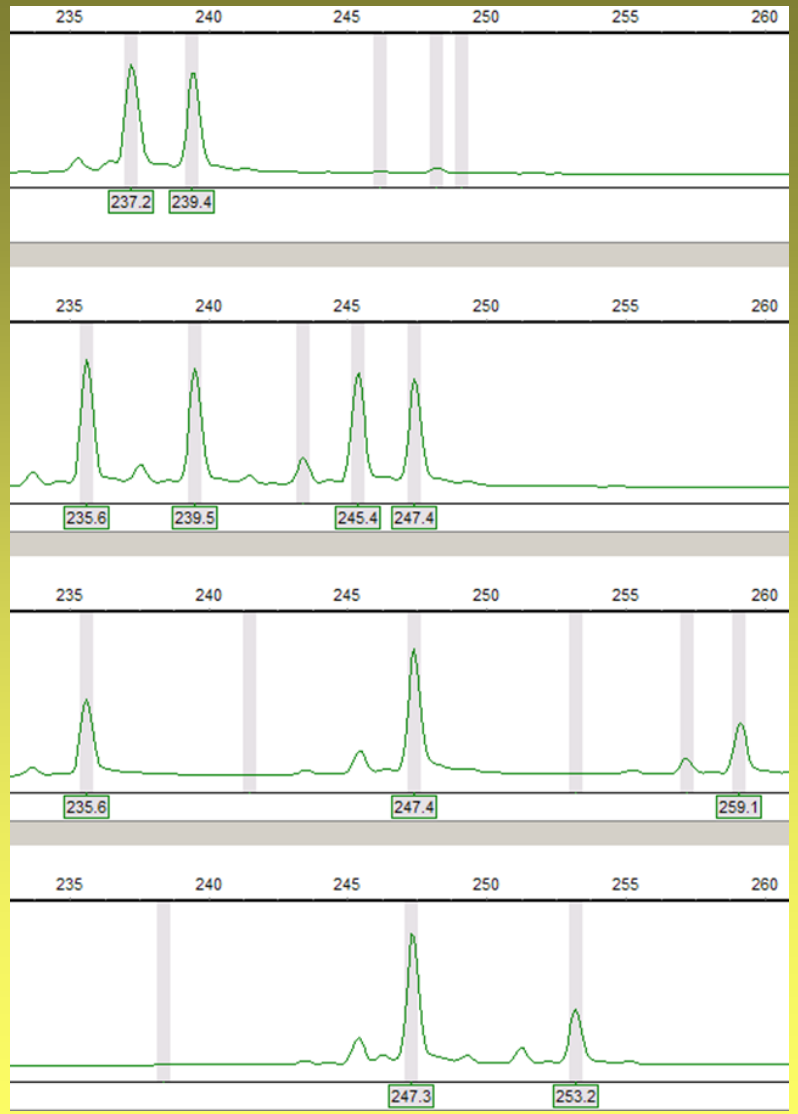
$$\frac{4}{86} = 4.65\%$$

Datová matice – AFLP

S002x1	1	0	0	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	1
S002x2	1	0	0	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	1
S002x3	1	0	0	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	1
S002x4	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1	1	0	0	1	0	0	1	1	0	0	0	1	0	1	0	0	1	0
S002x5	1	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0	1	1	1	1	0	1	0
S013x1	1	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1	1	1	0	0	1
S013x2	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	1
S013x3	1	0	0	0	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	1
S015x1	1	0	0	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	1
S015x2	1	0	0	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	1
S016x1	1	0	0	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	1
S016x2	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1
S016x3	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1
S016x4	1	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	0	1
S016x5	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	1	0	0	1	1	0	1	1	0	1	0	0	1	1	1	1	0	1
S016x6	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1
S016x7	1	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	1	0	1	0	1

Hodnocení kodominantních dat





diploidi vs. tetraploidi

Interpretace – předpoklady

- rozpoznáme všechny alely
  - známe ploidii
  - stutter bands
  - +A/-A PCR artefakty
  - rozpoznáme falešné peaky (artefakty)
  - allele drop-out
    - nulové alely
      - mutace v priming site
      - nekvalitní DNA zabrání amplifikaci jedné alely
    - delší alely jsou amplifikovány s nižší pravděpodobností
    - neskorované alely mimo obvyklý rozsah

Datová matice – SSRs

název lokusu

délka repetice

délka *flanking region*

jedno- nebo dvousloupcový formát

2			2		2		2		2		2	
			98		74		102		46		71	
			NLGA1		NLGA2		NLGA3		NLGA4		NLGA5	
A	d	1	160	160	86	96	142	142	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	160	166	86	86	152	152	198	198	100	100
B	d	1	166	166	86	96	150	150	196	198	100	100
B	d	1	160	166	84	84	150	150	196	198	100	104
B	d	1	160	166	92	92	150	152	196	198	100	100
B	d	1	160	166	92	92	150	152	196	198	100	100
B	d	1	166	166	90	92	150	150	198	198	100	100
B	d	1	166	166	82	82	150	152	200	200	100	100
B	d	1	160	162	86	96	nd	.	198	198	94	100
D	d	2	152	160	86	96	152	152	198	198	100	100
D	d	2	152	162	92	96	152	152	198	198	94	100
D	d	2	160	160	-1	-1	150	150			100	100

název populace

outbrední (d) nebo inbrední (h) individuum

číslo skupiny populací

chybějící data

Možnosti analýzy

- vztahy mezi jedinci – základní orientace ve struktuře
  - distanční stromy (NJ, UPGMA), sítě
  - mnohorozměrné analýzy (PCoA)
  - Bayesovské clusterování
- populačně-genetické parametry
  - diverzita (% polymorfních fragmentů, index diverzity)
  - divergence (% unikátních fragmentů, DW-index)
  - F-statistika, R-statistika (mikrosatelity)
- testování a zjišťování prostorové struktury
  - AMOVA
  - Bayesovské odhady, Mantelovy testy, prostorová autokorelace
- testování specifických hypotéz
  - podobnost a evoluční vztahy identifikovaných skupin
  - hybridizace
  - původ polyploidů
  - …

Práce s maticí, převody, exporty

A) AFLP data

- (RawGeno)
- AFLP Dat
- FAMD

B) mikrosatelity

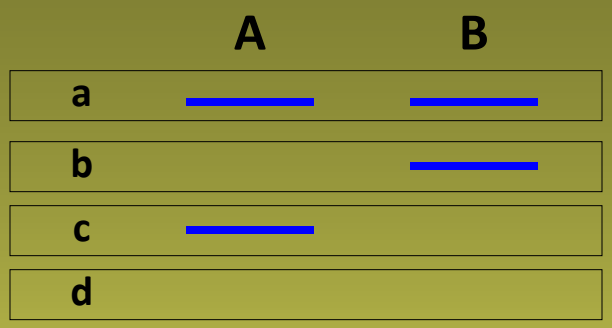
- MSA

Tvorba distančních stromů

- výpočet matice vzdáleností – index podobnosti
  - AFLP - Jaccard, Dice, Nei & Li, simple matching
  - SSRs – Nei´s genetic distance, Goldstein distance…
- algoritmus na tvorbu stromů
  - shlukování – UPGMA …
  - neighbour joining (NJ) – minimum evolution+minimalizace délky
  - maximální parsimonie (MP) – kladistika (binární data)
- strom
  - nezakořeněný × zakořeněný (midpoint, outgroup rooting)
  - kladogram × fylogram
- software
  - tvorba stromů – FAMD, PAST, PHYLIP, PAUP, R …
  - prohlížení stromů – TreeView, FigTree, …

Koeficienty podobnosti – AFLP

		jedinec A	
		presence 1	absence 0
jedinec B	presence 1	a	b
	absence 0	c	d



- Jaccardův koeficient (Jaccard 1908)

$$\frac{a}{a+b+c}$$

- Dice koeficient (Dice 1945) = Nei & Li 1979, Sørensen 1948

$$\frac{2a}{2a+b+c}$$

- „simple-matching“ koeficient (Sokal & Michener 1958)

$$\frac{a+d}{a+b+c+d}$$

Koeficienty podobnosti – SSRs

- X, Y – srovnávané populace
- L – počet lokusů, U – počet alel v lokusu
- X_u – frekvence u-té alely v populaci X

- Nei's distance 1983

$$D_A = 1 - \sum_u \sqrt{X_u Y_u}$$

- Nei's standard distance 1972

$$D_a = -\ln \frac{\sum_l \sum_u X_u Y_u}{\sqrt{(\sum_l \sum_u X_u^2)(\sum_l \sum_u Y_u^2)}}$$

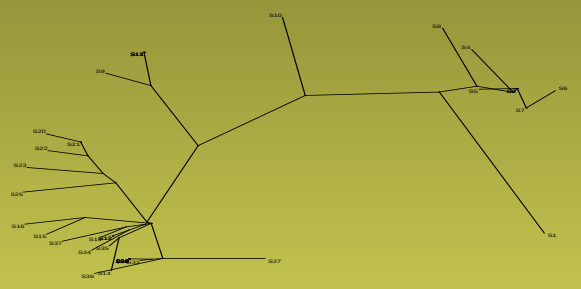
- Goldstein distance (1995)

$$(\delta\mu)^2 = (\mu_X - \mu_Y)^2$$

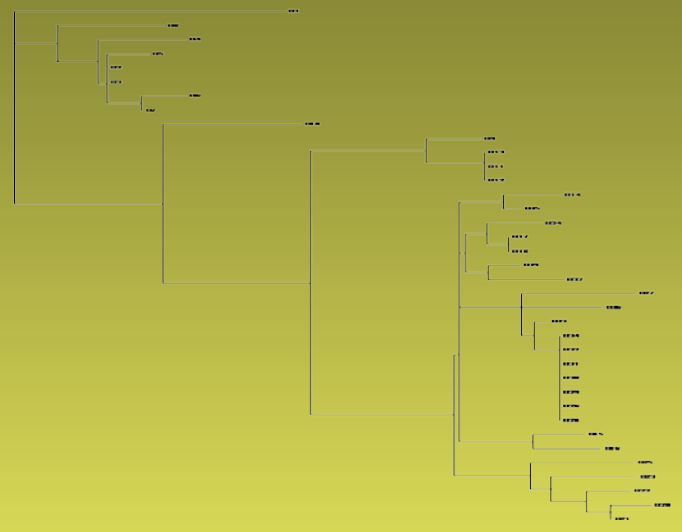
- proporce sdílených alel (Bowcock 1994) – D_{ps}

Typy stromů

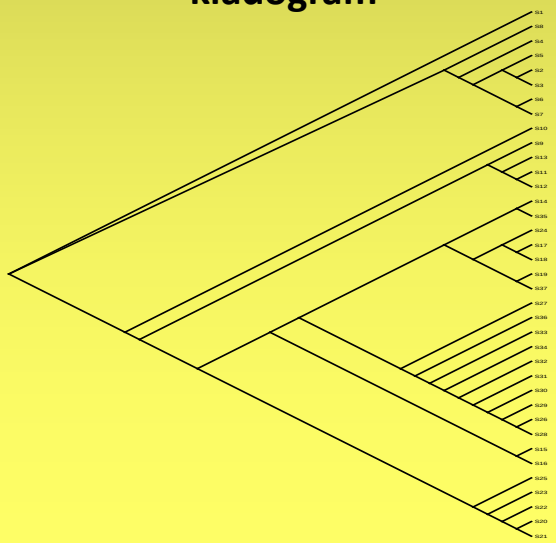
nezakořeněný



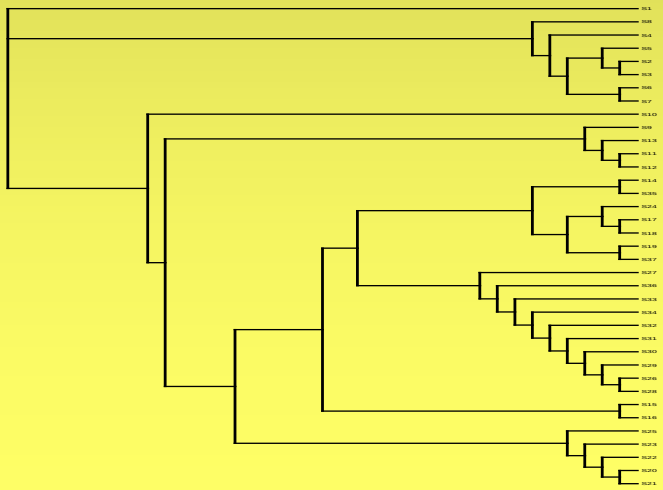
fylogram



kladogram

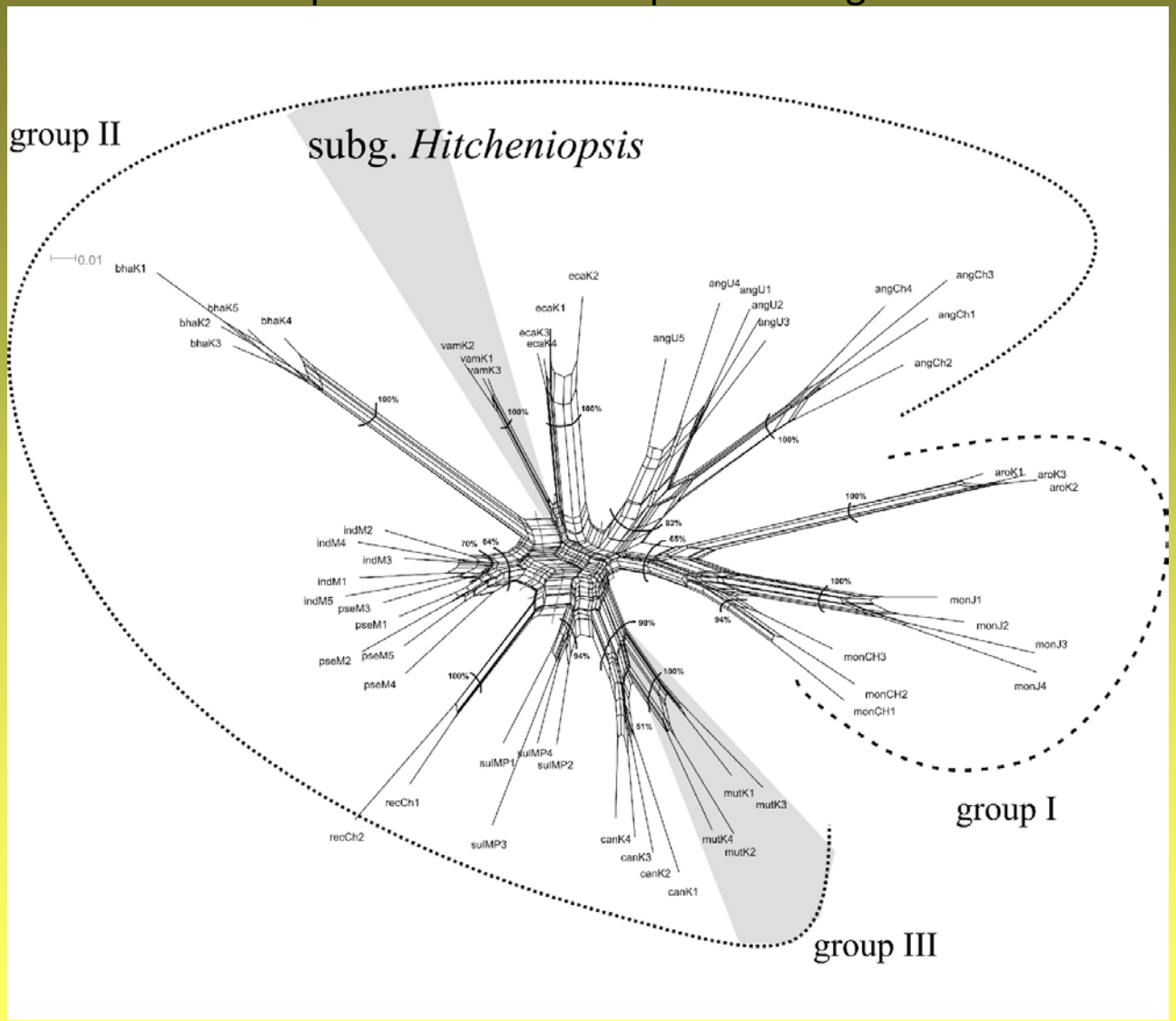


pravoúhlý kladogram



Neighbour-network

SplitsTree4 – www.splitstree.org



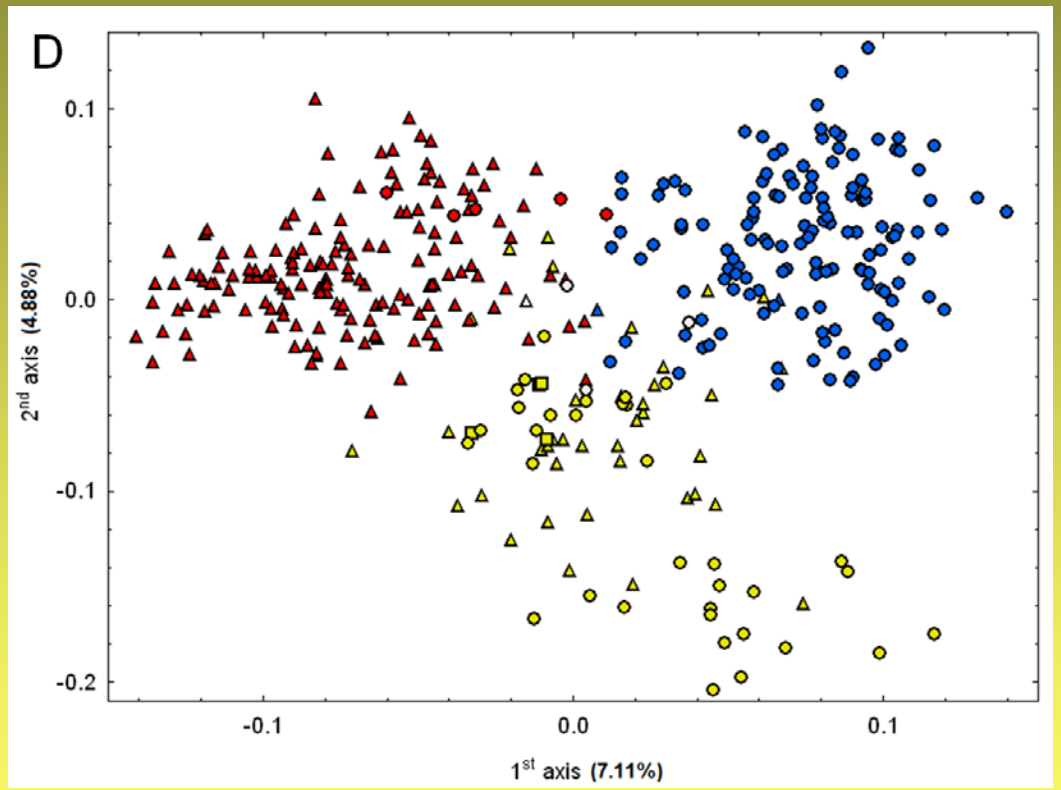
Curcuma – Záveská et al. 2011

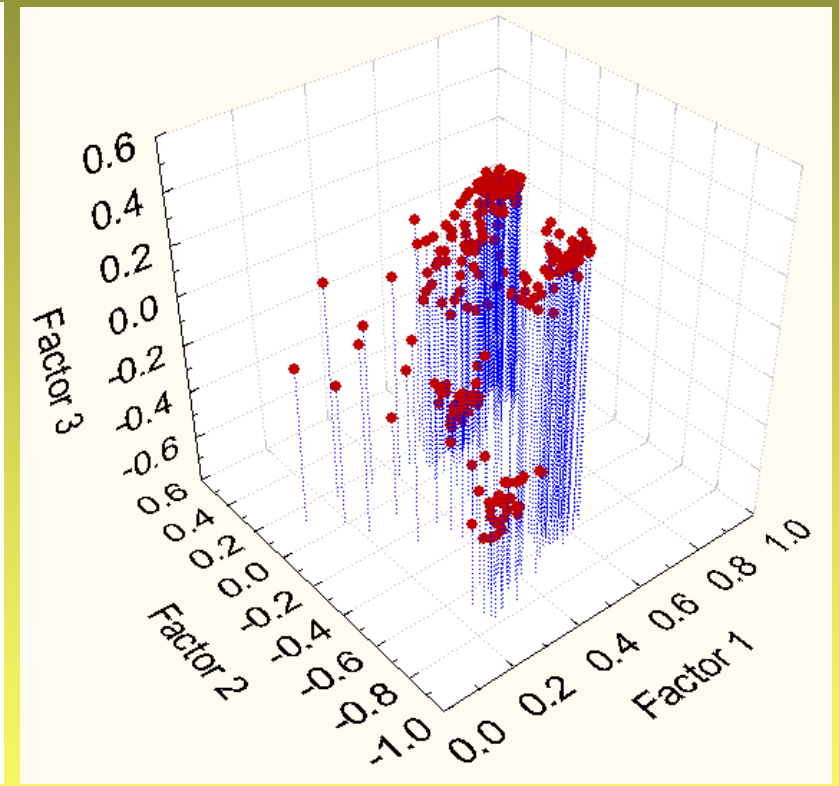
Principal coordinate(s) analysis – PCoA

- pro analýzu binárních znaků
- počet znaků může být větší než počet objektů
  → tj. typicky pro RAPD, AFLP data…

- nehierarchická vizualizace struktury dat
- výpočet matice vzdáleností mezi objekty (Jaccard)
- ordinační diagram – převedení objektů do nového prostoru na základě vzdálenosti mezi nimi a maximalizace vysvětlené variability
- software – FAMD, PAST, R, CANOCO, SYNTAX, …

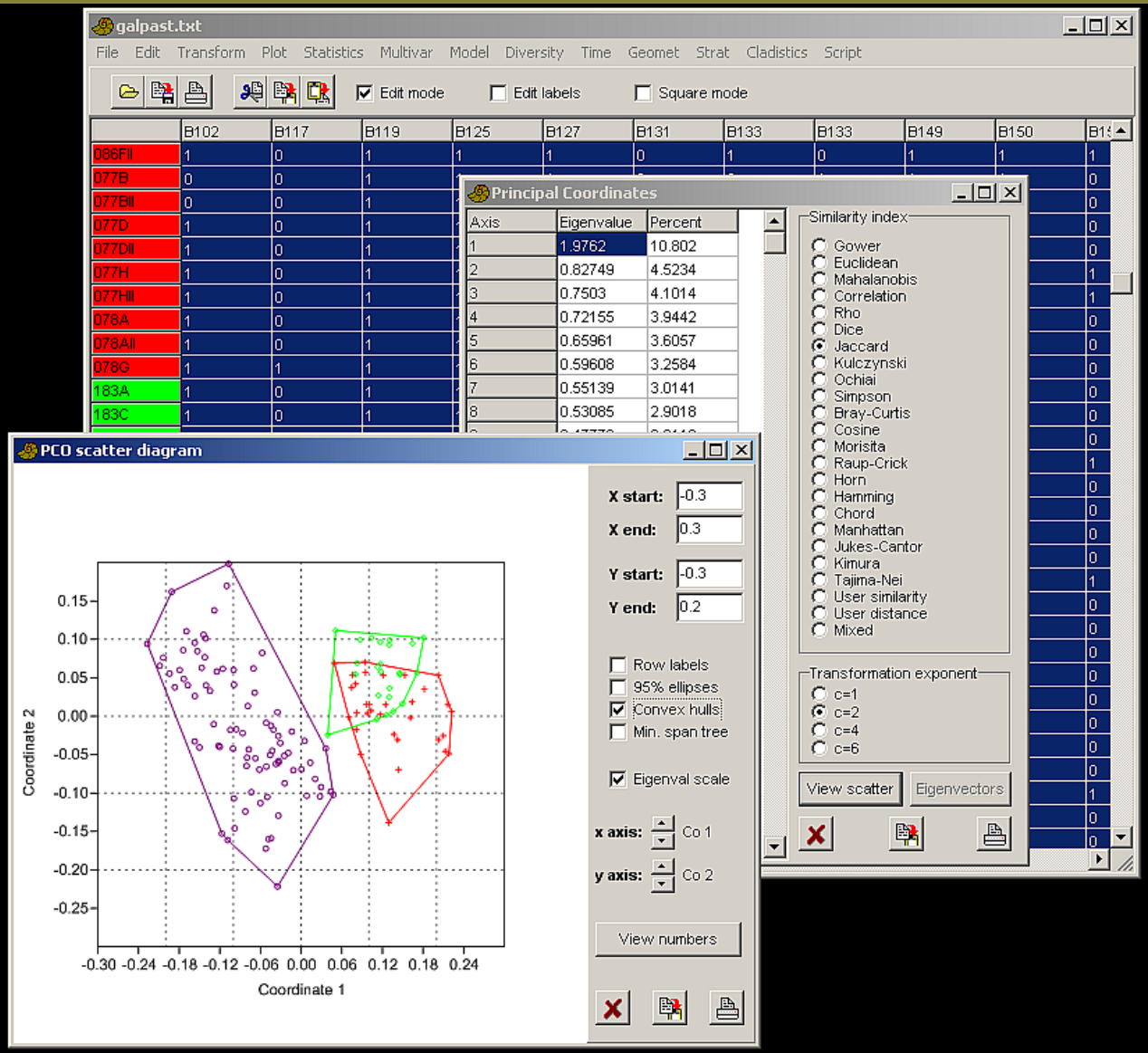
PCoA


D
0.1
0.0
-0.1
-0.2
2nd axis (4.88%)
-0.1
0.0
0.1
1st axis (7.11%)


0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
Factor 3
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Factor 2
0.0
0.2
0.4
0.6
0.8
1.0
Factor 1

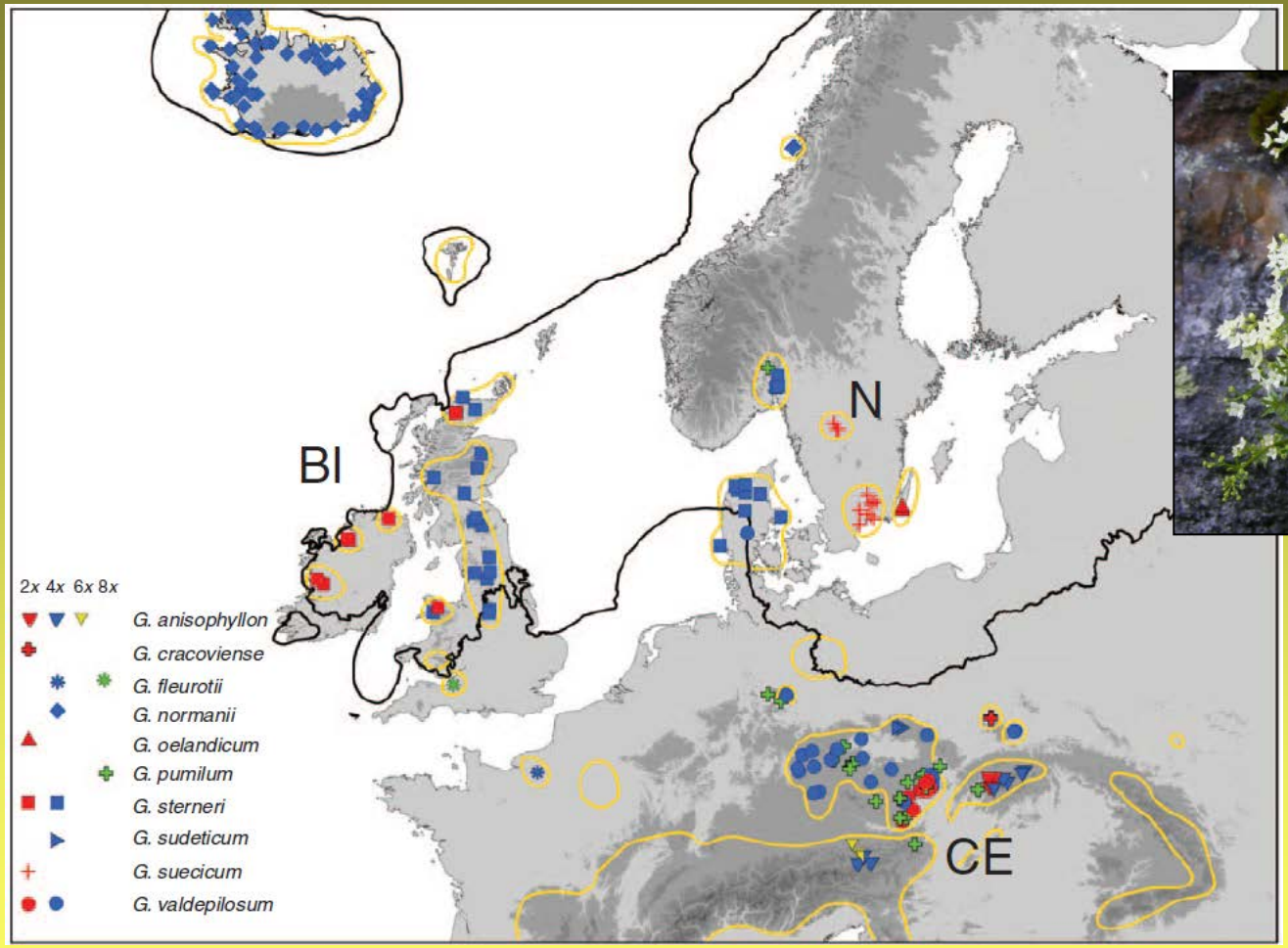
PAST

http://folk.uio.no/ohammer/past/



- PCO, stromy
- rychlé označení skupin
- vstup: matice prim. dat i distancí

AFLP vzorová data



Galium pusillum agg. – vybráni pouze diploidi

? liší se diploidi ze zaledněných (N+BI) a nezaledněných (CE) oblastí resp. mezi BI, N a CE regiony?

SSRs vzorová data

T. latifolia	176	176	278	278	176	190	269	269	179	179	93	93	278	278
T. angustifolia	210	210	286	286	196	196	287	287	193	193	101	101	280	280
T. x glauca	180	210	278	286	190	196	269	287	179	193	93	101	278	280
advanced hybrid	176	210	278	286	190	196	287	287	179	193	93	101	278	280

Typha latifolia

Typha ×glauca

Typha angustifolia

Hybridizace orobinců (*Typha*) v USA – T. ×glauca (F1) je invazní druh

Jaká je dynamika hybridizace? kříží se F1 dále mezi sebou (F2) a/nebo zpětně s rodič. druhy?

Snow et al. 2010

Praktické cvičení 1

1. vytvořte neighbour network síť z vašich dat (AFLP) nebo nezakořeněný NJ strom (AFLP, SSRs)

- network
  - importujte data do FAMD / AFLPDat a exportujte z něj Nexus
  - Nexus otevřete v SplitsTree
- NJ strom
  - vytvořte matici vzdáleností ve Phylip formátu (MSA nebo FAMD pomocí exportu)
  - spusťte program neigbor.exe (část balíku PHYLIP) ze stejného adresáře jako je matice
  - vzniklý strom (outtree) zobrazíte pomocí TreeView nebo FigTree

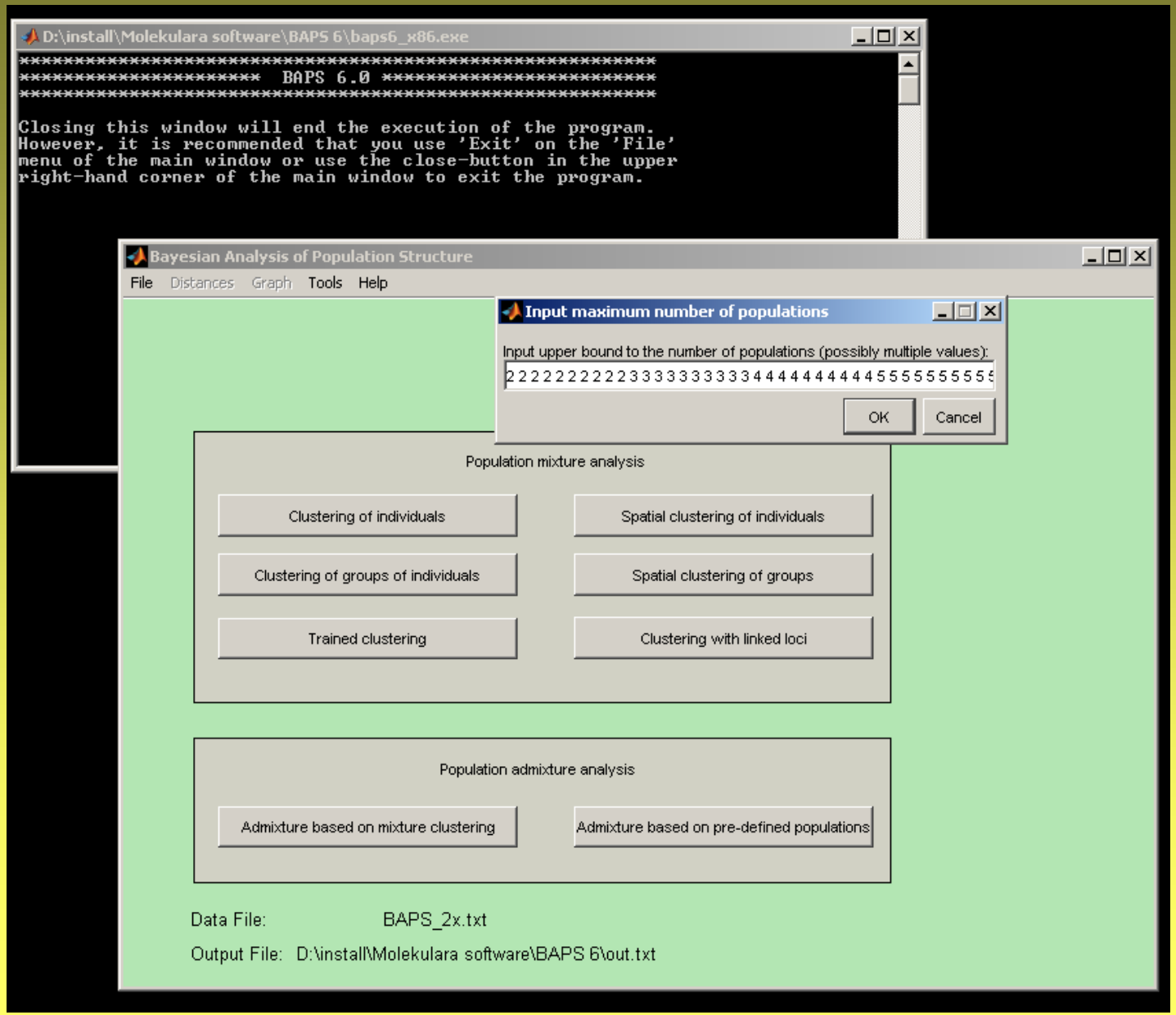
2. vytvořte PCoA diagram založený na Jaccardově distanci (AFLP) nebo D_{ps}/POSA (SSRs)

- AFLP
  - importujte 0/1 matici do PAST
  - obarvěte vzorky podle tří hlavních skupin, vytvořte PCoA diagram
  - alternativně udělejte PCoA ve FAMD (po definici skupin v Group Manager)
- SSRs
  - upravte matici z MSA do formátu CSV (text oddělený středníky – pomocí Excelu)
  - v R spusťte upravený „skript“, který udělá PCoA a nakreslí biplot
  - alternativa: načtěte distanční matici do PAST a vytvořte PCoA diagram za použití volby „User similarity“

Bayesovské clusterování

- hledání takového rozdělení individuí do K clusterů, které je nejvíce pravděpodobné (resp. má maximální záporný logaritmus marginální pravděpodobnosti)
- výsledkem je zjištění optimálního počtu clusterů, tj. „reálných populací“ a rozřazení všech individuí
- v rámci populací (clusterů) jsou lokusy v H-W a linkage ekvilibriu (resp. jedinci jsou rozřazováni do populací tak, aby tohoto bylo dosaženo)
  - mixture – každý jedinec do právě jedné populace
  - admixture – pravděpodobnostní rozřazení jedince do více populací
- software
  - BAPS 3.2 – Bayesian Analysis of Population Structure (Corander et al.) (stochastic optimization)
  - STRUCTURE (Pritchard et al.) (MCMC)

BAPS v6


D:\install\Molekulara software\BAPS 6\baps6_x86.exe
BAPS 6.0
Closing this window will end the execution of the program.
However, it is recommended that you use 'Exit' on the 'File'
menu of the main window or use the close-button in the upper
right-hand corner of the main window to exit the program.
Bayesian Analysis of Population Structure
File
Distances
Graph
Tools
Help
Input maximum number of populations
Input upper bound to the number of populations (possibly multiple values):
OK
Cancel
Population mixture analysis
Clustering of individuals
Spatial clustering of individuals
Clustering of groups of individuals
Spatial clustering of groups
Trained clustering
Clustering with linked loci
Population admixture analysis
Admixture based on mixture clustering
Admixture based on pre-defined populations
Data File: BAPS_2x.txt
Output File: D:\install\Molekulara software\BAPS 6\out.txt

Výstup z programu BAPS 6

```
RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:

Data file: BAPS_2x.txt
Number of clustered individuals: 141
Number of groups in optimal partition: 5
Log(marginal likelihood) of optimal partition: -10052.3776
Best Partition:

Cluster 1: {70, 103, 104, 105, 106, 107, 108, 109, 110, 111,
           113, 114, 115, 116, 117, 118, 119}
...

Cluster 4: {57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67}

Cluster 5: {120, 121, 122, 123, 124, 125, 126, 127, 128, 129,
           130, 131, 132, 133, 134, 135, 136, 137, 138, 139,
           140, 141}

  Changes in log(marginal likelihood) if indvidual i is moved to group j:

  ind      1        2        3        4        5
    1:    -36.9    -21.6       .0     -8.8    -44.0
    2:    -24.4    -12.3       .0    -21.2    -34.9
    3:    -36.0    -31.9       .0    -30.3    -49.6
    4:    -23.0    -16.5       .0    -29.5    -33.5
    5:    -42.6    -37.0       .0    -28.7    -51.8
    6:    -45.0    -39.5       .0    -29.1    -49.1
    7:    -19.4    -12.3       .0    -23.1    -32.2
    8:    -28.9    -26.2       .0    -24.6    -44.4
...

KL-divergence matrix in PHYLIP format:
5
Cluster_1  0.000  0.182  0.270  0.324  0.225
Cluster_2  0.182  0.000  0.214  0.256  0.216
Cluster_3  0.270  0.214  0.000  0.275  0.385
Cluster_4  0.324  0.256  0.275  0.000  0.381
Cluster_5  0.225  0.216  0.385  0.381  0.000

Probabilities for number of clusters

5   0.66458
6   0.33542
```

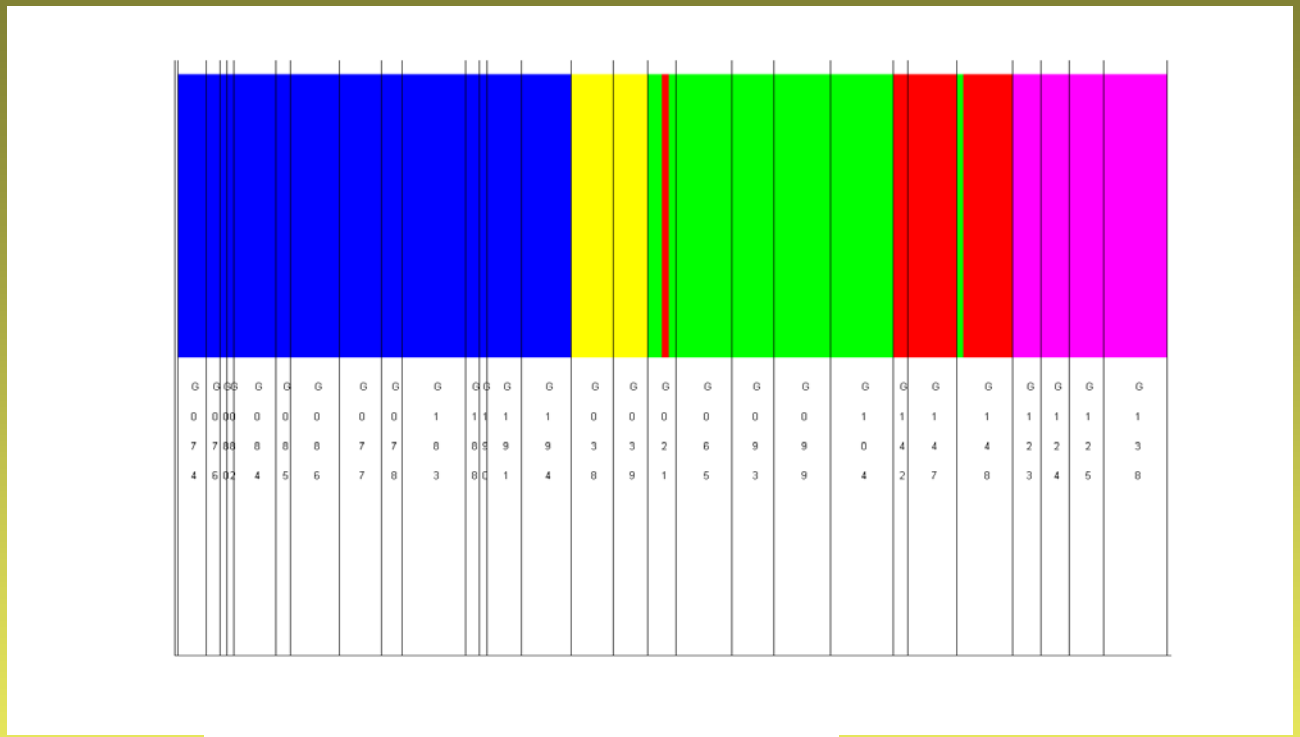
rozdělení jedinců do skupin

změna likelihood modelu při přesunu jedince do jiné skupiny

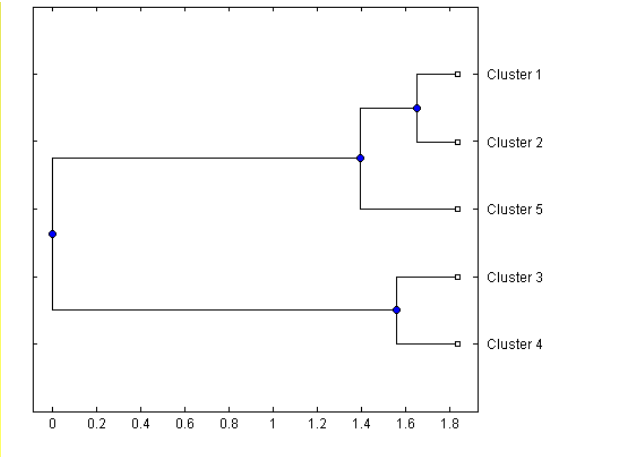
nepodobnosti mezi skupinami

pravděpodobnost modelu

Výstup z programu BAPS 6

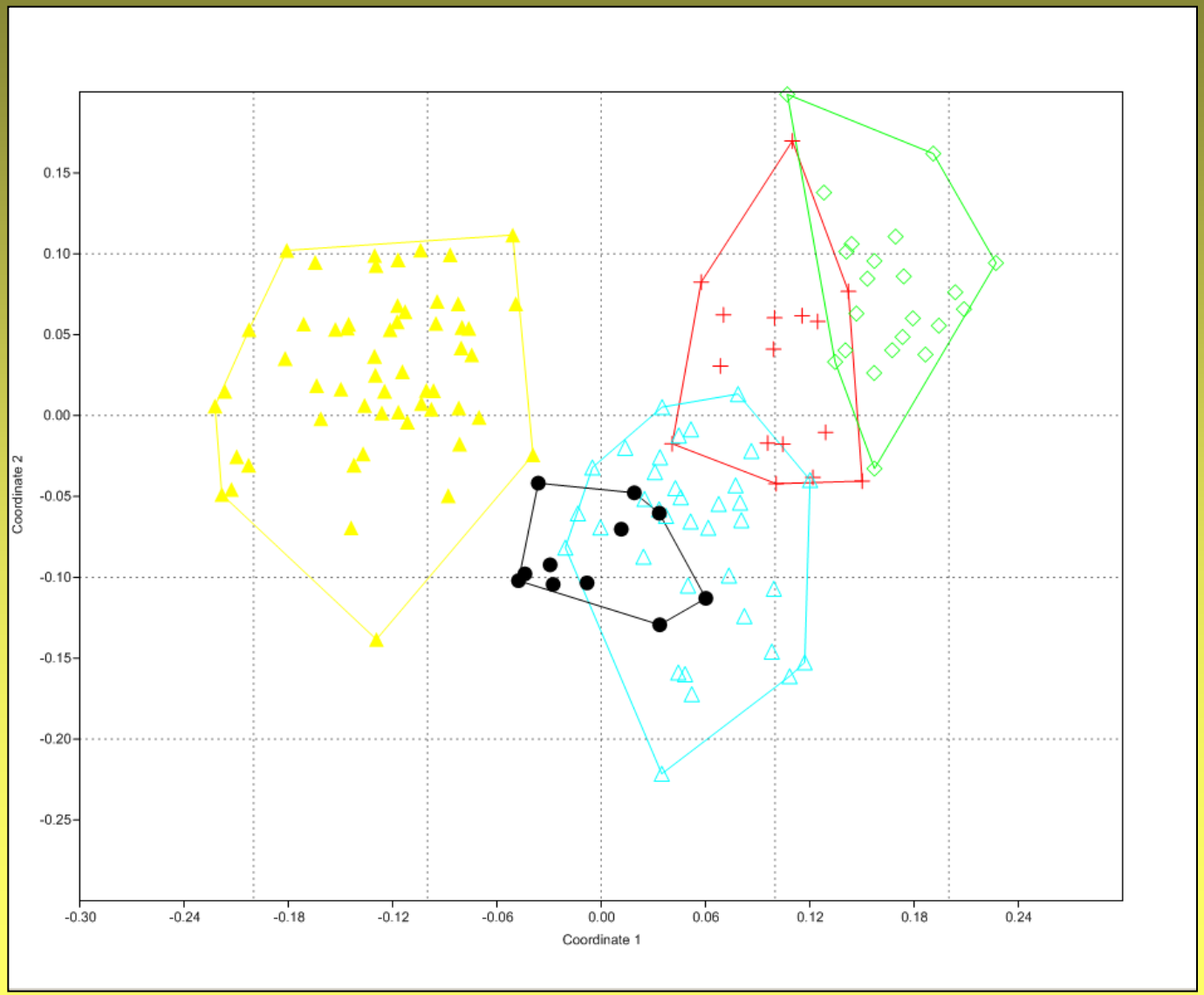


rozdělení jedinců do skupin



podobnosti mezi skupinami

Srovnání výsledků PCoA a bayesovského clusterování (BAPS 6)



Výstup z programu STRUCTURE

```
--------------------------------------------
Proportion of membership of each pre-defined
 population in each of the 6 clusters

Given    Inferred Clusters                          Number of
 Pop       1      2      3      4      5      6      Individuals

  1:     0.086  0.012  0.023  0.861  0.015  0.003       10
  2:     0.011  0.037  0.060  0.796  0.089  0.007       10
  3:     0.094  0.010  0.003  0.031  0.858  0.004       10
  4:     0.789  0.005  0.007  0.158  0.029  0.010       10
  5:     0.251  0.631  0.109  0.004  0.003  0.002        4
--------------------------------------------
Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.
       1       2         3           4         5         6
 1      -    -42.6055  -90.4091  -62.0519  -25.8868  -119.1667
 2   -42.6055     -    -58.3355  -83.5292  -56.3593  -100.9015
 3   -90.4091  -58.3355     -    -114.9450  -64.8990  -54.0048
 4   -62.0519  -83.5292  -114.9450     -    -54.0265  -139.2714
 5   -25.8868  -56.3593  -64.8990  -54.0265     -    -77.1662
 6   -119.1667  -100.9015  -54.0048  -139.2714  -77.1662     -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster  1  : 1091.5688
cluster  2  : 1109.7404
cluster  3  : 1127.4684
cluster  4  : 1034.3344
cluster  5  : 1047.3957
cluster  6  : 1094.7986

Inferred ancestry of individuals:
        Label (%Miss) Pop:  Inferred clusters
  1      110    (0)    1 :  0.002 0.003 0.005 0.985 0.002 0.002
  2      111    (0)    1 :  0.332 0.037 0.140 0.421 0.062 0.008
  3      112    (0)    1 :  0.452 0.036 0.015 0.462 0.031 0.003
  4      113    (0)    1 :  0.033 0.010 0.007 0.942 0.006 0.002
  5      114    (0)    1 :  0.009 0.003 0.002 0.962 0.022 0.002
  6      115    (0)    1 :  0.016 0.012 0.047 0.906 0.014 0.006
  7      116    (0)    1 :  0.009 0.005 0.003 0.972 0.009 0.001
  8      118    (0)    1 :  0.004 0.003 0.003 0.983 0.004 0.002
  9      119    (0)    1 :  0.002 0.004 0.003 0.986 0.002 0.002
 10      120    (0)    1 :  0.005 0.003 0.003 0.986 0.002 0.002
```

a priori populace vs. clustery

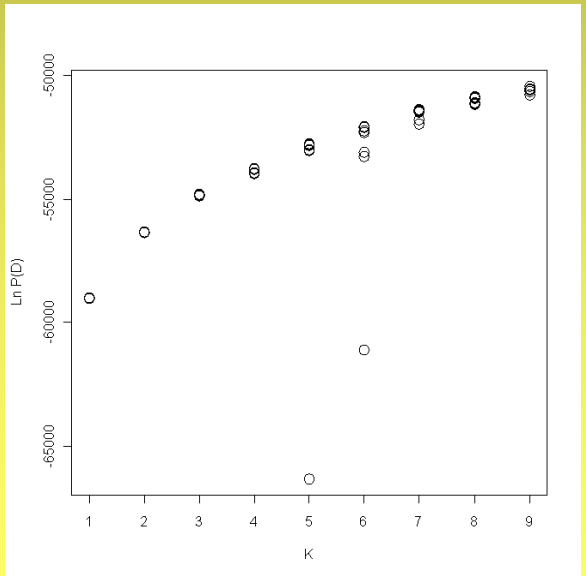
divergence mezi clustery

variabilita uvnitř clusterů

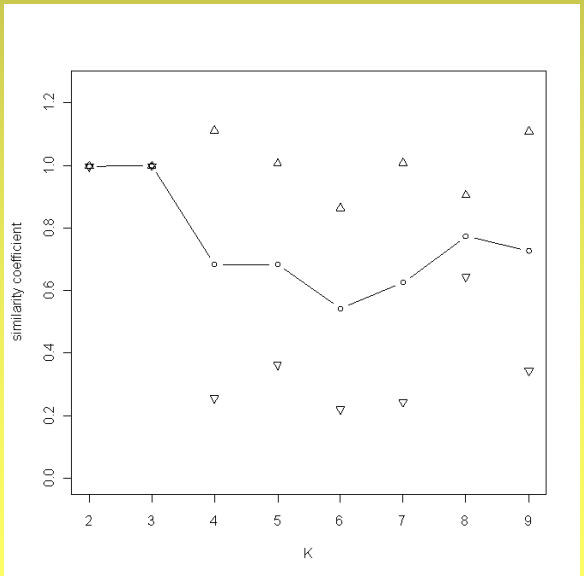
pravděpodobnost přiřazení jedince do clusteru

Hodnocení STRUCTURE výsledků

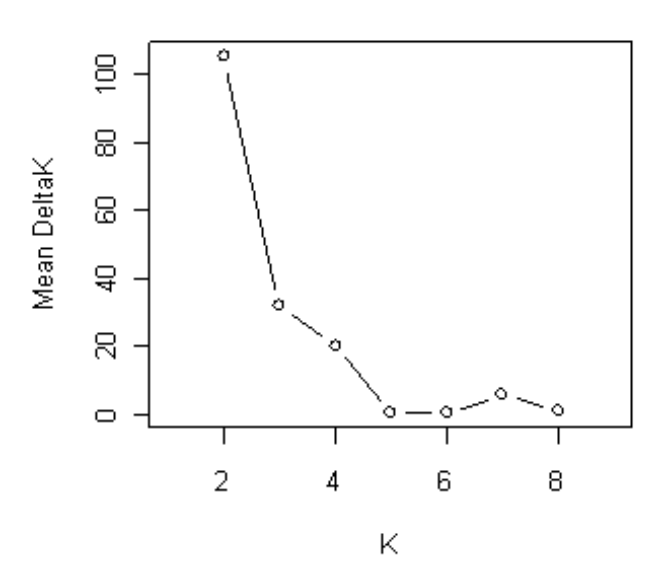
- Structure-sum (R script)
  - sumarizace jednotlivých běhů podle K
  - koeficienty podobnosti mezi opakováními pro dané K
  - určení deltaK (optimální počet clusterů)



logaritmus pravděpodobnosti



podobnost



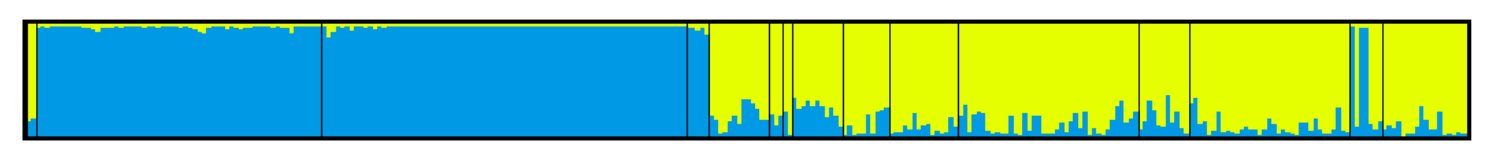
deltaK

Hodnocení STRUCTURE výsledků

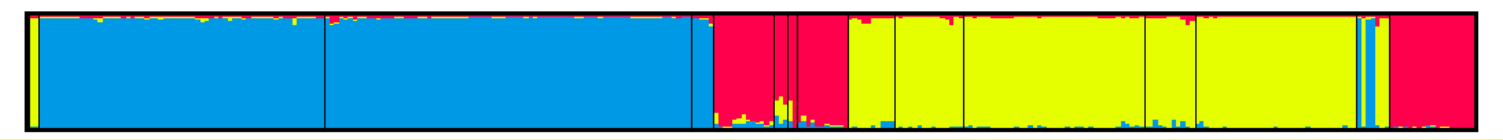
Distruct (Rosenberg 2004)

- grafické znázornění pravděpodobností příslušnosti jedinců do jednotlivých clusterů

K2

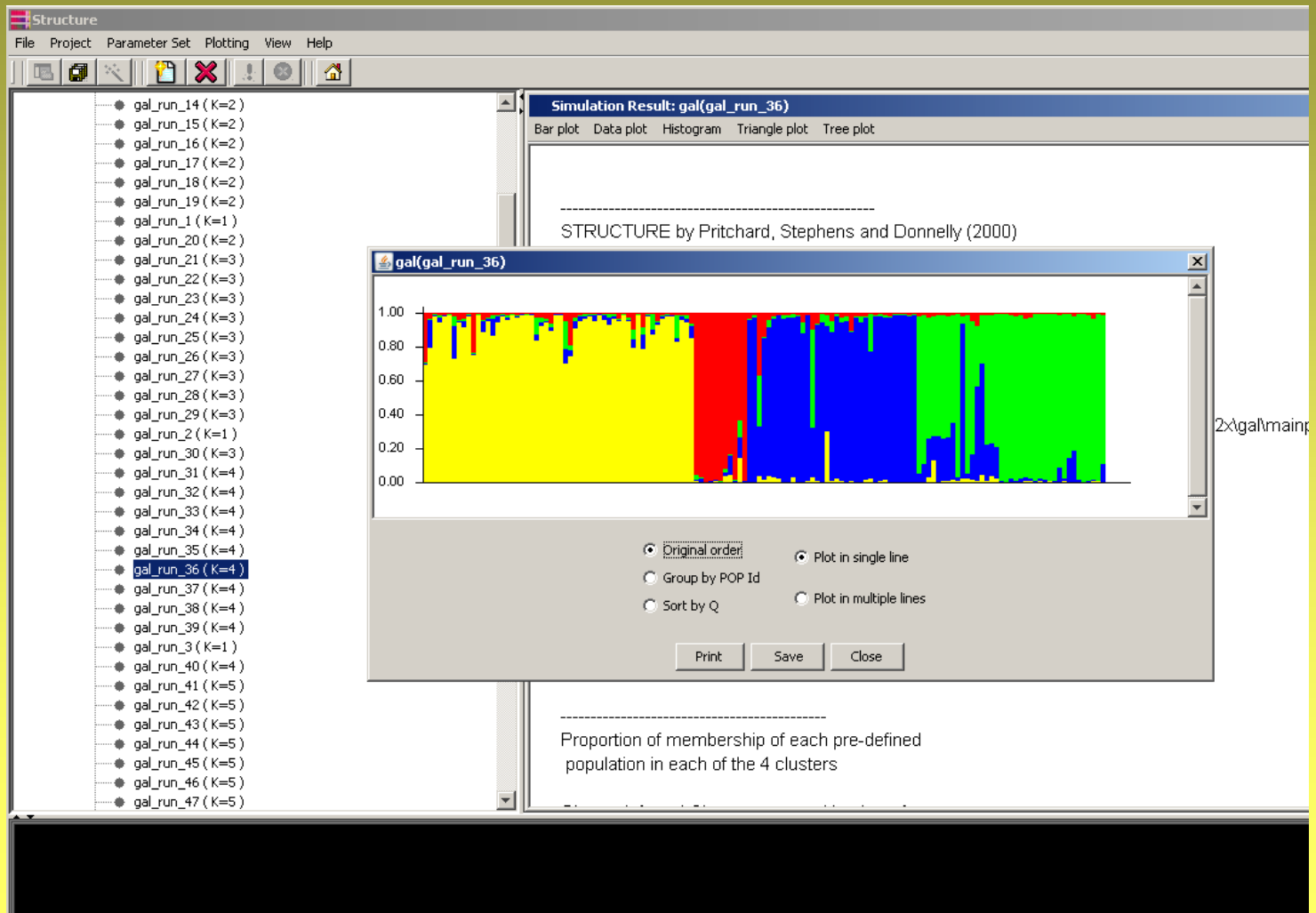


K3



Hodnocení STRUCTURE výsledků

grafické rozhraní



Praktické cvičení 2

1. vyhledejte optimální rozdělení do K skupin pomocí BAPS

- uložte vstupní data pro BAPS z Excelu
- spusťte BAPS analýzu pro K=2-7 a 5x opakování od každého maximálního K až do Kmax=7
- vyberte optimální K
- zobrazte vámi identifikované BAPS skupiny v PCoA diagramu

2. vyhledejte optimální rozdělení do K skupin pomocí STRUCTURE

- exportujte data z MSA do formátu Structure
- upravte matici tak, aby obsahovala čísla jednotlivých „druhů“ jako populační sloupec (Typha_US_Structure_populcodes.txt) nebo použijte připravenou matici
- spusťte Structure s parametry 10 000 burnin/20 000 run pro K=1-6 s pěti opakováními
- sumarizujte výsledky Structure pomocí R skriptů Structure-sum
  - jaké K má nejvyšší LnP(D) ?
  - která K mají vysoký *similarity coefficient*?
  - jaké K má nejvyšší deltaK?
- nakreslete barevný *barplot* pomocí programu Distruct pro konvergující K

Práce s programy pro hodnocení binárních dat

FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html

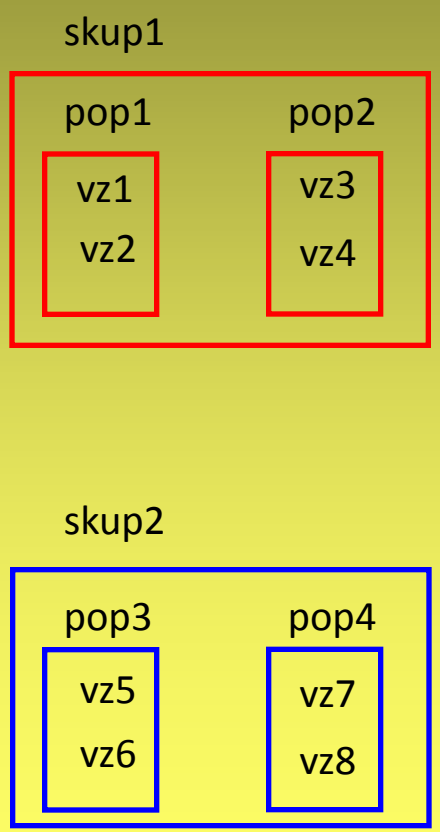
- různé indexy podobnosti (Jaccard, Dice, SMC, Euclidean…
- stromy (NJ, UPGMA) + bootstrap (prohlížení v TreeView)
- PCoA (+ grafické 3D zobrazení)
- AMOVA
- Shannonův index (+ testy signifikantních rozdílů)
- statistika proužků (polymorfní, fixované, privátní)
- zacházení s chybějícími daty
- exporty do jiných formátů dat
- řada dalších věcí…

FAMD

Fingerprint **A**nalysis with **M**issing **D**ata

http://www.famd.me.uk/famd.html

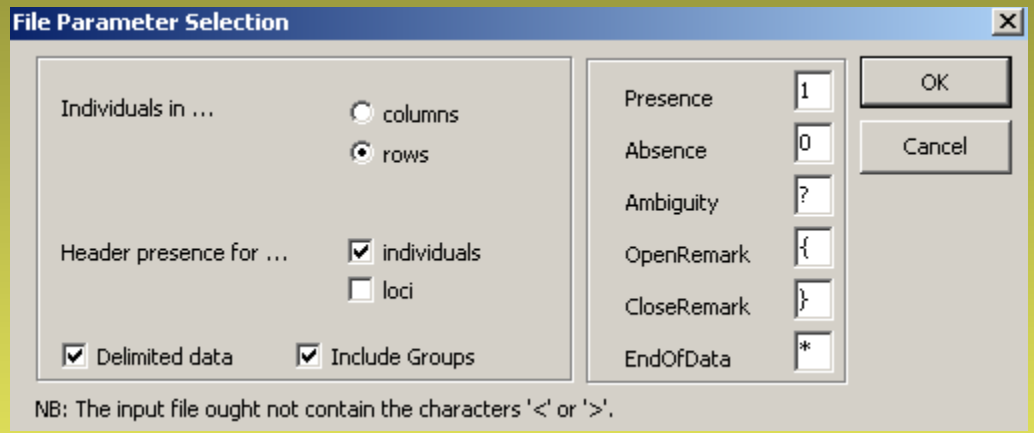
```
vz1 0 1 1 0 1 1 1 1 1
vz2 0 1 1 0 1 1 1 0 1
vz3 0 1 0 1 1 0 0 0 1
vz4 1 1 0 1 1 0 0 0 1
vz5 0 0 1 0 1 1 1 1 0
vz6 0 0 1 0 1 1 1 0 0
vz7 1 0 0 0 1 0 0 0 0
vz8 1 0 0 0 1 0 0 0 0
*
[Groups]
AllData/skup1= vz1, vz2, vz3, vz4;
AllData/skup2= vz5, vz6, vz7, vz8;
AllData/skup1/pop1= vz1, vz2;
AllData/skup1/pop2= vz3, vz4;
AllData/skup2/pop3= vz5, vz6;
AllData/skup2/pop4= vz7, vz8;
*
```



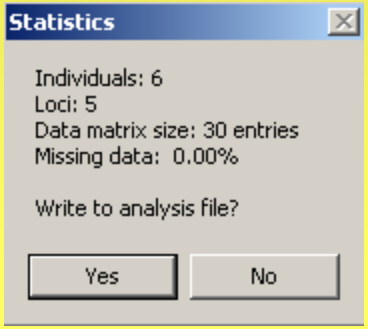
FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html

- File – Load



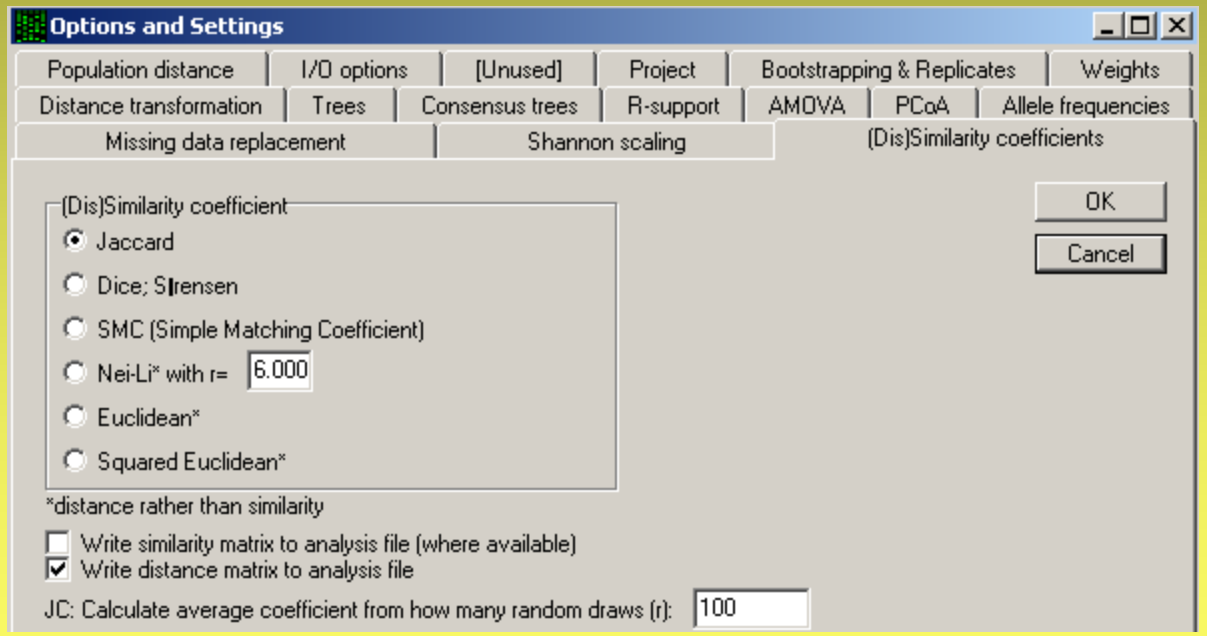
- DataMatrix – Matrix Statistics



FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html

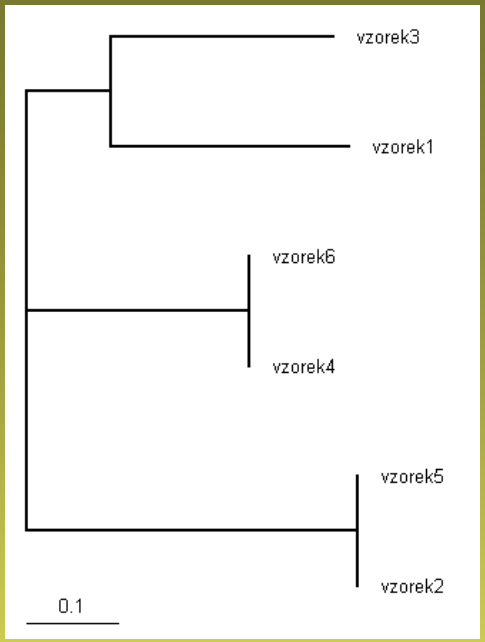
- Options – (Dis)Similarity Coefficients



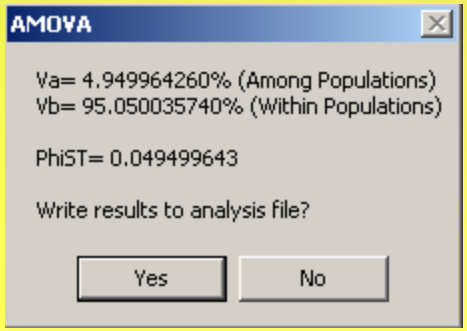
- Analysis – Standard Similarity

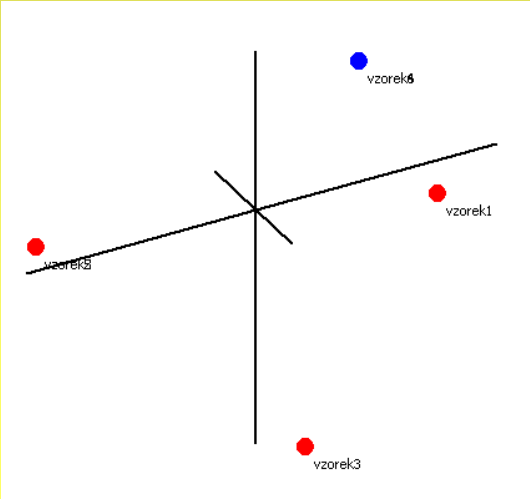
FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html



- Trees – Neighbour Joining
- View – Tree File (vyžaduje TreeView)
- Trees – Principal Coordinate Analysis
- Analysis – AMOVA (s Euclidean dist.)





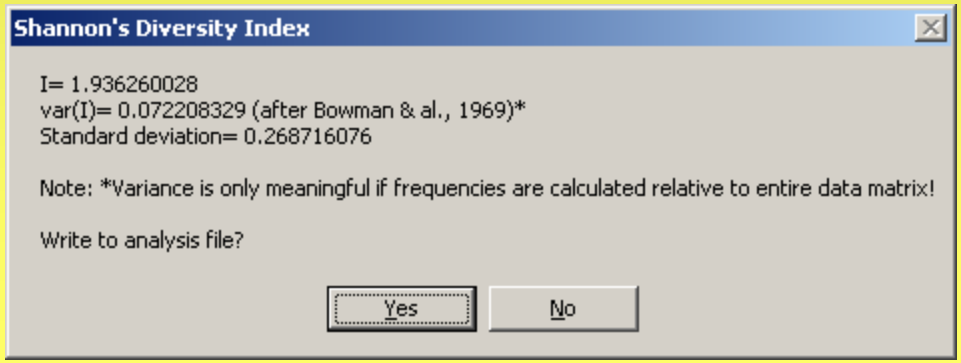
FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html

- DataMatrix – Count Bands… – Polymorphic bands (Fixed Bands/Private Bands/Fixed Private Bands)

```
Number of polymorphic bands found in group AllData: 4
Number of polymorphic bands found in group AllData/306: 4
Number of polymorphic bands found in group AllData/307: 3
```

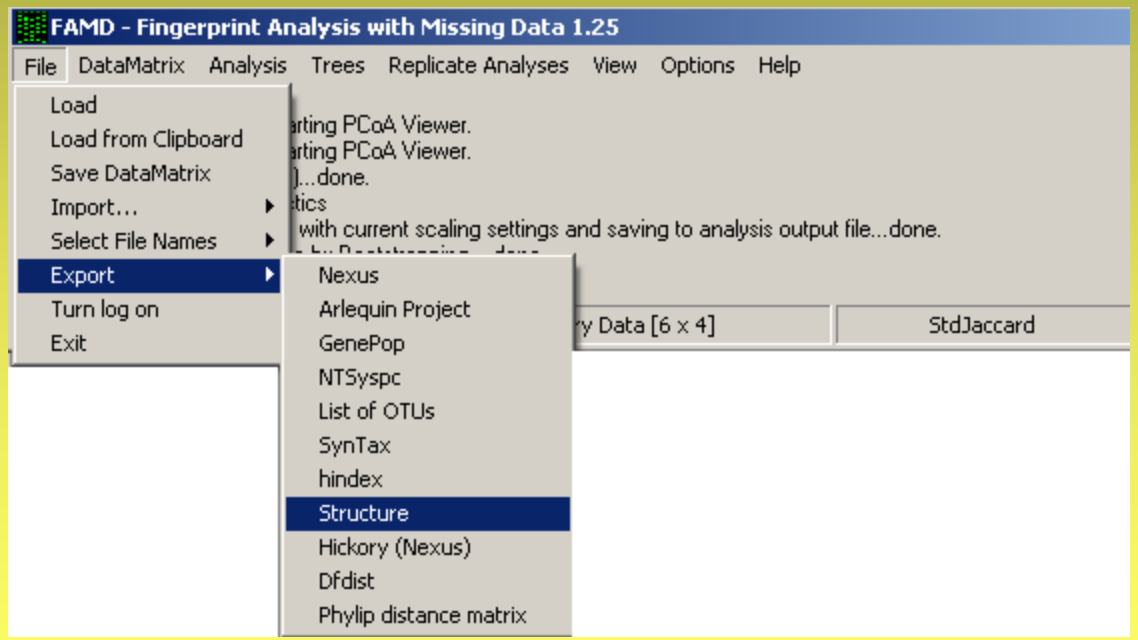
- Analysis – Shannon's index



FAMD

Fingerprint **A**nalysis with **M**issing **D**ata
http://www.famd.me.uk/famd.html

- File – Export



AFLPdat

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- spustit R
- File – Source R code… (vybrat AFLPdat.R)
- File – Change dir… (vybrat adresář s daty)

number	pop	g68	g75	g76	g78	g79	g83	g84
01-1	pop1	1	1	1	0	0	1	1
01-2	pop1	1	1	1	1	0	1	1
01-3	pop1	1	1	1	1	0	1	1
01-4	pop1	1	1	1	0	0	1	1
01-5	pop1	1	1	1	1	1	1	1
04-1	pop2	1	1	1	0	0	1	1
04-2	pop2	1	1	1	0	0	1	1
04-3	pop2	1	1	1	1	0	1	1
04-4	pop2	1	1	1	0	0	1	1
04-5	pop2	1	1	1	0	0	1	1

AFLPdat

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Diversity ("AFLPdat.txt") (Nei's gene diversity)

sample	n	proportion of variable markers	gene diversity
pop1	5	0.242718446601942	0.112621359223301
pop2	5	0.199029126213592	0.0941747572815534
pop3	5	0.135922330097087	0.0689320388349514
pop4	4	0.199029126213592	0.104368932038835

diversities.txt

- Diversity.boot ("AFLPdat.txt ", 1000)

sample	observed diversity	95% conf.int.(lower bound)	upper bound
pop1	0.112621359223301	0.0864077669902912	0.139805825242718
pop2	0.0941747572815534	0.0689320388349514	0.118446601941748
pop3	0.0689320388349514	0.0466019417475728	0.0932038834951456
pop4	0.104368932038835	0.0760517799352751	0.132686084142395

div-boot.txt

AFLPdat

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Rarity ("AFLPdat.txt")

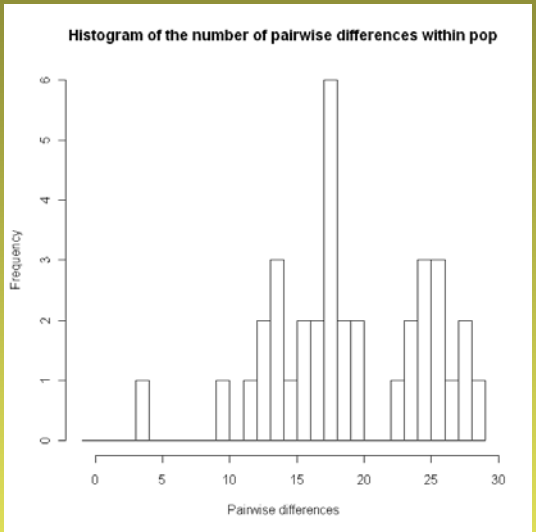
(DW index)

sample	n	rarity 1	rarity 2
pop1	5	5.37479533538357	26.25
pop2	5	4.39386146533205	22.0833333333333
pop3	5	4.73310202869026	21.25
pop4	4	5.37280146324264	24.4166666666667

rarity-pops.txt

AFLPdat

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip



- Clones.list ("AFLPdat.txt", *x*)
  (seznam klonů lišících se o *x* proužků)

List of possible clones

04-1 04-4 4

08-2 08-5 10

- Clones ("AFLPdat.txt", *x*)
  (počet různých klonů v populacích, Dg – genotype diversity, effective number of genotypes)
  Dg = n/(n-1) * [1 – sum (genotype frequencies^2)]
  Effective nb = 1 / sum (genotype frequencies^2)

sample	n	nb of genotypes	genotype diversity	eff. nb of genotypes	gene diversity
pop1	5	5	1	5	0.112621359
pop2	5	4	0.9	3.57142857142857	0.101941747
pop3	5	4	0.9	3.57142857142857	0.068770226
pop4	4	4	1	4	0.104368932

AFLPdat – exporty

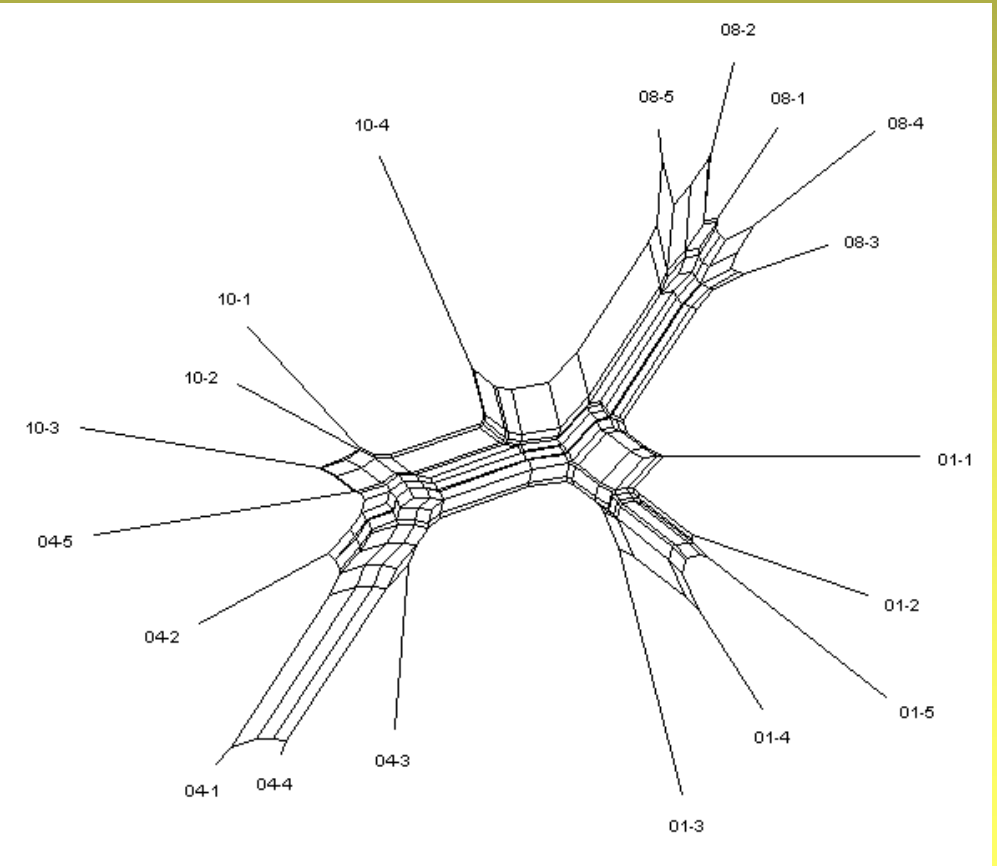
http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Arlequin ("AFLPdat.txt")
- Structure ("AFLPdat.txt")
- Baps ("AFLPdat.txt")
- Popgene ("AFLPdat.txt")
- Hickory ("AFLPdat.txt")
- Nexus ("AFLPdat.txt")

SplitsTree

http://www.splitstree.org

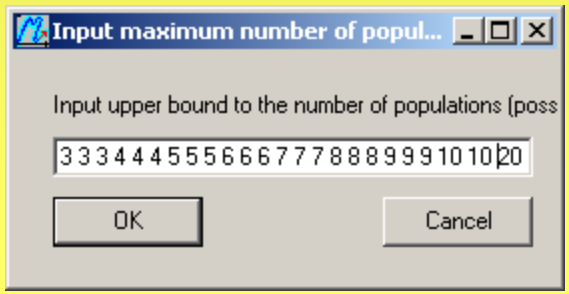
- File – Open (otevřít NEXUS file)



BAPS 3.2

http://www.helsinki.fi/bsg/software/BAPS/

- File – Output File – Set (nastavit výstupní soubor – založit nový!)
- Population Mixture Analysis – Clustering of individuals – BAPS format
- Specify populations (v souboru)
- Input maximum number of populations (možno zadat různá čísla oddělená mezerou – proběhne více „runů“ s max. počtem populací podle zadaného čísla



BAPS 3.2

http://www.helsinki.fi/bsg/software/BAPS/

1	1	0	1	1	0	1	1	1
1	1	0	1	1	0	1	0	2
0	1	0	1	1	0	1	0	3
0	1	0	1	1	0	1	0	4
0	1	0	1	1	0	1	0	5

datový soubor

1
3
9
14
16
20
25

lokality (číslo určuje, který jedinec je už z další lokality)

mixture analysis

Výstup z programu BAPS v3.2

```
RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Data file: S_preprocessed_BAPS.mat
Number of clustered individuals: 258
Number of groups in optimal partition: 8
Log(marginal likelihood) of optimal partition: -8062.66
Best Partition:
Cluster 1: {10, 11, 33, 40, 41, 58, 59, 60, 61, 101, 114, 123,
           124, 126, 131, 132, 137, 159, 161, 162, 164, 166,
           167, 168, 169, 243, 244, 245, 246, 251, 252, 253,
           254}
Cluster 2: {47, 49, 50, 51, 52, 56, 57, 64, 68, 69, 70, 75,
           77, 81, 85, 86, 91, 112, 113, 115, 117, 130, 155,
           173}
Cluster 3: {25, 26, 27, 28, 29, 138, 139, 140, 141, 146, 147,
           148, 149, 150, 151, 156, 160, 163, 165, 184, 185,
           186, 187, 188, 189, 190, 191, 192, 194, 196, 198,
           199, 200, 201, 202, 203, 204, 205, 206, 207}
```

rozdělení jedinců do skupin

```
Changes in log(marginal likelihood) if indvidual i is moved to group j:
  ind      1        2        3        4        5        6        7        8
    1:    -61.0    -20.1    -77.6    -94.0       .0    -67.1   -125.9    -71.3
    2:    -50.4    -18.2    -69.4    -85.4       .0    -60.5   -117.8    -70.9
    3:    -22.5    -60.5    -29.7    -83.0    -64.8       .0    -70.6   -130.4
    4:    -22.9    -58.9    -28.8    -83.2    -61.3       .0    -69.3   -126.6
    5:    -22.3    -53.1    -28.7    -78.5    -61.0       .0    -65.7   -123.4
```

změna likelihood modelu při přesunu jedince do jiné skupiny

```
KL-divergence matrix (Kullback-Leibler):
       1       2       3       4       5       6       7       8
  1
  2   0.415
  3   0.507   1.062
  4   0.407   0.637   1.515
  5   0.782   0.244   1.030   1.433
  6   0.327   0.817   0.269   1.353   0.826
  7   0.583   1.069   1.009   0.780   1.603   0.944
  8   1.050   0.613   1.833   0.910   0.953   1.751   0.776
```

podobnosti mezi skupinami

```
Probabilities for number of clusters
8   0.9984
9   0.001605
```

pravděpodobnost modelu

STRUCTURE 2.3.4

http://pritchardlab.stanford.edu/structure.html

- výpočetně náročné – lépe spouštět na clusteru, např. http://lifeportal.uio.no – verze 2.3.3?
- dva vstupní soubory (data + mainparams)
- *mixture*, *admixture*
- *recessive allele model*
- *independent allele frequencies*
- 10 opakování pro každé K (K=1-10…)
- burn-in: 100 000
- run: 1 000 000
- další zpracování dat – Structure-sum, CLUMPP, distruct

STRUCTURE 2.3.4

http://pritchardlab.stanford.edu/structure.html

datový soubor

```
0        0        0        0        0        0        0
SP-01-1  pop1     1        1        1        0        0        1        1
SP-01-1  pop1     1        1        1        0        0        1        1
SP-01-2  pop1     1        1        1        1        0        1        1
SP-01-2  pop1     1        1        1        1        0        1        1
SP-01-3  pop1     1        1        1        1        0        1        1
SP-01-3  pop1     1        1        1        1        0        1        1
SP-04-1  pop2     1        1        1        0        0        1        1
SP-04-1  pop2     1        1        1        0        0        1        1
SP-04-2  pop2     1        1        1        0        0        1        1
SP-04-2  pop2     1        1        1        0        0        1        1
SP-04-3  pop2     1        1        1        1        0        1        1
SP-04-3  pop2     1        1        1        1        0        1        1
```

Structure-sum

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- R script pro základní analýzu výsledků opakovaných běhů STRUCTURE
- spustit R
- File – Source R code… (vybrat Structure-sum-2009.R)
- File – Change dir… (vybrat adresář s daty)
- list.txt – je třeba vytvořit a umístit do adresáře s výsledky

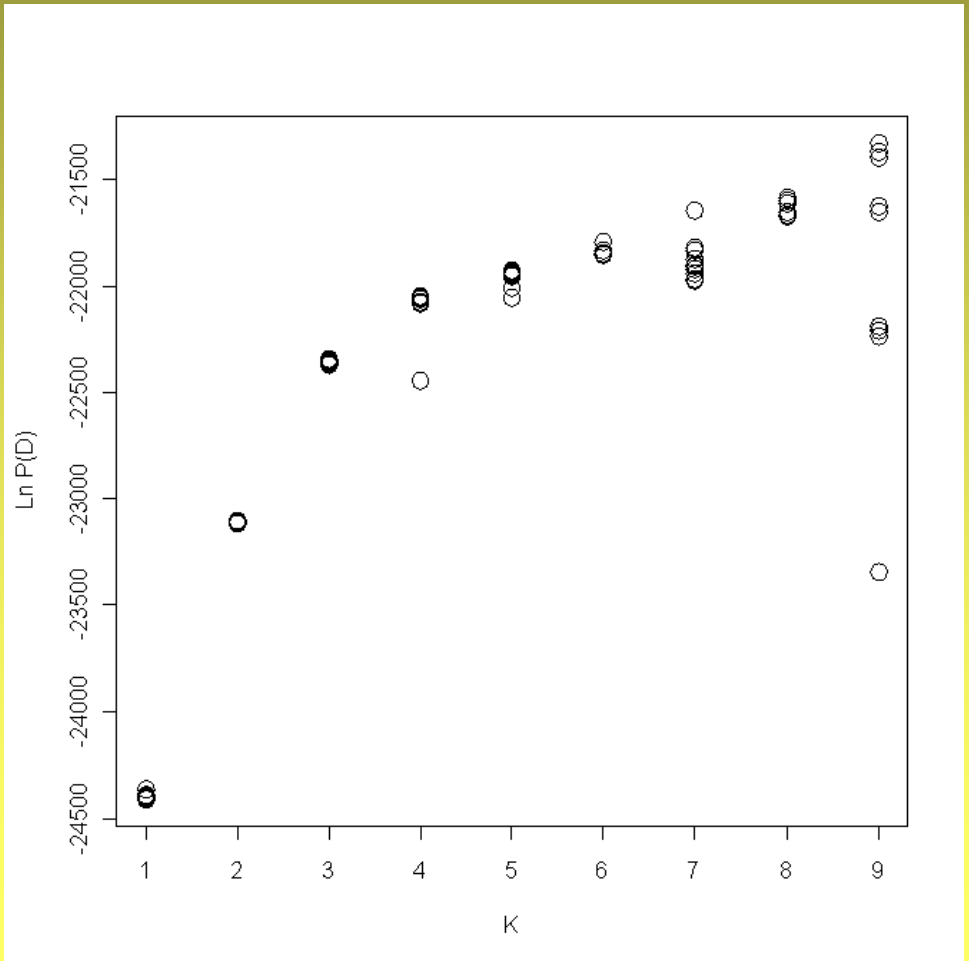
```
1    output_f.1
1    output_f.2
1    output_f.3
1    output_f.4
1    output_f.5
1    output_f.6
1    output_f.7
1    output_f.8
1    output_f.9
1    output_f.10
2    output_f.11
2    output_f.12
2    output_f.13
2    output_f.14
2    output_f.15
```

Structure-sum

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Structure.table ("list.txt", *x*)

x = počet populací



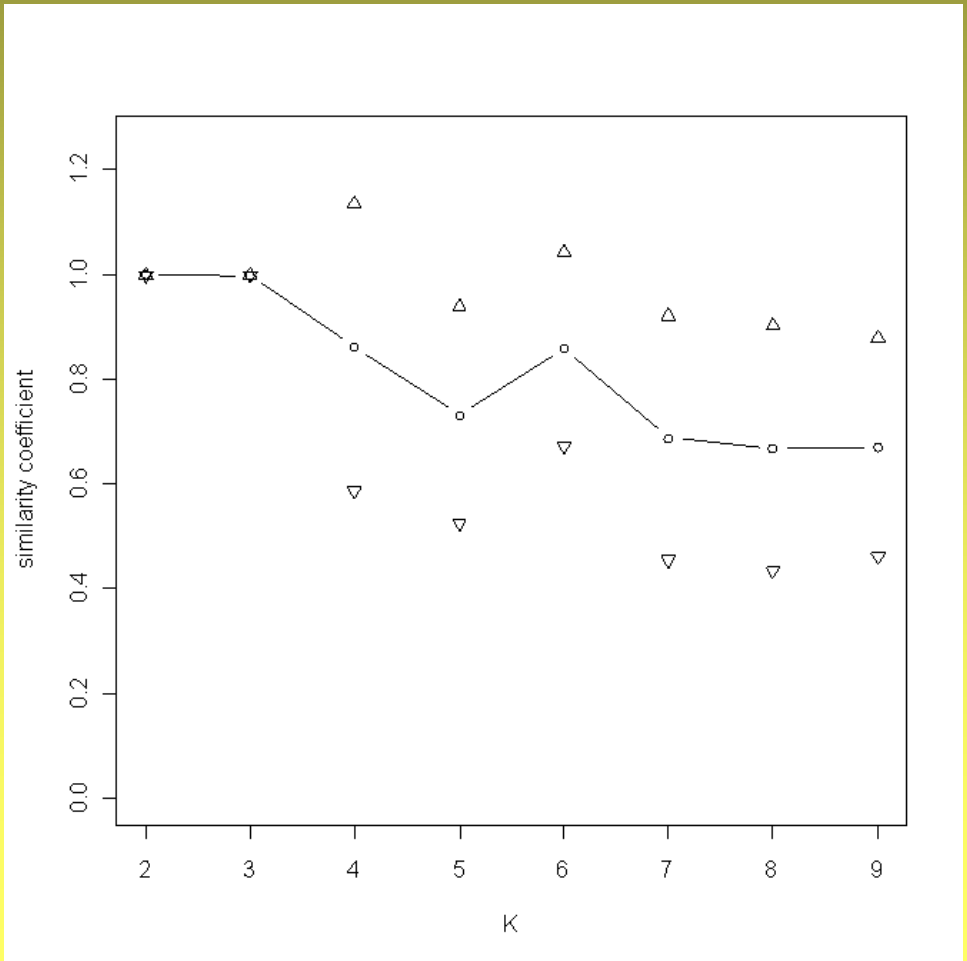
pravděpodobnost modelu
LnP(D)
vs
počet clusterů (K)

Structure-sum

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Structure.simil ("list.txt", *x*)

x = počet populací



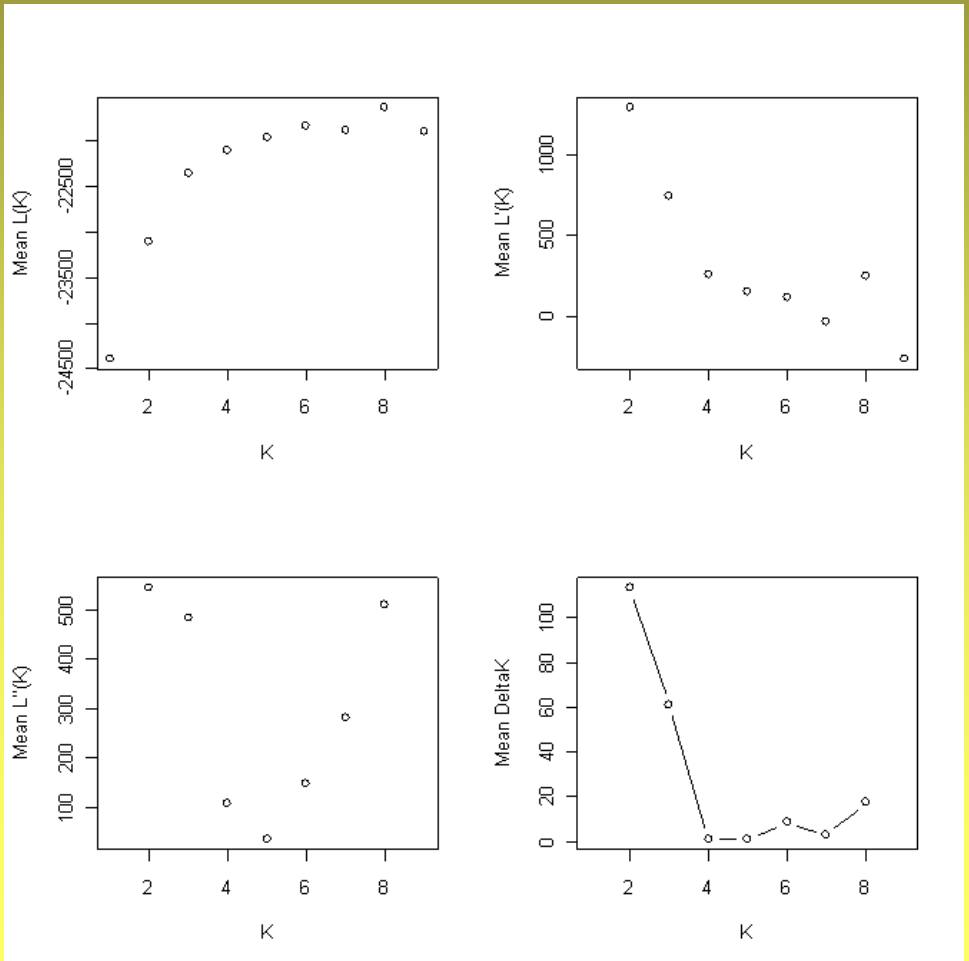
koeficient podobnosti
mezi opakováním pro dané K
(Nordborg et al. 2005)
vs
počet clusterů (K)

Structure-sum

http://www.nhm.uio.no/english/research/ncb/aflpdat/AFLPdat.zip

- Structure.deltaK ("list.txt", *x*)

x = počet populací



delta K
(Evano et al. 2005)
vs
počet clusterů (K)

Distruct

http://www.stanford.edu/group/rosenberglab/distruct.html

- grafické znázornění rozřazení jedinců do clusterů
- barevný *bar plot*
- několik vstupních souborů (umístěné v jednom adresáři společně s distructWindows1.1.exe
  - *.indivq (pravděpodobnosti pro jedince)
  - *.popq (pravděpodobnosti pro populace)

    výsledky STRUCTURE
  - *.names (čísla a jména populací)
  - *.perm (názvy barev pro *bar plot*)
  - drawparams (parametry pro vykreslení)
- výstup – *.ps (post-scriptový soubor) – čitelný např. pomocí Adobe Acrobat (Acrobat Distiller), Ghostscript+GSView nebo http://view.samurajdata.se/

Distruct

http://www.stanford.edu/group/rosenberglab/distruct.html

- *.indivq (pravděpodobnosti pro jedince)

```
1        1    (0)    2 :  0.083 0.917
2        2    (0)    2 :  0.218 0.782
3        3    (0)    2 :  0.236 0.764
4        4    (0)    2 :  0.152 0.848
5        5    (0)    2 :  0.138 0.862
```

- *.popq (pravděpodobnosti pro populace)

```
 2:     0.119  0.881              62
 3:     0.824  0.176              79
 5:     0.155  0.845               5
18:     0.564  0.436               3
```

- *.names (čísla a jména populací)

```
2   pop1
3   pop2
5   pop3
18  pop4
```

- *.perm (názvy barev)

```
1 blue
2 yellow
```

Distruct

http://www.stanford.edu/group/rosenberglab/distruct.html

- drawparams (parametry pro vykreslení)

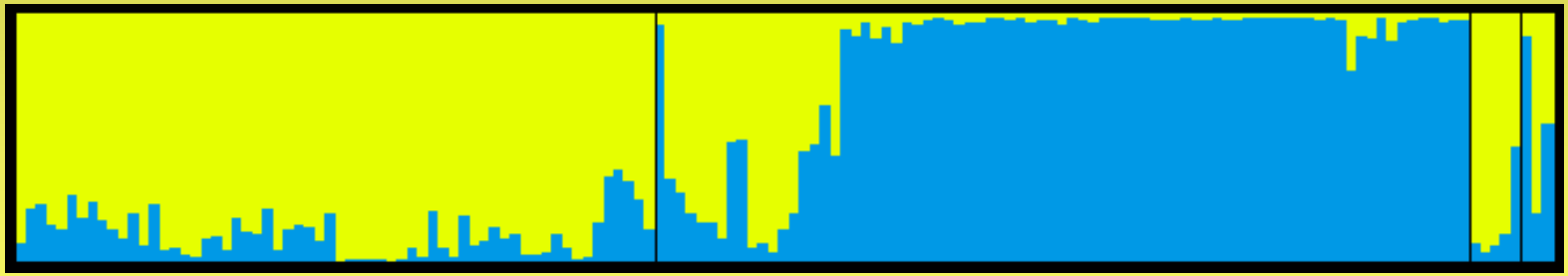
```
"(int)" means that this takes an integer value.
"(B)"   means that this variable is Boolean
        (1 for True, and 0 for False)
"(str)" means that this is a string (but not enclosed in quotes)
"(d)"   means that this is a double (a real number).

Data settings

#define INFILE_POPQ        data.popq      // (str) input file of population q's
#define INFILE_INDIVQ      data.indivq    // (str) input file of individual q's
#define INFILE_LABEL_BELOW data.names     // (str) input file of labels for below figure
#define INFILE_CLUST_PERM  data.perm      // (str) input file of permutation of clusters to print
#define OUTFILE            data.ps        // (str) name of output file

#define K     2       // (int) number of clusters
#define NUMPOPS 4    // (int) number of pre-defined populations
#define NUMINDS 149  // (int) number of individuals

Main usage options

#define PRINT_INDIVS      1  // (B) 1 if indiv q's are to be printed, 0 if only population q's
#define PRINT_SEP         1  // (B) print lines to separate populations

Figure appearance

#define FONTHEIGHT 6       // (d) size of font
#define DIST_ABOVE 5       // (d) distance above plot to place text
#define DIST_BELOW -7      // (d) distance below plot to place text
#define BOXHEIGHT  36      // (d) height of the figure
#define INDIVWIDTH 1.5     // (d) width of an individual
```

Distruct

http://www.stanford.edu/group/rosenberglab/distruct.html

- dvojklik na **distructWindows1.1.exe**
- je vytvořen *.ps soubor
- převést do *.pdf pomocí
  - Adobe Acrobat (Acrobat Distiller) nebo
  - Ghostscript+GSView (http://pages.cs.wisc.edu/~ghost) – freeware
  - http://view.samurajdata.se/

K=2



K=3