Vyhodnocování multilokusových dat I

verze 2013-11-28 (T. Fér, F. Kolář)

1. MSA – základní analýza mikrosatelitových dat
2. FAMD – úprava, export a analýzy AFLP dat
3. AFLPdat – úprava, export a analýzy AFLP dat
4. PAST – analýza hlavních koordinát (PCoA) a vytvoření distančních stromů
5. SplitsTree – distanční sítě
6. BAPS – rychlé určení genetické struktury
7. STRUCTURE – určení genetické struktury

1. MSA - základní analýza mikrosatelitových dat

Pomocí programu MSA (http://i122server.vu-wien.ac.at/MSA/MSA_download.html, freeware) můžeme (pro diploidy!) počítat mj. následující populačně-genetické parametry:

- popisné statistiky pro populaci a lokus (počet alel, H_O, H_E, H_{Sh}, F_{IS}…)
- vzdálenosti mezi jedinci a populacemi (D – standard genetic distance, $(\delta\mu)^2$, D_{ps} – proporce sdílených alel, D_{kf} – kinship koeficient…)
- F-statistika (F_{ST},F_{IS},F_{IT} – globálně i párově mezi populacemi)

Kromě toho program ohlásí varování, pokud by v datech mohla být chyba (alely neodpovídající násobku repetice, příliš velké vzdálenosti mezi alelami, odlehlé alely…)

Vstupní soubor pro MSA je nejlépe zformátovat např. v Excelu a uložit jako TAB-delimited text (a nebo překopírovat do poznámkového bloku). Vstupní soubor pro diploidní druh:

název lokusu

délka repetice

délka *flanking region*

jedno- nebo dvousloupcový formát

2			2		2		2		2		2	
			64		74		46		46		24	
			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

chybějící data

outbrední (d) nebo inbrední (h) individuum

číslo skupiny populací

název populace

Analyzovaný soubor uložíme do stejného adresáře jako je MSAnalyzerMr.exe, který potom spustíme. Otevře se okno pro zadání vstupního souboru a parametrů analýzy. Vložením (i) zadáme vstupní soubor (včetně koncovky!). Pod (d) a následně (p) můžeme nastavit nejrůznější genetické vzdálenosti (1)-(9) a také nastavit výpočet vzdáleností mezi populacemi (c) a jedinci (i), případně zapnout bootstrapové testování (n). Napsáním (b) se dostaneme zpátky do „distance menu“. Pod (s) lze nastavit parametry *F*-statistik, pomocí (c) zapneme výpočet *F*-statistik, pomocí (g) zvolíme, zda budou počítány globálně, párově nebo obojí. Pomocí (m) se vrátíme do hlavního menu, kde pomocí (c) lze nastavit např. vytvoření vstupního souboru pro program Arlequin a/nebo Structure (3). Analýzu lze kdykoliv spustit pomocí (!).

Pokud nebyl nikde zadán bootstrap, proběhne analýza velmi rychle. Na základě nastavených parametrů program vytvoří řadu Excelových tabulek a textových souborů a uloží je do přehledné adresářové struktury:

- Allelecount – počty a frekvence alel pro jednotlivé lokusy a populace
- Distance_data – textové soubory s maticemi vzdáleností mezi jedinci
- Formats&Data – vstupní soubory pro Arlequin a další software
- F-Statistic – *F*-statistika globální i párová
- Group_data – informace o parametrech podle zadaných skupin populací
- Single_data – informace o parametrech pro jednotlivé populace

Pro další informace o parametrech, metodách výpočtu a interpretaci výstupních souborů viz manuál (http://i122server.vu-wien.ac.at/MSA/info.html/MSA_info.html).

2. FAMD – úprava, export a analýzy AFLP dat

Program FAMD (Fingerprinting Analysis with Missing Data) umožňuje jednoduché analýzy binárních dat jako je tvorba matic podobností, UPGMA a NJ stromů, PCoA, AMOVA, výpočty Shannonova indexu ad. Výhodou je, že data nemusí být nijak specielně formátována, stačí 0-1 matici překopírovat z tabulkového procesoru (např. Excelu) do textového souboru, příp. pouze přenést přes schránku. Program je k dispozici na http://www.famd.me.uk/famd.html. Vstupní soubor = binární matice může vypadat např. takto (je uložen jako textový soubor):

```
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;
*
```

V sekci [Groups] jsou definovány čtyři populace (pop1-pop4) a dvě skupiny populací (skup1, skup2). Tuto sekci nemusíme zadávat, program obsahuje „Group manager“, pomocí kterého můžeme populační strukturu „naklikat“ a uložit.

Program spustíme poklepáním na ikonu 

1. Otevření datového souboru: *File* → *Load*, zvolíme *Individuals in rows* a *Header presence for… individuals*, zaškrtneme *Delimited data* a *Include Groups*.
2. Výběr koeficientu podobnosti: *Options* → *Similarity Coefficient Selection*.
3. Vytvoření matice podobností (*Analysis* → *Standard Similarity*). Program ji zapíše do souboru `analysis.txt` (do stejného adresáře odkud jsme načetli data).
4. Strom vytvoříme pomocí *Trees* → *Neighbour Joining*. Strom bude zapsán do souboru `outtree.ph`. Strom zobrazíme pomocí *View* → *Tree File* (je potřeba mít nainstalovaný program TreeView – pokud ho nemáme, můžeme strom otevřít např. v programu FigTree, http://tree.bio.ed.ac.uk/software/figtree).
5. PCoA: *Trees* → *Principal Coordinate Analysis*. Zobrazí se další okno s grafickým výstupem, který je možno modifikovat (výsledky analýzy jsou opět zapsány do souboru `analysis.txt`). Pokud jsme ve vstupním souboru definovali skupiny, je možné si různé skupiny v koordinačním diagramu zobrazit různou barvou nebo symbolem (pod *Group appearence* zvolíme skupinu a přiřadíme jí barvu a symbol).
6. Bootstrap: *Replicate Analyses* → *Bootstrap Std. Tree*), konsenzuální strom lze spočítat pomocí *Tree* → *Strict Consensus* a zobrazit pomocí *View* → *Consensus Tree File*.
7. AMOVA: nejprve zvolíme Euklidovskou vzdálenost *Options* → *Similarity Coefficient Selection*, poté *Analysis* → *AMOVA*. Je nutné mít ve vstupním souboru definovanou příslušnost každého jedince do právě jedné populace, případně skupiny.
8. Statistika proužků pro skupiny: *Data Matrix* → *Count Bands* → *Polymorphic bands / Fixed Bands / Private Bands / Fixed Private Bands*

   Polymorphic bands – % polymorfních proužků
   Fixed Bands – počet proužků, které jsou u všech jedinců ve skupině
   Private Bands – proužky, které se vyskytují pouze v dané skupině
   Fixed Private Bands – proužky, které se vyskytují u všech jedinců v dané skupině a nikde jinde
9. Převedení datové matice na vstupní soubory pro jiné programy: *File* → *Export*

3. AFLPdat - úprava, export a analýzy AFLP dat

Program AFLPdat souborem funkcí spustitelných v balíku R. Umožňuje práci s datovou maticí binárních dat a výpočty populační genetické diverzity, indexů vzácnosti (*rarity*, tzv. DW index), identifikaci identických nebo velmi podobných profilů (možných klonů) a také exportovat data do dalších programů. Program je k dispozici na http://www.nhm.uio.no/english/research/ncb/aflpdat. Vstupní soubor = binární matici můžeme vytvořit např. v Excelu a vypadá takto (textový soubor):

```
number      pop         f1    f2    f3    f4    f5    f6    f7
01x1        pop1        1     1     1     0     0     1     1
01x2        pop1        1     1     1     1     0     1     1
01x3        pop1        1     1     1     1     0     1     1
01x4        pop1        1     1     1     0     0     1     1
01x5        pop1        1     1     1     1     1     1     1
04x1        pop2        1     1     1     0     0     1     1
04x2        pop2        1     1     1     0     0     1     1
```

Před prací s programem AFLPdat je třeba spustit program . Pomocí *File → Source R code…* vyhledáme a vybereme soubor s funkcemi `AFLPdat.R`). Pomocí *File → Change dir…* vybereme adresář, kde se nachází vstupní soubor s daty (např. `AFLPdat.txt)`. Dále píšeme na příkazovou řádku následující funkce:

1. Výpočet vnitropopulační genetické diverzity (pozor soubor nesmí obsahovat populace jen s jedním jedincem):
   `Diversity ("AFLPdat.txt")`
2. Bootstrap genetické diversity:
   `Diversity.boot ("AFLPdat.txt",x)` – 'x' je počet opakování, např. 1000
3. Výpočet vzácnosti (tzv. DW indexu)
   `Rarity ("AFLPdat.txt")`
4. Počet genotypů v populaci, genotypová diverzita
   `Clones ("AFLPdat.txt",x)` - 'x' je počet fragmentů o které se můžou klony lišit (*a priori* rozhodnutí např. na základě spočítané *error rate*)
5. Seznam genotypů lišících se o maximálně zadaný počet fragmentů ('x')
   `Clones.list ("AFLPdat.txt",x)`
6. Převedení datové matice na vstupní soubory pro jiné programy
   - `Structure ("AFLPdat.txt")` – zformátuje data pro program STRUCTURE
   - `Nexus ("AFLPdat.txt")` – vytvoří NEXUS formát, např. pro program SplitsTree
   - `Arlequin ("AFLPdat.txt")` – vytvoří speciální formát pro program Arlequin

4. PAST - analýza hlavních koordinát (PCoA) a vytvoření distančních stromů

Pozn.: Tento návod je k verzi 2.17, nyní uvolněná nová verze 3.1 vypadá částečně jinak (a lépe), ale zatím se v ní nepodařilo plně spustit některé funkce.

Program PAST je distribuován jako samostatně spustitelný *.exe soubor (http://folk.uio.no/ohammer/past/). Umožňuje rychlé provedení analýzy hlavních koordinát i vytvoření distančních stromů (neighbour-joining, UPGMA), kromě toho však nabízí celou škálu dalších analýz např. při zpracování (geometrických) morfometrických dat. Data = binární matici do programu můžeme nejsnáze nahrát přes schránku (clipboard). Pokud chceme s maticí vložit i záhlaví (jména jedinců a jména markerů), je vhodné nejprve zaškrtnout nahoře pole *Edit labels* a pak data vložit. Pokud budeme chtít označovat skupiny, je dobré si jedince nejprve v Excelu seřadit tak, aby byli členové jednotlivých skupin pokud možno u sebe. Alternativně můžeme vkládat přes textový soubor následné podoby:

nr	f1	f2	f3	f4	f5	f6	f7	f8
01x1	1	1	1	0	0	1	1	0
01x2	1	1	1	1	1	1	0	1
01x3	1	1	1	1	0	1	1	1
01x4	1	1	1	0	0	1	1	1
04x1	1	1	1	1	1	1	1	0
04x2	1	1	1	1	0	0	1	1

Alternativně je možné jako data vložit přímo symetrickou distanční matici (typicky pro mikrosatelitová data, např. upravený výstup z programu MSA). Pokud chceme vložit i jména jedinců, je potřeba zaškrtnout nahoře pole *Edit labels*.

```
      01x1  01x2  01x3  01x4
01x1  1     0.225 0.75  0.44
01x2  0.225 1     0.33  0.98
01x3  0.75  0.33  1     0.55
01x4  0.44  0.98  0.55  1
```

1. Definice skupin (budou patřičně obarveny ve všech následujících výstupech): Shift + myší označit patřičné jedince (řádky) z dané skupiny → *Edit* → *Row color/symbol* → vybrat barvu a symbol (Pozn. v této podobě můžeme soubor uložit s libovolnou příponou a kódování barev se uloží i s daty)
2. PCoA: Označit všechny jedince (klik do levého horního rohu) nebo požadovaný výběr (Shift + myš) → *Multivar* → *Principal coordinates* → vybrat distanci (v případě AFLP dat *Jaccard* nebo *Dice*) → tlačítkem zkopírovat do schránky  a uložit si % vysvětlené variability → *View scatter*

   ! Pozn.: pokud máme jako vstupní data přímo symetrickou distanční matici, zaškrtneme při výběrů distancí *User similarity*.
3. Úprava a uložení PCoA diagramu: kliknutím do grafu měníme velikost symbolů písmo apod. a uložit obrázek (*Save picture*). *View numbers* nám zobrazí ordinační skóre jedinců na osách, opět možné přes schránku uložit a vytvořit si hezčí XY obrázek jinde. Tvar a barvu symbolů můžeme měnit pouze ve zdrojové tabulce přes výše zmiňované *Row color/symbol*.
4. Distanční strom (NJ): *Multivar* → *Neighbour joining* → vybrat distanci (v případě AFLP dat *Jaccard* nebo *Dice*) → do okénka *Boot N* napsat počet Bootstrapových opakování a stisknout *Enter*. Analogicky se vytváří i UPGMA, která se nachází v *Multivar* → *Cluster Analysis* (ujistěte se, že *Paired group* algoritmus je označený puntíkem)

   ! Pozn.: pokud máme jako vstupní data přímo symetrickou distanční matici, zaškrtneme při výběrů distancí *User similarity*.

5. SplitsTree – distanční sítě

Program SplitsTree je freeware stažitelný z adresy http://www.splitstree.org. Program umožňuje mimo jiné sestrojení sítí, které ukazují komplexní vztahy mezi jedinci. Jednou z nejvíce využívaných typů sítě je tzv. neighbour network. Vstupním formátem pro program je NEXUS formát, je možné ho vytvořit například v programu AFLPdat (viz výše).

Program otevřeme poklepáním na ikonu 

1. *File* → *Open* – otevřeme vstupní NEXUS soubor
   program zobrazí síť zobrazující vztahy mezi jedinci založenou na uncorrcted p-distances
2. Pokud si přejeme např. některého jedince z analýzy odstranit jednoduše na něj klikneme pravým tlačítkem myši a zvolíme *Exclude selected taxa*. Pokud chceme filtrovat počet spojnicových čar („splits“) učiníme tak v menu *Data* → *Filter Splits* → v záložce *Splits* nastavíme např. podle zobrazené distribuce vah jednotlivých splits vhodné *Weight treshold*.

6. BAPS (v. 6.0) – rychlé určení genetické struktury

Pozn. k instalaci (blíže viz web programu!): Před vlastní instalací programu je třeba nainstalovat Matlab runtime component (MCR) v7.17, dostupný také ze stránek BAPS, popř. ještě jeden Windowsový doplněk (Windows SDK a Net framework). Po instalaci všech komponent raději restartujte počítač.

Program BAPS (http://www.helsinki.fi/bsg/software/BAPS/) hledá takové rozdělení jedinců do K skupin (clusterů), které je na základě molekulárních dat nejlepší (nejpravděpodobnější, má největší *likelihood*). Jeho výhodou je velmi rychlá analýza (řádově minuty max. hodiny) a výsledky srovnatelné se Structure (viz níže), alespoň v případě modelu nepředpokládajícího míšení genotypů v rámci jedinců (mixture = non-admixture).

Vstupní soubor = matici můžeme vytvořit např. v Excelu a je tvořena pouze vlastními daty a posledním sloupcem, který obsahuje řadu čísel („indices“ jednotlivých jedinců). Pokud vládáme kodominantní data, musí být každý jedinec tvořen dvěma řádky

Pro AFLP data vypadá takto (textový soubor; případná chybějící data se značí -9. zde zobrazeno 6 jedinců):

```
1    1    1    0    0    1    1    0    1
1    1    1    1    1    1    0    1    2
1    1    1    1    0    1    1    1    3
1    1    1    0    0    1    1    1    4
1    1    1    1    1    1    1    0    5
1    1    1    1    0    0    1    1    6
```

Pro mikrosatelitová data vypadá takto (textový soubor; případná chybějící data se značí -9; zde zobrazeni 2 jedinci se 4 lokusy):

```
88   148  -9   56   1
86   148  -9   59   1
88   146  23   56   2
88   148  25   59   2
```

Pro následnou lepší vizualizaci je užitečné (ale ne povinné) vytvořit další dva *.txt soubory. První obsahuje jména populací (seznam v jednom sloupci pod sebou) a druhý ta čísla („indices“) jedinců, kteří jsou v daném datovém rámci v každé populaci jako první.

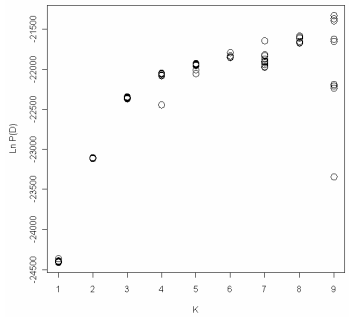
1. BAPS spustíme poklepáním na *.exe soubor  . Po nepříjemné chvilce strávené s černou obrazovkou se objeví vlastní okno BAPSu v němž nejprve zadáme výstupní soubor, kam se nám budou ukládat výsledky. *File* → *Output file* → *Set*.
2. Spustíme *Clustering of individuals* → *BAPS format* → vybereme datový soubor = matici. → Nyní můžeme nebo nemusíme zadat výše zmíněné doplňkové soubory označující rozdělení jedinců do populací. → Nepřejeme si ukládat pre-processed data.
3. V *Input maximum number of populations* zadáme nejvyšší počet skupin (K), do nichž bude program hledat optimální rozdělení. Pokud tedy zadáme např. 5, bude program hledat likelihood rozdělení do 2, 3, 4 a 5 skupin. Obecně je však doporučováno běhy programu (nezávislé starty analýz) opakovat a navíc je vhodné začínat prohledávat možnosti pro méně i více maximálních K. Reálné zadání tak může vypadat např. 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 (čísla oddělena pouze mezerou).
4. Po skončení analýzy se otevře okno zobrazující optimální rozdělení do optimálního počtu skupin (obrázek je možné exportovat např. do pdf). Numerické výsledky včetně

přiřazení jednotlivých jedinců do skupin nalezneme v předem definovaném výstupním souboru. Pokud poslechneme nabídku a uložíme mixture soubor (*.mat) můžeme ho použít v následující analýze *Admixture based on mixture clustering*.

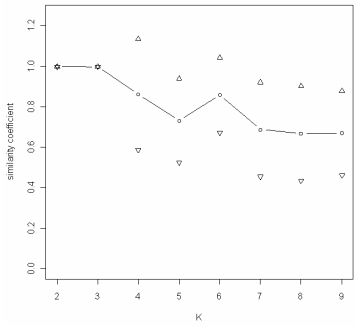
5. V některých případech se hodí podívat se pouze na předem definovaný počet skupin (např. víme, že máme v datech jen dva druhy a chceme se podívat, jací jedinci připadnou k prvnímu a jací k druhému druhu). V takovém případě před započetím analýzy nastavíme *Tools* → *Enable Fixed K Clustering*. Spustíme *Clustering of individuals* a v příslušném okně nastavíme naše preferované K a počet opakování = nezávislých běhů

7. STRUCTURE – určení genetické struktury

Program STRUCTURE hledá takové rozdělení jedinců do K skupin (clusterů), které je na základě molekulárních dat nejlepší (nejpravděpodobnější, má největší *likelihood*) a současně je toto rozdělení nalézáno i při opakovaných bězích programu. K hledání ideálního modelu je využíváno Markovových řetězců (*Markov chain*, MC), které postupně konvergují k nějakému optimálnímu řešení. Protože Markovovy řetězce se zprvu nacházejí zcela mimo ideální řešení, je potřeba stanovit počet kroků řetězce, které předchází stabilní fázi (tzv. *burn-in*, zahoření řetězce), Při běhu programu (tzv. *runu*) postupně hledáme nejlepší model pro K=1 až např. pro K=10 a navíc pro každé K třeba 10 opakování. Jednou ze zásadních otázek potom je, které K je to optimální, protože model s vyšším K mívá často i vyšší likelihood. Následující obrázek ukazuje likelihood modelu [LnP(D)] ve vztahu k vzrůstajícímu K. Vidíme, že po strmém nárůstu hodnoty likelihood od K=1 až po K=4 se křivka začíná zplošťovat. Ideální K je tedy to, kde dochází k narovnávání křivky. Metoda, která zjišťuje, kde k tomu přesně dochází, se jmenuje deltaK.



Dalším parametrem, na který je třeba se dívat je to, zda jednotlivá opakování pro dané K konvergují k podobnému řešení. Pro všechny páry řešení pro jedno K lze spočítat takzvaný Nordborgův koeficient podobnosti, který ukazuje na podobnost mezi jednotlivými *runy*. Za použitelné K lze považovat pouze to, které má tyto koeficienty vysoké. Následující obrázek ukazuje průměry a rozptyl hodnot koeficientů podobnosti pro jednotlivá K. Pro K=2 a K=3 všechny runy konvergují k jednomu řešení (koeficient podobnosti je jedna), všechna vyšší K už mají řešení hodně odlišná.



Pro dominantní data se využívá tzv. *recessive allele model* a předpokládá se, že jednotlivé lokusy nejsou korelované (nejsou ve vazbě nebo jen ve slabé – *independent allele frequencies*).

Běh programu je výpočetně velmi náročný, proto je lepší rozsáhlé datové soubory spouštět na nějakém počítačovém clusteru, kde je program Structure nainstalovaný (např. Lifeportal v Oslu, http://lifeportal.uio.no – na registraci pro přístup se pracuje). K běhu programu na tomto portálu je třeba připravit dva soubory – data a tzv. *mainparams* soubor, který obsahuje parametry analýzy.

A. AFLP data = <u>binární matici</u> je potřeba zformátovat takto (umí např. program AFLPdat, viz výše):

```
0     0     0     0     0     0     0
vz1   01    1     1     1     0     0     1     1
vz1   01    1     1     1     0     0     1     1
vz2   01    1     1     1     1     0     1     1
vz2   02    1     1     1     1     0     1     1
vz3   02    1     1     0     1     0     0     1
vz3   02    1     1     0     1     0     0     1
```

Na prvním řádku je tolik nul, kolik máme lokusů (fragmentů). Na dalších řádcích je '01' informace, které předchází označení jedince a populace. Pozor, kód populace nesmí obsahovat text, pouze celá čísla! Každý vzorek je uveden dvakrát na dvou samostatných řádcích (je diploidní).

B. Mikrosatelitová data <u>= matici alel</u> je potřeba zformátovat následovně:

```
vz1   01    88    148   -9    56
vz1   01    86    148   -9    59
vz2   01    88    146   23    56
vz2   01    88    148   25    59
vz3   02    88    148   -9    56
vz3   02    88    148   -9    56
vz4   02    88    146   23    56
vz4   02    88    146   25    59
```

Na řádcích jsou čísly označeny kódy alel jednotlivých lokusů (v našem případě 1. lokus zahrnuje alely 86 a 88, 2 lokus alely 146 a 148 atd.), kterým předchází označení jedince a populace. Pozor, kód populace nesmí obsahovat text, pouze celá čísla! Každý vzorek je uveden dvakrát na dvou samostatných řádcích (je diploidní), všimněte si, že v případě heterozygotních lokusů/jedinců jsou alely různé.

Soubor s parametry (*mainparams*) vypadá takto (tučně vyznačená části je potřeba změnit a jsou vpravo okomentovány):

```
#define OUTFILE outfile                    název výstupního souboru
#define INFILE data-structure.txt          název vstupního souboru
#define NUMINDS 6                          počet jedinců
#define NUMLOCI 7                          počet lokusů
#define LABEL 1
#define RECESSIVEALLELES 1                 u AFLP dat zde 1, u SSR je 0
#define POPDATA 1
#define POPFLAG 0
#define PHENOTYPE 0
#define MARKERNAMES 0
#define MAPDISTANCES 0
```

```
#define ONEROWPERIND 0
#define PHASEINFO 0
#define PHASED 0
#define EXTRACOLS 0
#define MISSING -9
#define PLOIDY 2
#define MAXPOPS 1
#define BURNIN 100000
#define NUMREPS 1000000
#define LINKAGE 0
#define NOADMIX 0
#define USEPOPINFO 0
#define FREQSCORR 1
```

```
počet generací burn-in
počet generací runu
```

Spuštění STRUCTURE lokálně s grafickým klikacím rozhraním („Front end“)

1. Ikonou spustíme program a nejprve založíme nový projekt *File* → *New project*. Nastavíme adresář, vybereme náš soubor s daty (identický jako v případě spouštění přes portál). V dalším okně napíšeme počet jedinců (ne počet řádek!), lokusů a napíšeme -9 jako hodnotu pro chybějící data. V dalším okně zaškrtneme, že data obsahují *Row of recessive alleles* v případě AFLP dat nebo necháme vše nezaškrnuté v případě mikrostaelitů. Dále zaškrtneme, že naše data obsahují jak *Individual ID for each individual*, tak i *Putative population origin*... Odklepneme ostatní a zkontrolujeme, zda se matice načetla dobře.
2. V *Parameter set* → *New* nastavíme *Length of Burnin Period* (např. 100 000 pro AFLP nebo 10 000 pro mikrosatelity) a *Number of MCMC Reps after Burnin* (např. 1 000 000 pro AFLP nebo 100 000 pro mikrosatelity); v dalších oknech zkontrolujeme, že máme nastaven *Admixture ancestry model* a *Correlated Allele frequencies*. Pojmenujeme a uložíme.
   Pozn. pro mikrosatelity stačí výrazně menší počty MCMC kroků pro *burn-in* i následných replikací, protože jde o analýzu výrazně menšího počtu lokusů, zpravidla 10-30 oproti stovkám lokusů u AFLP, tj. data konvergují rychleji. Také pokud je v datech jasná stuktura, konvergují data mnohem rychleji
3. Vlastní soubor analýz pustíme přes *Project* → *Start a Job*. Vybereme náš právě uložený parametrový soubor a v *Set K from ... to* zadáme rozpětí počtu skupin (K), které chceme analyzovat (např. 1 a 10) a *Number of Iterations*, tedy počet nezávislých běhů pro každé K (např. také 10). Start!
4. Po doběhnutí analýz můžeme v tomto grafickém rozhraní snadno zobrazovat výsledky jednotlivých běhů: *File* → *Open Project* otevřeme soubor s projektem (*.spj) a ve složce „Parameter Sets/jmenoprojektu/Results“ vlevo v adresářovém stromu vybereme patřičný běh a poklepáním myší ho zobrazíme v pravém okně. V menu pravého okna zvolíme *Bar plot* → *Show*. Program však neumožňuje rychlé porovnání více běhů najednou.

7.A. Zpracování výsledků STRUCTURE pomocí programu Structure-sum

Program STRUCTURE-sum je soubor funkcí spustitelných v balíku R. Umožňuje zpracovat výstupy všech *runů* a výsledky vypsat formou tabulek a obrázků. Program počítá také deltaK pro určení optimálního množství clusterů (populací) – viz výše.

Před prací s programem AFLPdat je třeba vytvořit textový soubor, který popisuje, který výstupní soubor je pro jaké K. Soubor nazveme `list.txt` a vypadá např. takto:

```
1     output_f.1
1     output_f.2
2     output_f.3
2     output_f.4
3     output_f.5
3     output_f.6
…
```

Soubor `list.txt` uložíme do stejného adresáře, kam nám program Structure uložil výsledky jednotlivých runů (adresář jsme vybrali při zakládání projektu, v něm se nyní objevily podadresáře s názvem parametru a v něm podadresář „Results“. V něm jsou výsledky runů, které končí číslem a písmenem „f“. Potom už můžeme spustit program R. Pomocí *File* → *Source R code…* vyhledáme a vybereme soubor s funkcemi `Structure-sum-2009.r`). Pomocí *File* → *Change dir…* vybereme adresář „Results“ – viz výše). Dále píšeme na příkazovou řádku následující funkce:

1. `Structure.table ("list.txt", x)`
   - *x* je počet populací ve vstupním souboru
   - vygeneruje obrázek K vs. logaritmus likelihood modelu [LnP(D)] – viz výše u programu Structure
2. `Structure.simil ("list.txt", x)`
   - vygeneruje obrázek K vs. koeficient podobnosti mezi opakováními pro dané K – viz výše u programu Structure
3. `Structure.deltaK ("list.txt", x)`
   - vygeneruje čtyři obrázky, ten vpravo dole ukazuje K vs. deltaK, tj. určení optimálního počtu clusterů (K)

7.B. Zpracování výsledků STRUCTURE pomocí programu Distruct

Program Distruct slouží ke grafickému znázornění pravděpodobnostního rozřazení jedinců do jednotlivých skupin (clusterů). Jeho výstupem je *.ps soubor (postscript), který po zkonvertování do formátu PDF vytvoří známý barevný sloupcový diagram (*bar plot*) s oddělením jednotlivých populací/druhů a popisky.

Program pro svůj běh potřebuje několik vstupních souborů:

- `*.indivq` (pravděpodobnosti pro jedince – z *output filu* Structure pro dané K)
```
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
```
- `*.popq` (pravděpodobnosti pro populace – z *output filu* Structure pro dané K)
```
 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 pro bar plot)
  ```
  1 blue
  2 yellow
  ```
- `drawparams` (parametry pro vykreslení – samovysvětlující… – tučně a šedě vyznačené části je třeba upravit). Pozn. název tohoto souboru není možné měnit a nesmí mít koncovku!

```
"(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
```

Poté co umístíme výše uvedené do stejné složky s programem `distructWindows1.1.exe`, spustíme tento program a zdánlivě se nic nestane. Pokud vše proběhlo správně, ve složce přibyl soubor s koncovkou *.ps, který je nutné překonvertovat do formátu PDF např. pomocí programů Ghostscript+GSView (http://pages.cs.wisc.edu/~ghost), případně on-line na stránce http://view.samurajdata.se/. Výsledkem může být např.: