Plant Biosystematics Group

Masaryk University
Brno, Czech Republic

PŘÍRODOVĚDECKÁ FAKULTA · MASARYKOVA UNIVERZITA · BRNO

ÚSTAV BOTANIKY A ZOOLOGIE

Bi6589

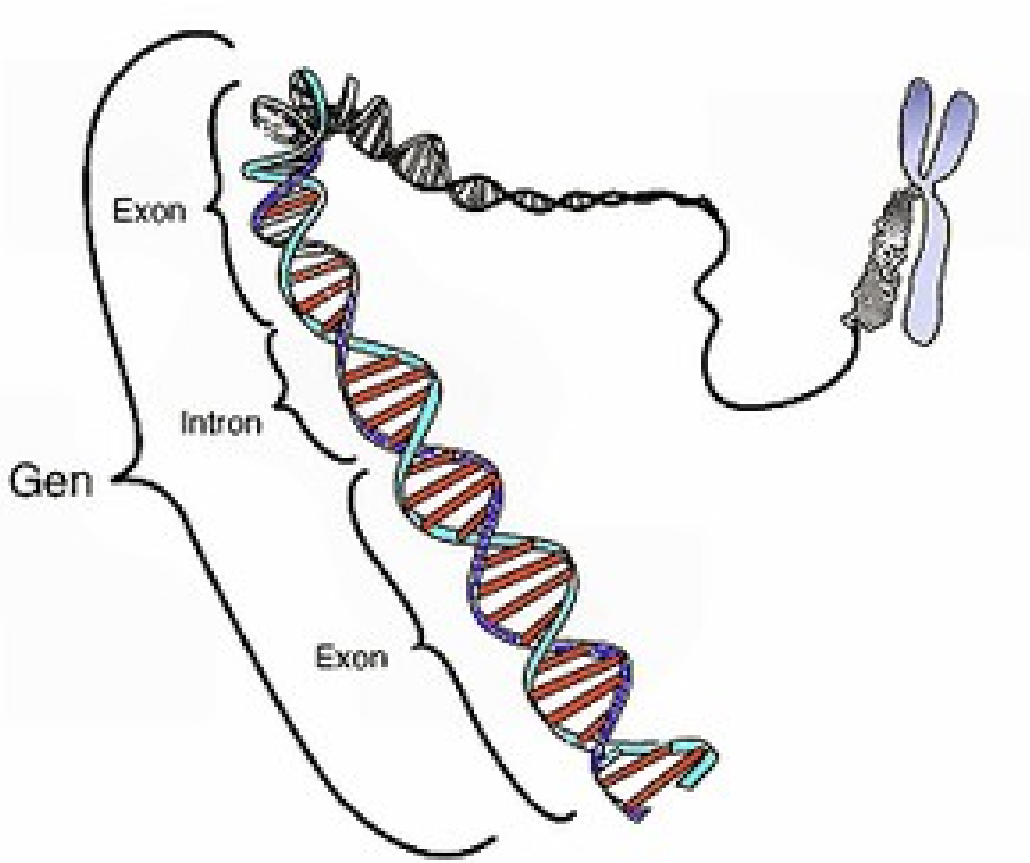
Laboratorní a bioinformatické metody rostlinné biosystematiky

Sekvenační data

Jednotlivé sekvence – obvykle známe historii a původ

Mikrosatelity – délkový vs. sekvenční polymorfismus

NGS – mnoho sekvencí náhodně vybraných z celého genomu



Next Generation Sequencing (NGS)

Next Generation Sequencing (NGS)

Souhrnný termín zahrnující principiálně odlišné technologie, které však umožňují masivní paralelní sekvenování celého genomu nebo jeho částí.

Kritické kroky NGS

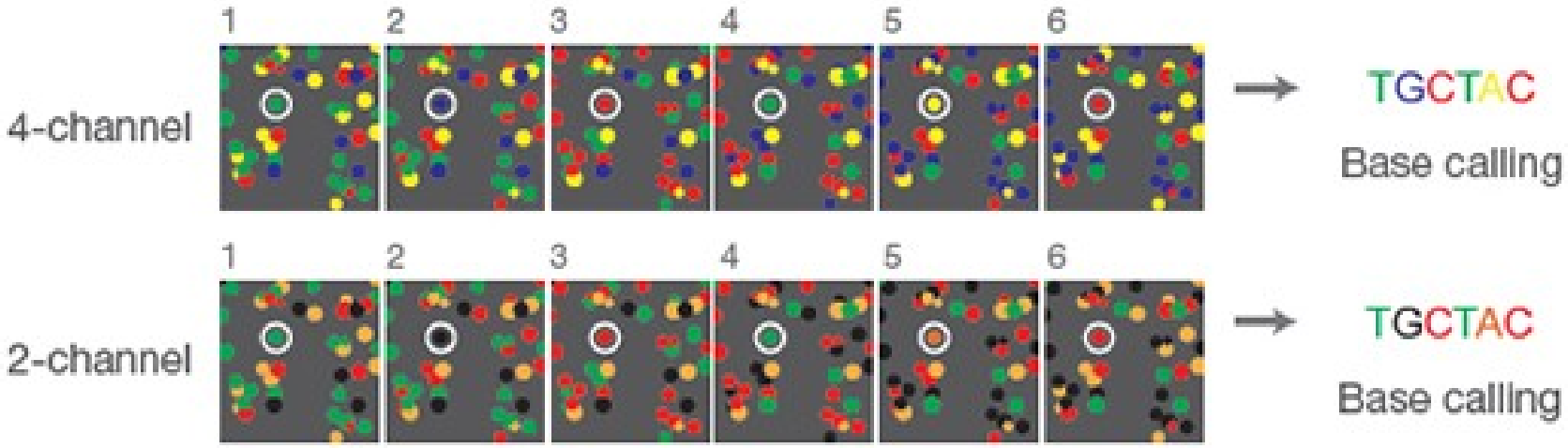
1) příprava vzorků/knihoven (amplifikace DNA; přidání sekvenačních adaptorů/linkerů/bar kódů)
2) generování clusterů (typicky zmnožení DNA fragmentu poté, co linker hybridizuje s pevnou složkou pro můstkovou PCR; mostová amplifikace)
3) sekvenování (různé přístupy, které generují sekvence s různou délkou čtení, chybovostí a finální cenou
4) analýza dat (generováno obrovské množství informací; 1 vzorek > milion sekvencí)

Next Generation Sequencing (NGS)

Přehled sekvenačních přístupů (technologií)

Illumina (mostová amplifikace klastrů, dNTP s reverzibilními blokátory)
Roche/454 (emulzní PCR, pyrosekvenování)
SMRT (Singel Molecule Real Time)
Ion Torrent
SOLiD
cPAL
Oxford Nanopore

a další

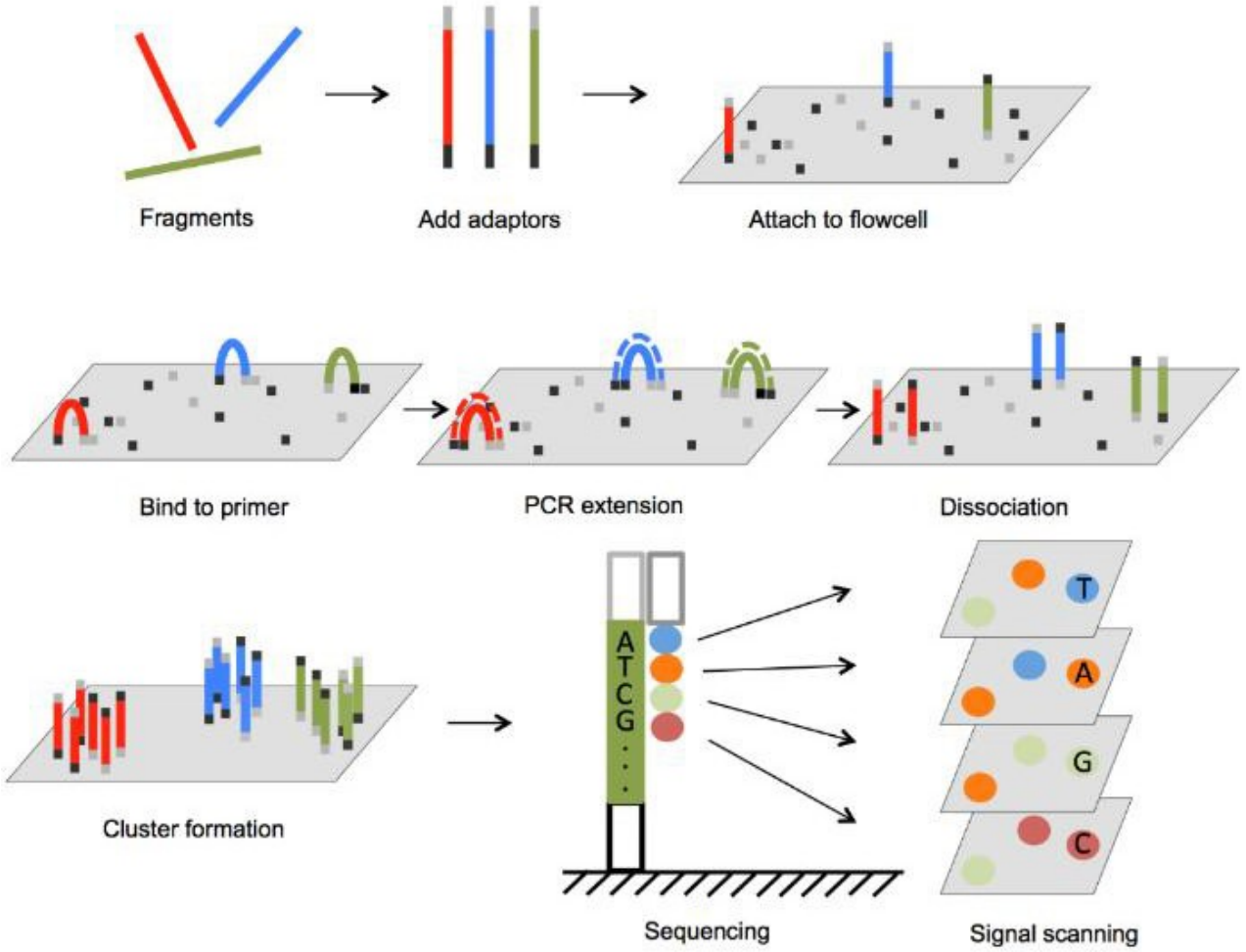


Next Generation Sequencing (NGS)

Princip illuminy

1) Příprava vzorků
2) Generování clusterů
3) Sekvenování

a) doplnění dNTP s reverzibilním fluorescenčním blokátorem;
b) odečtení fluorescenčního signálu;
c) odštěpení blokátoru;
d) pokračování syntézy řetězce.



https://youtu.be/fCd6B5HRaZ8

Virtuální organizace Metacentrum

Gridové výpočetní a úložné centrum vzniklo propojením individuální výpočetních clustrů jednotlivých institucí.

ɔs://wiki.metacentrum.cz/

Virtuální organizace Metacentrum

Je otevřené všem akademickým pracovníků, zaměstnancům a studentům vědeckovýzkumných institucí v České republice.

Hlavní cíl: využití dostupných výpočetních zdrojů pro řešení velmi náročných výpočetních úloh, jejichž zvládnutí je nad možností samostatného pracoviště v ČR.

Výhody
Disponuje špičkovou výpočetní kapacitou
Počítače se samovolně nevypínají/restartují!!!
Spolehlivější uložení a sdílení dat
Umožňuje mezinárodní spolupráci vědeckých týmů

https://wiki.metacentrum.cz/

Virtuální organizace Metacentrum

cesnet
metacentrum

- O MetaCentru
- Aktuality
- Dokumentace a služby
- Přihláška
- Můj účet
- Stav zdrojů
  - Osobní pohled
  - Sestavovač qsub
  - Fyzické stroje
  - Stav PBS uzlů
  - Virtuální stroje
  - Fronty úloh
  - Úlohy
  - Čekající úlohy
  - Uživatelé
  - Vlastnosti strojů
  - Seznam hardware
  - Cloud
  - Statistiky využití

Uživatel jsmerda

Data z PBS server meta: 5. 12. 2023 12:11:13
Data z PBS server cerit: 5. 12. 2023 12:11:20
Data z PBS server elixir: 5. 12. 2023 12:11:22
Data z PBS cache : 5. 12. 2023 12:11:10
Data z MetaCentrum Cloud : 5. 12. 2023 12:07:16
Zobrazeno: 5. 12. 2023 12:11:55

Uživatel(ka) **Ing. Jakub Šmerda PhD.** z organizace **Masarykova univerzita** patří do výzkumné skupiny **žádná skupina - no group** a má účet ve VO **MetaCentrum** platný do **2. 2. 2024**. Spočítal(a) v MetaCentru **3 228 úloh** s celkovou spotřebou **6729,8 dnů** procesorového času.

rok	počet úloh	CPUdny úloh
2021	737	1 717,8
2022	1200	1 989,1
2023	1291	3 028,3

Využití cloudu

žádné VM v cloudu nenalezeny

Úlohy v PBS

fronta	Počet úloh					Počet CPU úloh				
	celkem	ve frontě	běžících	dokončených	ostatních	celkem	ve frontě	běžících	dokončených	ostatních
q_2w@meta-pbs.metacentrum.cz	31	0	21	10	0	31	0	21	10	0
q_4d@meta-pbs.metacentrum.cz	3	0	1	2	0	3	0	1	2	0
celkem	**34**	**0**	**22**	**12**	**0**	**34**	**0**	**22**	**12**	**0**

úloha	server	CPU	vyhraz. paměť	použitá paměť		jméno	CPU čas		čas běhu		stav	stroj/cpu	fronta	čas vytvoření
19011426.meta-pbs.metacentrum.cz	meta	1	10gb	10gb	100%	ipyrad-CIR-HYB.sh	16:31:52	97%	17:01:46	18%	F - dokončena (exit 0)		q_4d@meta-pbs.metacentrum.cz	03.12.23 19:25
19025432.meta-pbs.metacentrum.cz	meta	1	1gb	117mb	11%	STRUCTURE-K1a.sh	17:21:14	100%	17:21:43	7%	F - dokončena (exit 0)		q_2w@meta-pbs.metacentrum.cz	04.12.23 13:26
19025451.meta-pbs.metacentrum.cz	meta	1	1gb	129mb	13%	STRUCTURE-K2a.sh	21:56:32	100%	21:57:11	9%	F - dokončena (exit 0)		q_2w@meta-pbs.metacentrum.cz	04.12.23 13:27
19025523.meta-pbs.metacentrum.cz	meta	1	1gb	118mb	12%	STRUCTURE-K2b.sh	22:38:10	100%	22:40:46	9%	R - běží	kirke1/43	q_2w@meta-pbs.metacentrum.cz	04.12.23 13:27
19025546.meta-pbs.metacentrum.cz	meta	1	1gb	129mb	13%	STRUCTURE-K2c.sh	22:19:16	100%	22:20:16	9%	F - dokončena (exit 0)		q_2w@meta-pbs.metacentrum.cz	04.12.23 13:27
19025562.meta-pbs.metacentrum.cz	meta	1	1gb	129mb	13%	STRUCTURE-K2d.sh	22:03:35	100%	22:04:39	9%	F - dokončena (exit 0)		q_2w@meta-pbs.metacentrum.cz	04.12.23 13:27

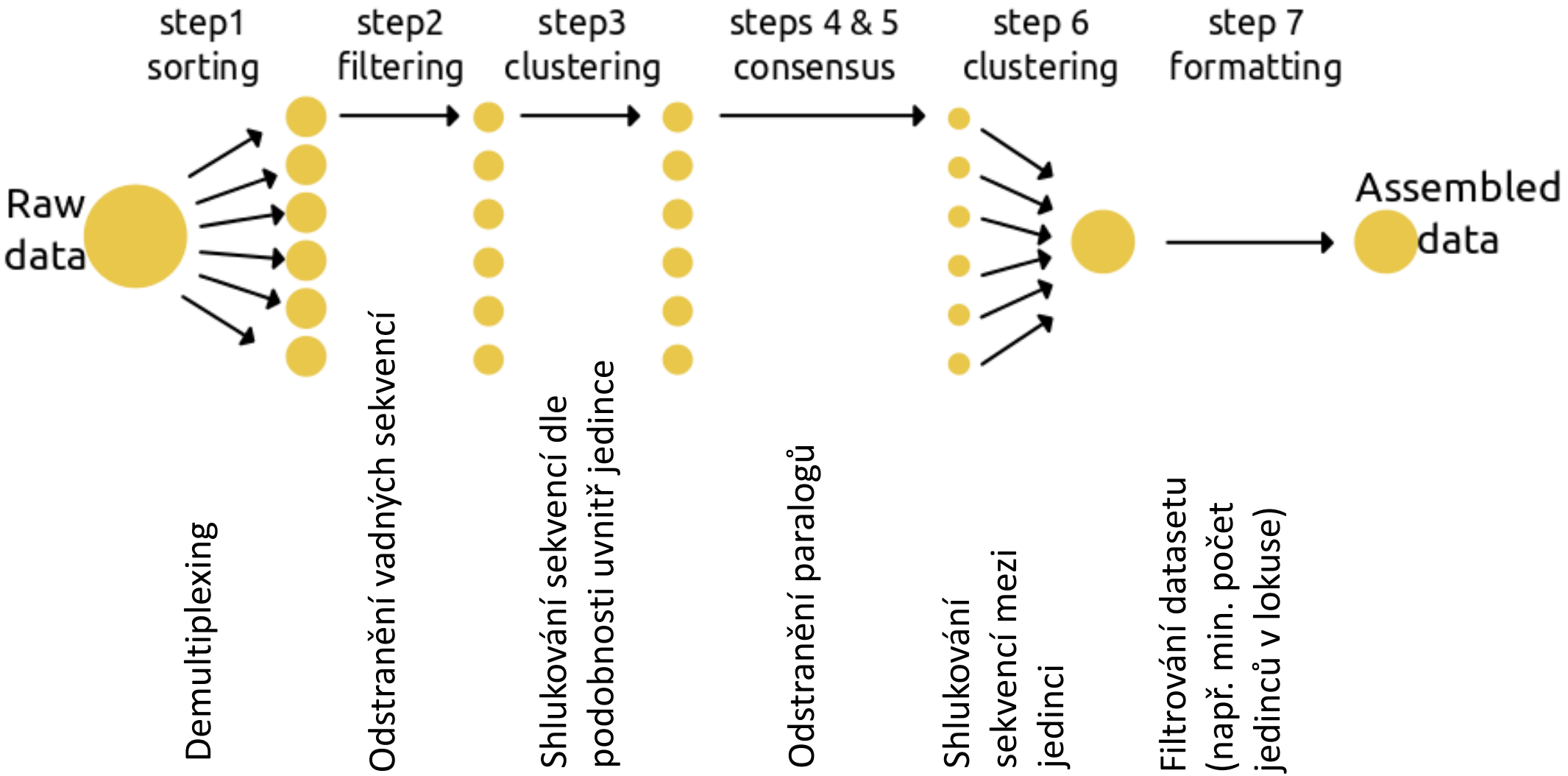
Analýza dat - software

ipyrad (s výhodou využívaný na fylogenetické analýzy; dokáže pracovat s různými ploidními úrovněmi)
STACKS (vhodný na populační analýzy)

Liší se v algoritmu shlukování, uživatelským komfortem/podporou, časovou náročností, atd.

Způsoby shlukování sekvencí
Mapování na referenční genom
De novo analýza dat (bez referenčního genomu)

Analýza dat - ipyrad



Analýza dat - ipyrad

Final Sample stats summary

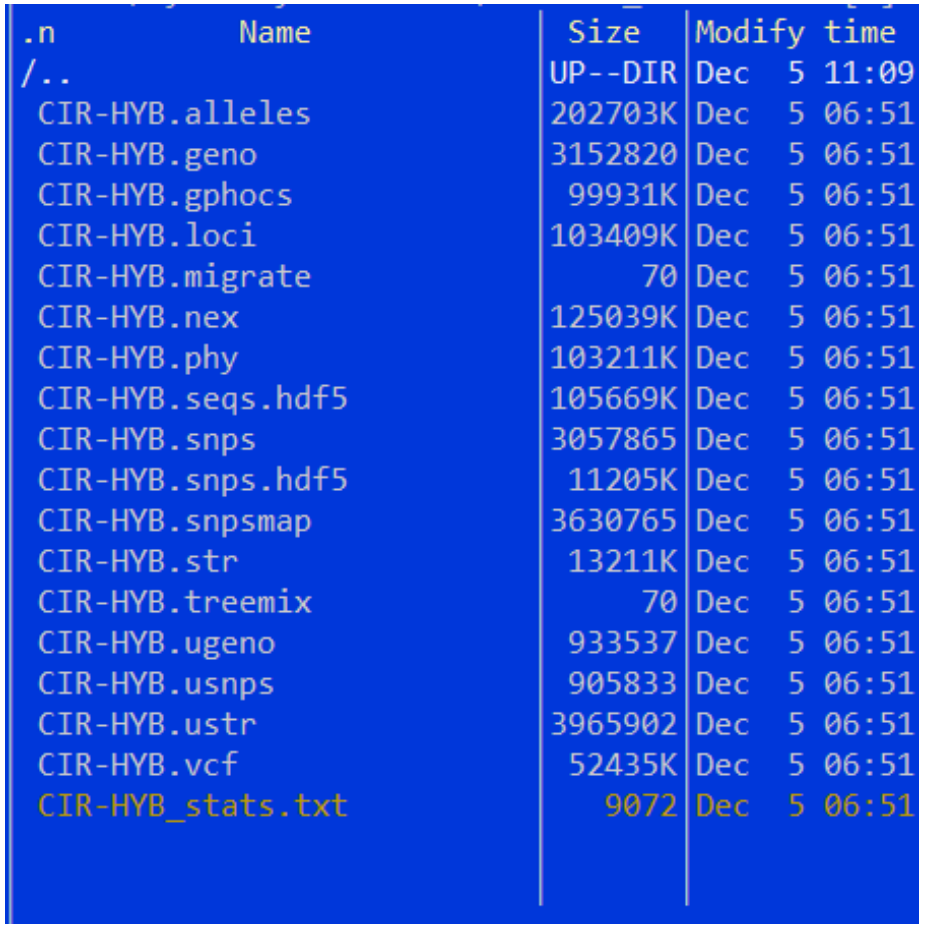
	state	reads_raw	reads_passed_filter	clusters_total	clusters_hidepth	hetero_est	error_est	reads_consens	loci_in_assembly
CIR-D20-4.1	7	4784110	4784110	293796	95878	0.020352	0.004058	80234	30947
CIR-D26-1.1	7	2912505	2912505	104064	51594	0.013374	0.003090	46549	30725
CIR-D27-2.1	7	3678920	3678920	101844	55848	0.014446	0.002890	50086	30016
CIR-D27-3.1	7	4310214	4310214	110747	50121	0.015235	0.003147	44457	30655
CIR-D29-1.1	7	2168088	2168088	133580	48112	0.008350	0.003584	44513	29238
CIR-D29-4.1	7	2221559	2221559	137936	54666	0.015695	0.003334	48604	30670
CIR-D34-3.1	7	2313852	2313852	118996	52579	0.017403	0.003649	45703	29829
CIR-D42-2.1	7	2431954	2431954	87105	46997	0.014361	0.003258	41929	29873
CIR-D57-1.1	7	3564144	3564144	368846	89589	0.019309	0.004410	75329	30278
CIR-D68-3.1	7	1444013	1444013	88739	40438	0.017024	0.004209	35115	26302
CIR-DB13-3.1	7	2289503	2289503	160152	55745	0.016433	0.003825	48786	30122
CIR-DB9-3.1	7	2784283	2784283	140133	52744	0.014855	0.003578	47051	31014
DCAN_1.1	7	3566350	3566350	311868	48600	0.011114	0.004374	43782	29624
DCAN_10.1	7	2714318	2714318	128868	55312	0.010050	0.002639	50572	29130
DCAN_2.1	7	1910191	1910191	132405	45086	0.012578	0.004301	40402	27375
DCAN_3.1	7	2304634	2304634	118430	45069	0.013457	0.003892	40158	28303
DCAN_4.1	7	3422460	3422460	232417	77794	0.019626	0.005980	65798	28242
DCAN_5.1	7	7620944	7620944	304735	99809	0.018457	0.003952	85098	28799
DCAN_6.1	7	2193707	2193707	112661	45432	0.013149	0.003535	40838	28464
DCAN_7.1	7	2374574	2374574	109446	42799	0.010867	0.003562	38985	28350
DCAN_8.1	7	3295664	3295664	164723	56038	0.015527	0.003847	49282	29553
DCAN_9.1	7	3107826	3107826	94780	47614	0.012189	0.003049	43029	29648
DPAL_10.1	7	3177885	3177885	154396	50604	0.009499	0.003437	46371	29922
DPAL_11.1	7	4238926	4238926	118410	50378	0.009096	0.002963	46605	30164
DPAL_2.1	7	2853446	2853446	77245	42327	0.009283	0.003154	39078	28714
DPAL_3.1	7	2160806	2160806	93048	41992	0.009637	0.003274	38701	28603
DPAL_4.1	7	2857725	2857725	118563	50056	0.007617	0.003272	46637	29588
DPAL_5.1	7	2955469	2955469	120640	50358	0.008169	0.003351	46695	29500
DPAL_6.1	7	3358485	3358485	155313	62932	0.011655	0.003424	57204	30683
DPAL_7.1	7	2947075	2947075	98782	47283	0.008864	0.003039	43822	29737
DPAL_8.1	7	2882979	2882979	260854	76593	0.011366	0.003260	69360	29835
DPAL_9.1	7	3701838	3701838	166655	58896	0.011189	0.002775	53715	30168

Alignment matrix statistics:

snps matrix size: (32, 95540), 21.23% missing sites.

sequence matrix size: (32, 3302726), 20.59% missing sites.

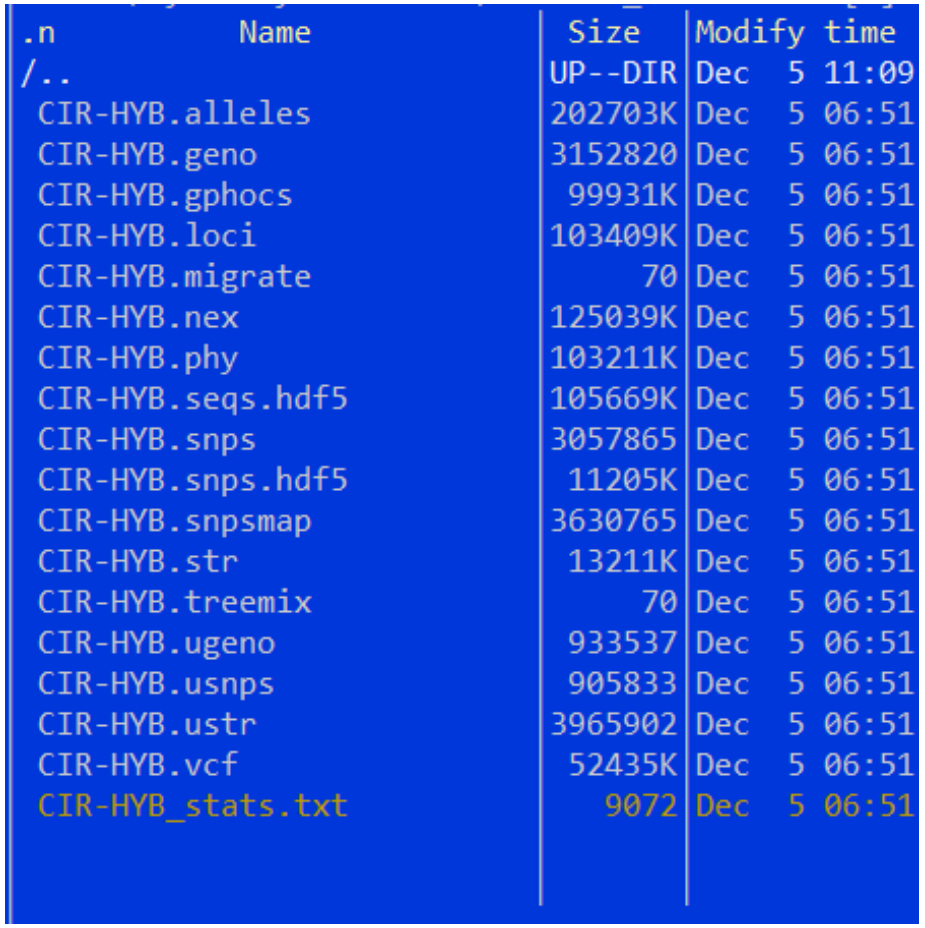
Analýza dat - ipyrad



.n Name	Size	Modify time
/..	UP--DIR	Dec 5 11:09
CIR-HYB.alleles	202703K	Dec 5 06:51
CIR-HYB.geno	3152820	Dec 5 06:51
CIR-HYB.gphocs	99931K	Dec 5 06:51
CIR-HYB.loci	103409K	Dec 5 06:51
CIR-HYB.migrate	70	Dec 5 06:51
CIR-HYB.nex	125039K	Dec 5 06:51
CIR-HYB.phy	103211K	Dec 5 06:51
CIR-HYB.seqs.hdf5	105669K	Dec 5 06:51
CIR-HYB.snps	3057865	Dec 5 06:51
CIR-HYB.snps.hdf5	11205K	Dec 5 06:51
CIR-HYB.snpsmap	3630765	Dec 5 06:51
CIR-HYB.str	13211K	Dec 5 06:51
CIR-HYB.treemix	70	Dec 5 06:51
CIR-HYB.ugeno	933537	Dec 5 06:51
CIR-HYB.usnps	905833	Dec 5 06:51
CIR-HYB.ustr	3965902	Dec 5 06:51
CIR-HYB.vcf	52435K	Dec 5 06:51
CIR-HYB_stats.txt	9072	Dec 5 06:51

Následné analýzy a vizualizace

Analýza dat - ipyrad



.n Name	Size	Modify time
/..	UP--DIR	Dec 5 11:09
CIR-HYB.alleles	202703K	Dec 5 06:51
CIR-HYB.geno	3152820	Dec 5 06:51
CIR-HYB.gphocs	99931K	Dec 5 06:51
CIR-HYB.loci	103409K	Dec 5 06:51
CIR-HYB.migrate	70	Dec 5 06:51
CIR-HYB.nex	125039K	Dec 5 06:51
CIR-HYB.phy	103211K	Dec 5 06:51
CIR-HYB.seqs.hdf5	105669K	Dec 5 06:51
CIR-HYB.snps	3057865	Dec 5 06:51
CIR-HYB.snps.hdf5	11205K	Dec 5 06:51
CIR-HYB.snpsmap	3630765	Dec 5 06:51
CIR-HYB.str	13211K	Dec 5 06:51
CIR-HYB.treemix	70	Dec 5 06:51
CIR-HYB.ugeno	933537	Dec 5 06:51
CIR-HYB.usnps	905833	Dec 5 06:51
CIR-HYB.ustr	3965902	Dec 5 06:51
CIR-HYB.vcf	52435K	Dec 5 06:51
CIR-HYB_stats.txt	9072	Dec 5 06:51

Heatmapa
NeigborNet
STRUCTURE

Distanční matice a degenerované báze

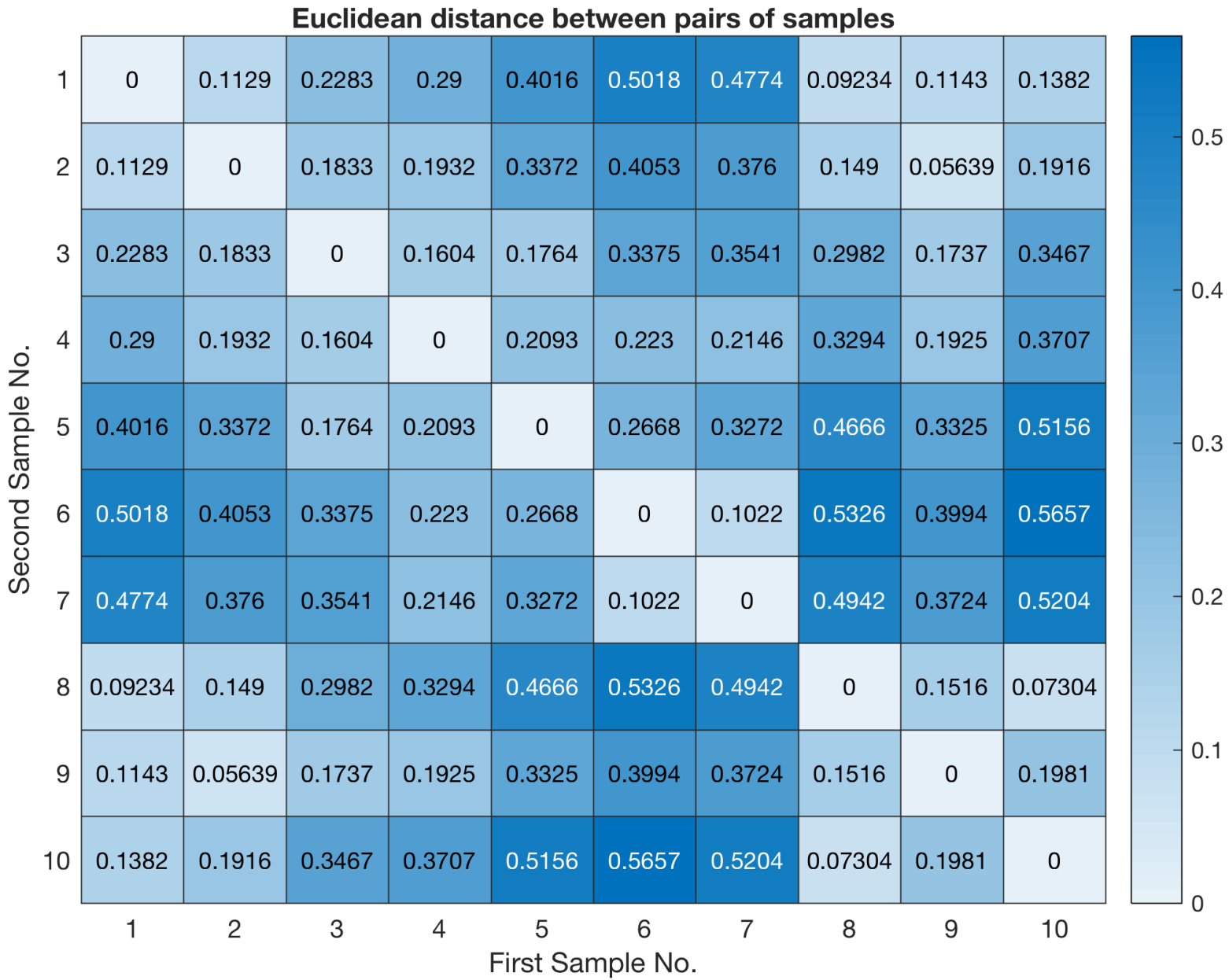
Degenerované báze (ambiguous character/bases) představují problém při výpočtu distanční matice

Několik možností, jak s nimi naložit:

1) berou se jako úplně jiný znak (tzn.: Y není ani C, ani T, ale samostatný znak). Např. P distance
2) mohou být jedna i druha báze (tzn.: Y může být jak C, tak T). Tento přístup snižuje diskriminaci a stahuje hybridy k rodičům (= hybrid je jak rodič-1, tak současně rodič-2). Toto dělá v R např.: dist.ml
3) jsou důsledkem sekvenčního polymorfizmu=konsenzu mezi dvěma sekvenčně odlišnými alelami. Degenerovaná báze znamená, že nese příslušný podíl informace (v Y je 1/2 C a 1/2 T). Hybridi jsou z poloviny rodic-1 a z druhé poloviny rodic-2. Toto dělá v R např.: dist.p
4) degenerovaný kód může být i ignorován, ale pak vznikají fatální chyby (ztráta variability).

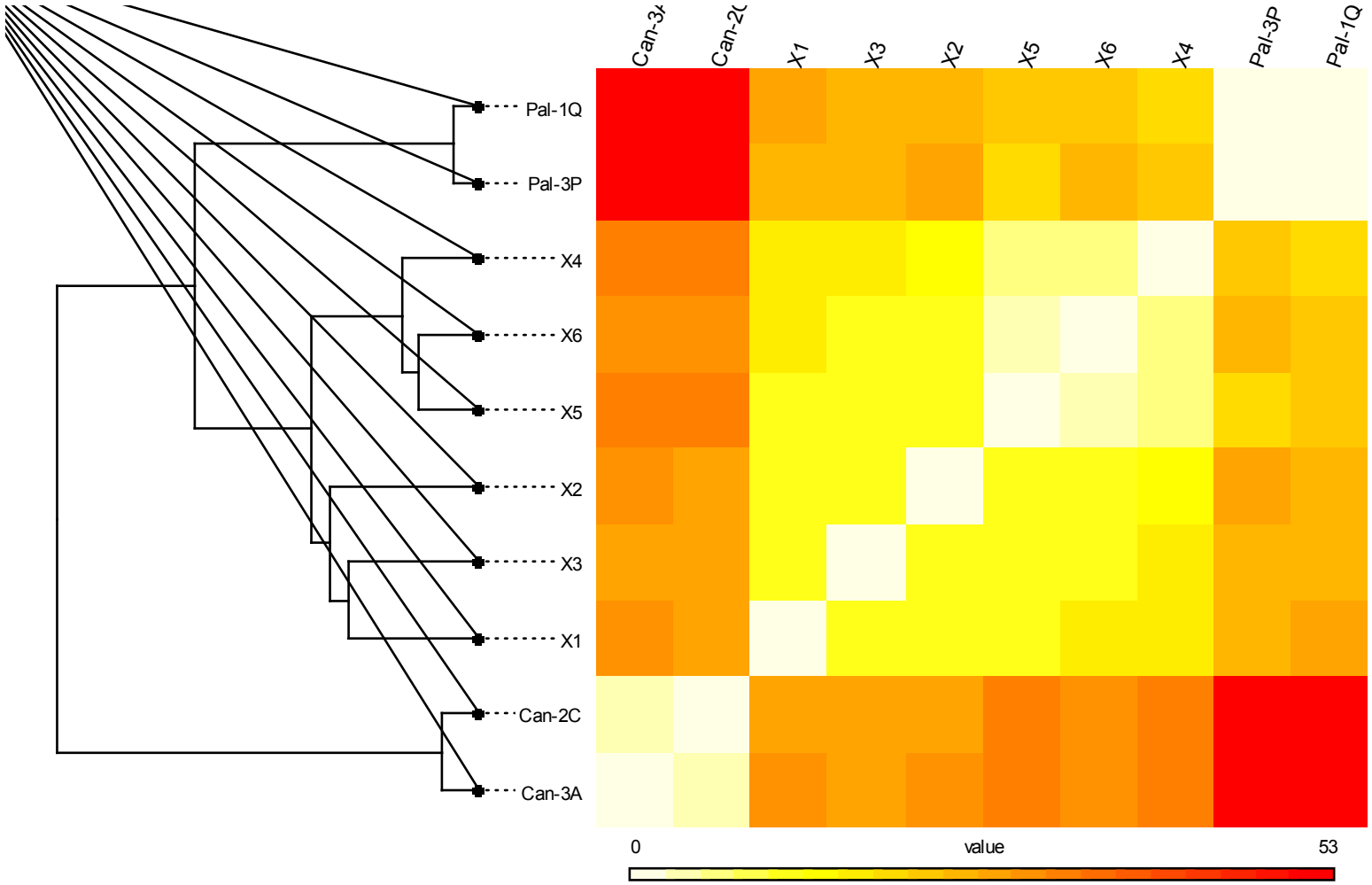
Heatmapa

Vizualizace distanční matice



Heatmapa

Vizualizace distanční matice



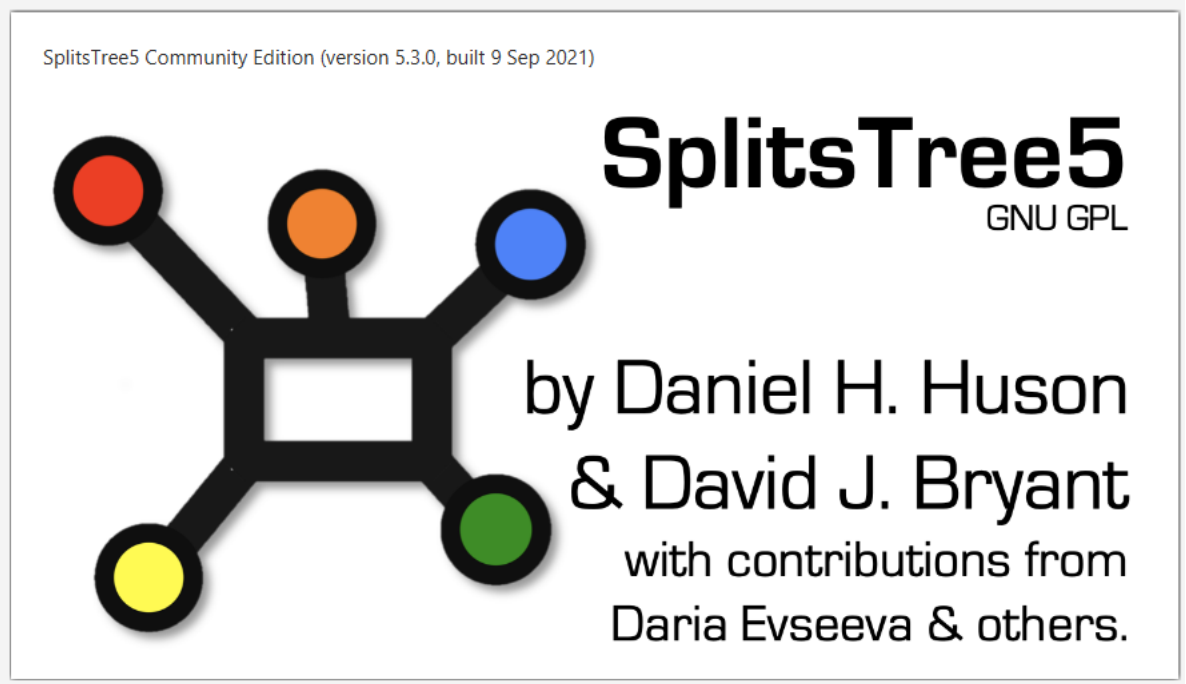
NeigborNet

Algoritmus založený na spojování sousedů do fylogenetické sítě.

Umožňuje vizualizaci konfliktních nebo alternativních evolučních scénářů, které zahrnují např. genové rekombinace, hybridizace a horizontální přenos genů.

Vstupní soubor: matice vzdáleností

Software: **SplitsTree**



Jak číst NeigborNet zobrazení

Velmi odlišní jedinci téže populace = intenzivní genový tok mezi populacemi nebo důsledky hybridizace

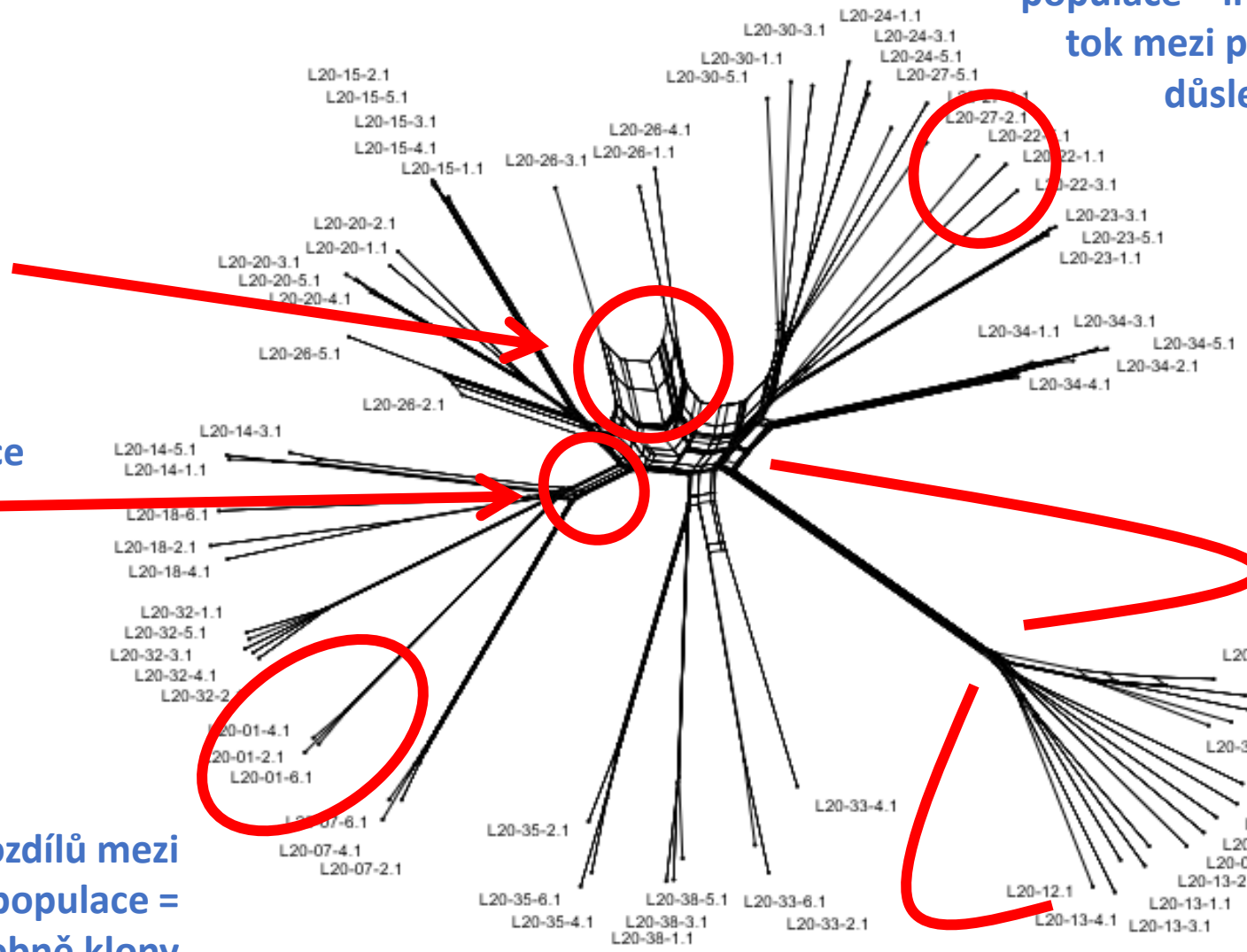
Čím víc se čáry od sebe vzdalují, tím více si alternativní scénáře odporují

Čím více čar, tím více alternativních scénářů

Čím delší společná větev, tím více shodných znaků; může to být důsledek evolučního stáří skupiny nebo, že se nekříží s jinými taxony

Minimum rozdílů mezi jedinci téže populace = pravděpodobně klony

Čím delší individuální větev, tím více jedinečných znaků



STRUCTURE

Software umožňující studovat genetickou strukturu populace/populací:

- zda jsou populace odlišné
- kde se nacházejí hybridní zóny (které populace jsou smíšené)
- kteří jedinci patří do které populace
- kteří jedinci jsou migranti/hybridi apod.

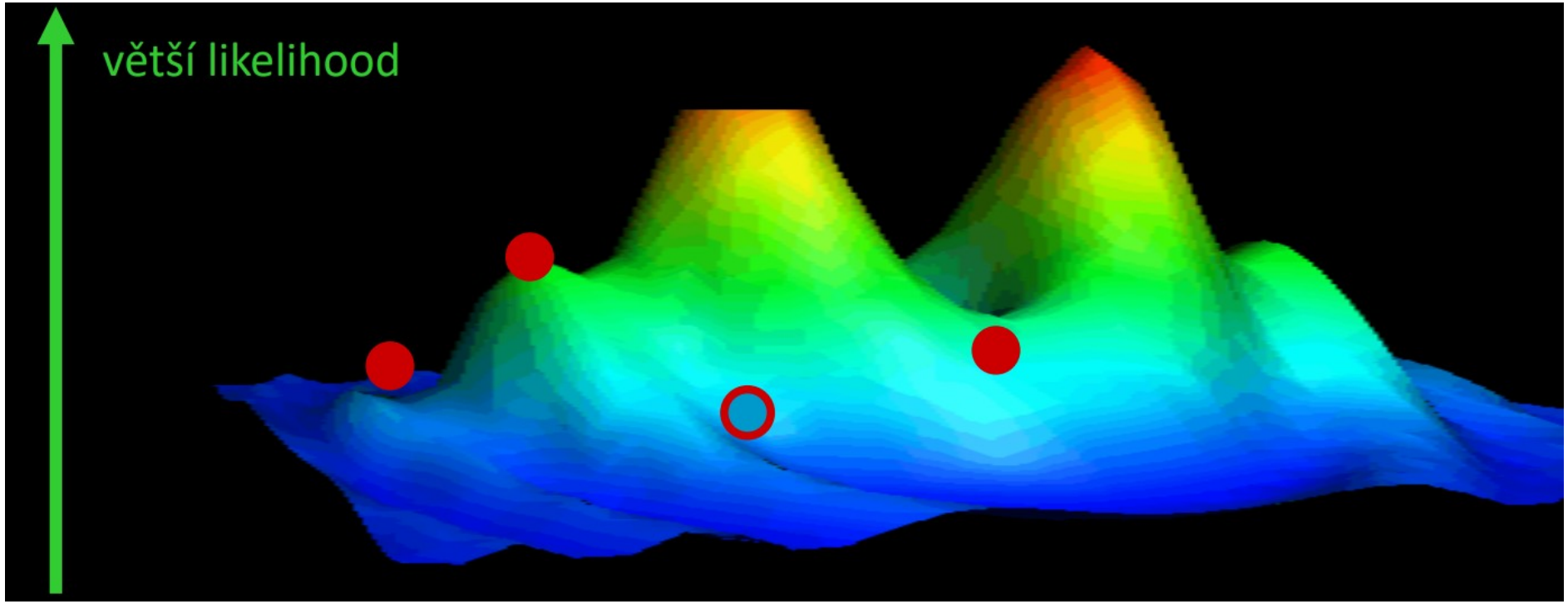
Na základě detekce rozdílů ve frekvenci alel v datech, je jedinec s určitou pravděpodobností přiřazen k nějaké skupině (clusteru).

STRUCTURE neukazuje, který výsledek je správný, ale které rozdělení jedinců do clusterů je nejpravděpodobnější.

Vstupní soubor: lokusy s alelami
Burnin/Iterace (rozrůznění počátečního datasetu/vlastní MCMC výpočet): 100 000/100 000

STRUCTURE

Bayesovský přístup (MCMC: Markov Chain Monte Carlo)



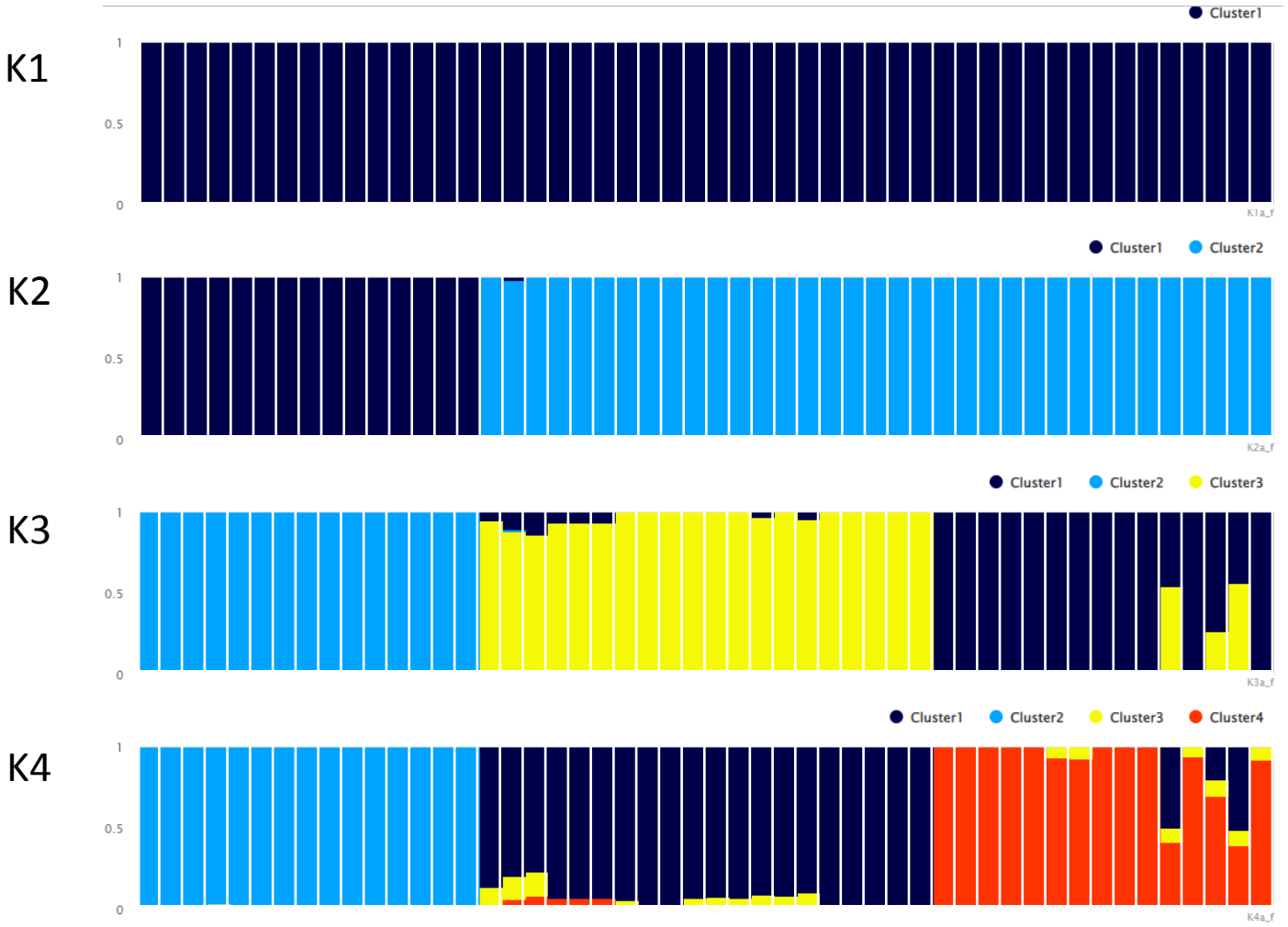
Prohledávání adaptivní krajiny:

čím výše se algoritmus dostane, tím lépe výsledek odpovídá datům

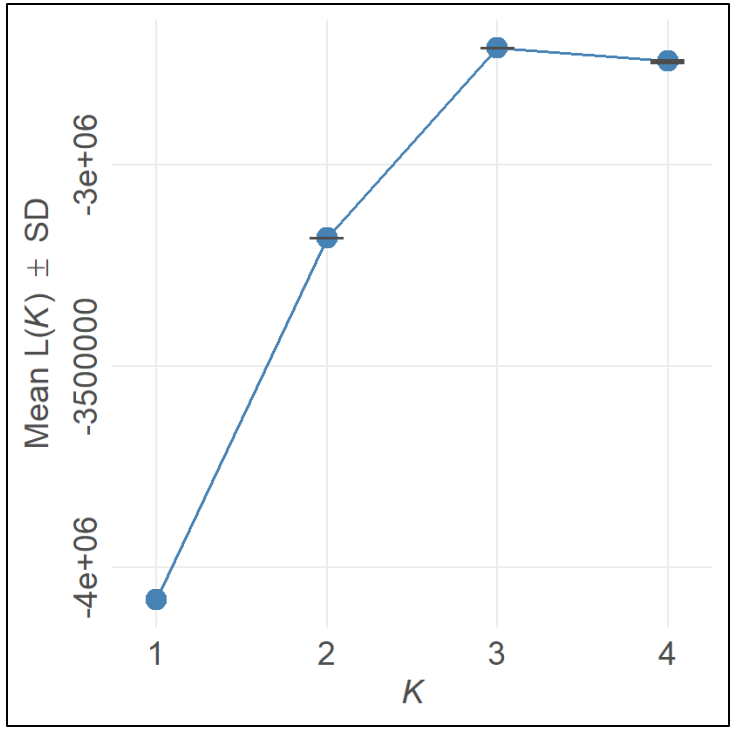
Příklad: 3 teplé řetězce prohledávají krajinu (skáčou z místa na místo) a zavolají 1 studený řetězec v případě, že je výsledek lepší (vyšší místo) než je aktuální poloha studeného řetězce.
= náhodné rozřazení jedinců do clusterů; odhadnuty dílčí genetické frekvence pro každou skupinu; přerozdělení jedinců do clusterů atd.

Jak číst STRUCTURE analýzu

Počítač se snaží rozdělit vzorky do tolika skupin (K), kolik po něm chceme. Poté spočítá k jednotlivým rozdělením (K) i jejich pravděpodobnosti s jakou mohou nastat.

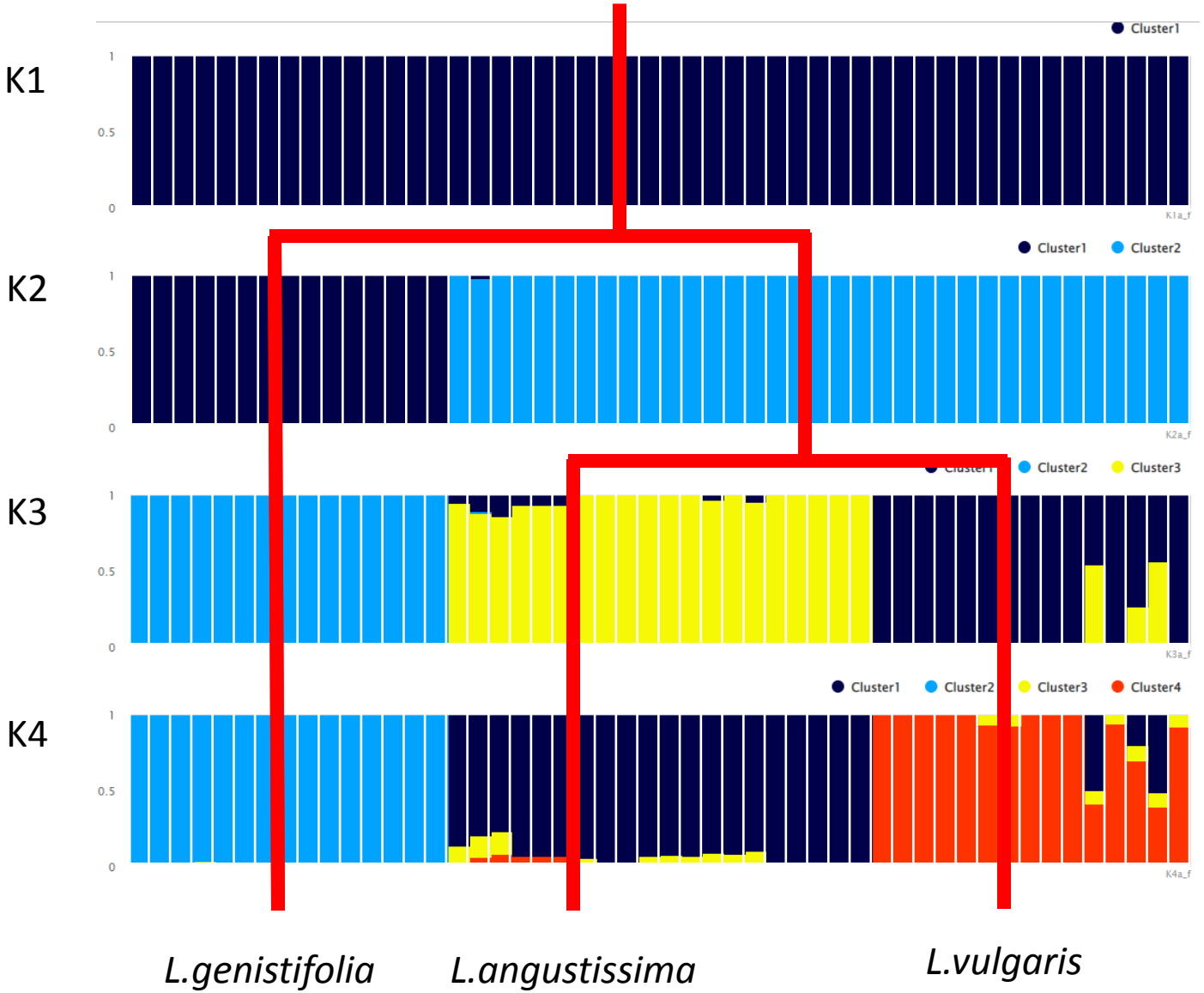


Co sloupec, to jeden vzorek. Barva znázorňuje procentické přiřazení (0-100%) do některé ze skupin (clusteru); hybridi, pak mají barvy kombinované. Malé barevné podíly jsou dost často artefakty metody. Změny barev u různých K nic neznamenají.



Pravděpodobnost s jakou mohou jednotlivé rozdělení (K) nastat; zde je nejpravděpodobnější K3

Jak číst STRUCTURE analýzu



Ve STRUCTURE analýze se „odlupují“ nejdříve skupiny, které jsou nejlépe definované (s charakteristickými znaky)

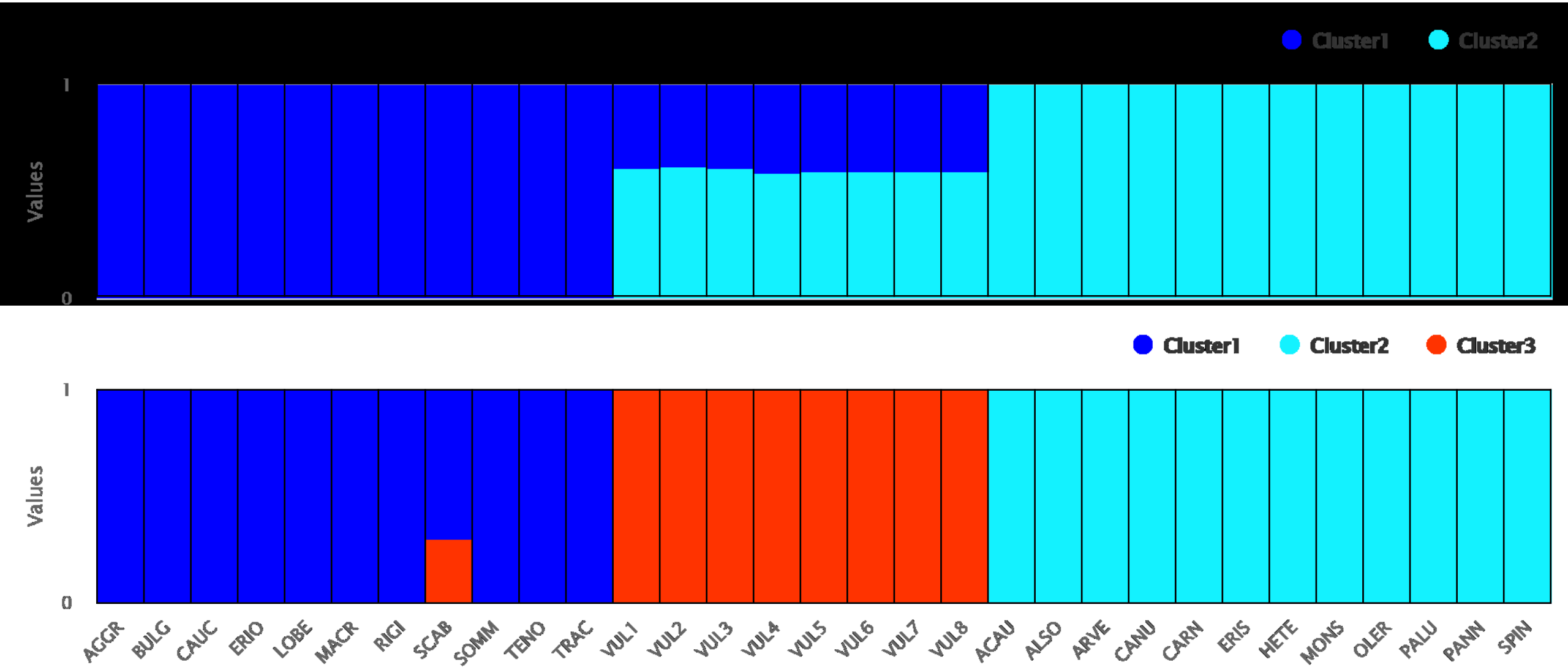
Analýza pak naznačuje i fylogenetické dělení.

Nicméně! Analýza nedává jednoznačné odpovědi – někdy mohou být stejně pravděpodobné různá K, tedy např. pravděpodobnost K2=K3. Někdy v rámci jednoho K vyjde více alternativních rozdělení. Je to Bayesovská analýza a je potřeba být při interpretaci opatrný. Vždy je potřeba přihlížet k tomu, zda výsledek není biologický/genetický nesmysl.

Analýzu mohou ovlivnit nestejně početné skupiny taxonů, málo početné skupiny taxonů, přítomnost hybridů, absence rodiče/ů hybrida.

Alignment: ISCT95 (2n+4n); BSCT93; sample filter: 25
Model: admixture; allele: corelated; without population info
K: 1-4, 10 opakování
Burn-in: 100 00; MCMC iterations: 100 000

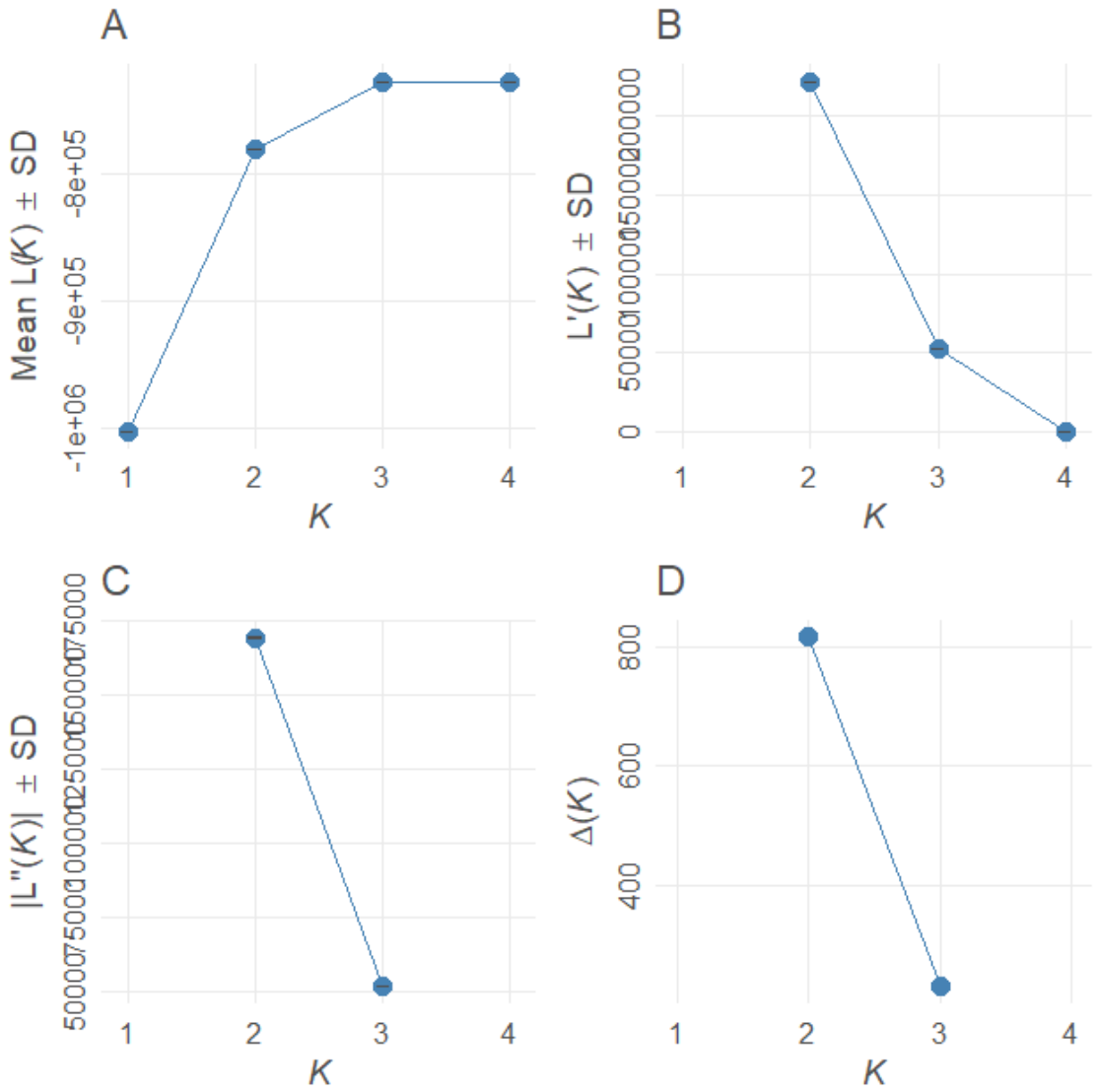
Za určitých okolností mohou dávat smysl (být biologicky pochopitelné) více K



K3a_f

Alignment: ISCT95 (2n+4n); BSCT93; sample filter: 25
Model: admixture; allele: independent; without population info
K: 1-4, 10 opakování
Burn-in: 100 000;
MCMC iterations: 100 000

Evanno analýzy STRUCTURE výstupů VUL

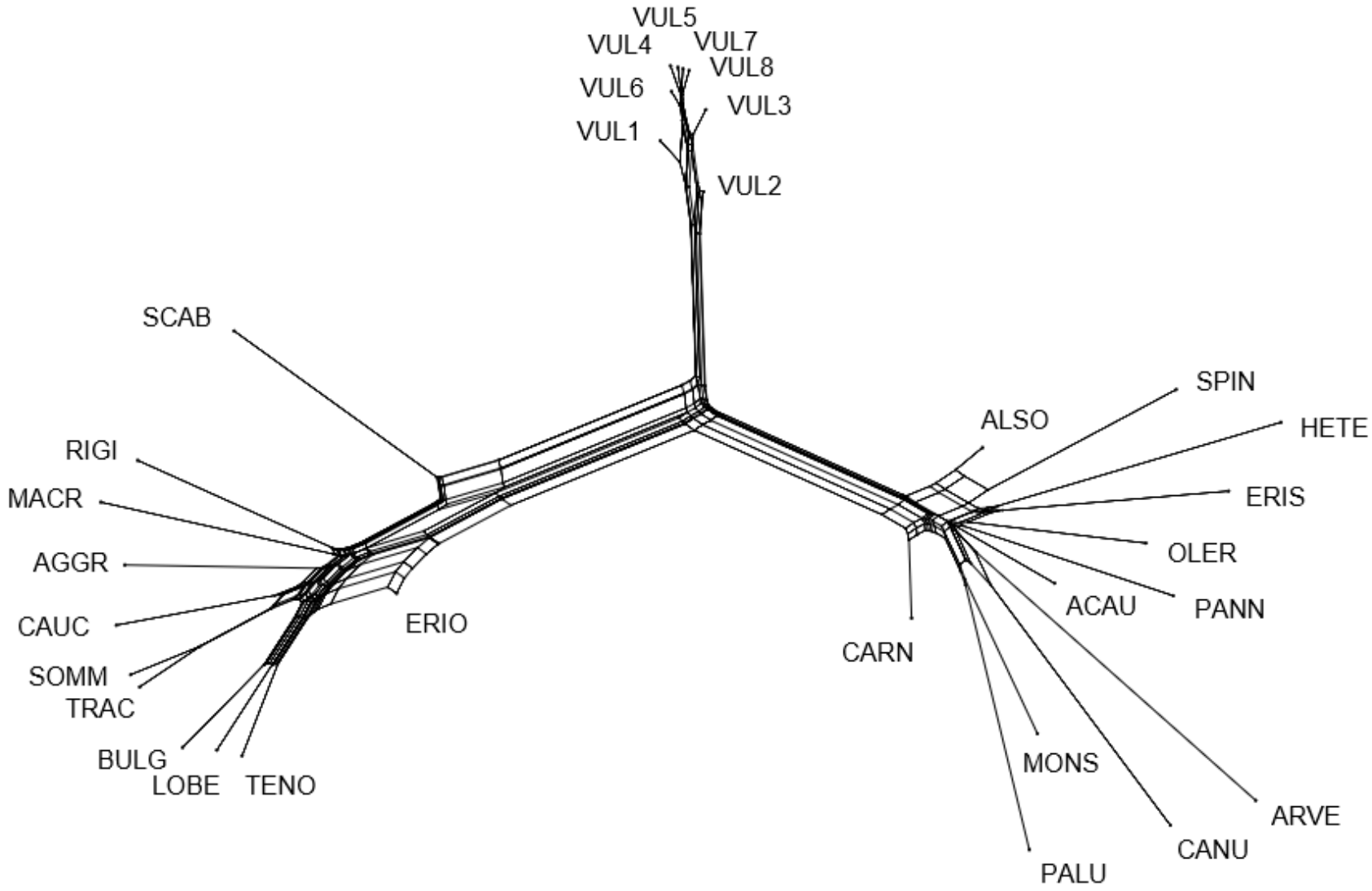


Může být K2, ale i K3
K2 je o trochu pravděpodobnější

NeighborNet zorbazení (software Splitstree)
Alignment: ISCT95 (2n+4n); BSCT93; sample filter: 25
Matrix calculation: dist.p (R package phangorn)

Jaderné sekvence

Fit: 99.95



Alignment: ISCT95 (2n+4n); BSCT93; sample filter: 25
Matrix calculation: dist.p (R package phangorn)
Heatmapa: phylo.heatmap (R package phytools)
Paleta: inferno

Jaderné sekvence