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
- **alellic** – 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 apparat for data analysis

Drawbacks

- anonymous marker
- asymetry 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

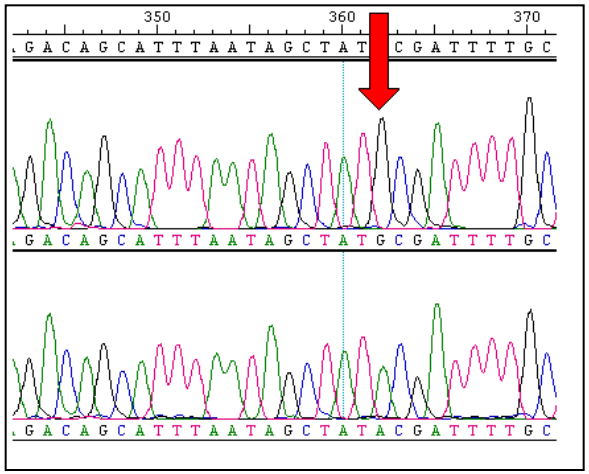
- usually high variation – many alleles
- codominant – distinguish among homozygotes and heterozygotes, allelic frequencies
- models of allele evolution – „known“ relationships among alleles
- more objective evaluation
- statistical apparat for data analysis
- possible to add further samples

Drawbacks

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

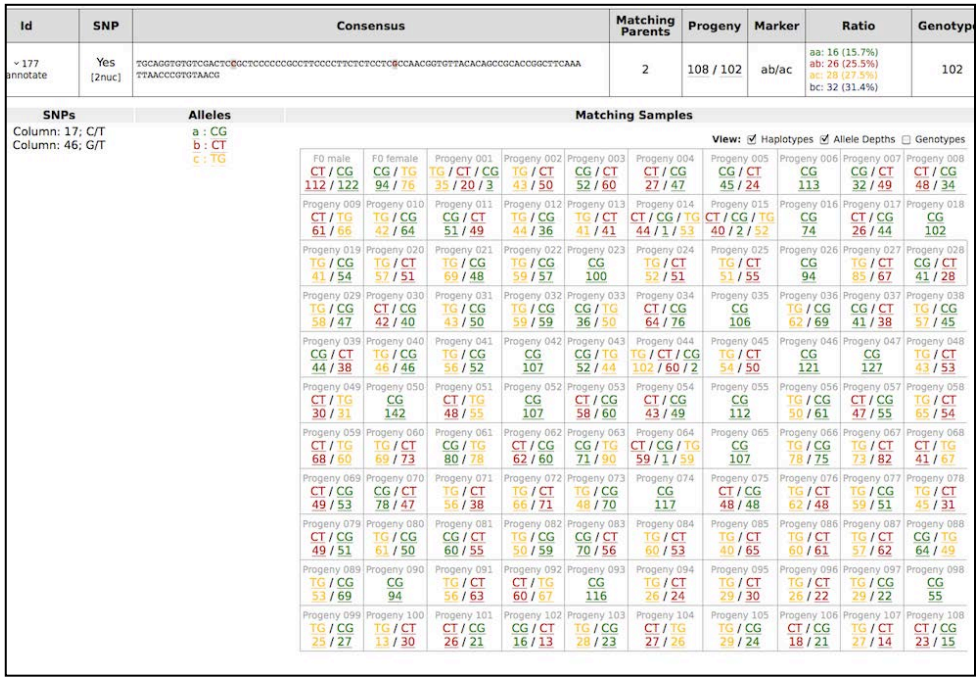
SNPs

= single nucleotide polymorphisms



Advantage

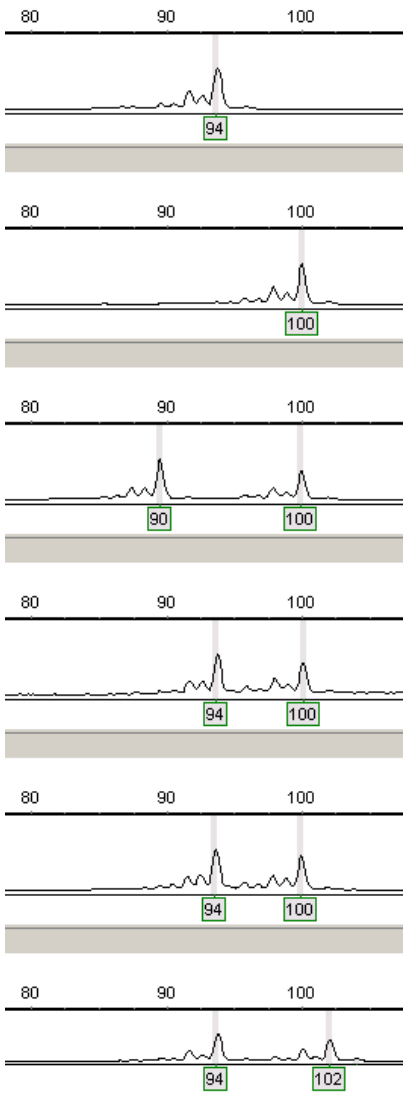
- combined advantage of AFLP and SSR
- codominant – mainly biallelic
- sequence-based (NGS, e.g., RADseq)
- non-anonymous
- multilocus
- substitution changes => evolutionary models
- up to tens of thousands of loci
- study of neutral and adaptive variability (selection)



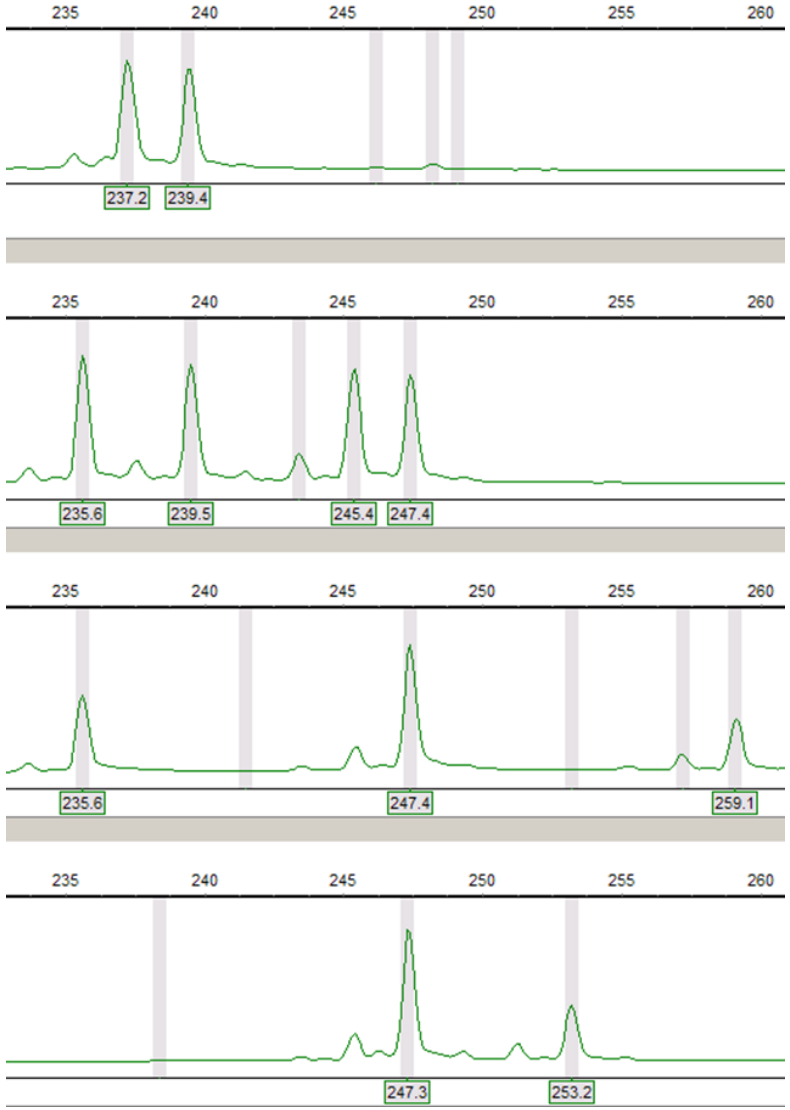
Drawbacks

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

Evaluation of codominant data



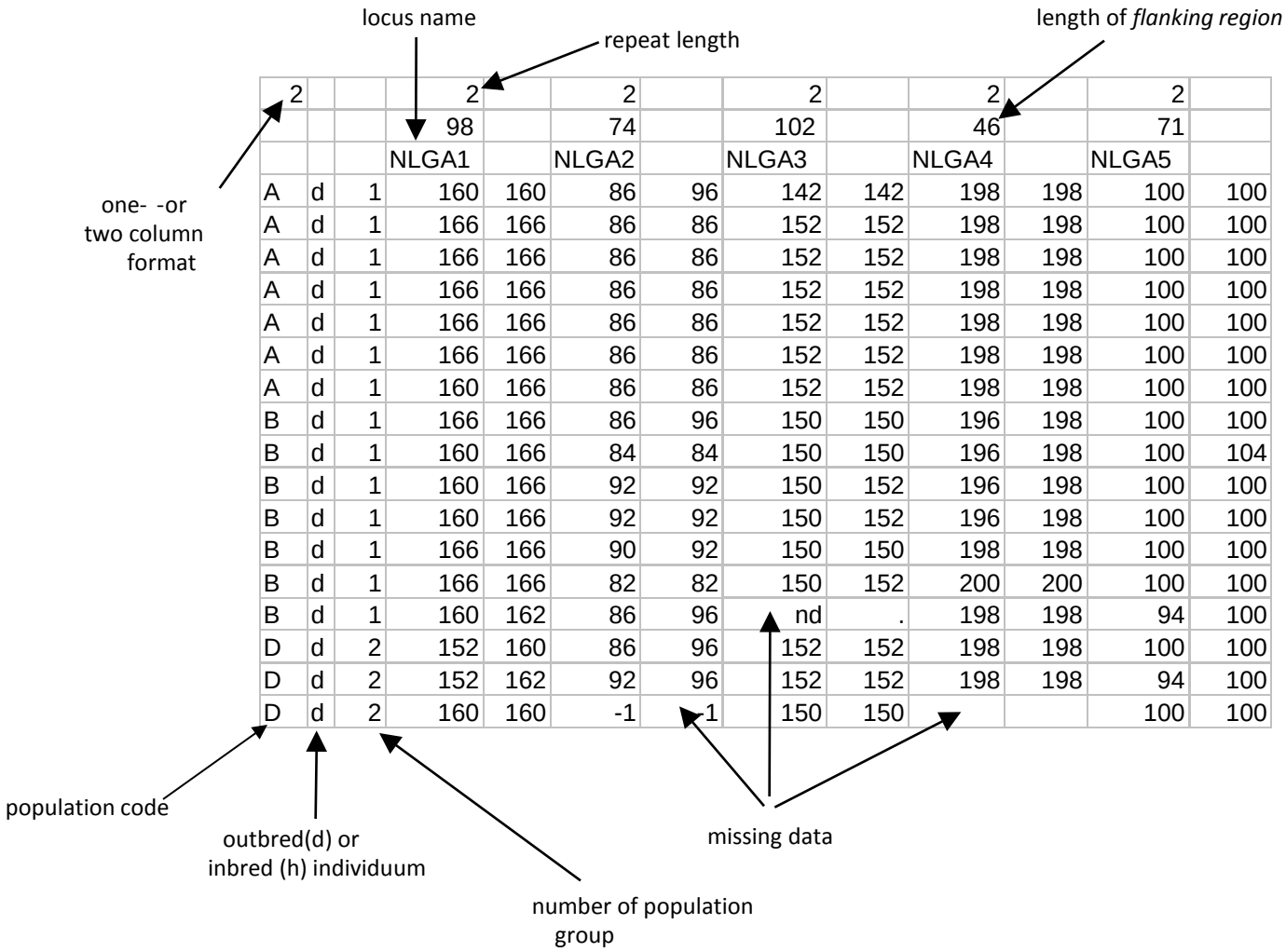
diploids vs. tetraploids



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



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

Analysis options

- relationships among individuals – basic orientation in the structure
  - distance trees (NJ, UPGMA), networks
  - multidimensional analysis (PCoA)
  - Bayesian clustering
- population-genetic parameters
  - diverzity (% 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. ×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 ekvilibria are minimized (the individuals are assign to cluster in the way to reach this goal)
  - mixture – each samples 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    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 populations vs. clusters

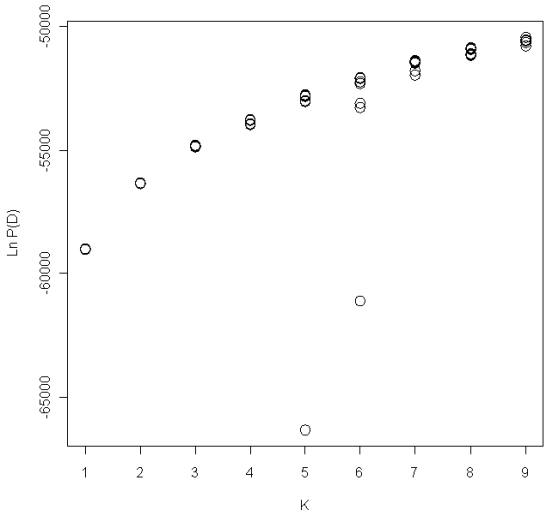
divergence among clusters

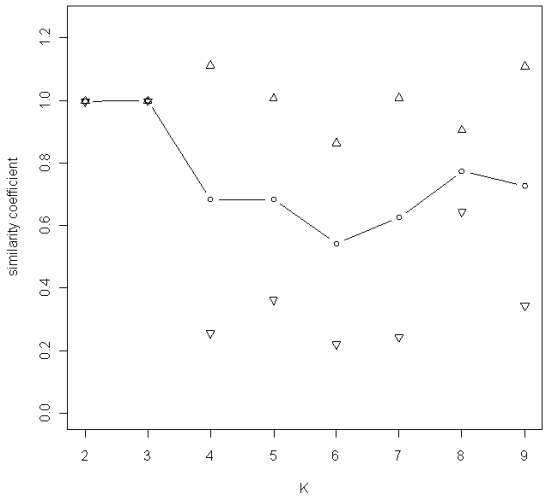
variability inside clusters

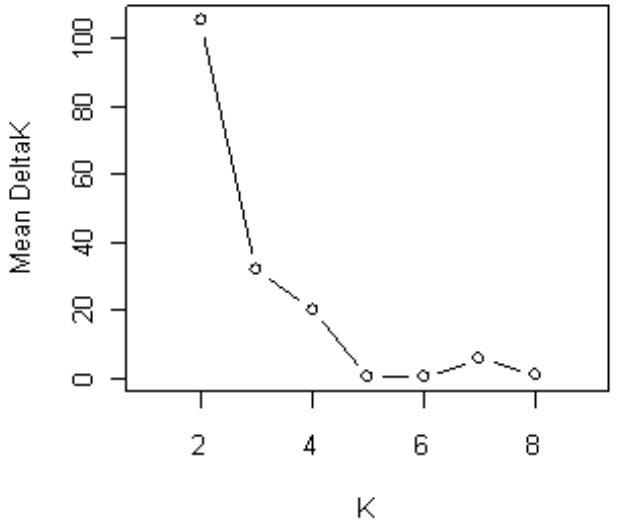
sample assignment probability to all clusters

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 approcha (optimal number of clusters)







logarithm of the model likelihood

similarity

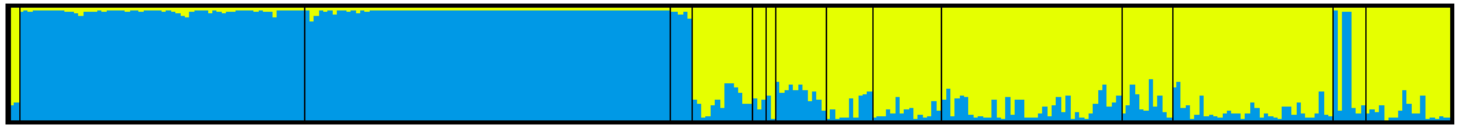
deltaK

STRUCTURE results evaluation

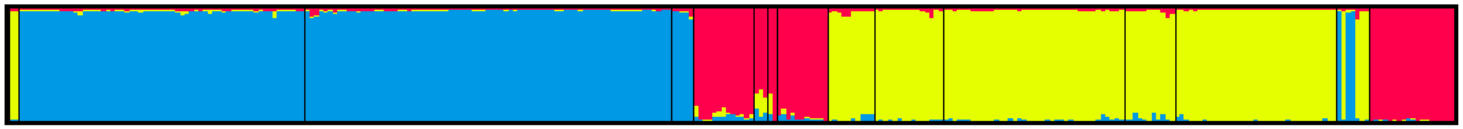
Distruct (Rosenberg 2004)

- graphical representation of sample assignment to individual clusters

K2

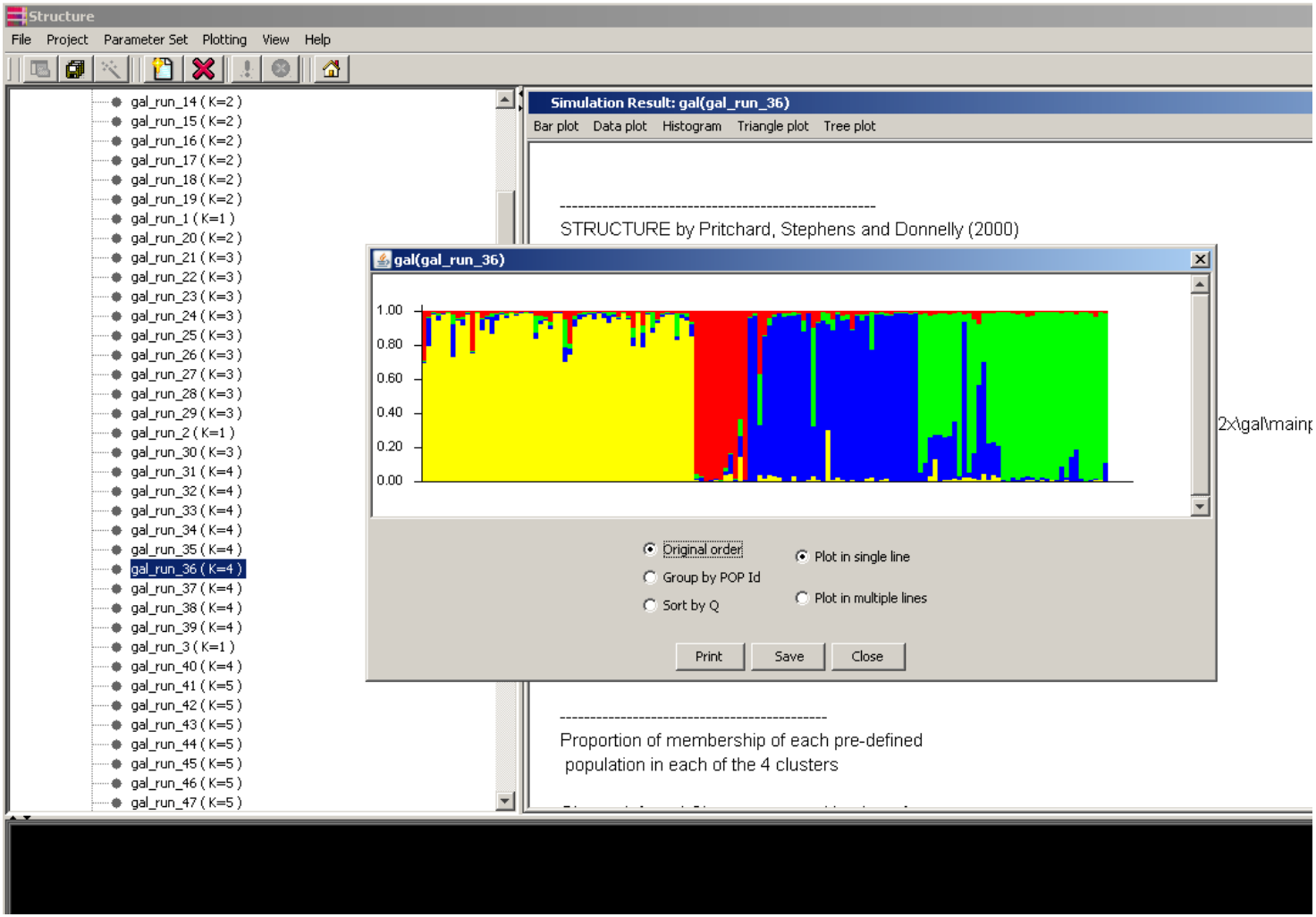


K3



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 LnP(D) ?
  - which Ks have high *similarity coefficient*?
  - which K has the highest deltaK?
- draw colour *barplot* using Distruct for converging K