

Protocols and manuals for the practical course

**Molecular markers in systematics and population
biology of plants**
(B120C45)

Part II. DNA sequencing, AFLP and microsatellites

**in DNA laboratory of the Department of Botany, Faculty of Science, Charles University,
Benátská 2, Prague, Czech Republic**

A. PROTOCOLS

Protocol D: DNA sequencing

By sequencing we detect the order of DNA bases (A, C, T, G) in a particular part of DNA. The procedure can be divided into several parts:

1. *PCR amplification* of particular DNA region using a pair of specific primers. Primers anneal to a single concrete site in nuclear or chloroplast DNA. In nuclear DNA internal transcribed spacer (ITS) is the easiest and mostly used region to sequence because there are up to ten thousands of (presumably) identical copies in the genome. Chloroplast regions are usually easily amplified as well because there are many circular cpDNA molecules in the extracted DNA. We usually sequence non-coding regions (introns or spacers), e.g., *trnL-trnF*, *psbA-trnH* etc. (see Table 1)
2. *PCR product purification* – it is usually done with commercial kits (e.g., GenElute PCR Clean-Up Kit (Sigma) or PCR Extraction Kit (Geneaid)). During purification reagents of primers and non-incorporated dNTPs are removed. This is an important step because during the subsequent sequencing reaction only one primer is used and also dNTP concentration must be balanced towards other compounds, especially ddNTP (see the next step). It is necessary to achieve the optimal primer : template (PCR product) ratio. Hence, it is good to measure PCR product concentration (e.g., photometrically using Nanodrop).
3. *Cycle sequencing* – it is a modified PCR reaction where we used only one sequencing primer and PCR product as a template. There is only linear template amplification not exponential. Apart from classical dNTPs we also add ddNTPs (2', 3'-dideoxynukleotidtriphosphates). These are fluorescence-labelled (each base with a different colour) and they prevent the newly arising DNA chain from further elongation when incorporated. The result of sequencing reaction is a mixture of DNA chains of different lengths and each is fluorescence-labelled according to the last base. These products are later electrophoretically separated using an automated sequencer and it is possible to read the sequence. Sequencing is always done with one primer only (which allows to read ca. 450-850 bases; it depends on PCR product quality and the sequence of bases). We have to do two sequencing reactions with two different primers (forward and reverse) if we want to sequence region longer than ca. 500 bases. PCR product will be sequenced from both sides and the sequences will be merged into single contiguous sequence (contig) using a specific software.

4. *Precipitation of sequencing reaction products* – before we proceed to sequencing on automated sequencer the product of cycle sequencing reaction has to be purified and precipitated with NaOAc (sodium acetate) and ethanol.
5. *sequencing analysis on an automated sequencer* – the products of cycle sequencing reaction are electrophoretically separated according to their lengths using a capillary sequencer ABI 3130xl Avant (in the sequencing facility of the Faculty of Science, Charles University).

D1. Protocol for PCR amplification

Prepare PCR mix for desired number of samples to 1.5 ml tube. It is possible to change the amount of polymerase and template DNA. If this is the case, adjust the volume of water to ensure that the total volume is 20 μ l.

- sterile Milli-Q water..... 15.60 μ l
- 10 \times PCR buffer (Sigma).....2.00 μ l
- dNTP (10 mM).....0.40 μ l
- *forward* primer (25 pmol/ μ l)0.25 μ l
- *reverse* primer (25 pmol/ μ l)0.25 μ l
- (JumpStart) RedTaq DNA Polymerase (Sigma) (1U/ μ l)0.50 μ l
- + template DNA (5 ng/ μ l) 1.00 μ l

Work on ice with all chemicals. Take the polymerase out of the freezer directly before pipetting and return it back immediately. Use primers for amplification of target region according to the Table 1. Distribute 19 μ l of the mixture to individual PCR tubes and add 1 μ l of template DNA.

Put PCR tubes to the thermocycler with the following program:

| | | |
|----------|-----------|------------------|
| 94 °C | 1:00 | } 25-35 \times |
| 94 °C | 0:45 | |
| 52-62 °C | 0:45 | |
| 72 °C | 1:00-2:00 | |
| 72 °C | 10:00 | |
| 10 °C | hold | |

Change the annealing temperature and elongation time according to the Table 1. It is good to perform a gradient test for optimum temperature before we amplify more samples.

Test the PCR product by loading ca. 5 μ l to the 1 \times TAE agarose gel (1%). Load 1.5 μ l of ladder (100 bp DNA Ladder Plus) to the first well. We should see one clear band on the gel

with the expected size according to the Table. We could estimate its size with KODAK 1D Image Analysis Software by comparing with the ladder.

If we see single and clear band of expected size our PCR was successful and we can continue with product purification (protocol D2). In other case we have to optimise the PCR reaction (increase annealing temperature, increase MgCl_2 concentration, add additives... see general comments to PCR).

D2. Protocol for PCR product purification with GenElute PCR Clean-Up Kit (Sigma)

The PCR product will be purified using GenElute PCR Clean-Up Kit (Sigma), which uses „*spin column technique*“, i.e., columns with membrane that binds the DNA. Alternatively we can use PCR Extraction Kit (Geneaid) or purified the fragment excised from the agarose gel using MinElute Gel Extraction Kit (Qiagen).

- add 5× volume amount of Binding solution to the remaining 15 µl of the PCR product and vortex thoroughly
- put column to labelled (!) 2 ml tube, pipette the mixture from the previous step into it
- centrifuge 1 min with maximum rotation
- discard everything what passed through the column (DNA remained on the membrane)
- put the column back to 2 ml tube and add 500 µl of Wash Solution (with added ethanol)
- centrifuge 1 min with maximum rotation
- discard everything what passed through the column (DNA on the membrane is washed)
- put the column back to 2 ml tube and centrifuge 2 mins with maximum rotation (to remove rest of the ethanol)
- put the column to a new (and labelled!) 1.5 ml tube and pipette 20 µl of 1× TE (pre-warmed to 65°C, e.g., in a tube on thermoblock) directly onto the centre of the membrane
- let it stand for 15-30 mins
- centrifuge 1 min with maximum rotation

See the kit manual for thorough procedure and other remarks to the PCR fragment purification.

Measure DNA concentration using Nanodrop.

D3. Cycle sequencing

Sequencing reaction is done using BigDye Terminator v3.1 Cycle Sequencing Kit from the ABI company. Pipette sequencing reaction according to the table (see comments under the table):

| | whole | ½ | ¼ |
|----------------------|----------|----------|----------|
| Ready Reaction Mix | 8 µl | 4 µl | 2 µl |
| primer | 3.2 pmol | 3.2 pmol | 3.2 pmol |
| template DNA | | | |
| Sequencing Buffer 5× | 0 | 2 µl | 4 µl |
| sterile water | | | |
| total volume | 20 µl | 20 µl | 20 µl |

- the whole volume of the reaction is 20 µl: 7 µl is Ready Reaction Mix, Buffer and a primer; water and PCR product is remaining 13 µl
- in case of ½ a ¼ reaction we save sequencing premix and its part is substituted by a sequencing buffer; use this in the case of non-problematic template
- provided primer is already diluted to 3.2 pmol/µl, i. e., add 1 µl
- use the amount of template DNA according to the following table (we usually sequence fragments around 1,000 bp and use 15 ng of DNA); the amount of PCR product in µl depends on its concentration:

| template (PCR product) | amount |
|------------------------|----------|
| 100-200 bp | 1-3 ng |
| 200-500 bp | 3-10 ng |
| 500-1000 bp | 5-30 ng |
| 1000-2000 bp | 10-40 ng |

- calculate how much to add (in µl) for each PCR product according to its concentration (we need, e.g., 15 ng of DNA in the reaction); adjust the amount of water accordingly
- prepare a mixture from Ready Reaction Mix, Buffer and primer for desired number of samples (work on ice, mixture contains polymerase), vortex briefly, centrifuge shortly

- distribute the mixture to PCR tubes or stripes and add the calculated amount of water and PCR product to each of the tube
- vortex briefly, spin down in minifuge and put to the cycler with the following program:

| | | |
|-------|------|--------|
| 96 °C | 0:10 | } 25 × |
| 50 °C | 0:05 | |
| 60 °C | 4:00 | |
| 10 °C | hold | |

D4. Product precipitation

- pipette cycle sequencing product to labelled 0.5 ml tubes
- add 2 µl 3M NaOAc (sodium acetate) and 50 µl 96% ethanol (from the freezer) to 20 µl of the product
- vortex and let it stand for 10-15 mins.
- centrifuge in refrigerated centrifuge for 30 mins at 13,200 rpm and 22 °C
- carefully (!) remove the supernatant by pipetting (precipitate is invisible), pipette on the other side of the tube than is the pellet
- add 250 µl 70% ethanol (from the freezer), vortex and centrifuge in refrigerated centrifuge for 15 mins at 13,200 rpm and 22 °C
- repeat the two previous steps
- remove the supernatant and let stand the opened tubes 5 mins laminar box
- dry out at 65 °C (on thermoblock) for 5-10 mins, remove the rest of the ethanol from the upper part of the tube with paper tissue
- dried product is ready for sequencing in DNA sequencing facility; product in this form should be OK for 1-2 weeks at -20 °C

Protocol E: AFLP (Amplified Fragment Length Polymorphism)

AFLP method is used to assess the polymorphism in total DNA. It is based on restriction of DNA with two restriction endonucleases and subsequent selective amplification of only subset of the fragments. The method can be divided to the following steps:

1. *Restriction* of the total DNA – with a pair of restriction endonucleases (typically *EcoRI* and *MseI*) is total DNA digested to many fragments which now have *EcoRI* or *MseI* restriction site at their ends. Restriction is specific and reproducible under standard conditions (temperature, pH).
2. *Ligation* – using the enzyme T4 ligase are adaptors (i.e., short double-stranded DNA fragments with known sequence and with one end compatible with the fragments generated by restriction endonucleases) ligated to the ends of all fragments. The results of this step are fragments with known sequences at both ends.
3. *Preselective amplification* (preamplification) – using two primers with a sequence complementary to adaptor sequence +1 selective base only 1/16 of restriction fragments are PCR amplified. Only fragments with one *EcoRI* and one *MseI* adaptor are amplified.
4. *Selective amplification* – PCR reaction with selective primers (i.e., primers with two to four selective bases) further reduces the number of fragments. After this step the number of fragments should be optimal for evaluation. We often use primers with three selective bases (Table 2). The selective amplification is usually done three or four times for each sample; always with a different primer combination and with a different fluorescence labelling. Automated sequencer could recognize four or five fluorescence colours at once (the fourth (fifth) colour is a standard added to each sample, see next step).
5. *Selective PCR product pooling and precipitation with NaOAc* – each sample is made by pooling of 1 µl of each 3-4 selective PCR products with differently labelled primer. Usually the ratio has to be optimized for each species and primer pair. Both, the used fluorescence dye and the selected combination of *EcoRI* and *MseI* primers in combination with concrete DNA influence the intensity. Precipitation of the product in ethanol with salts (sodium acetate) is used to clean the PCR products. Finally, fluorescence-labelled ladder (e.g., GeneScan-ROX-500) is added as an internal standard to each of the samples. It allows identification of each AFLP fragment and unambiguous comparison with fragments in any other sample.

6. *Fragment analysis on automated sequencer* – the AFLP products are electrophoretically separated in a capillary sequencer ABI 3130xl (in the sequencing facility, Faculty of Science, Charles University).

E1. DNA restriction

AFLP® Core Reagent Kit I from the company Invitrogen (Life Technologies) is used for restriction and ligation. See the kit booklet for exact composition of particular parts of the kit. Pipette the restriction premix for the desired number of samples to 1.5 ml tube according to the recipe. Work on ice, remove the mixture of restriction enzymes from the freezer directly before pipetting and put it back to -20 °C immediately.

- 5× Reaction Buffer..... 1.0 µl
- *EcoRI/MseI* 0.4 µl
- sterile H₂O 1.1 µl

Vortex the mixture briefly, centrifuge shortly and distribute 2.5 µl to 0.2 ml PCR tubes. Add 2.5 µl DNA (20-50 ng/µl) and incubate for 2-6 hours (need optimisation for each species) at 37 °C in thermocycler.

E2. Adaptor ligation

Prepare the ligation mix for desired number of samples in 1.5 ml tube (also using AFLP® Core Reagent Kit I). Work on ice, both ligation solution and T4 ligase are susceptible to higher temperatures.

- Adapter/Ligation Solution..... 4.8 µl
- T4 DNA Ligase..... 0.2 µl

Add 5 µl to the samples after restriction (work very carefully, pipette with white tip tightly under the liquid surface; steadily control the correct pipetting and release of the volume; hold the PCR tube in hand when pipetting!). Incubate DNA with ligation mix at 37 °C for 2-12 hours or overnight (best in thermocycler). Then freeze the samples in -20 °C if we do not directly continue with next step.

Load the restriction/ligation product to 1.8% agarose gel in 1× TBE buffer at 60V to test if restriction was successful. Load 3-5 µl of the product (after mixing with 1 µl of 6×loading dye

on parafilm). A smear from digested DNA, especially between 50-1000 bp, should be visible on the gel.

E3. Preselective amplification (preamplification)

We use AFLP® Pre-amp Primer Mix I from the Invitrogen company (Life Technologies) for preamplification. Prepare a mixture following the recipe to 1.5 ml tube. Again work on ice.

- PA mix.....4.0 µl
- 10× Buffer for JumpStart RedTaq0.5 µl
- JumpStart RedTaq Polymerase (Sigma)0.1 µl

Label new PCR tubes and distribute 4.6 µl of the preamplification mixture to them. Add 0.5 µl of DNA after restriction and ligation from the previous step (work precisely with 1 µl pipette, visually control correct pipetting and release of small liquid amounts; hold PCR tubes in hand when pipetting!). Close the tubes carefully and put to the thermocycler with the following program:

| | | |
|-------|-------|--------|
| 72 °C | 2:00 | } 20 × |
| 94 °C | 0:30 | |
| 56 °C | 0:30 | |
| 72 °C | 2:00 | |
| 60 °C | 30:00 | |
| 10 °C | hold | |

Test the correct preamplification on 1.8% agarose gel in 1× TBE buffer at 60V. Load 2 µl of undiluted (!) preamplification product mixed with 1 µl 6× loading dye. Again a smear with several more intense fragments should be visible.

Dilute 10× the preamplification reaction with sterile H₂O (dilute less if bands on control gel are weaker), vortex briefly and centrifuge shortly. Freeze in -20 °C if we do not directly continue with next step.

E4. Selective amplification

Selective amplification is done three or four times for each sample, always with a different combination and different fluorescently labelled primer. Automated sequencer can recognize either four (dye set D-30 with 6-FAM, HEX, NED dyes and ROX as a standard) or five (dye set D-33 with 6-FAM, VIC, NED, PET dyes and LIZ as a standard) colours. It is only possible to combine dyes from the same set! Prepare three (or four) PCR mixtures to three (or four) different 1.5 ml tubes that will only differ by primers:

- sterile H₂O 5.1 µl
- 10× Buffer for JumpStart RedTaq 1.0 µl
- dNTP 0.2 µl
- *Eco*RI primer (1 pmol/µl) 0.5 µl
- *Mse*I primer (5 pmol/µl) 0.5 µl
- JumpStart RedTaq DNA Polymerase (1U/µl) 0.2 µl

Briefly vortex the mixture and centrifuge shortly. Label PCR tubes for each combination of sample and primer combination (or we can work with stripes or plates). Distribute 7.5 µl of the mixture to each tube (best with multipipette) and add 2.3 µl of DNA after preamplification (mix by pipetting). Centrifuge shortly tubes, stripes or plate and put to the thermocycler with following program:

| | | | |
|-------|---------------------------|---|------|
| 94 °C | 2:00 | | |
| 65 °C | 0:30 | | |
| 72 °C | 2:00 | | |
| 94 °C | 0:30 | } | 8 × |
| 64 °C | 0:30 (-1 °C in each step) | | |
| 72 °C | 2:00 | | |
| 94 °C | 0:30 | } | 23 × |
| 56 °C | 0:30 | | |
| 72 °C | 2:00 | | |
| 60 °C | 30:00 | | |
| 10 °C | hold | | |

There is a touch-down protocol used for selective amplification. It allows increasing the specificity of amplification without considerable decrease of yield. In the first nine cycles with higher annealing temperature (65-57 °C) only specific products are amplified (may be with lower effectivity but with much higher probability than non-specific products). Lower annealing temperatures in the following 23 cycles increase the yield and because the concentration of specific products is increased in previous phase of the reaction these “correct” fragments are effectively amplified.

Test the correct selective amplification on 1.8% agarose gel in 1× TBE buffer at 60V. Load 3 µl of the selective PCR product (dye is already included in the polymerase buffer).

E5. Product precipitation with sodium acetate and adding the standard (ROX, LIZ)

- pipette 3 µl of each of the selective product to clean PCR tubes (stripes) (or optimized ratio of products), close, vortex briefly, centrifuge shortly
- pipette 2 µl NaOAc (sodium acetate) to the wall of 1.5 ml tube (using dispenser)
- add 3 µl of the mixture of selective PCR products

- add 25 µl of 96% ethanol (from the freezer), again with dispenser
- vortex briefly, centrifuge shortly
- let it stand 20 mins at –20 °C
- centrifuge in pre-chilled centrifuge for 30 mins at 4 °C
- discard supernatant carefully, turn over the tubes and put on cellulose and centrifuge at 40 rcf (ethanol runs out, precipitate remains at the tube wall – invisible)
- add 100 µl of 70% ethanol (from the freezer) with dispenser
- centrifuge in pre-chilled centrifuge for 5 mins at 4 °C, discard supernatant and again centrifuge at 40 rcf upside down
- dry open tubes for 5-10 mins at 65 °C (on thermomixer)
- the samples can be stored 1-2 weeks at -20 °C

Directly before taking the samples to the sequencing facility we continue with following:

- prepare mixture for desired number of samples in 1.5 ml tube
 - deionized formamide..... 10.00 µl
 - GeneScan-ROX or LIZ (according to the dye set!)0.25 µl
- vortex briefly and centrifuge shortly
- add 10 µl of the mixture to each sample
- denature at 95 °C for 3 mins. in thermomixer or in thermocycler
- pipette samples (in a given order) to the optical plate (provided by sequencing facility) and cover with a foil

Protocol F: Microsatellite analysis (SSRs)

Microsatellites („*simple sequence repeats*“) are repetitive DNA sequences consisting of mono-, di-, tri- or tetranucleotide repeats, e.g., (A)_n, (AT)_n, (ATA)_n etc. Number of repeats in a concrete DNA region (locus) defines the allele. Two alleles can be detected in diploids, up to four alleles in tetraploids etc. The length of allele(s) is detected by PCR amplification of particular locus using primers compatible to flanking regions. The length of PCR fragments is detected using, e.g., capillary electrophoresis in automated sequencer. We have to know the sequences of flanking regions for microsatellite analysis.

Microsatellites occur in nuclear as well as in chloroplast DNA. *Nuclear microsatellites* are usually species specific and we need to work with a species for which the primers are known. They are generally highly variable and are used at within-species and population level. Di- and trinucleotide loci are mostly analysed. *Chloroplast microsatellites* occur in regions between conserved sequences, i.e., using consensual (universal) primers (compatible to those conserved parts) we can amplify loci for species from many families. They are mostly mononucleotide repeats. Its variability corresponds to the cpDNA variability and they are used, e.g., for evaluating the variability among species but also for phylogeography.

The protocol can be separated to the two parts:

1. *PCR with a pair of specific primers* – two primers are used to amplify concrete DNA locus containing a repetitive sequence pro, one primer is fluorescence-labelled (we use dyes according to the selected dye set, see comments at AFLP). For cpDNA microsatellite analysis we select primers according to the Table 3. In case of nuclear microsatellites it is best to use the protocol described in the original publication. The list of selected plant species for which microsatellites primers were published is at <http://botany.natur.cuni.cz/dna> (Primery - Mikrosatelitové primery - Seznam publikovaných mikrosatelitových primerů).

If we need to analyze several microsatellite loci (often 8-10 loci is analyzed) it is advantageous to build up multiplex PCR. This is a normal PCR reaction but with several primer pairs to amplify more loci (even five or more). This substantially decreases finance and time needs for the analysis. For successful amplification of all loci we need to ensure that: (1) primers have roughly the same annealing temperature, (2) primers do not build dimers, (3) loci with overlapping fragment lengths are labelled with different fluorescence dye. The software FastPCR (<http://primerdigital.com/fastpcr.html>, freeware) can be used to test whether primers

build dimers. The example of multiplex PCR is in Table 4. Multiplex PCR Kit (Qiagen) is great for successful amplification.

2. *Product precipitation with sodium acetate and adding the standard (ROX, LIZ)* – it is necessary for purification of PCR products. Finally, fluorescence-labelled ladder (e.g., GeneScan-ROX-500) is added as an internal standard to each of the samples. It allows identification of each AFLP fragment and unambiguous comparison with fragments in any other sample.
3. *Fragment analysis on automated sequencer* – the microsatellite alleles are electrophoretically separated in a capillary sequencer ABI 3130xl (in the sequencing facility, Faculty of Science, Charles University).

F1. PCR amplification

Pipette the mixture for desired number of samples to 1.5 ml tube. Work on ice with all chemicals. Take the polymerase out of the freezer directly before pipetting and return it back immediately. This is a general example; in case of multiplex PCR primers for particular compatible loci are added and water volume is adjusted to have the total reaction volume 20 µl (including template DNA!).

- sterile H₂O 13.60 µl
- 10× Buffer for (JumpStart) RedTaq..... 2.00 µl
- dNTP 0.40 µl
- forward primer (25 pmol/µl)..... 0.25 µl
- reverse primer (25 pmol/µl)..... 0.25 µl
- (JumpStart) RedTaq DNA Polymerase (1U/µl) 0.50 µl

Vortex briefly, centrifuge shortly and distribute 17 µl to PCR tubes (or stripes). Add 3 µl of diluted DNA (5 ng/µl), vortex briefly and put to the thermocycler with following program:

| | | |
|-------------------|-------|--------|
| 94 °C | 1:00 | } 35 × |
| 94 °C | 0:30 | |
| T _a °C | 0:30 | |
| 72 °C | 0:35 | |
| 72 °C | 15:00 | |
| 10 °C | hold | |

Test successful amplification on 1% agarose gel in TAE buffer. We should see one (or multiple in case of multiple PCR) short band(s) with the length(s) corresponding to supposed allele size (e.g., according to the Table 4 or original publication). Dilute PCR product 1:2 to

1:20 according to the band intensity. Dilution should be adjusted according to the peak intensity in the data from sequencer.

F2. Product precipitation and adding a standard

Product precipitation and addition of standard (GeneScan-ROX or LIZ) is similar to the AFLP protocol (see protocol E5).

Protocol G: Next-generation sequencing (NGS) library preparation (Illumina)

Next-generation sequencing (NGS) allow high-throughput sequencing of whole genomes or their large portions. One of the approaches is so-called genome skimming which is a low-pass sequencing (i.e., with very low average coverage) of the unmodified genomic library. The most common NGS technology is Illumina which require to prepare a specific library prior to sequencing. The following steps are necessary to prepare and sequence the samples:

1. DNA sonication (fragmentation) using ultrasonicator – a specific instrument is required to shear the genomic DNA into pieces under specifically defined conditions. The success of sonication is checked on agarose gel.
2. Library preparation using NEBNext Ultra II DNA Library Prep Kit for Illumina – the sequencing library is prepared from sheared DNA using this kit following the instructions of manufacturer. Optionally (if we want to multiplex several samples prior to sequencing) the barcodes are introduce using NEBNext Multiplex Oligos (single- or dual labelling).
3. Measuring NGS library concentration using Qubit and sending to core facility for sequencing.

G1. DNA Shearing with M220 Focused-ultrasonicator (Covaris)

The M220 Focused-ultrasonicator is based on Covaris Adaptive Focused Acoustics (AFA) technology.

Turn on the instrument – The switch is located at its back next to the power cord, turn on the computer. Double click on the SonoLab icon to start the software.

Filling the water bath – Fill the wash bottle with AFA-grade water, slide the blue cover on the nozzle up. Open the safety cover of the instrument, place the tube holder insert into the water bath housing. Place the tip of the wash bottle in the water sense aperture (in the holder) and fill the water bath by squeezing the wash bottle until the water reaches the top surface of the tube holder. Approximately 15 ml of water will be required.

Loading the sample – Lift the sliding weight and rotate it into the loading position. Take the specific sonicator tube (micro TUBE-AFA Fiber Screw-Cap Case-50µl) and pipette 50 µl of genomic DNA into it. Place it to the insert (the yellow part) and together place it to the holder (the blue part). Drop the sliding weight to the running position on top of the sample tube. Close the safety cover.

Processing the sample – Select an existing method in SonoLab. Use, e.g., these parameters: Peak Incident Power: 50W, Duty Factor: 20%, Cycles per Burst: 200, Treatment Time: 45

sec, Temperature: 20 °C. Click on “Run” to start the sonication. After the process is finished, remove the sample tube.

Checking sonication results on the gel – prepare 0.8 agarose gel and load 2 µl of sonicated sample mixed with 1 µl loading dye. Load 0.7 µl of 1kb ladder in the first well and 0.4 µl of 100kb ladder in the last well. The DNA should be fragmented to the pieces of desired length (i.e., 400-600 bp). If the majority of fragments is too long sonication has to be repeated.

G2. Library preparation using NEBNext Ultra II DNA Library Prep Kit for Illumina

For all the steps follow the most recent version of the NEBNext Ultra II protocol (cat. nr. E7645L). The library preparation consists of following steps:

End repair and A-tailing – this correct the sonicated fragments to fully double-stranded DNA and adds A overhangs. End Repair Reaction Buffer and End Prep Enzyme Mix is mixed together with the sonicated sample and incubated for 30 mins at 20 °C and additional 30 mins at 65 °C in thermocycler.

Adaptor ligation – this ligates Illumina-specific adaptors to the fragments. Ligase Master Mix, NEBNext Adaptor for Illumina and Ligation Enhancer is mixed with the repaired sample and incubated for 15 mins at 20 °C. Then, USER enzyme is added and the sample incubated at 37 °C for additional 15 mins.

Sample clean-up – the sample is purified using QIAquick PCR Purification Kit following the protocol. The sample is eluted with 32 µl ddH₂O.

Size selection – the desired fragment lengths are selected either by cutting from the gel (and purified) or using Pippin Prep instrument.

Final concentration measurement – the DNA concentration is precisely measured using fluorometry with Qubit instrument.

B. Analysis of primary data

1. Inspecting sequences using FinchTV

Sequence data (in format *.ab1) are saved at the server (<ftp://botany.natur.cuni.cz/>, user: sequences, accessible with, e.g., Total Commander) in the folder „sekvence“ and then in the folder according to the analysis data (in the format yy-mm-dd). Individual downloaded sequences can be viewed, e.g., with the program FinchTV (freeware, <http://www.geospiza.com/Products/finchtv.shtml>). Sequence should be readable from the base ca. 30-50 bases to ca. 550-650 bases (or longer). There is a problem (probably with sequencing reaction) if we do not see unambiguously defined peaks for individual bases. Only quality sequences can be further analyzed. In FinchTV we can edit sequences, i.e., change „base calling“, delete and insert bases. The software further allows to search in sequences, export to FASTA format, print the chromatogram and also viewing the primary (*raw*) data.

2. Making contigs in SeqMan (DNASTAR)

So-called contig originates by connecting *forward* and *reverse-complement* (i.e., complementary bases in the opposite order) sequence from the same region (PCR fragment). This can be done automatically, e.g., in SeqMan (part of the commercial software package DNASTAR Lasergene, <http://www.dnastar.com/t-seqmanpro.aspx>, you can ask for evaluation version). Also here obviously wrong „base calling“ (transformation of peaks to individual bases) can be edited.

- add sequences with *File* → *New* → *Add Sequences*
- add *forward* and *reverse* sequences from the same sample (double-click or select + „Add“)
- click to „*Assemble*“ then double-click to *Contig 1* (if the sequences are correct and overlap in the middle just one contig is created)
- control if forward sequence is from the left to the right; if not select *Contig* → *Complement Contig*
- expand chromatograms of all sequences by clicking on the triangle left to their name
- displaying beginnings and ends of each sequence can be changed by shifting with mouse
- check if „base calling“ is correct in the *Translate* row (consensus sequence), edit (overwrite) obvious errors
- save edited contig in the native SeqMan format (*.sqd), it can be useful for later check

- select with mouse and copy (Ctrl+C) the top row (*Translate*), or its part
- paste the sequence to a text file (e.g., to Notepad) and name the sequence in the fasta format, e.g.,

```
>sequence1
ATTGCGTATGCCTGTCAT
```

- repeat this for all sequences and save the text file with *.fas suffix

The contigs can be alternatively prepared in the software Geneious (<http://www.geneious.com>), or with the web application of the software CAP3 (<http://doua.prabi.fr/software/cap3>).


3. Evaluation of AFLP data in GeneMarker (by F. Kolář)

AFLP data from sequencer can be analysed in the software GeneMarker (<http://www.softgenetics.com/GeneMarker.html>, commercial software, it is possible to download evaluation or demo version where you can't save your projects).




Importing and checking input files

- import data using *File* → *Open Data*, in next window *Add...* (we can select more file when pressing Ctrl or Shift), then *Open* and *OK*
- data analysis (calibration with internal standard) is done by *Project* → *Run* (window with this choice automatically appears also after import of files), select ABI as *Template*, „GS500_1“ as *Size Standard* (accepts all standard peaks of GS-500 ROX except the peak with 250 bases), AFLP as *Analysis type* → different parameters of the automatic analysis can be set in the next window; this will be important when analysing with panel (see below), now only select *Allele call / Auto Range (bps)* the range for allele call – usually 100-500 bp → nothing to be changed in the next window → *OK*
- analysis of all samples proceeds; in the window we should check if there is „*Completed*“ for all the samples. If there is „*Failed*“ the sample cannot be used. Close the window with the analysis report.
- there is a list of samples on the left side of the screen; fully analysed samples are in green, problematic samples in yellow, not analysed samples in red. In the top middle part there is a virtual gel showing all the samples, below is one (or several) sample(s) displayed as peaks.
- virtual gel (together with peaks in the lower middle part) can be zoomed in by left mouse click and drawing to the right (a rectangle of desired size appears). To zoom out back to the full view we use left mouse click and draw to the left. In zoomed-in gel we can move


to the left and to the right when right mouse click is pressed. Analogous we work also with samples displayed as histograms (peaks).

- check if standard peaks are aligned: select a colour of standard (usually red or orange) by repeated clicking on  and have a look at virtual gel (or also at histograms) if peaks of particular size are one above each other.
- it is also good to control how the individual samples look like – if we want to disable a sample (e.g., due to weak signal) click on it and press Ctrl+Del (by repeated pressing Ctrl+Del the sample is re-activated)



Panel construction

- now we prepare a panel, i.e., a framework for our semi-automated scoring of our samples. The final panel construction is usually done after we have the majority samples available. All manual changes in the data after analysis with the panel will be missing after repeated analysis with the panel.
- Click *Tools* → *Panel Editor*, with button  select the colour for which we want make the panel, click on  to open the window for creating a new panel; keep all pre-set values (create panel automatically using all samples) → *OK*
- this creates a raw version of the panel which needs to be adjusted. Click on  to switch between display of histograms of all samples (one over each other) and virtual gel where we see automatically created *bins*, i.e., “racks” for individual AFLP loci. In the bins there will be automatically scored the presence or absence of fragments (peaks). Histograms or gel can be also zoomed in and zoomed out (see above). Numbers, position and range of *bins* can be manually changed (best in virtual gel) – click on particular *bin* and you can delete it (right mouse click – *Delete Allele*), move (Shift+left mouse click and move) or change the range (move to bin edge then Shift+left mouse click and move). New *bin* can be created by pressing Shift and clicking on proper place. Save the file after finishing of all changes and close the Panel Editor window.
- we can export created panel as a file with *.xml suffix and backup for further analyses in the future

Analysis with the panel and subsequent adjustments

- panel constructed above is now used for automating scoring of our data (we do *allele call*). Run the analysis again by clicking on  → in the first window under *Panel* select our new panel (anything else should be same as in the previous analysis) → set parameters of automatic analysis in next window: change range for scoring fragments

(*Auto Range (bps)* to usually 100-500 bp), we can lower Intensity at *Peak Detection Threshold* to 50 (also lower peaks will be scored) and unclick *Stutter peak filter* and *Plus-A Filter* (their rules are usually too stringent for AFLP data) → in the third window we can set a “reliability score” (slightly mystic, automatically computed value) which influences whether the particular peak will be accepted (pass) / set as uncertain (has to be checked) / reject: it might be good to lower the value between *Check* and *Pass* to ca. 3 → *OK* → leave field at *Call size again* empty and click on *Apply to All* (we don’t need to re-calibrate standards again). Note: For final analysis it is good to try to analyse the data with different settings and select an optimal solution.


- after the analysis (and closing the small window) we have automatically scored fragments (= peaks) in all loci (*bins*): some peaks are accepted (grey), uncertain (yellow) or rejected (red).
- even if we thoroughly played with parameter settings it is necessary to go through the data carefully and make manual tuning (most time consuming part of the work). Yellow (or red) marked peaks can be confirmed (click and Ctrl+M), grey peaks can be deleted (click and Del) and we can even add completely new alleles (right mouse click and *Insert Allele*). For easier viewing we can switch off the gel display by clicking on  and open more samples at once by clicking on particular samples in the list in the left window (maximum number of simultaneously displayed samples, and other preferences, can be changed under *View – Preferences – Display Settings*). Between neighbouring samples we can easily switch with PgUp and PgDn. Attention! All manual changes are cancelled if we re-run the analysis with panel.
- data export (only available in a full version) can be done by saving the table in the right window as *.txt. or *.xls. If we want just to have simple „0-1-?“ matrix, click on  and deselect *Show Peak Area*.

Adding further samples

- to the project at work we can add additional samples during analysis with *Project – Add Samples to Project*. The samples are automatically analysed with the panel in use and we can start to modify them.

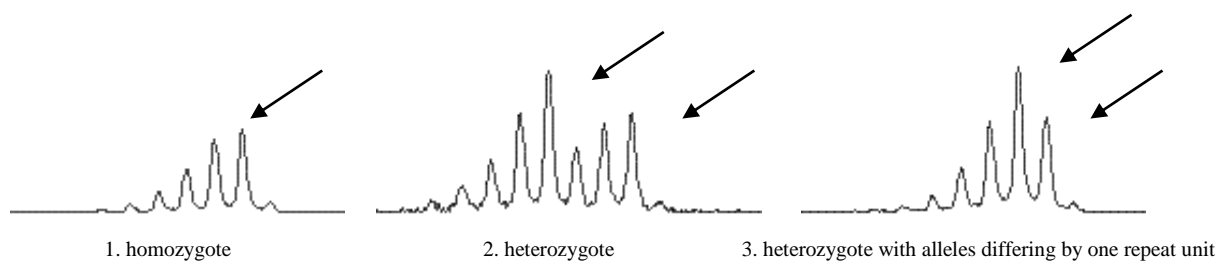
4. Microsatellite data evaluation in GeneMarker

Microsatellite data from sequencer can be analysed and visualized in the software GeneMarker (<http://www.softgenetics.com/GeneMarker.html>, commercial software, it is possible to download evaluation or demo version where you can't save your projects).

- import data using *File* → *Open Data*, in next window *Add...* (we can select more file when pressing Ctrl or Shift), then *Open* and *OK*
- data analysis (calibration with internal standard) is done by *Project* → *Run* (window with this choice automatically appears also after import of files), select ABI as *Template*, „GS500_1“ as *Size Standard* (accepts all standard peaks of GS-500 ROX except the peak with 250 bases), nothing to be changed in the next window and click OK
- analysis of all samples proceed; in the window we should check if there is „*Completed*“ for all the samples. If there is „*Failed*“ the sample cannot be used. Close the window with the analysis report.
- there is a list of samples on the left side of the screen; fully analysed samples are in green, problematic samples in yellow, not analysed samples in red.
- the best way how to evaluate the microsatellite data is to display samples in individual loci one under each other and read the alleles. Right click on the first sample and select „*Select Max*“. Up to 32 first samples are displayed in lower middle window. By several clicks on  we select one out of three (four) colours which we want to evaluate. By clicking on icon “magnifying glass” and using sliders we navigate to particular locus with enough high resolution.
- we can read the allelic composition for each sample and each locus and write it, e.g., to Excel table

Comments to microsatellite patter evaluation

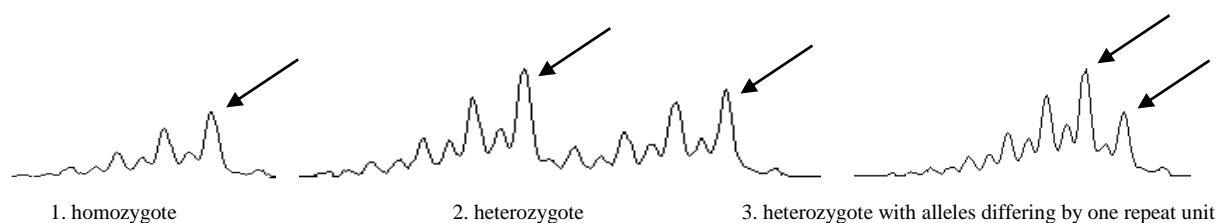
1. Diploids can have no more than 2 alleles at the locus; the individual is homozygote if we see one allele. Tetraploids can have up to 4 alleles at single locus; if we see one or four we can unambiguously identify the genotype. When we score, e.g., three alleles one should be twice but it is hard to say which one. We can score allelic presence but not the exact genotype. It is even more complicated at higher ploidy levels.
2. For some loci is typical the presence of so-called „stutter bands“ (i.e., peaks that are longer by two, three, four etc. multiple of repeat unit than the correct allele). They originate by *in vitro DNA slippage* during replication in PCR. Typical pattern look like this (correct alleles are marked by arrow):



The scoring is simple if the sample is homozygote (1) – we consistently score the highest peak only. In case of heterozygotes with alleles differing by at least repeat units (2) is the situation also simple – we score the two highest peaks; the peaks after them are much lower. If the alleles differ by just one repeat unit (3) we score the two most right alleles. The height of the second most right peak is higher because there is also a stutter from the other allele.

Stutter bands do not prevent from correct allele identification; they can even sometimes help to identify correct alleles and recognize unspecific product (correct alleles will always have similar stutter band pattern).

3. The other phenomenon which complicates microsatellite pattern interpretation are so-called „-A“ products. *Taq* polymerase adds adenine to the end of each product (terminal transferase activity) and thus enlarges all fragments by 1 base. For this reason we put the final elongation step in length of 10-15(-30) mins to ensure that adenine is added to all products and the results are consistent. Regardless of longer final elongation step some loci still have „-A“ products. It is usually not a problem but we should consistently score the right (i.e., „+A“) allele. Examples of patterns with „-A“ products (correct alleles are marked by arrow):



C. Evaluation of codominant data

1. Basic analysis of microsatellite data using MSA

Using software MSA (http://i122server.vu-wien.ac.at/MSA/MSA_download.html, freeware) we can (for diploids!) calculate following population-genetic parameters:

- descriptive statistics for populations and loci (number of alleles, H_O , H_E , H_{Sh} , F_{IS} ...)
- distance matrices among individuals and populations (D – standard genetic distance, $(\delta\mu)^2$, D_{ps} – proportion of shared alleles, D_{kf} – kinship coefficient...)
- F -statistics (F_{ST} , F_{IS} , F_{IT} – both global and pair-wise)

The software also warns us if there might be an error in the dataset (alleles do not correspond to the multiple of repeat unit, too big distance among alleles, distant alleles etc.).

The input format for MSA is best to prepare in, e.g., Excel and save as TAB-delimited text (or better copy it to a simple text editor, e.g., Notepad++). Input file for diploid species:

| | | locus name | | repeat length | | | | | | length of flanking region | | | |
|---------------------------|---|------------|---|---------------|-----|-------|----|-------|-----|---------------------------|-----|-------|-----|
| | | 2 | | 2 | | 2 | | 2 | | 2 | | 2 | |
| | | | | 64 | | 74 | | 46 | | 46 | | 24 | |
| | | | | NLGA1 | | NLGA2 | | NLGA3 | | NLGA4 | | NLGA5 | |
| one- or two-column format | 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 |

outbred (d) or inbred (h) individuum

population name

number of population group

missing data

Input file has to be saved to the same folder where we have `MSAnalyzerMr.exe`, which we then run. The window is opened and we can specify the input file and parameters of the analysis. Typing (i) we enter the input file (including its suffix!). Typing (d) and later (p) we can set diverse genetic distances to be calculated (1)-(9) and also set calculation of distance among populations (c) and individuals (i), and possibly also switch on bootstrapping (n). By typing (b) we go back to „distance menu“. Typing (s) allows to set parameters of F -statistics, typing (c) switch on the calculation of F -statistics, typing (g) we select if fit will be calculated globally, pair-wise or both. Typing (m) we go back to „main menu“, where by typing (c) we can select building input file for the software Arlequin and/or Structure (3). The analysis can be run by typing (!).

If no bootstrap was used the analysis is really fast. Depending on selected parameters MSA creates many Excel tables and text files and saves them to synoptic folder structure:

- Allelecount – allele numbers and frequencies for individual loci and populations
- Distance_data – text files with distance matrices among individuals and populations
- Formats&Data – input files for Arlequin, Structure and other software
- F-Statistic – F -statistics global and pair-wise
- Group_data – information about parameters according to predefined population groups
- Single_data – information about parameters for individual populations

For further information about parameters, calculation methods and interpretation of output files see the MSA manual (http://i122server.vu-wien.ac.at/MSA/info.html/MSA_info.html).

2. Analysis of molecular variance (AMOVA) in Arlequin

In the software Arlequin v 3.5 (<http://cmpg.unibe.ch/software/arlequin35>, freeware) analysis of molecular variance (AMOVA) can be calculated. We used a file generated by the software MSA (see previous part) as an input. The file with *.arp suffix can be opened using „Open project“, at the tab „Project“ there are particular populations (*Samples*) and groups (*Groups*). At the tab „Settings“ we click what we want to calculate, for AMOVA this is *Genetic structure* → *AMOVA*. We can select the method for calculating the matrix of inter-individual pair-wise distances: *number of different alleles* (i.e., F_{ST} analogue) or *sum of squared differences* (i.e., R_{ST} analogue). Start the analysis by clicking on the button „Start“. The software generates a web page with results which can be navigated with menu on the left side

(it is necessary to allow running of active content!). Individual analyses are in the folder according to the date and time of the run; at the beginning of each result file there are all the settings, geographic structure, method for distance calculation etc. Interpretation of the AMOVA results is similar to the analysis with dominant marker data.


3. PAST – principal coordinate analysis (PCoA) and distance-based trees

Comment: This tutorial is valid for the version 2.17, the newest version 3.1x looks slightly different.



Click on the icon to run the program (downloaded from <http://folk.uio.no/ohammer/past/>). It allows performing principal coordinate analysis and construction of distance-based trees (neighbour-joining, UPGMA), but also many other analyses. This tutorial describes how to insert symmetrical distance matrix (e.g., as output from MSA). If we also want to insert sample labels it is necessary to tick *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. Groups definition (groups will be coloured in all following outputs): Shift + select corresponding samples (rows) with mouse click → *Edit* → *Row color/symbol* → select colour/symbol (Comment: in this phase we can save the file and the group definition will be saved as well)
2. PCoA: Select all samples (click to the upper left corner) or desired selection (Shift + mouse click) → *Multivar* → *Principal coordinates* → select *User distance* → click to  copies % of explained variation to the clipboard → *View scatter*
3. Modification and saving of PCoA diagram: clicking to the graph allows to change symbol size, font etc. and also to *Save picture*. *View numbers* displays ordination scores of individuals and again it is possible to save it via clipboard and make (nicer) picture elsewhere. Symbol colour and size is possible to change only in the source table via *Row color/symbol* (see above).
4. Distance tree (NJ): *Multivar* → *Neighbour joining* → select *User distance* → insert number of bootstrap replicates to the field *Boot N* and press *Enter*. UPGMA tree is made similarly under *Multivar* → *Cluster Analysis* (be sure that *Paired group* algorithm is selected).

D. Working with sequences

1. Alignment using ClustalX

To do an alignment means to put homologous bases under each other (*positional homology*). It can be done automatically, e.g., in the software ClustalX (<http://www.clustal.org/clustal2>).

- open the file with sequences in FASTA format with *File* → *Load Sequences*
- select *Alignment* → *Do Complete Alignment*
- the resulting alignment is automatically saved to the file with the suffix *.aln

Automatically generated alignment can be edited either directly in a text editor (e.g., Notepad) or use all the advantages of the software BioEdit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) or PhyDE (<http://www.phyde.de>). For the second one we need to have the alignment in FASTA or NEXUS sequential format. Alternatively (and preferentially in case of more diversified sequences) it is better to use iterative algorithm in, e.g., the software MAFFT (<http://mafft.cbrc.jp/alignment/server>).

2. Conversion between different sequence formats

Different sequence formats can be converted, e.g., using the software FORCON (<http://bioinformatics.psb.ugent.be/webtools/ForCon>), which is a part of the software TREECON (<http://bioinformatics.psb.ugent.be/software/details/3>).

- ForCon → Enter
- *Input format*: CLUSTAL, *Output format*: FASTA, OK
- change type of files to „all“, open the file with alignment (e.g., with *.aln suffix)
- *Cut-off after how many characters ?* – write more than there are alignment positions
- save with the suffix *.fas
- Select *All*, *OK*

Another option is , e.g., the software program PGDSpider, which can read/write many formats, e.g., FASTA, NEXUS, MSA, Arlequin etc. (<http://www.cmpg.unibe.ch/software/PGDSpider>).

We can also use on-line tool at the address

<http://sequenceconversion.bugaco.com/converter/biology/sequences/index.html>

3. Automated indel coding in SeqState

If the alignment contain *gaps*, i.e., insertions/deletions (indels), which can be phylogenetically informative, they can be coded as separate data (usually 0/1), which are add at the end of the alignment. This can be automatically done, e.g., in the software SeqState (<http://bioinfweb.info/Software/SeqState>). Details about indel coding methods can be found in the paper Müller (2006): Incorporating information from length-mutational events into phylogenetic analysis. *Molec. Phyl. Evol.*, 38, 667-676.

- *File* → *Load NEXUS file* (or *Load FASTA file*)
- select file with *.nex suffix (must not be in interleave format but in sequential), or with *.fas suffix
- *IndelCoder* → *modified complex coding* (see above mentioned paper for detailed information about indel coding)
- it generates NEXUS file (add _mcic.nex to the original file name) with the original matrix with added indel coding. This file can be directly used, e.g., in the software PAUP.

4. FaBox – manipulation with sequences in FASTA format

(<http://users-birc.au.dk/biopv/php/fabox>)

- using this web we can manipulate with sequences in FASTA format (work with headers, trim and join alignments, format for other programs etc.)
- the choice „*Create TCS input file from fasta (fasta2tcs)*“ creates correctly formatted file for the analysis in the software TCS

5. Haplotype network in the software TCS

- the software TCS 1.21 (<http://darwin.uvigo.es/software/tcs.html>) collapses sequences to particular haplotypes and construct a haplotype network using statistical parsimony
- you need Java to run it (Java Runtime Environment – download JRE, <http://www.oracle.com/technetwork/java/javase/downloads/index.html>)
- run TCS1.21.jar, select „*Start New TCS Analysis*“
- *File* → *Select NEXUS/PHYLIP Sequence file*
- leave everything as default, only change „*Gaps = missing*“, *RUN*
- the software draws a haplotype network which can be edited (saved in the file *.graph) or exported (as PostScript or PICT file)
- the information about analysis are saved to the file *.graph.log

Table 1: PCR and sequencing primers for non-coding cpDNA and nrDNA ITS regions (with primer localization in the genome).

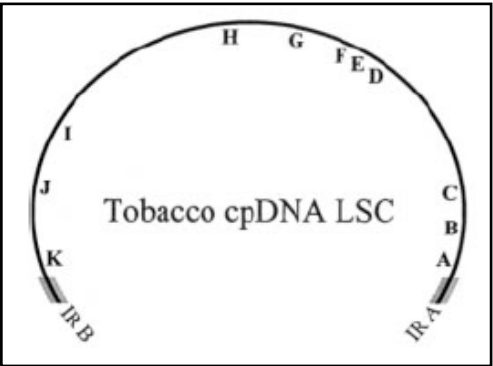
Shaw J. et al. (2005): The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am. J. Bot. 92, 142-166.

Primers for sequencing variable non-coding cpDNA regions

| Name | Forward primer | Sequence | Reverse primer | sequence | Size (bp) | T _a | Reference |
|------------------|----------------|----------------------------|----------------|--------------------------------|-----------|----------------|---|
| <i>trnT-trnL</i> | <i>trnT2F</i> | CAAAATGCGATGCTCTAACCT | <i>trnL-b</i> | TCTACCGATTTCGCCATATC | 527–1023 | 50-55 | Taberlet et al. 1991, Cronn et al. 2002 |
| <i>trnL-trnF</i> | <i>trnL-c</i> | CGAAATCGGTAGACGCTACG | <i>trnF-f</i> | ATTTGAACTGGTGACACGAG | 602–1076 | 50 | Taberlet et al. 1991 |
| <i>trnH-psbA</i> | <i>trnH</i> | CGCGCATGGATTCCACAATCC | <i>psbA</i> | GTTATGCATGAACGTAATGCTC | 198–1077 | 50-56 | Tate & Simpson 2003, Sang et al. 1997 |
| <i>trnS-trnG</i> | <i>trnS</i> | AGATAGGGATTTCGAACCTCGGT | <i>trnG2S</i> | TTTTACCACTAAACTATACCCGC | 619–1035 | 50 | Shaw et al. 2005 |
| <i>trnG-trnG</i> | <i>trnG2G</i> | GCGGGTATAGTTTAGTGGTAAAA | <i>trnG</i> | GTAGCGGGAATCGGACCCGCATC | 697–1008 | 50 | Ohsako & Ohnishi 2000, Shaw et al. 2005 |
| <i>rpoB-trnC</i> | <i>rpoB</i> | CKACAAAAYCCYTCRAATTG | <i>trnC</i> | CACCCRGATTYGAACCTGGGG | 914–1309 | 50-57 | Shaw et al. 2005 |
| <i>rps16</i> | <i>rpS16F</i> | AAACGATGTGGTARAAAGCAAC | <i>rpS16R</i> | AACATCWATTGCAASGATTTCGATA | 784–946 | 50-55 | Shaw et al. 2005 |
| <i>ycf6-psbM</i> | <i>ycf6F</i> | ATGGATATAGTAAGTCTYGCTTGGGC | <i>psbMR</i> | ATGGAAGTAAATATTCTYGCATTTATTGCT | 406-1283 | 50-55 | Shaw et al. 2005 |

Primers for sequencing of ITS rDNA region

| | | | | | | | |
|------|------|------------------------|------|----------------------|---------|--|-------------------|
| ITS1 | ITS5 | GGAAGTAAAAGTCGTAACAAGG | ITS2 | GCTGCGTTCTTCATCGATGC | 200-300 | | White et al. 1990 |
| ITS2 | ITS3 | GCATCGATGAAGAACGCAGC | ITS4 | TCCTCCGCTTATTGATATGC | 180-240 | | White et al. 1990 |



trnH-psbA

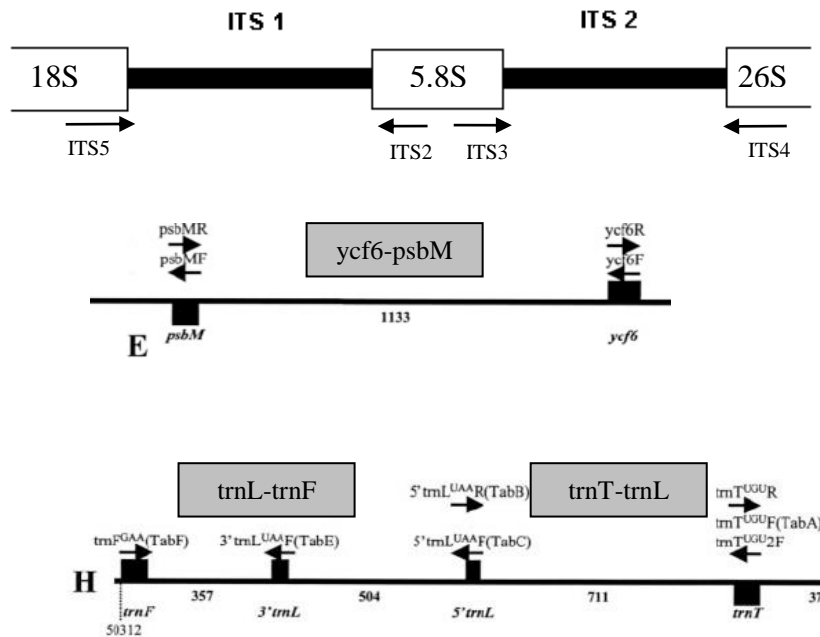
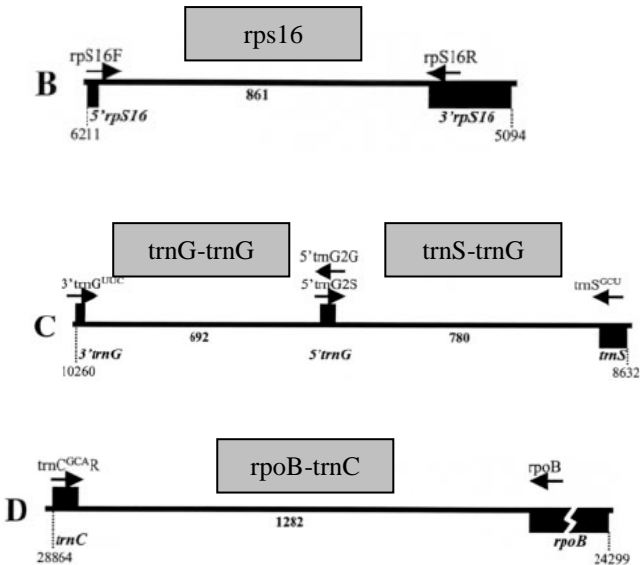


Table 2: Coding of the combination of AFLP primers (with three selective nucleotides) and fluorescence labelling of *EcoRI* primers (dye set DS-30 with marker ROX).

| | | | | <i>MseI</i> primer | | | | | | | |
|---------------------|---|-------|-----|--------------------|-----|-----|-----|-----|-----|-----|-----|
| | | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | | | | CAA | CAC | CAG | CAT | CTA | CTC | CTG | CTT |
| <i>EcoRI</i> primer | Y | NED | AAC | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
| | G | HEX | AAG | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 |
| | B | 6-FAM | ACA | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 |
| | Y | NED | ACC | Y9 | Y10 | Y11 | Y12 | Y13 | Y14 | Y15 | Y16 |
| | G | HEX | ACG | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 |
| | B | 6-FAM | ACT | B9 | B10 | B11 | B12 | B13 | B14 | B15 | B16 |
| | Y | NED | AGC | Y17 | Y18 | Y19 | Y20 | Y21 | Y22 | Y23 | Y24 |
| | G | HEX | AGG | G17 | G18 | G19 | G20 | G21 | G22 | G23 | G24 |
| | B | 6-FAM | ATC | B17 | B18 | B19 | B20 | B21 | B22 | B23 | B24 |

MseI primer

5' - GATGAGTCCTGAGTAAxxx -3'

EcoRI primer

5' - GACTGCGTACCAATTCxxx -3'

Table 3: List of primers for amplification of cpDNA microsatellite loci (primers fluorescence-labelled for dye set DS-30 with marker ROX).

| fluoresc. dye | | code | position (bp) | localization | repetition in tobacco | primer sequence | T _a | | Allele range in tobacco (bp) |
|---------------|-------|--------|---------------|----------------------|-------------------------------------|--------------------------|----------------|------|------------------------------|
| | | | | | | | Wal | Cpr | |
| G | HEX | ccmp1 | 3801 | <i>trnK</i> intron | (T) ₁₀ | CAGGTAAACTTCTCAACGGA | 58 | 53.3 | 139 |
| | | | | | | CCGAAGTCAAAAGAGCGATT | 58 | 57.3 | |
| Y | NED | ccmp2 | 8609 | 5' to <i>trnS</i> | (A) ₁₁ | GATCCCGGACGTAATCCTG | 60 | 58.0 | 189 |
| | | | | | | ATCGTACCGAGGGTTTCAAT | 60 | 58.6 | |
| B | 6-FAM | ccmp3 | 10075 | <i>trnG</i> intron | (T) ₁₁ | CAGACCAAAAGCTGACATAG | 58 | 51.3 | 112 |
| | | | | | | GTTTCATTTCGGCTCCTTTAT | 56 | 53.9 | |
| G | HEX | ccmp4 | 12872 | <i>atpF</i> intron | (T) ₁₃ | AATGCTGAATCGAYGACCTA | 60 | 55.3 | 126 |
| | | | | | | CCAAAATATTBGGAGGACTCT | 58 | 56.1 | |
| Y | NED | ccmp5 | 16950 | 3' to <i>rps2</i> | (C) ₇ (T) ₁₀ | TGTTCCAATATCTTCTTGTCATTT | 62 | 55.0 | 121 |
| | | | 16977 | | (T) ₅ C(A) ₁₁ | AGGTTCCATCGGAACAATTAT | 58 | 55.4 | |
| B | 6-FAM | ccmp6 | 45119 | ORF 77-ORF 82 | (T) ₅ C(T) ₁₇ | CGATGCATATGTAGAAAGCC | 58 | 52.7 | 103 |
| | | | | | | CATTACGTGCGACTATCTCC | 60 | 52.5 | |
| G | HEX | ccmp7 | 57339 | <i>atpB-rbcL</i> | (A) ₁₃ | CAACATATAACCACTGTCAAG | 56 | 45.1 | 133 |
| | | | | | | ACATCATTATTGTATACTCTTTC | 58 | 44.5 | |
| Y | NED | ccmp8 | 71563 | <i>rpl20-rps12</i> | (T) ₆ C(T) ₁₄ | TTGGCTACTCTAACCTTCCC | 60 | 52.9 | 77 |
| | | | | | | TTCTTTCTTATTTTCGCAGTGAA | 58 | 53.9 | |
| B | 6-FAM | ccmp9 | 74060 | ORF 74b- <i>psbB</i> | (T) ₁₁ | GGATTTGTACATATAGGACA | 54 | 45.1 | 98 |
| | | | | | | CTCAACTCTAAGAAATACTTG | 56 | 44.2 | |
| G | HEX | ccmp10 | 86694 | <i>rpl2-rps19</i> | (T) ₁₄ | TTTTTTTTTAGTGAACGTGTCA | 56 | 53.3 | 103 |
| | | | | | | TTCGTCGDCGTAGTAAATAG | 58 | 53.7 | |

Weising K & Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome* 42, 9-19.

Table 4: An example of multiplex PCR design for 10 microsatellite loci for *Nymphoides peltata*.

| allele range | bp | 152-162 | 147-175 | 149-173 | 176-204 | 160-174 | 220-234 | 147-169 | 143-167 | 135-137 | 127-129 |
|--------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| bp | locus | Np152 | Np207 | Np274 | Np306 | Np382 | Np641 | Np694 | Np730 | Np274-2 | Np382-2 |
| 152-162 | Np152 | | y | y | | y | | y | y | | |
| 147-175 | Np207 | | | y | x,y | y | | x,y | y | | |
| 149-173 | Np274 | | | | y | y | | y | x,y | | |
| 176-204 | Np306 | | | | | y | x | y | y | | |
| 160-174 | Np382 | | | | | | | y | y | | x |
| 220-234 | Np641 | | | | | | | | | | |
| 147-169 | Np694 | | | | | | | | y | y | |
| 143-167 | Np730 | | | | | | | | | x,y | |
| 135-137 | Np274-2 | | | | | | | | | | |
| 127-129 | Np382-2 | | | | | | | | | | |

x = primer dimers

y = allelic range overlap

Multiplex PCR:

| multiplex I | allele range |
|-------------|--------------|
| Np152 | 152-162 |
| Np207 | 147-175 |
| Np382 | 160-174 |
| Np274-2 | 135-137 |
| Np641 | 220-234 |

| multiplex II | allele range |
|--------------|--------------|
| Np274 | 149-173 |
| Np382-2 | 127-129 |
| Np694 | 147-169 |
| Np306 | 176-204 |

| multiplex III | allele range |
|---------------|--------------|
| Np730 | 143-167 |