

Microsatellite data analysis

Tomáš Fér & Filip Kolář

Multilocus data

- **dominant** – heterozygotes and homozygotes cannot be distinguished
- **binary** – biallelic data (fragments)
 - presence (dominant allele/heterozygote)
 - absence (recessive allele)
 - i.e., 0-1 scoring
- **anonymous** – unknown genomic origin
- **multilocus** – simultaneous analysis of hundreds of loci, i.e. analysis covers „whole genome“
- RAPD, ISSR, AFLP...
- **codominant** – heterozygotes and homozygotes can be distinguished
- **allelic** – known allelic frequencies in loci, populations...
- **anonymous** – unknown genomic origin
- **multilocus** – usually analysis of few loci (5-20)
- microsatellites (SSRs), isozymes

AFLP

Advantage

- high variability – many loci
- many independent loci (*multilocus method*)
- covering „whole“ genome
- statistical apparatus for data analysis

Drawbacks

- anonymous marker
- asymmetry in probability of loss and gain of fragments – yes/no?
- dominant – impossible to distinguish homozygotes and heterozygotes
- evaluation subjectivity
- unknown rate of mutation accumulation (impossible to use molecular clock)
- problematic (impossible) addition of further samples

microsatellites

Advantage

- usually high variation – many alleles
- codominant – distinguish among homozygotes and heterozygotes, allelic frequencies
- models of allele evolution – „known“ relationships among alleles
- more objective evaluation
- statistical apparatus for data analysis
- possible to add further samples

Drawbacks

- species-specific markers
- simultaneous analysis of limited number of loci
- more limited representation of the „whole“ genome

SNPs

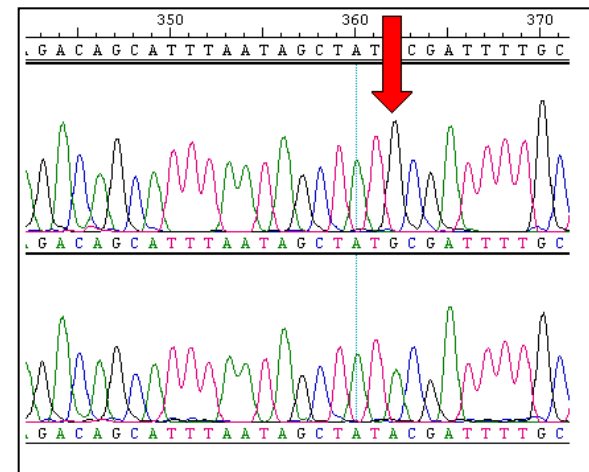
= single nucleotide polymorphisms

Advantage

- combined advantage of AFLP and SSR
- codominant – mainly biallelic
- sequence-based (NGS, e.g., RADseq)
- non-anonymous
- multilocus
- substitution changes => evolutionary models
- up to tens of thousands of loci
- study of neutral and adaptive variability (selection)

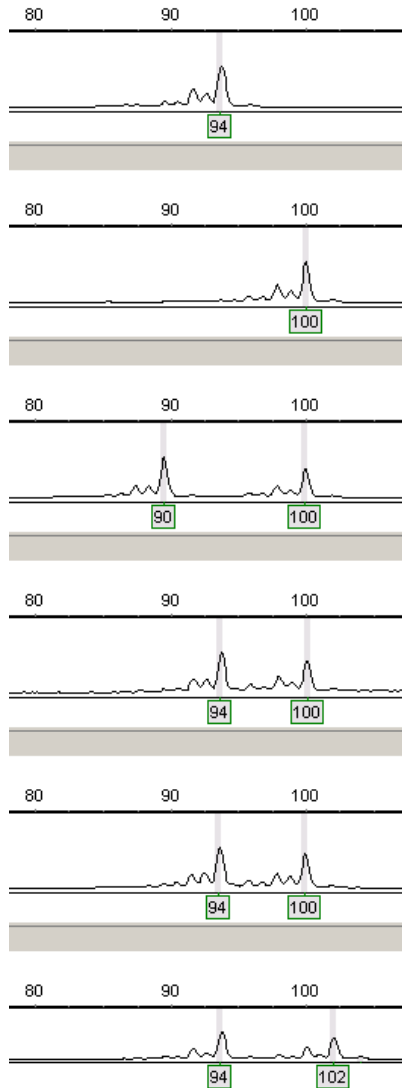
Drawbacks

- not yet stabilized statistical apparatus
- not-stabilized laboratory techniques
- RADseq: null alleles and coverage bias



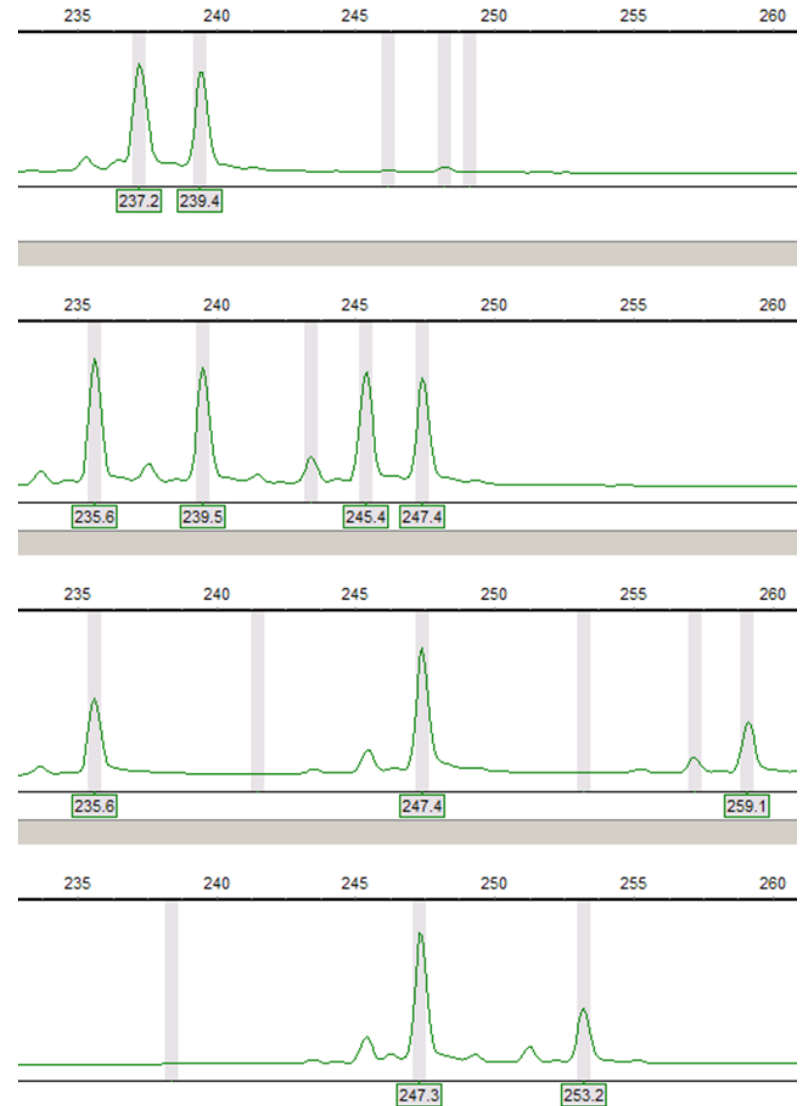
Id	SNP	Consensus	Matching Parents	Progeny	Marker	Ratio	Genotype		
~ 177 pannotate	Yes [2nuc]	TGCAGDGTGTGCACTCCCTCCGCCCGDCTCCCTCTCTCTCTCCCAAGDGTGTAGACACGCCACCGGCTCCAAA TTAACCCCGTAAACG	2	108 / 102	ab/ac	aa: 16 (15.7%) ab: 26 (25.5%) ac: 28 (27.5%) bc: 32 (31.4%)	102		
SNPs		Alleles	Matching Samples						
Column: 17; C/T Column: 46; G/T		a : CG b : CT c : TG	View: <input checked="" type="checkbox"/> Haplotypes <input checked="" type="checkbox"/> Allele Depths <input type="checkbox"/> Genotypes						
FO male	FO female	Progeny 001	Progeny 002	Progeny 003	Progeny 004	Progeny 005	Progeny 006	Progeny 007	Progeny 008
CT / CG 112 / 122	CG / TG 94 / 76	TG / CT / CG 35 / 20 / 3	TG / CT 43 / 50	CG / CT 52 / 60	CT / CG 27 / 47	CG / CT 45 / 24	CG 113	CG / CT 32 / 49	CT / CG 48 / 34
Progeny 009	Progeny 010	Progeny 011	Progeny 012	Progeny 013	Progeny 014	Progeny 015	Progeny 016	Progeny 017	Progeny 018
CT / TG 61 / 66	TG / CG 42 / 64	CG / CT 51 / 49	TG / CG 44 / 36	TG / CT 41 / 43	CT / CG / TG 44 / 1 / 33	CT / CG / TG 40 / 2 / 52	CG 74	CT / CG 26 / 44	CG 102
Progeny 019	Progeny 020	Progeny 021	Progeny 022	Progeny 023	Progeny 024	Progeny 025	Progeny 026	Progeny 027	Progeny 028
TG / CG 41 / 54	TG / CT 57 / 51	TG / CG 69 / 48	TG / CG 59 / 57	CG 100	TG / CT 52 / 51	TG / CT 33 / 55	CG 94	TG / CT 85 / 67	CG / CT 41 / 28
Progeny 029	Progeny 030	Progeny 031	Progeny 032	Progeny 033	Progeny 034	Progeny 035	Progeny 036	Progeny 037	Progeny 038
TG / CG 58 / 47	CT / CG 42 / 40	TG / CG 43 / 50	TG / CG 59 / 59	CG / TG 36 / 50	CT / CG 64 / 76	CG 106	TG / CG 62 / 69	CG / CT 41 / 38	TG / CG 57 / 45
Progeny 039	Progeny 040	Progeny 041	Progeny 042	Progeny 043	Progeny 044	Progeny 045	Progeny 046	Progeny 047	Progeny 048
CG / CT 44 / 38	TG / CG 46 / 46	TG / CG 56 / 52	CG 107	CG / TG 52 / 44	TG / CT / CG 102 / 60 / 2	TG / CT 54 / 50	CG 121	CG 127	TG / CT 43 / 53
Progeny 049	Progeny 050	Progeny 051	Progeny 052	Progeny 053	Progeny 054	Progeny 055	Progeny 056	Progeny 057	Progeny 058
CT / TG 30 / 31	CG 142	CT / TG 48 / 55	CG 107	CT / CG 58 / 60	CT / CG 43 / 49	CG 112	TG / CG 50 / 61	CT / CG 47 / 55	TG / CT 65 / 54
Progeny 059	Progeny 060	Progeny 061	Progeny 062	Progeny 063	Progeny 064	Progeny 065	Progeny 066	Progeny 067	Progeny 068
CT / TG 68 / 60	TG / CT 69 / 73	CG / TG 80 / 78	CT / CG 62 / 60	CG / TG 71 / 90	CT / CG / TG 59 / 1 / 59	CG 107	TG / CG 78 / 75	TG / CT 73 / 82	CT / TG 41 / 67
Progeny 069	Progeny 070	Progeny 071	Progeny 072	Progeny 073	Progeny 074	Progeny 075	Progeny 076	Progeny 077	Progeny 078
CT / CG 49 / 53	CG / CT 78 / 47	TG / CT 56 / 38	TG / CT 66 / 71	TG / CG 48 / 70	CG 117	CT / CG 48 / 48	TG / CT 62 / 48	TG / CG 59 / 51	TG / CT 45 / 31
Progeny 079	Progeny 080	Progeny 081	Progeny 082	Progeny 083	Progeny 084	Progeny 085	Progeny 086	Progeny 087	Progeny 088
CT / CG 49 / 51	TG / CG 61 / 50	CG / CT 60 / 55	TG / CG 50 / 59	CG / CT 70 / 56	TG / CT 60 / 53	TG / CT 40 / 65	TG / CT 60 / 61	TG / CT 57 / 62	CG / TG 64 / 49
Progeny 089	Progeny 090	Progeny 091	Progeny 092	Progeny 093	Progeny 094	Progeny 095	Progeny 096	Progeny 097	Progeny 098
TG / CG 53 / 69	CG 94	TG / CT 56 / 63	CG 60 / 67	TG / CT 116	TG / CT 26 / 24	TG / CT 29 / 30	TG / CT 26 / 22	TG / CT 29 / 22	CG 95
Progeny 099	Progeny 100	Progeny 101	Progeny 102	Progeny 103	Progeny 104	Progeny 105	Progeny 106	Progeny 107	Progeny 108
TG / CG 25 / 27	TG / CT 13 / 30	TG / CT 26 / 21	CG / CT 16 / 13	CG / TG 28 / 23	CT / TG 27 / 26	TG / CG 29 / 24	CT / CG 18 / 21	TG / CT 27 / 14	CT / CG 23 / 15

Evaluation of codominant data



diploids

vs.



tetraploids

Interpretation – assumptions

- alleles can be recognized
 - known ploidy
 - stutter bands
 - +A/-A PCR artefacts
 - artefact peaks peaks recognized
 - allele drop-out
 - null alleles
 - mutation in priming site
 - poor-quality DNA prevents amplification of some alleles
 - longer alleles are amplified with lower probability
 - non-scoring alleles out of the usual range

Data matrix – SSRs

locus name

repeat length

length of flanking region

one- -or two column format

		2			2			2			2			
			98			74			102			46		
			NLGA1			NLGA2			NLGA3			NLGA4		
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	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		

population code

outbred(d) or inbred (h) individuum

number of population group

missing data

Analysis options

- relationships among individuals – basic orientation in the structure
 - distance trees (NJ, UPGMA), networks
 - multidimensional analysis (PCoA)
 - Bayesian clustering
- population-genetic parameters
 - diversity (% polymorphic alleles, diversity indices)
 - divergence (% of unique alleles, DW-index)
 - F-statistics, R-statistics
- testing and detecting spatial structure
 - AMOVA
 - Bayesian estimates, Mantel tests, spatial autocorrelation
- testing specific hypotheses
 - similarity and evolutionary relationships of identified groups
 - hybridization
 - origin of polyploids
 - ...

SSRs example 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



L

Typha × *glauca*



G

Typha angustifolia



A

cattail (*Typha*) hybridization in the USA – *T. x glauca* (F1) is invasive species

Hybridization dynamics? Are there crosses among F1 and are F2 produced? Backcrossing?

Snow et al. 2010

Practical part 1

1. Make PCA/PCoA diagram and unrooted tree based on selected distance matrix from MSA

- run MSA
- modify matrix from MSA to CSV format (text separated by semicolons – using Excel)
- run script in R to make PCoA biplot and unrooted tree
- alternative: copy distance matrix to PAST and make PCoA/NJ tree using „User similarity“

2. Make AMOVA in R

- run script in R

Bayesian clustering

- searching for an optimal partitioning of individuals to K clusters, i.e., with maximum negative logarithm of the marginal likelihood
- result is an optimal number of clusters (i.e., „real populations“) and assignment of all individuals to that clusters
- within populations (clusters) deviation from H-W and linkage equilibrium are minimized (the individuals are assigned to cluster in the way to reach this goal)
 - mixture – each sample to just one population
 - admixture – probabilistic assignment of a sample to more populations
- software
 - BAPS 3.2 – Bayesian Analysis of Population Structure (Corander et al.) (stochastic optimization)
 - STRUCTURE (Pritchard et al.) (MCMC)

STRUCTURE output example

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

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

a priori populations vs. clusters

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	-

divergence among clusters

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

variability inside clusters

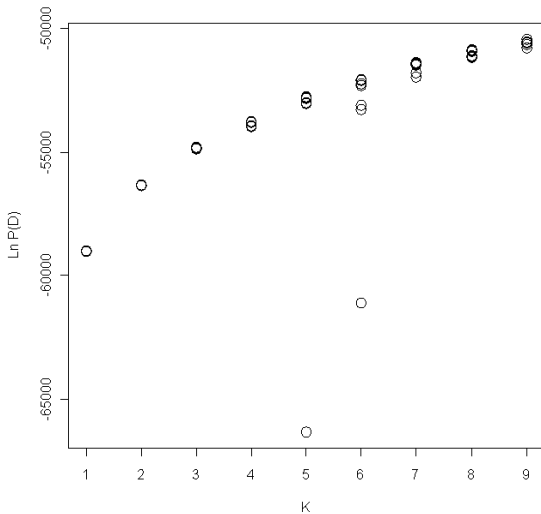
Inferred ancestry of individuals:

	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

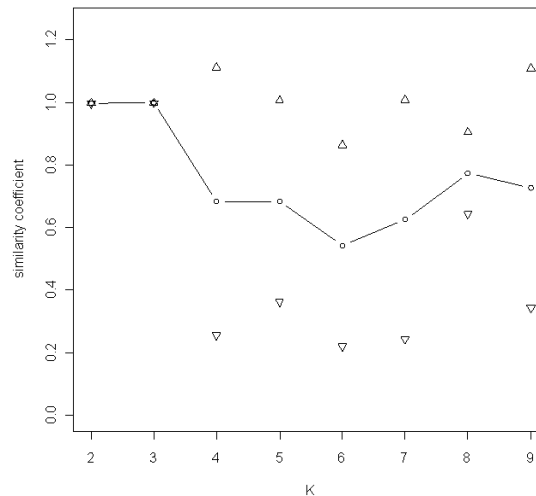
sample assignment probability
to all clusters

STRUCTURE results evaluation

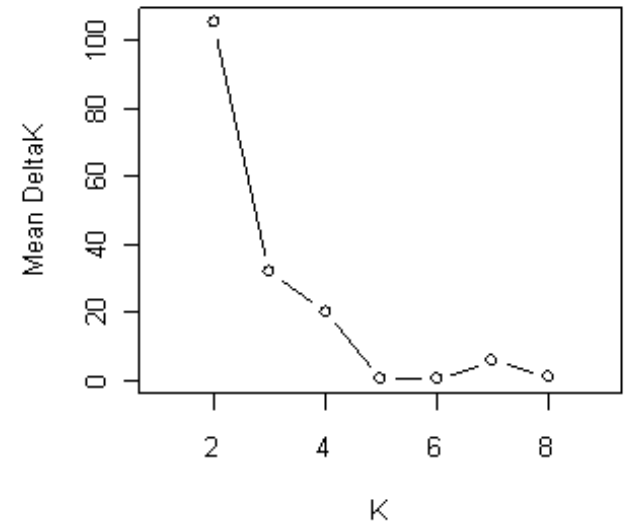
- Structure-sum (R script)
 - summary of all runs with a respect to K
 - similarity coefficient among runs for the same K
 - estimation of best K using deltaK approach (optimal number of clusters)



logarithm of the model likelihood



similarity



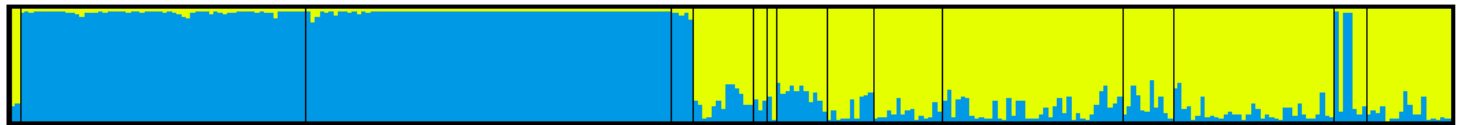
deltaK

STRUCTURE results evaluation

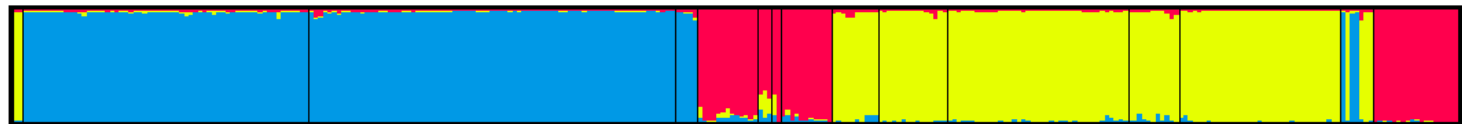
Distruct (Rosenberg 2004)

- graphical representation of sample assignment to individual clusters

K2

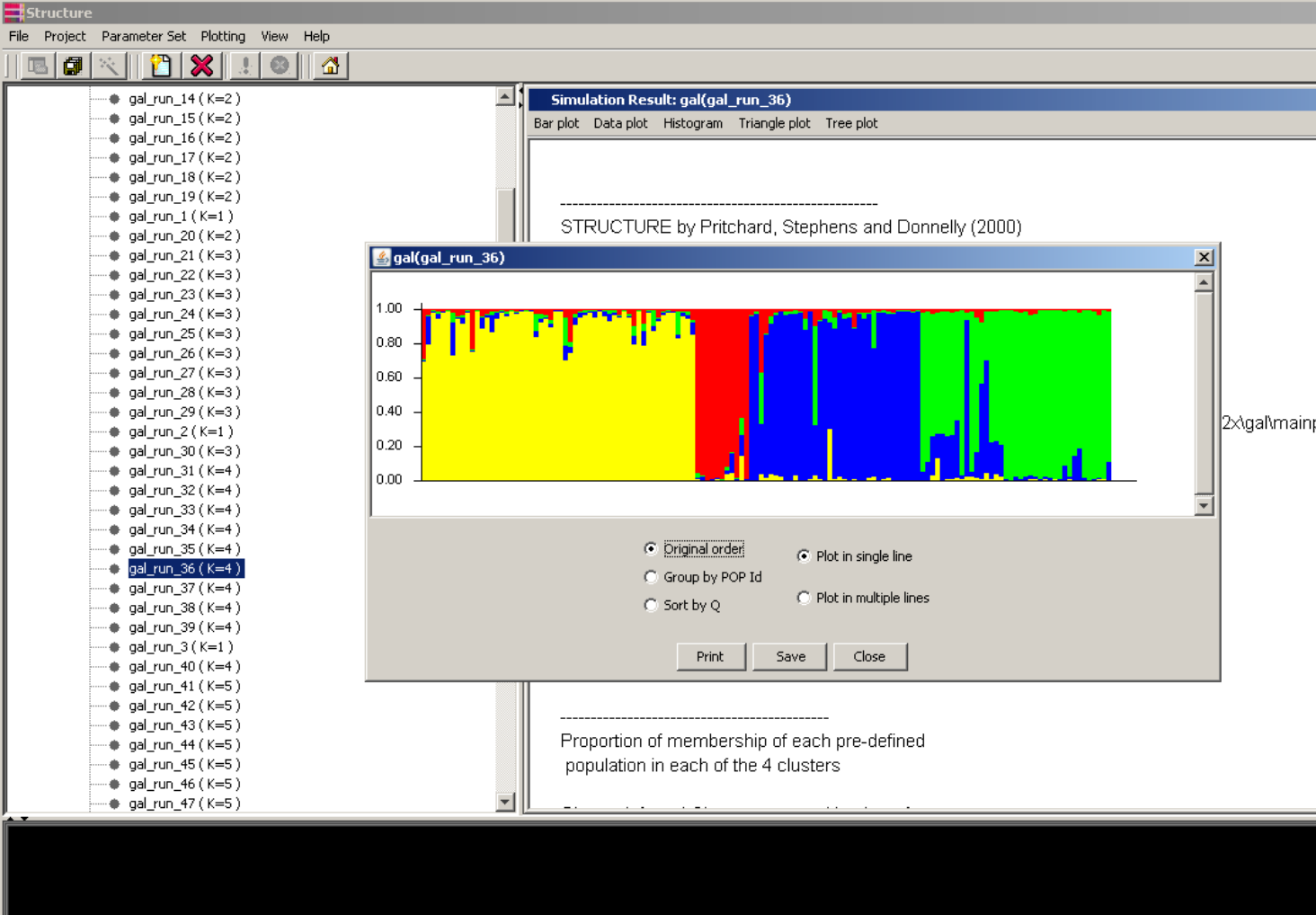


K3



STRUCTURE results evaluation

graphical interface



Practical part 2

Find optimal partitioning to K clusters with STRUCTURE

- export data from MSA to the STRUCTURE format (or use PGDSpider)
- modify the matrix that it now includes column with numbers indicating the particular species (look at Typha_US_Structure_populcodes.txt)
- run STRUCTURE with parameters 10 000 burnin/20 000 run for K=1-6 with five runs for each K
- summarize STRUCTURE results with R scripts Structure-sum
 - which K has the highest $\text{LnP}(D)$?
 - which Ks have high *similarity coefficient*?
 - which K has the highest ΔK ?
- draw colour *barplot* using Distruct for converging K