Analysis of microsatellite data

Tomáš Fér
Department of Botany, Charles University, Prague
tomas.fer@centrum.cz
https://botany.natur.cuni.cz/fer/markers/practicals/SSRs.htm

What are microsatellites ?

- *simple sequence repeats* (SSRs)
- *short tandem repeats* (STRs)
- tandem repeats, shorter than 6 bp, usually 2, 3 or 4 bp

…GTTCTGTCATATATATATATAT----CGTACTT…

…GTTCTGTCATATATATATATATATATCGTACTT…

- alelles are defined by repeat number
- PCR – length polymorphism

Types of microsatellites

- *simple*
  …CACACACACACACACACACACACA…
- *compound*
  …CACACACACATGTGTGTGTGTG…
- *interrupted*
  …CACACATTCACACATTCACA…

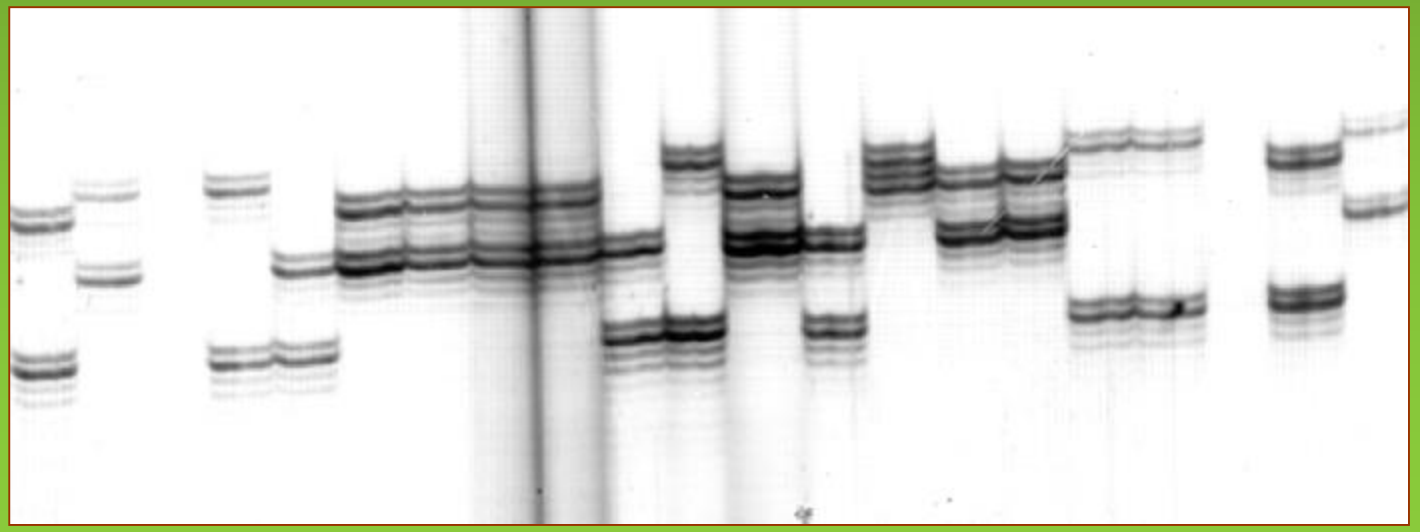
Microsatellite characteristics

- *single locus* – highly specific
- frequent occurrence throughout the genome
- high polymorphism – many alleles
- codominant

- BUT – need for primers (sequences of flanking regions)

Gel interpretation

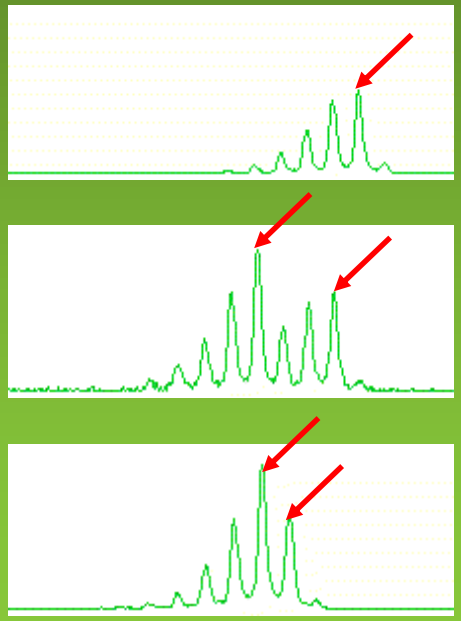
- *„stutter bands“* – additional bands around the right one (most intensive) – *in vitro DNA slippage*
- *„terminal transferase activity“* – tendency of *Taq* polymerase to add A to the 3´-end



Gel interpretation II.

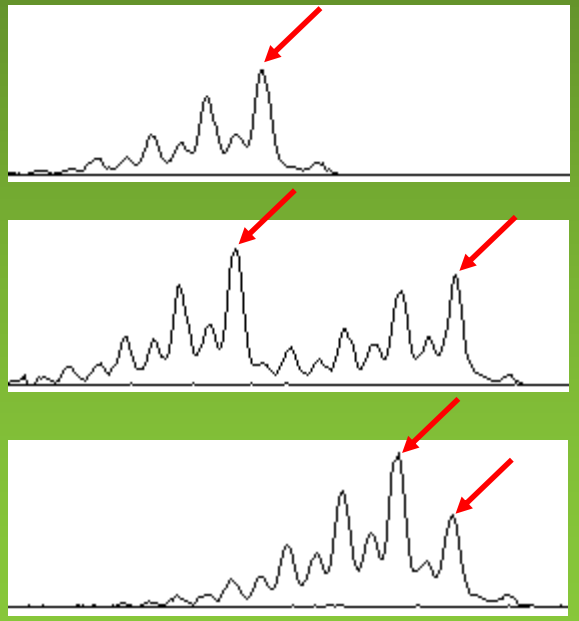
stutter bands

- products of 2, 4, 6 etc. bp shorter
- highest *peak* is the longest – real allele



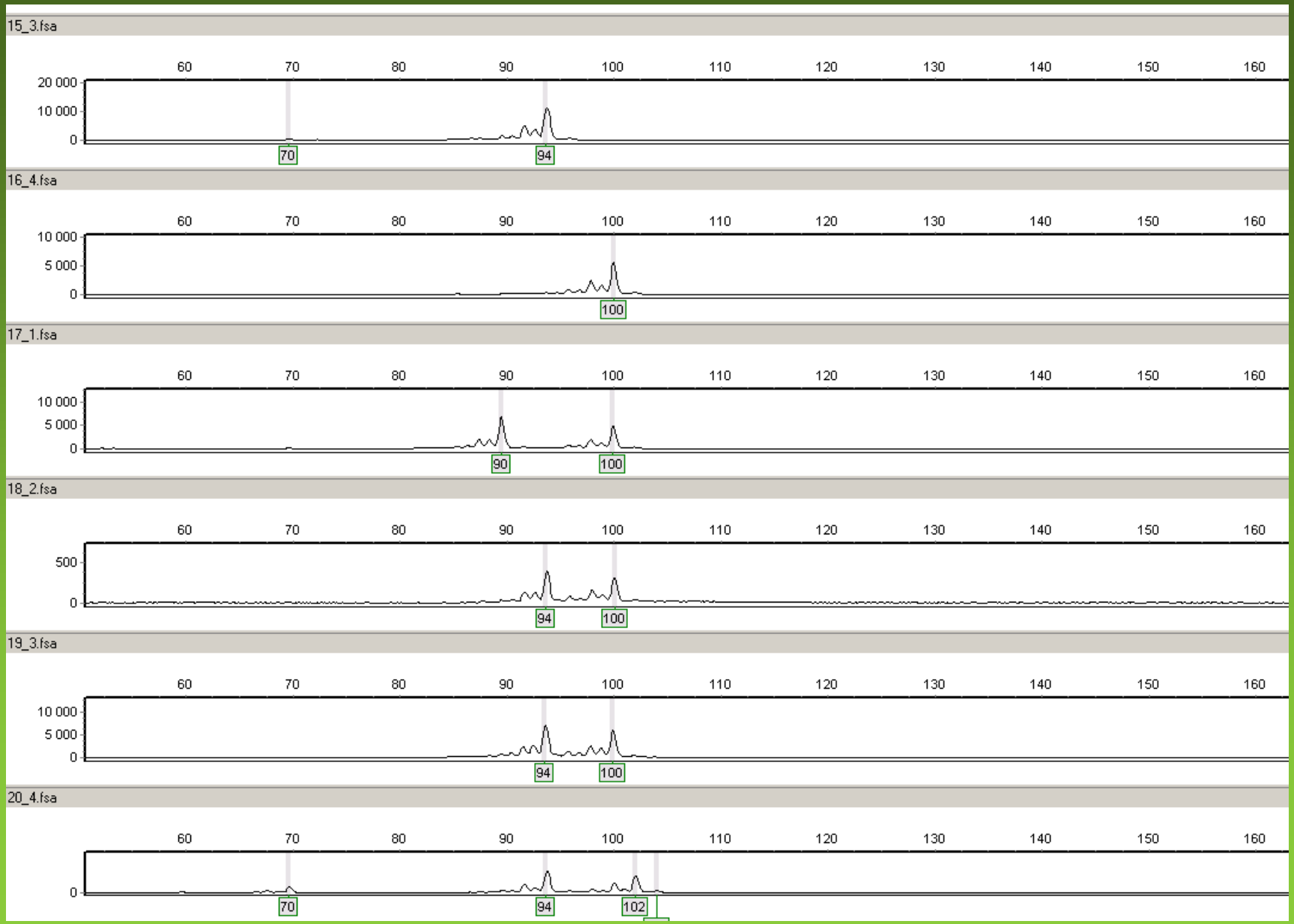
stutter bands and **-A** products

- *sttuter bands* of 2, 4, 6 etc. bp shorter
- each band has also -A product



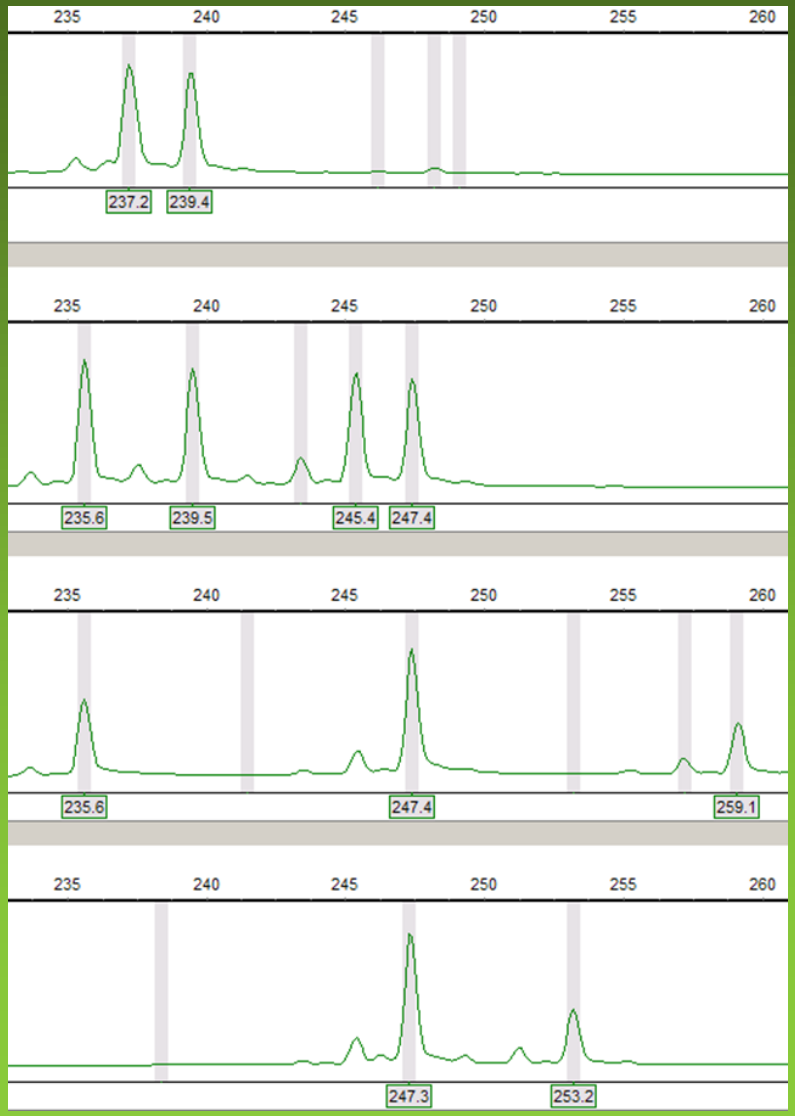
real alleles

Automated analysis (GeneMarker)

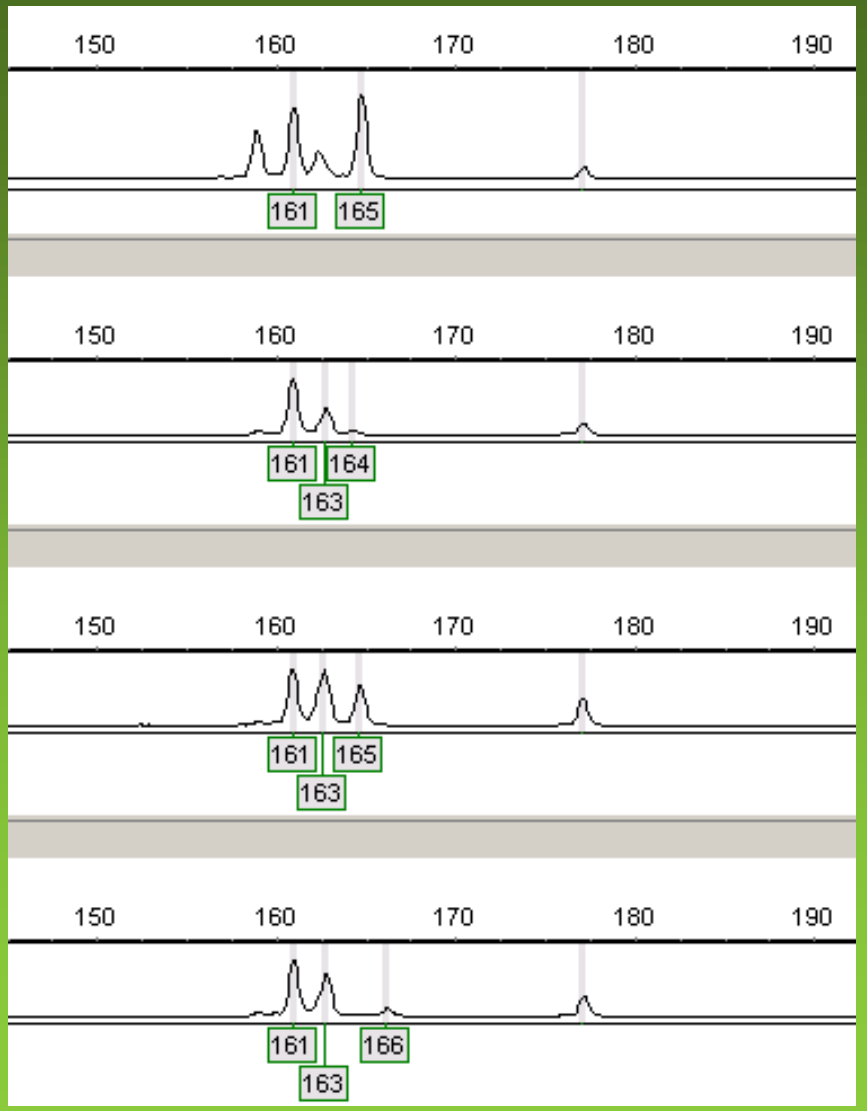


Tetraploid data

Betula



Phragmites



How to evaluate tetraploids

- treat as dominant data – presence-absence of alleles
  - allele frequency-based approaches cannot be used
- codominant – problem with *allele dosage* (we see alleles, i.e. phenotype, but what is the genotype?)
  - three alleles – one is twice, but which one? (i.e., scored as 3 allles + missing)
  - two alleles – each twice or one thrice? (i.e., scored as 2 alleles + 2 missing)
  - problem – huge amount of missing data
  - alternative – allele dosage estimated from peak height/area (MAC-PR; Esselink et al. 2004)
- autopolyploids/allopolyploids – polysomic/disomic inheritance
- null alleles?
- software for analysis of polyploid data – SPAGeDi, TETRASAT, BAPS, STRUCTURE, GenoDive, PolySat…

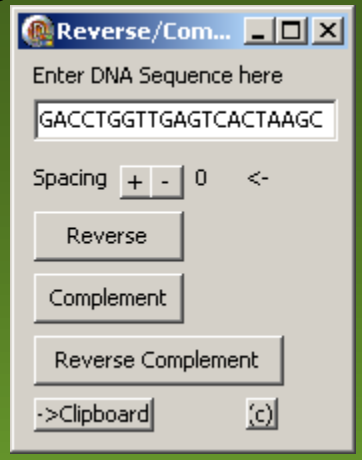
Data evaluation

- codominant marker – allelic evaluation (similar to allozymes)
  - heterozygosity (observed, expected)
  - F-statistics (F_{IS})…
  - distances (among populations, individuals)
    - proportion of shared alleles (D_{ps})
    - Nei´s chord distance (D_a)
    - Nei´s standard distance (D)
- specific microsatellite coefficients
  - R_{ST} – F_{ST} analogue (Slatkin 1995)
    - includes SMM logic (stepwise mutation model – variance in allele length)
    - estimates – ρ_{ST} (Rousset 1996)
  - distances
    - delta mu – $(dm)^2$, D_{dm} (Goldstein et al. 1995)
    - D_{sw} – stepwise weighted genetic distance
  - …
- software
  - MICROSAT (Minch 1996)
  - MSA – Microsatellite Analyser (Dieringer & Schlötterer 2003)

Data evaluation

- relationships among individuals/populations
  - trees
  - PCoA
  - Bayesian model clustering (BAPS, STRUCTURE)
- estimating and testing spatial population structure
  - AMOVA
  - *isolation by distance* (relationship between pair-wise F_{ST} and distance)
  - Mantel tests, spatial autocorrelation
- influence of mutations
  - difference between F_{ST} and R_{ST}

Length of *flanking regions*



- find sequence in GenBank (http://www.ncbi.nlm.nih.gov) according to accession number
- switch to FASTA
- copy sequence
- find forward primer sequence (e.g., in Word - CTRL+F)
- make reverse-complement of reverse primer sequence (e.g., using RC.exe – http://www.famd.me.uk/AGL/RC.zip) and find sequence
- find repetitive sequence (microsatellite motif)
- subtract length of repetition from total amplicon length (incl. primers)

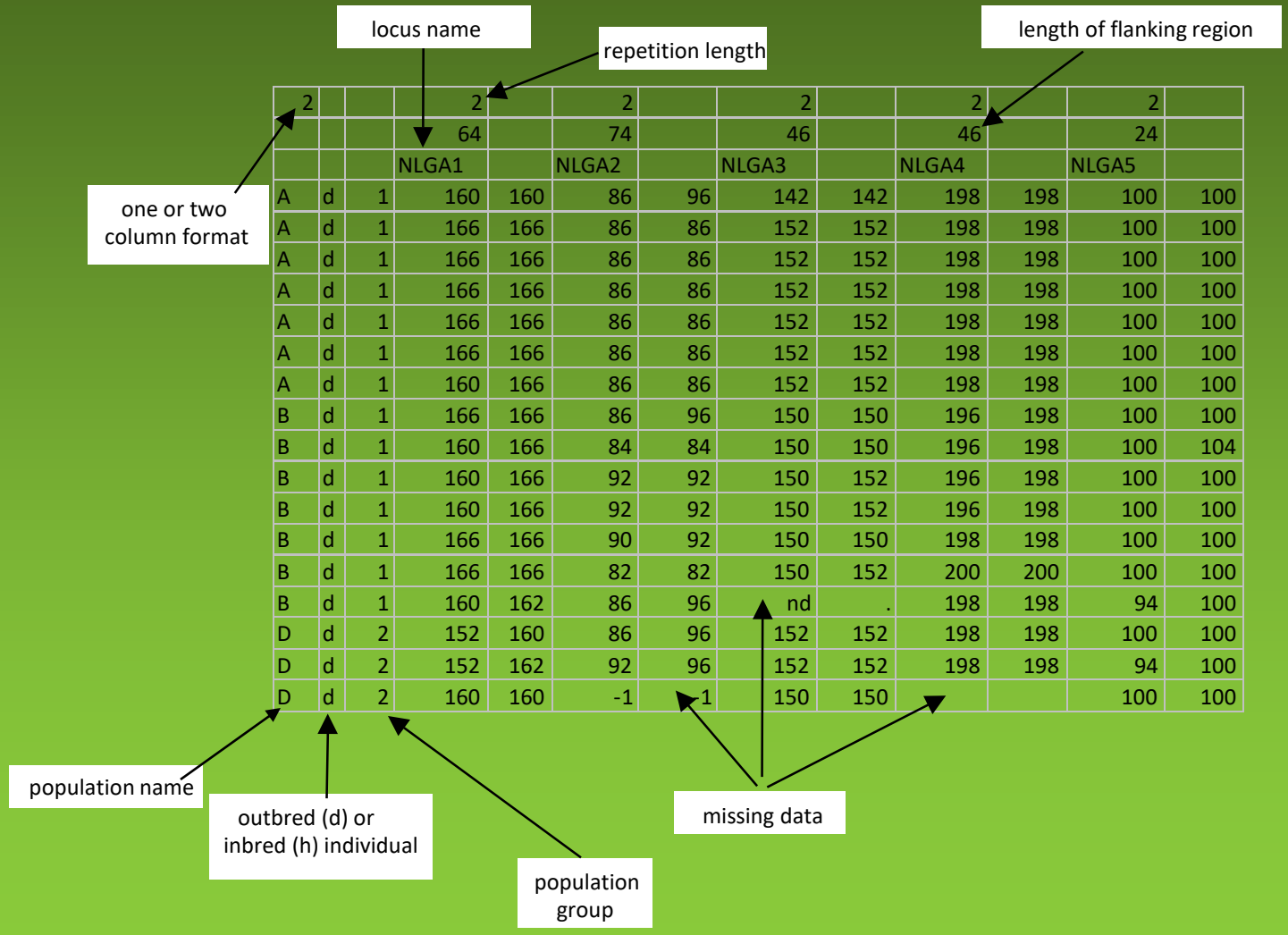
NLGA1

GATCCATTTATTACAGATTTATGGAGTATTTGTTACAATTTAATAGATAGATATTGGTAATGTGCAGATGATTCTTTTTGGCTTTCCCTTGTGTCCT
CTCTCTCTCTCTCCCCTATTTNTTGTTTCAGGTGAATAATCATGGTAGCTTCCTGATGCCATCTTTGTCCATTTTCATCCTTGGCTACAAATGCAAG
TNTTTCAGTGGACCCCACCAATTTCTCTAGGCTGAATATTAACCGATTAGCTCCAACCAATGACCCTTTGGAGTTGGACTATTCCCCTTTTGTTGAG
AGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGATGGCCCATGGTCTTGACCTGGTTGAGTCACTAAGCA
ACTTGGGGCGCCTTGTTGCTTGCGTTGGTCTACTCATCACCGCAACCAATTATTGGAGCTTGACTTGAGAATGNNCACGAGCTCCCCTCCATTTGCA
TTTGNAGACAACA

accession length	498
amplicon length	162
repetition length	64
flanking regions	**98**

MSA – Microsatellite Analyser

http://i122server.vu-wien.ac.at/MSA/MSA_download.html

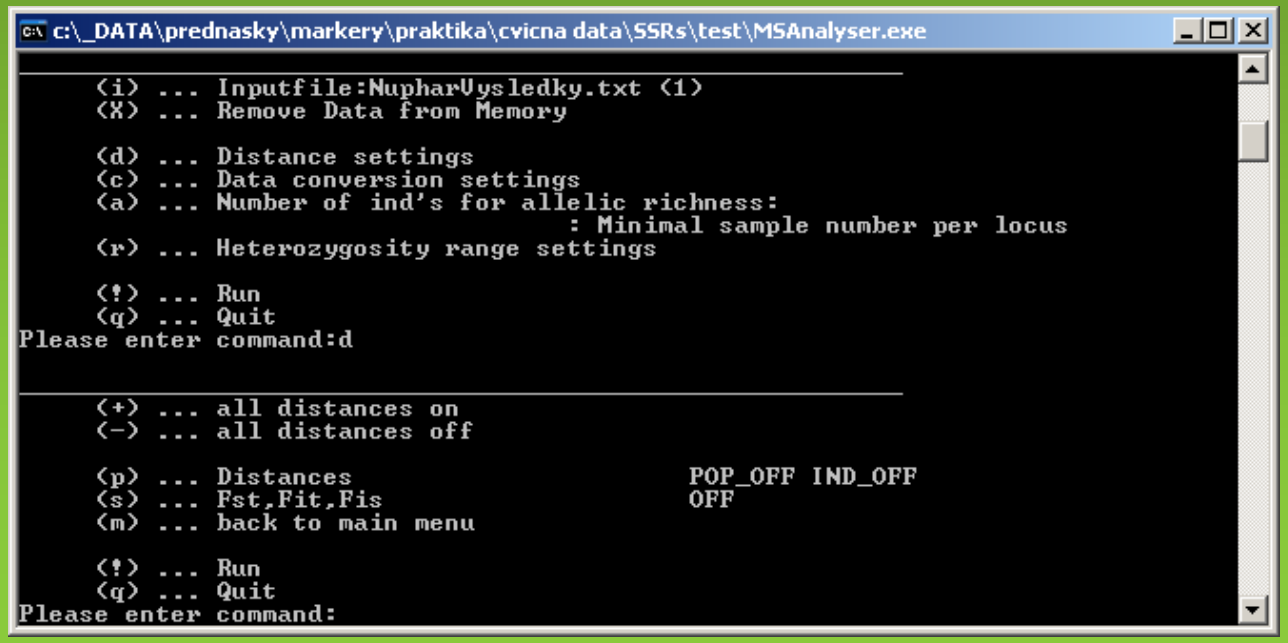


2			2		2		2		2		2		
			64		74		46		46		24		
			NLGA1		NLGA2		NLGA3		NLGA4		NLGA5		
A	d	1	160	160	86	96	142	142	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
A	d	1	166	166	86	86	152	152	198	198	100	100
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

MSA – Microsatellite Analyser

http://i122server.vu-wien.ac.at/MSA/MSA_download.html

- input file in the same folder as MSAnalyser.exe
- double click on MSAnalyser.exe
- i + ENTER – write input file name (incl. suffix!)
- d + ENTER – distance settings



MSA – Microsatellite Analyser

http://i122server.vu-wien.ac.at/MSA/MSA_download.html

- p distance settings (coefficients)
  - ci – switch on distances among both individuals and populations
  - number – switching on calculation of particular distances
  - b – back to distance menu
- s – setting *F*-statistics parameters
  - c – switch on *F*-statistic calculation
  - g –calculations global/pair-wise /both (three-switch)
  - m – back to main menu
- c (from main menu) –conversion settings
  - Arlequin, GENEPOP

MSA – Microsatellite Analyser

http://i122server.vu-wien.ac.at/MSA/MSA_download.html

output files

- Allelecount – number and frequencies of alleles for individual loci and populations
- Distance_data – text files with distance matrices among individuals/populations (trees can be constructed using software PHYLIP)
- Formats&Data – input file for Arlequin and other software
- F-Statistic – *F*-statistics global and pair-wise
- Group_data – results according to groups of populations
- Single_data – results for particular populations

Arlequin 3.5

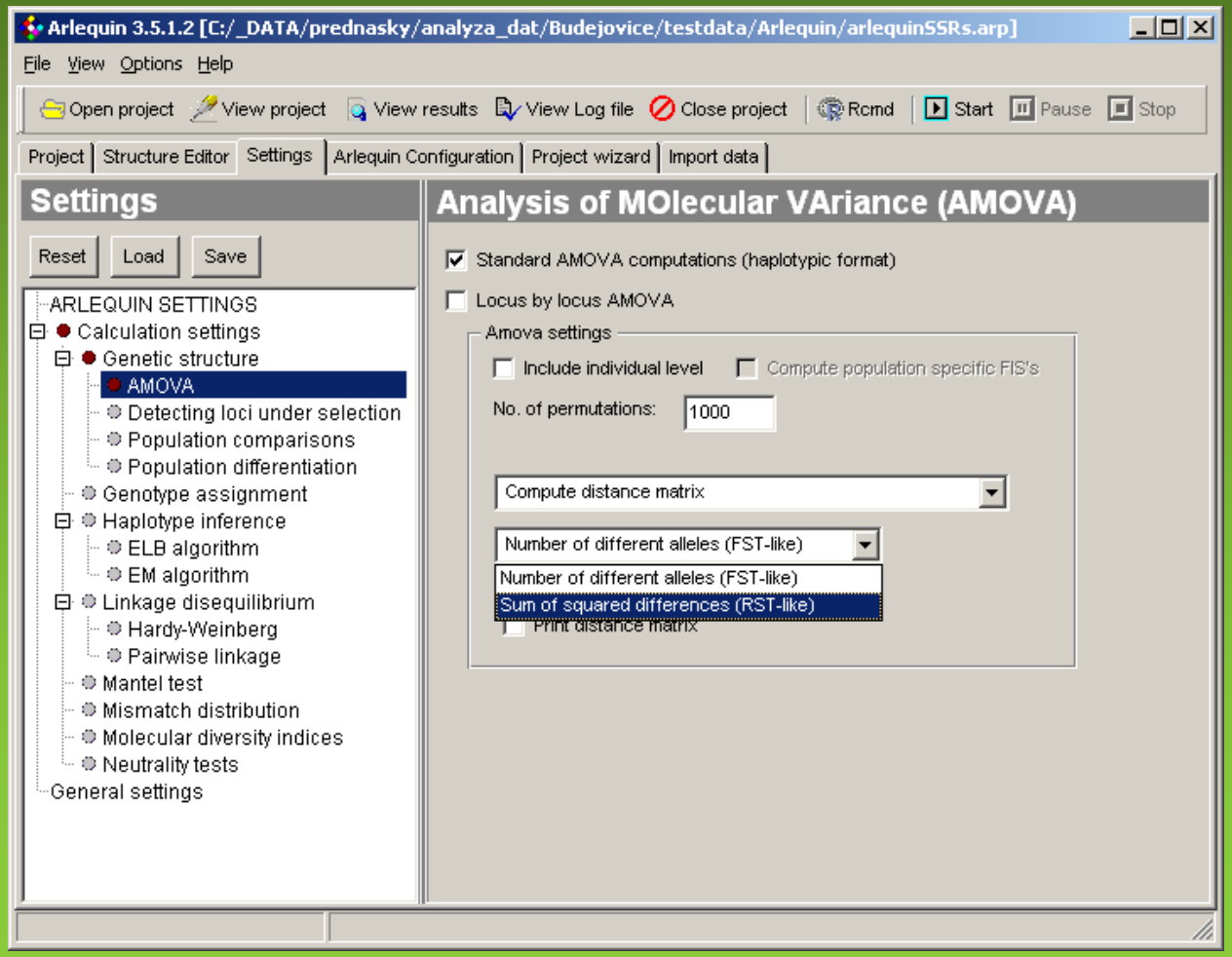
http://cmpg.unibe.ch/software/arlequin35

- *.arp file generated by MSA
- AMOVA
  - F_{ST} or R_{ST}-like analyses
- population comparisons – pair-wise F_{ST}
  - Slatkin's linearized [$F_{ST}/(1-F_{ST})$]
  - Reynold's linearized [$-\ln(1-F_{ST})$]
  - $(\delta\mu)^2$
- HW-equillibrium (exact test)
- Start
- Rcmd (R must be installed) – inserts graphs to output file

Arlequin 3.5

http://cmpg.unibe.ch/software/arlequin35

AMOVA – F_{ST} or R_{ST}-like analyses



PHYLIP

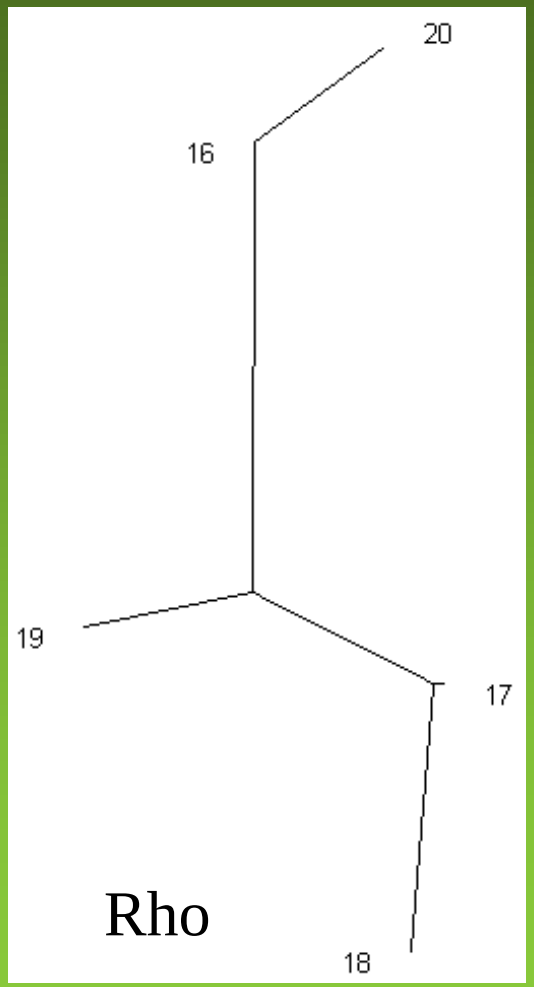
http://evolution.genetics.washington.edu/phylip.html

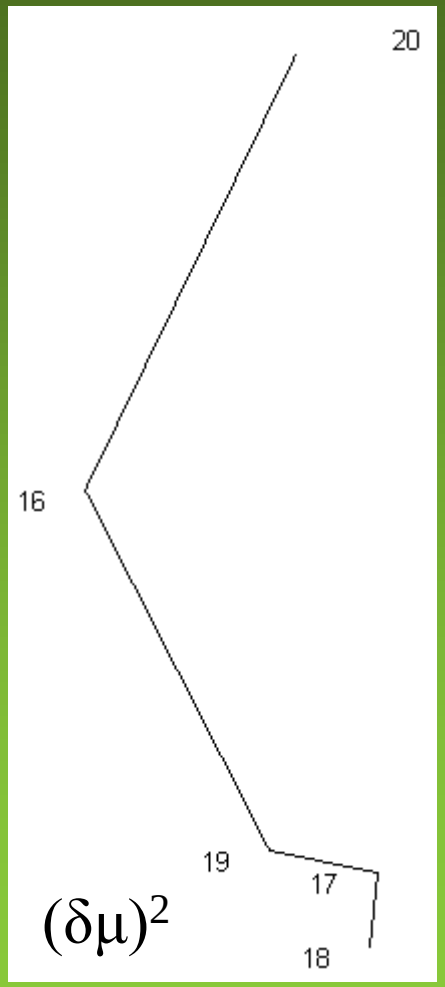
- allows to construct NJ and UPGMA trees from distance matrices generated by MSA, RSTcalc…
- neighbor.exe in the folder "exe"

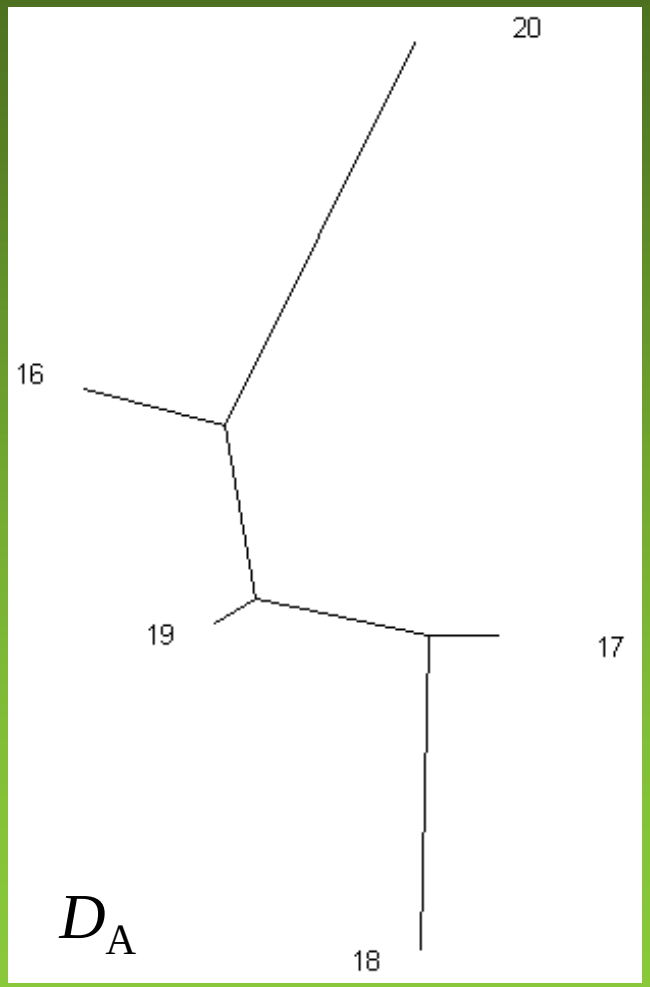
```
5
 16              0         0.3997    0.4763    0.3069    0.0606
 17              0.3997    0         0.1641    0.1792    0.5039
 18              0.4763    0.1641    0         0.4125    0.5655
 19              0.3069    0.1792    0.4125    0         0.4706
 20              0.0606    0.5039    0.5655    0.4706    0
```

PHYLIP

http://evolution.genetics.washington.edu/phylip.html

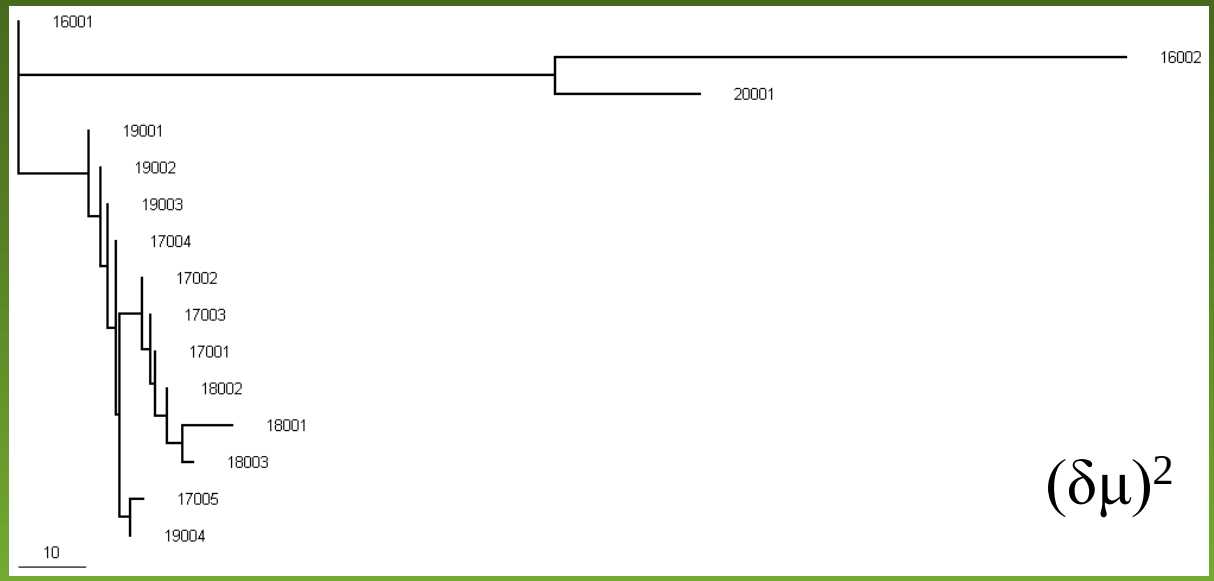


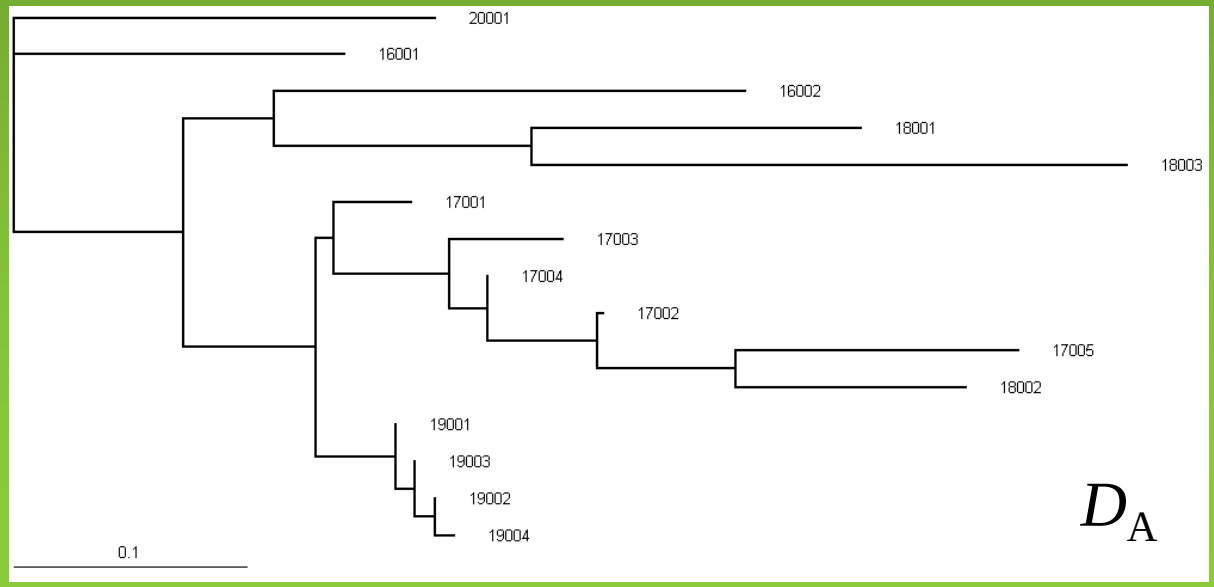




PHYLIP

http://evolution.genetics.washington.edu/phylip.html





polysat in R

https://github.com/lvclark/polysat/

- polyploidy microsatellite analysis
- handles also mix ploidy samples
- pairwise distances – SMM, IAM models
- indexes of genotype diversity
- estimates allele frequencies in autopolyploids

polysat in R

https://github.com/lvclark/polysat/

data in GenoDive format

- 1st line – name of the dataset
- 2nd line – nr. indiv, nr. pops, nr. loci, max ploidy, digits per allele
- next p lines – names of pops
- header line – pop, ind, loci names
- data lines – pop nr, ind name, genotype per every locus

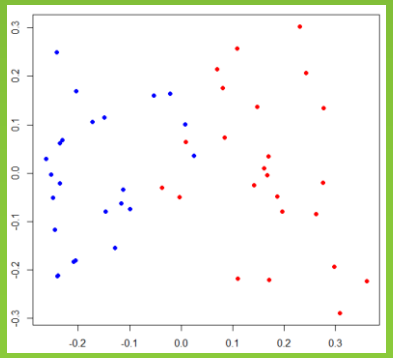
```
Betula
6       2       4       4       3
pop1
pop2
pop     ind     locL01          locL02          locL03          locL04
1       indA    237243243243    216218218218    175177187       147147149153
1       indB    243243243247    204216218218    173175175185    147147149153
1       indC    243247249271    212216218230    165173187       149151151157
2       indD    239241243       212216216230    165177185195    143147147151
2       indE    243243247259    218218218230    165165179181    147149151153
2       indF    243243243243    216216230230    165165185193    147149151151
```

polysat in R

https://github.com/lvclark/polysat/

```
#Basic data import & checking
library (polysat)
GDdata <- read.GenoDive("Betula.txt") #read data in GenoDive format
summary (GDdata) #show data summary
Samples (GDdata) #show samples
Loci (GDdata)  #show loci
viewGenotypes(GDdata, loci="locL54") #show genotypes for locus 'L54'
find.missing.gen(GDdata) #show samples with missing data
ploidy <- estimatePloidy(GDdata) #estimates ploidy level, edit the table!
Usatnts (ploidy) <- c(2, 2, 2, 2) #set repeat length for loci (i.e., all dinucl.)
summary (ploidy)
```

```
#Bruvo distance + PCA
testmat <- meandistance.matrix(ploidy) #pairwise distance matrix (Bruvo as default)
pca <- cmdscale(testmat) #PCA
plot(pca[,1], pca[,2], col=c("red", "blue")[PopInfo(ploidy)]) #make PCA plot
```



polysat in R

https://github.com/lvclark/polysat/

```
#Binary coding & PCoA
binary <- genambig.to.genbinary(ploidy) #recode microsatellite data to binary
Genotypes (binary, loci="locL54") #show binary data for locus L54
write.table(Genotypes(binary), file="Betula_binary.txt") #save table
bm <- as.matrix(Genotypes(binary)) #make binary 0-1 matrix
library (ade4)
distbin <- dist.binary(bm, method = 1) #Jaccard dist. matrix (2=SMC, 5=Sorensen)
library (ape)
res <- pcoa(distbin) #PCoA based on the distance matrix
axes <- res$vectors #save PCoA coordinates to "axes"
x <- axes[,1] #save coordinates of 1st axis to x
y <- axes[,2] # save coordinates of 2nd axis to y
plot(x,y, col=c("red", "blue")[PopInfo(ploidy)]) #make PCoA plot
```

```
#Allele numbers in populations
allelediv <- alleleDiversity(ploidy) #calculate number of alleles
allelediv$counts #show number of alleles in populations and total
```

FSTAT

http://www2.unil.ch/popgen/softwares/fstat.htm

- allele numbers and frequencies per sample and locus
- Nei's gene diversity
- F_{IS}
- Nei's *F* statistics
- Weir & Cockerham (1984) estimations
  - Capf (F_{IT}), theta (F_{ST}) and smallf (F_{IS})
- HW equilibrium testing (significance of F_{IS})
  - within populations , global

FSTAT

http://www2.unil.ch/popgen/softwares/fstat.htm

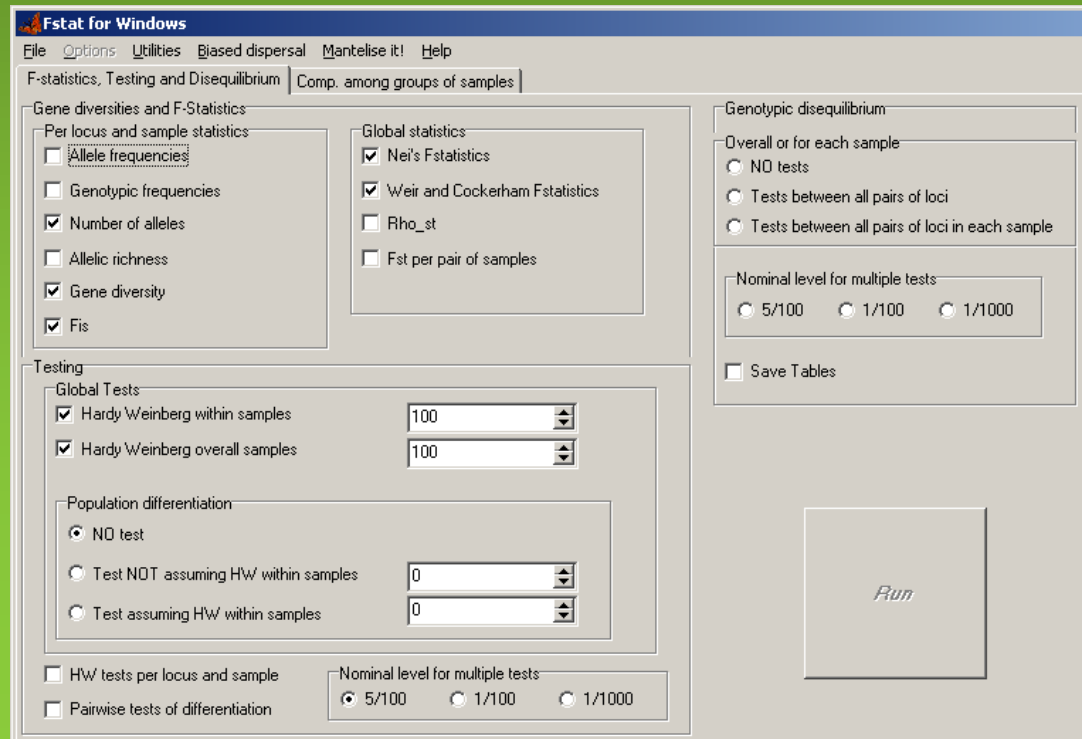
- Utilities – File Conversion – Genepop->Fstat
- open Genepop.gen generated by MSA

```
5 4 4 2 (number of samples=populations, number of loci, highest allele number, 2=allele have two characters)
NLGA1
NLGA3
NLGA5
NLGA6
   1    303    303    202    202
   1    103    202    101    202
   2    303    202    204    202
   2    304    202    203    202
   2    304    202    204    202
   2    304    202    202    202
   2    404    202    202    102
   3    203    202    303    102
   3    204    202    203    202
   3    303    101    303    202
   4    303    202    202    202
   4    303    202    202    202
   4    303    202    202    202
   4    303    202    202      0
   5    303    404    102    202
```

FSTAT

http://www2.unil.ch/popgen/softwares/fstat.htm

- File – Open
- Gene diversities and F-Statistics
  - Per locus and sample
  - Global statistics
  - Testing
- Run – results are written to the output file



RSTcalc

http://www.biology.ed.ac.uk/research/institutes/evolution/software/rst/rst.html

- Rho – estimation of R_{ST} (global and pair-wise)
- significance, bootstrap variance
- $(\delta\mu)^2$

```
Nuphar (file name)
4 (loci)
5 (size of the biggest population)
5 (number of populations)
2 (number of individuals in pop1)
5 (number of individuals in pop2 etc.)
3
4
1
NLGA1 (locus name)
2 (repetition length)
98 (length of flanking region)
NLGA3
2
102
NLGA5
2
71
NLGA6
2
240
16 (name of pop1)
60 160 154 154 94 94 266 266 (individual 1)
50 160 152 152 70 70 266 266
17
60 160 152 152 94 102 266 266
```

- double click on RST22.exe
- press “a” to set “Basepairs”
- press “p” to se number of permutations
- press “b” to set bootstrap
- press “i” to set input file
- press “r” to start calculations
- results are in output file RSTOUT.txt

RSTcalc

http://www.biology.ed.ac.uk/research/institutes/evolution/software/rst/rst.html

********** TOTAL POPULATION COMPARISON **********

RHO VALUES OVER ALL POPULATIONS:

LOCUS	MEAN ALLELE	TOTAL SAMPLE	MEAN SAMPLE	SA	SW	RHO
NLGA1	159.53334	30	5.66667	0.76261	9.19555	0.07658
NLGA3	153.06667	30	5.66667	21.27425	0.48000	0.97794
NLGA5	93.33334	30	5.66667	7.42546	99.76888	0.06927
NLGA6	265.42856	28	5.28571	-0.10878	3.41333	-0.03292

RHO (AVERGING VAR COMP)= 0.20641 : Nm= 0.96119 P= 0.21690

RHO (AVERAGED OVER LOCI)= 0.27272 : Nm= 0.66670 P= 0.01980

NUMBER OF PERMUTATIONS= 10000

RSTcalc

http://www.biology.ed.ac.uk/research/institutes/evolution/software/rst/rst.html

********** BOOTSTRAP RESULTS FOR RHO & Nm VALUES **********

RESULTS FOR RHO & Nm CALCULATED OVER ALL POPULATIONS

	OBS RHO	95% CI		MEAN	VARIANCE	STD
		L	U	RHO		ERROR
RHO (VAR COMP):	0.20641	0.1373	0.6096	0.2993	0.00184	0.0429
RHO (LOCI) :	0.27272	0.2297	0.4982	0.3414	0.00044	0.0209

NUMBER OF BOOTSTRAPS : 100

********** PAIRWISE POPULATION COMPARISIONS **********

RHO ESTIMATES FOR PAIRWISE POPULATION COMPARISONS:

POPS	LOCUS	MEAN ALLELE	TOTAL SAMPLE	MEAN SAMPLE	SA	SW	RHO
16 x 17	NLGA1	31.000	14	5.714	0.736	3.264	0.18403
16 x 17	NLGA3	25.143	14	5.714	0.083	0.167	0.33333
16 x 17	NLGA5	10.571	14	5.714	19.044	25.606	0.42653
16 x 17	NLGA6	12.714	14	5.714	-0.000	0.800	-0.00000

RHO (AVERGING VAR COMP)= 0.39968

RHO (AVERGING OVER LOCI)= 0.23597

RSTcalc

http://www.biology.ed.ac.uk/research/institutes/evolution/software/rst/rst.html

RHO VALUES AVERAGING OVER VARIANCE COMPONENTS AND LOCI, ESTIMATED Nm & (DELTA-MU)^2 DISTANCE:

POPS	RHO (VAR COMP)	Nm	P	:	RHO (LOCI)	Nm	P	(DELTA-MU)^2
16 x 17	0.39968	0.3755	0.04000	:	0.23597	0.8094	0.04000	13.47063
16 x 18	0.47626	0.2749	0.12000	:	0.19096	1.0592	0.16000	18.34896
16 x 19	0.30688	0.5647	0.21000	:	0.16667	1.2500	0.14000	9.45313
16 x 20	0.06061	3.8750	0.69000	:	0.08186	2.8041	0.60000	14.45313
17 x 18	0.16410	1.2734	0.06000	:	0.13826	1.5582	0.06000	1.23000
17 x 19	0.17921	1.1450	0.03000	:	0.16531	1.2623	0.01000	0.40500
17 x 20	0.50393	0.2461	0.01000	:	0.40055	0.3741	0.08000	28.70500
18 x 19	0.41250	0.3561	0.02000	:	0.27424	0.6616	0.03000	2.04167
18 x 20	0.56548	0.1921	0.06000	:	0.35548	0.4533	0.11000	35.87500
19 x 20	0.47059	0.2813	0.20000	:	0.25000	0.7500	0.20000	25.00000

NUMBER OF PERMUTATIONS= 100

MATRIX OF RHO VALUES (AVERAGING VARIANCE COMPONENTS):

5

16				
17	0.3997			
18	0.4763	0.1641		
19	0.3069	0.1792	0.4125	
20	0.0606	0.5039	0.5655	0.4706

Literature

Jarne P. & Lagoda P.J.L. (1996): *Microsatelites, from molecules to populations and back*. Trends in Ecology & Evolution 11(10):424-429

Goldstein D.B. & Schlötterer Ch. (1999): *Microsatellites. Evolution and Applications*. Oxford University Press

Lulkart G. & England P.R. (1999): *Statistical analysis of microsatellite DNA data*. Trends in Ecology & Evolution 14(7):253-256

Robinson J.P. & Harris S.A. (1999): *Amplified Fragment Length Polymorphisms and Microsatellites: A phylogenetic perspective*. In: Gillet E.M.[ed.]: Which DNA Marker for Which Purpose? http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm

Provan J. et al. (2001): *Chloroplast microsatellites: new tools for studies in plant ecology and evolution*. Trends in Ecology & Evolution 16(3):142-147

Balloux F. & Lugon-Moulin N. (2002): *The estimation of population differentiation with microsatellite markers*. Molecular Ecology 11:155-165

Jones A.G. & Ardren W.R. (2003): *Methods of parentage analysis in natural populations*. Molecular Ecology 12:2511-2523

Selkoe K.A. & Toonen R.J. (2006): *Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers*. Ecology Letters 9: 615-629.

Guichoux E. et al. (2011): *Current trends in microsatellite genotyping*. Molecular Ecology Resources 11: 591-611.

Dufresne F. et al. (2014): *Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools*. Molecular Ecology 23:40-69.