Molecular markers in plant systematics and population biology

5. Microsatellites

Tomáš Fér

tomas.fer@natur.cuni.cz

What are microsatellites ?

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

```
…GTTCTGTCATATATATATATAT----CGTACTT…
…GTTCTGTCATATATATATATATATATCGTACTT…
```

- alleles are defined by different number of repetitions
- PCR – length polymorphism

Types of microsatellites

- *simple*
  …CACACACACACACACACACACA…

- *compound*
  …CACACACACATGTGTGTGTGTG…

- *interrupted*
  …CACACATTCACACATTCACA…

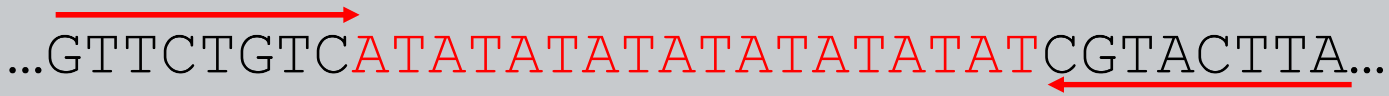
Repetitive sequences

- dinucleotides
    - AT repeat most common in plants
    - every 30-50 kb
    - number of repeats up to 30
- trinucleotides
    - occurrs also in exons (do not break the reading frame) – especially GC-rich repeats
    - AT-rich trinucleotides distributed roughly evenly
    - GTG – subtelomeric localization on chromosome
- tetranucleotides
    - GATA/GACA only
    - localization near centromeres, highest occurrence in UTRs
    - often compound or interrupted

Characteristics of microsatellites

- *single locus* – highly specific
- common occurrence in the genome
- distributed throughout the whole genome
- highly polymorphic – many alleles
- codominant inheritance

- BUT – primers must be known (i.e., sequences of *flanking regions*)

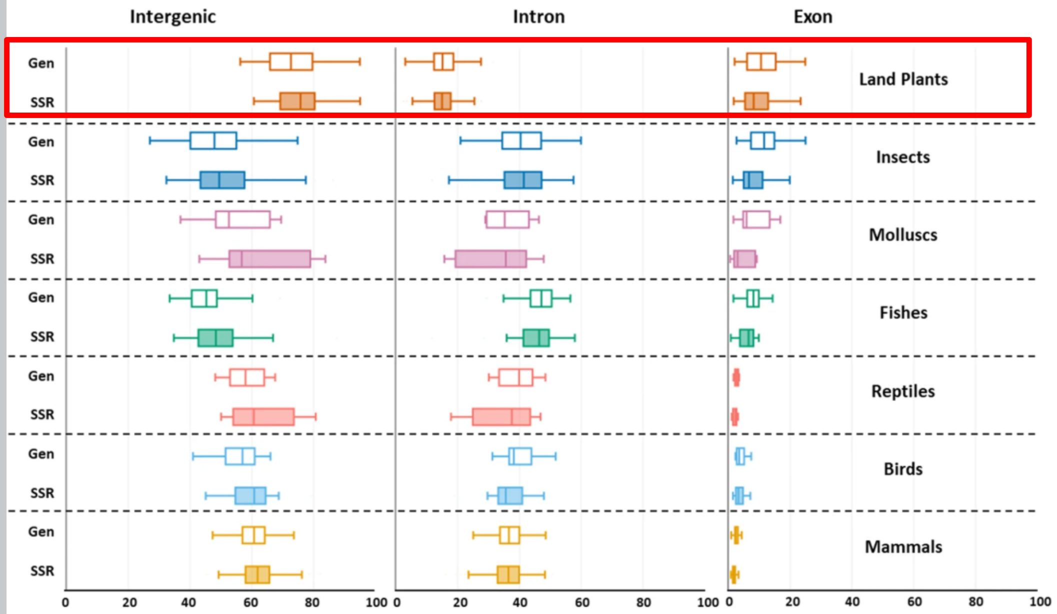


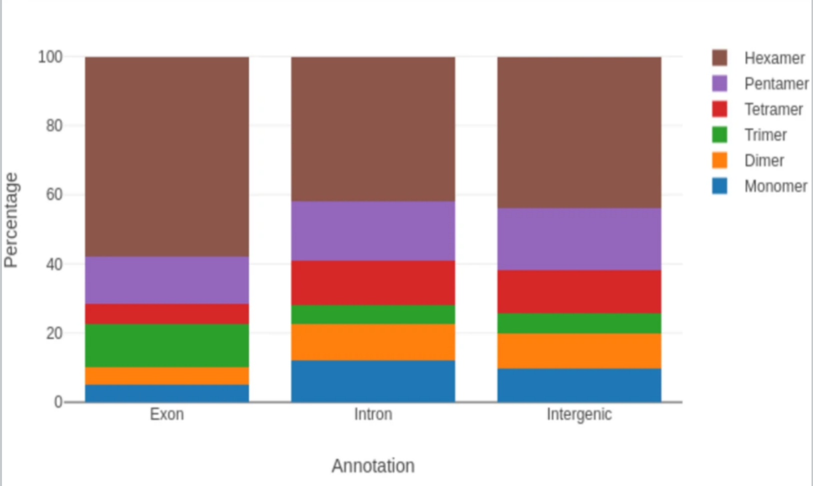
Distribution in the genome

- distributed throughout the whole genome
  (BUT – reflects variability of the studied loci,
  i.e., limited number of loci)
- rather in non-coding regions, tri- and hexanucleotide repeats also in exons
- high frequency in UTRs (variations in 5'-UTRs could regulate gene expression)

- nuclear microsatellites
  - species specific
- chloroplast microsatellites
  - usually repeats of single base – i.e., $(T)_{12}$
  - *flanking regions* – less variable – possible to design consensual primers

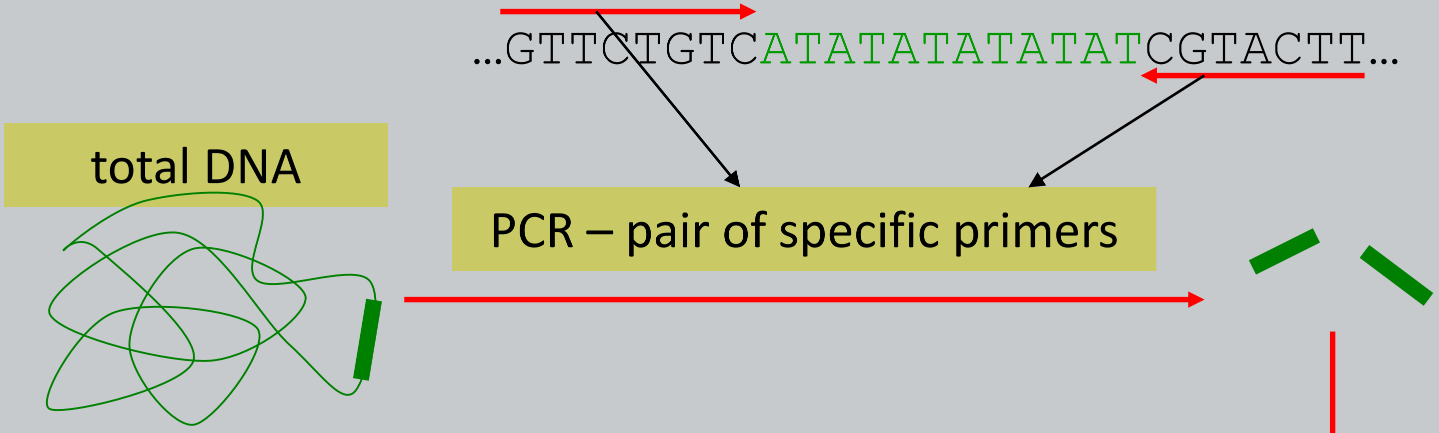
Distribution in the genome





Srivastava S, Avvaru AK, Sowpati DT & Mishra RK (2019): Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics 20: 153.

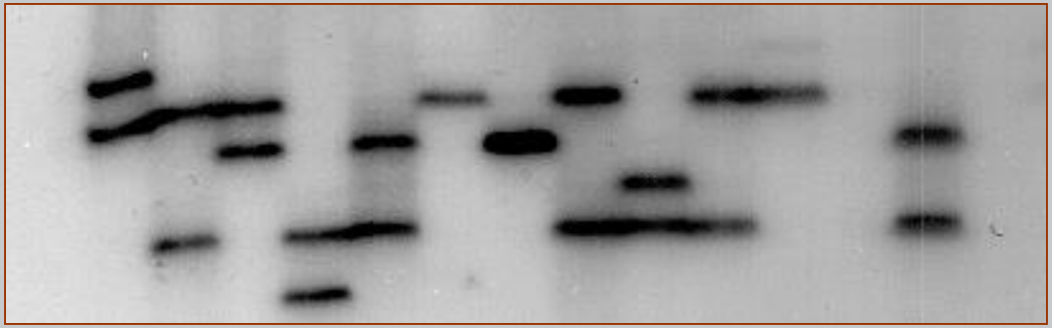
Polymorphism detection



...GTTCTGTCATATATATATATATCGTACTT...

total DNA

PCR – pair of specific primers

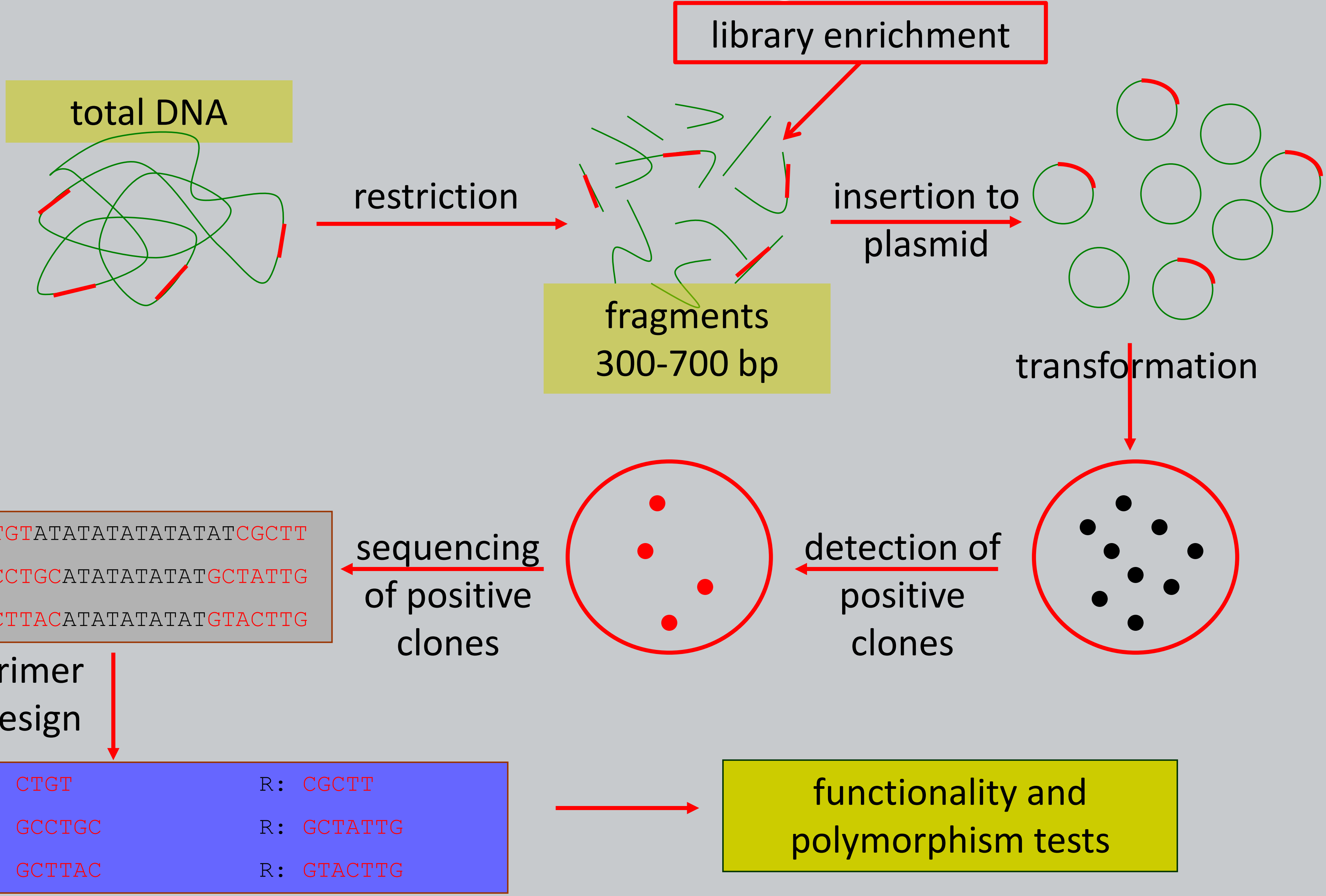


Microsatellite primers

- locus specific – only once in the genome
- species specific

- do exists for the target species (published)
  - web search (SSRs or microsatellites)
  - mined from onekp.com project (Matasci et al. 2014, Hodel et al. 2016)
- search the GenBank – SRA (target enrichment, genome skimming, transcriptomes…)
- test of primers from related species (same genus) – *cross-amplification* – does not work in most cases or problem with null alleles
- necessary to design
  - classical cloning
  - NGS – search for *reads* with microsatellites

Primer development



library enrichment

total DNA

restriction

fragments
300-700 bp

insertion to plasmid

transformation

detection of positive clones

sequencing of positive clones

CTGTATATATATATATATATCGCTT

GCCTGCATATATATATATGCTATTG

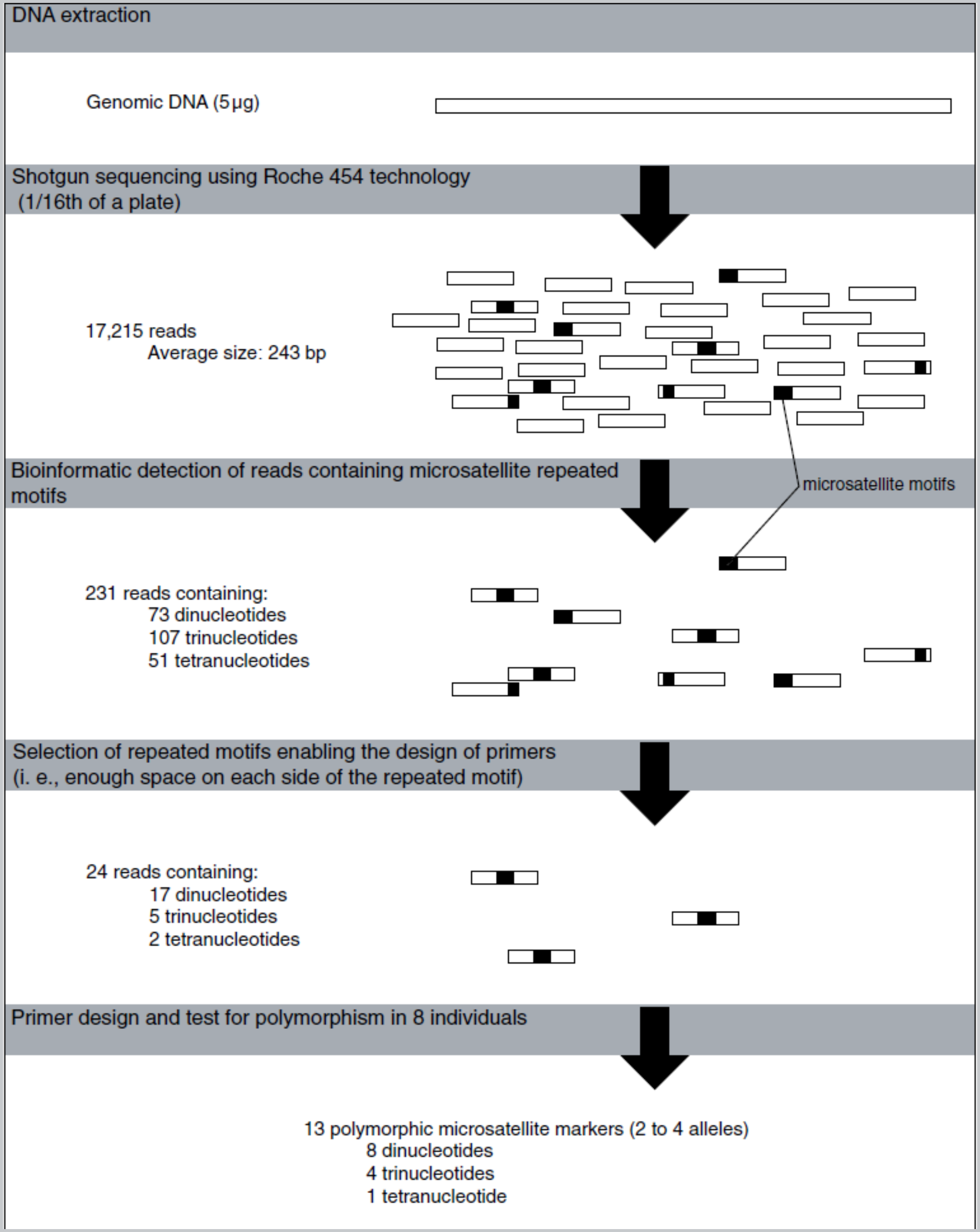
GCTTACATATATATATATGTACTTG

primer design

F: CTGT	R: CGCTT
F: GCCTGC	R: GCTATTG
F: GCTTAC	R: GTACTTG

functionality and polymorphism tests

Primer development – NGS



DNA extraction

Genomic DNA (5µg)

Shotgun sequencing using Roche 454 technology
(1/16th of a plate)

17,215 reads
Average size: 243 bp

Bioinformatic detection of reads containing microsatellite repeated motifs

microsatellite motifs

231 reads containing:
73 dinucleotides
107 trinucleotides
51 tetranucleotides

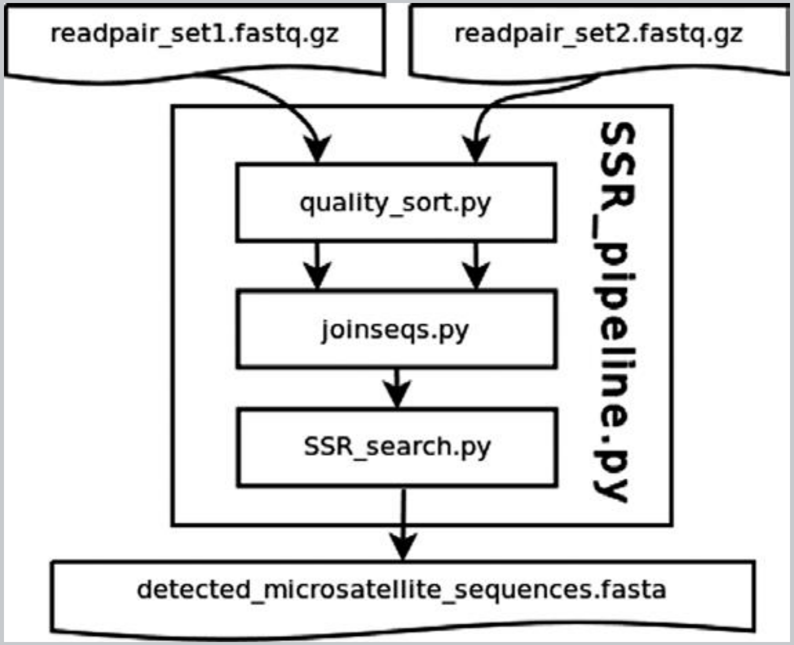
Selection of repeated motifs enabling the design of primers
(i. e., enough space on each side of the repeated motif)

24 reads containing:
17 dinucleotides
5 trinucleotides
2 tetranucleotides

Primer design and test for polymorphism in 8 individuals

13 polymorphic microsatellite markers (2 to 4 alleles)
8 dinucleotides
4 trinucleotides
1 tetranucleotide

Abdelkrim J, Robertson BC, Stanton JL, Gemmell NJ (2009): Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Bio Techniques* 46: 185-192.



readpair_set1.fastq.gz

readpair_set2.fastq.gz

SSR_pipeline.py

quality_sort.py

joinseqs.py

SSR_search.py

detected_microsatellite_sequences.fasta

Miller PM, Knaus BJ, Mullins TD & Haig SM (2014): *SSR_pipeline*: A Bioinformatic Infrastructure for Identifying Microsatellites From Paired-End Illumina High-Throughput DNA Sequencing Data. Journal of Heredity

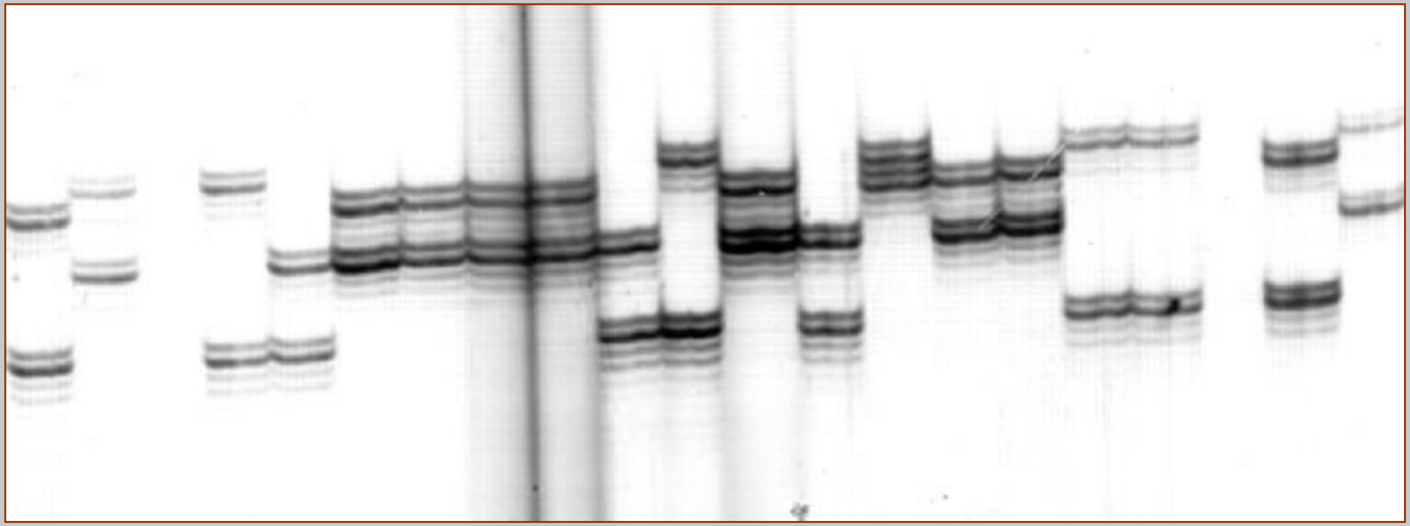
Software for primer development

- identification of potential loci
    - minimum number of repeat unit
    - minimum length of flanking regions
    - (primer design)

- Geneious (+ Phobos, Primer3, MISA plugins)
- GMATo (Wang et al. 2013)
- HighSSR (Churbanov et al. 2012)
- MISA (Thel et al. 2003)
- MSATCMMANDER (Faircloth 2008)
- PAL_FINDER (Castoe et al. 2012)
- QDD3 (Meglécz et al. 2014)
- SSR_pipeline (Miller et al. 2013)

Gel interpretation

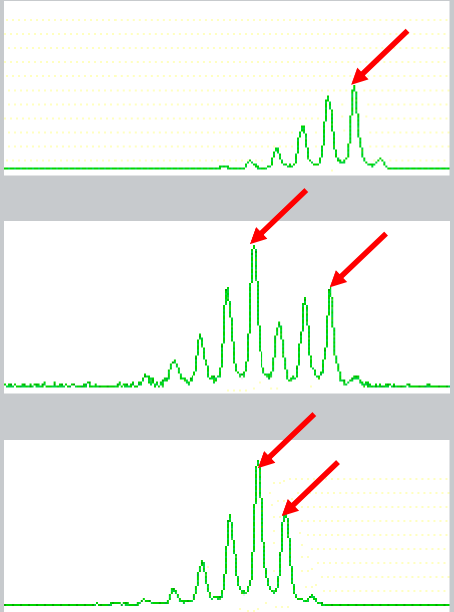
- *„stutter bands“* – additional bands around the band with the right length (most intense) – *in vitro DNA slippage*
- *„terminal transferase activity“* – tendency of *Taq* polymerase to add A at 3´-terminus



Gel interpretation II.

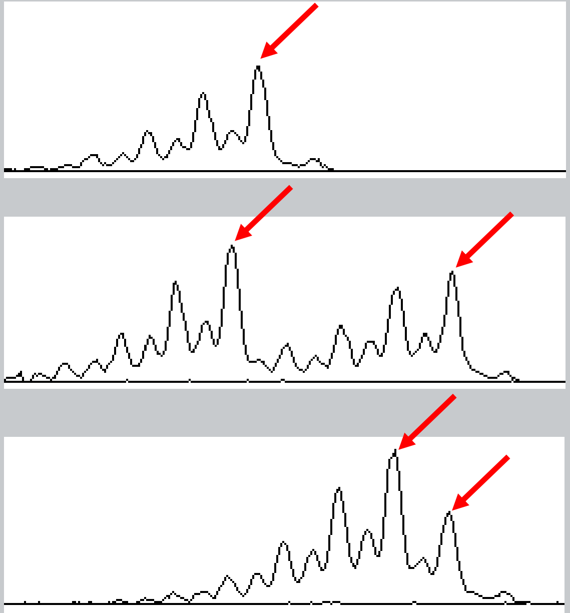
stutter bands

- products by 2, 4, 6 etc. bp shorter
- highest *peak* the longest – the right allele



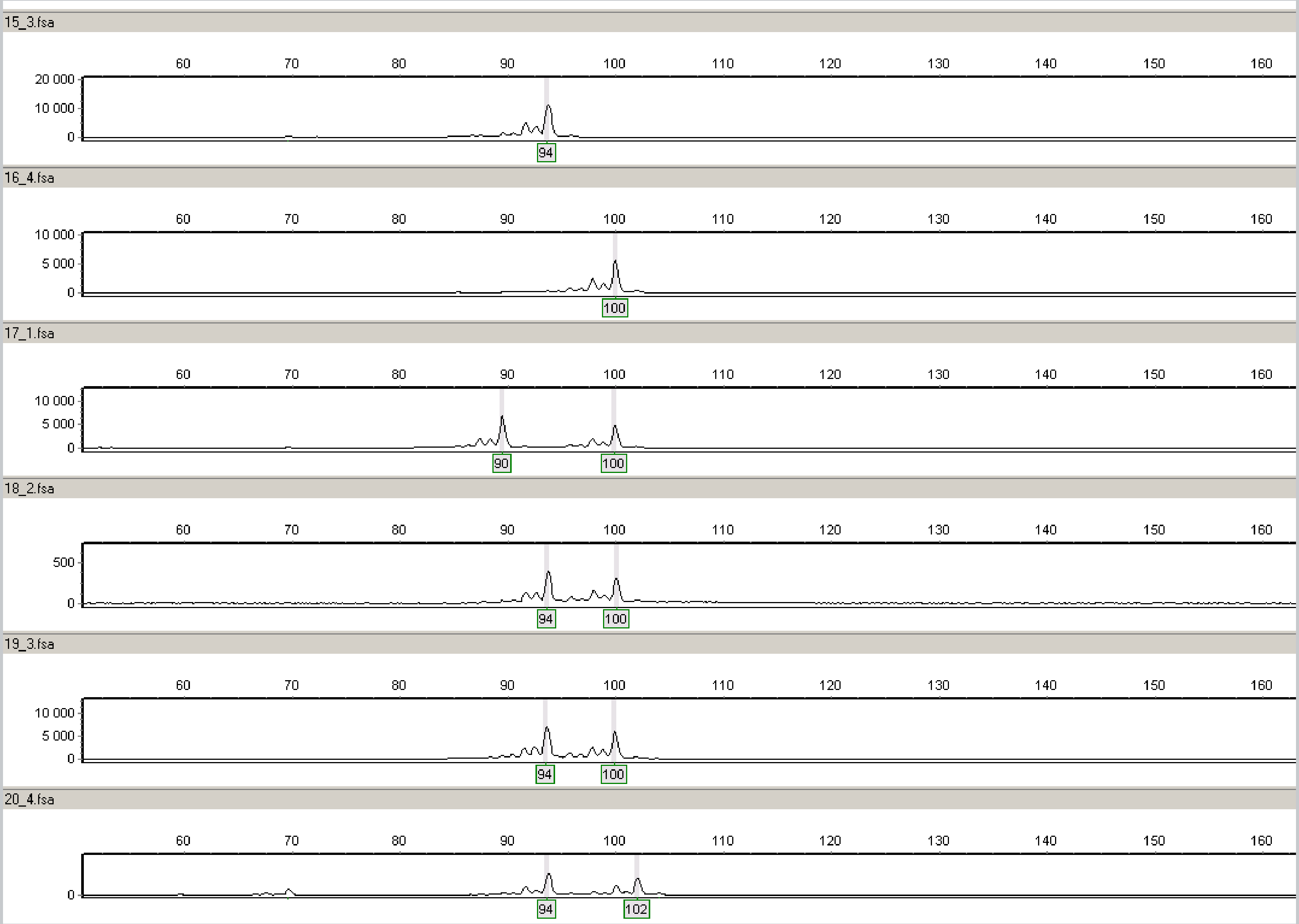
stutter bands and **-A** products

- *stutter bands* by 2, 4, 6 etc. bp shorter
- -A product to each band as well



↙ correct allele

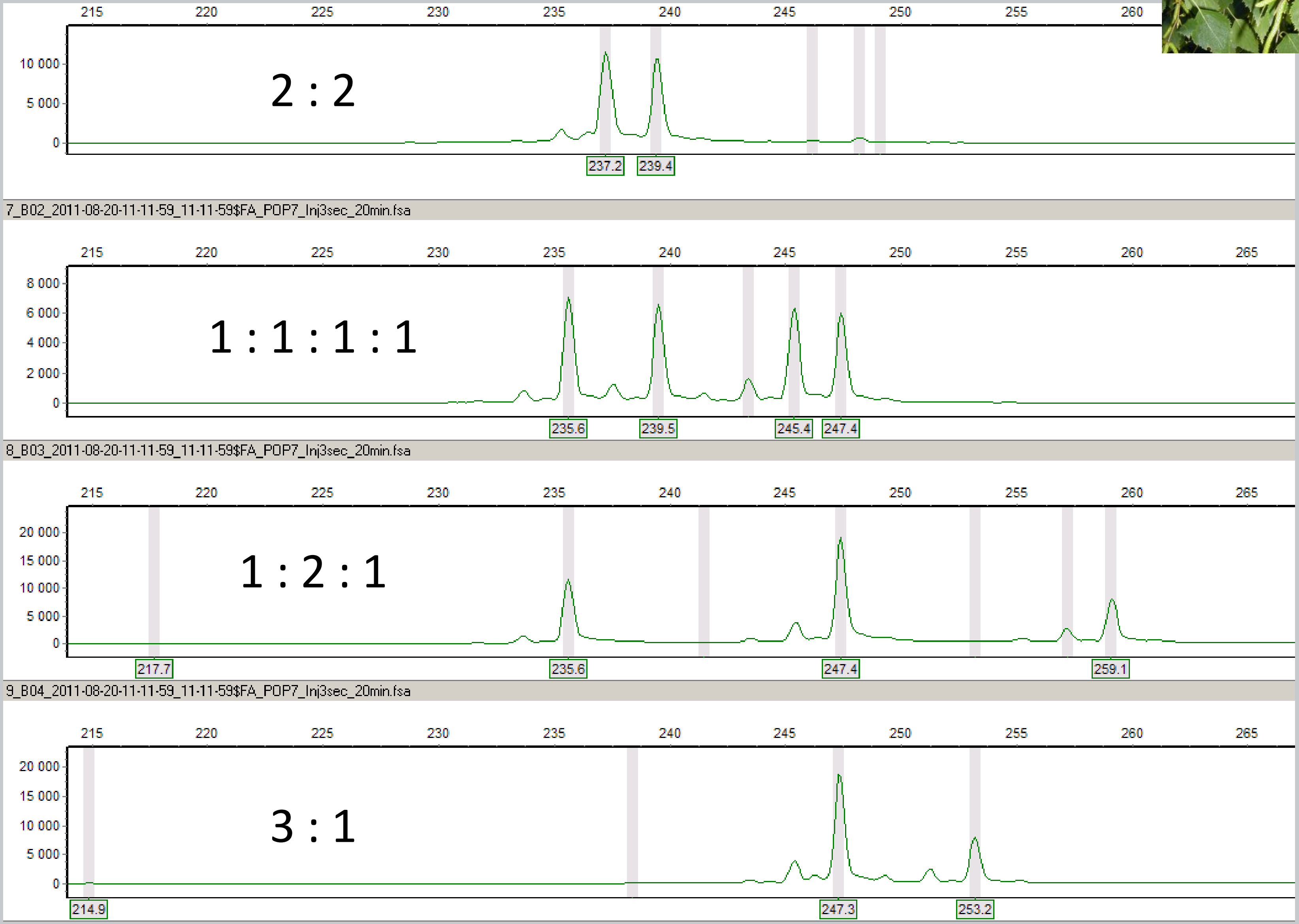
Automatic analysis (GeneMarker)



How to assess tetraploid data

- as dominant data – presence/absence of alleles
- codominantly (we see alleles, but what is the genotype?)
  - three alleles – one is twice but which one? (i.e., treated as 3 alleles + missing)
  - two alleles – each twice ore one of them thrice? (i.e., treated as 2 alleles + 2 missing)
  - problem – large amount of missing data
  - alternative – number of alleles determined from the peak area
- autopolyploids/allopolyploids ?
- software for different ploidy level data analysis – POLYSAT, SPAGeDi, TETRASAT, BAPS, STRUCTURE…

Tetraploid data (*Betula*)



2 : 2

7_B02_2011-08-20-11-11-59_11-11-59$FA_POP7_Inj3sec_20min.fsa

1 : 1 : 1 : 1

8_B03_2011-08-20-11-11-59_11-11-59$FA_POP7_Inj3sec_20min.fsa

1 : 2 : 1

9_B04_2011-08-20-11-11-59_11-11-59$FA_POP7_Inj3sec_20min.fsa

3 : 1

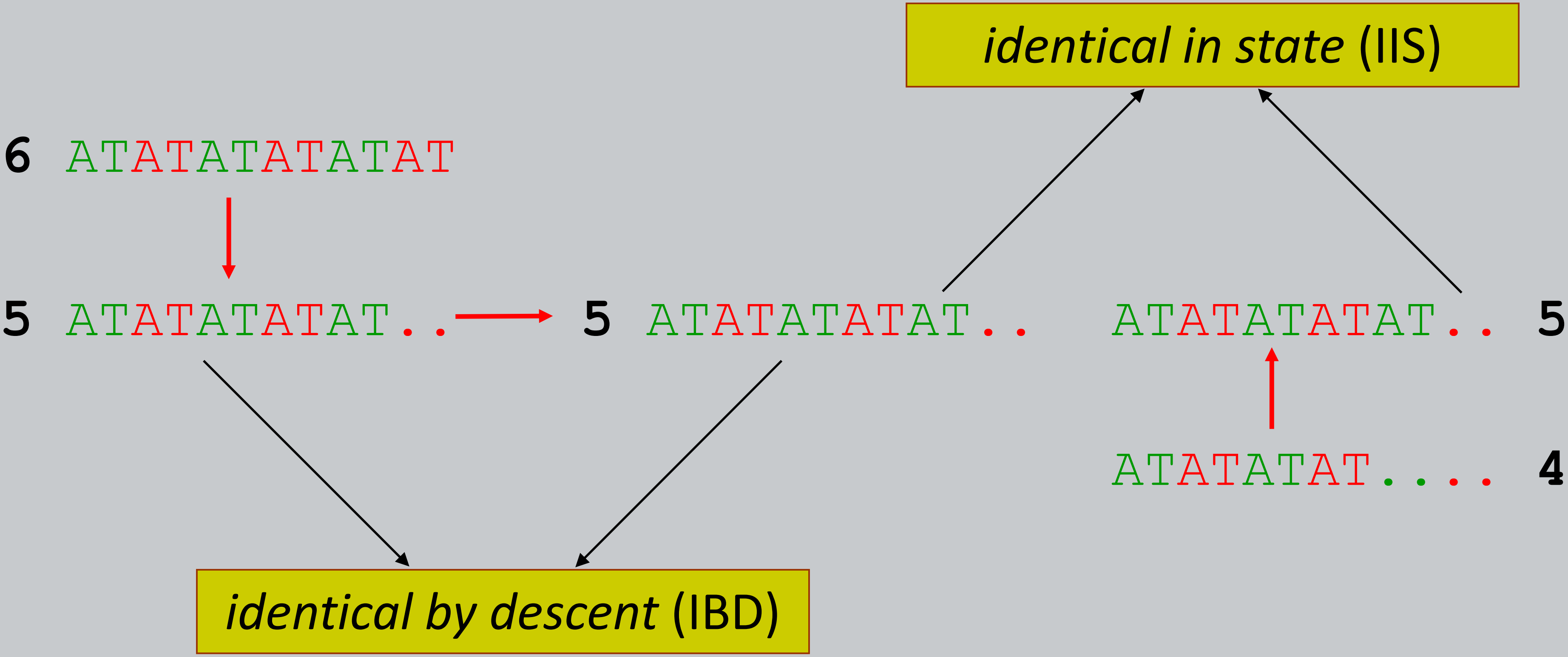
Polymorphism origin

- *DNA „slippage“*
  - DNA polymerase „slips“ during replication
  - extension or reduction the length by one repeat
- *„unequal crossing over“*
  - more extensive changes
- high mutation rate – 10^{-3} - 10^{-5}

Mutation of microsatellites

- mutation rate is estimated to be 10^{-3} – 10^{-5}
    - differs in 2, 3 and 4 bp repeats
    - according to microsatellite type
    - different in different species …
- mutation rate – balance between mutation and their reparation
- mostly – loss or gain of one repeat
- loci with more repeat units and with purer repeats – higher mutation rate

Allele homology



identical in state (IIS)

6 ATATATATATAT

5 ATATATATAT..

5 ATATATATAT..

ATATATATAT.. 5

ATATATAT.... 4

identical by descent (IBD)

Mutation models

- *infinite alleles model* (IAM) – *Kimura & Crow 1964*
  - new allele with mutation rate *u*
  - homoplasy not allowed
  - identical alleles are IBD
- *stepwise mutation model* (SMM) – *Kimura & Ohta 1978*
  - new allele as an addition or loss of just one repeat
  - same probability of gain and loss (*u*/2)
  - generates homoplasy (alleles are not IBD, only IIS)
  - alelles of similar lengths are more related
- *two-phase model* (TPM) – *DiRienzo et al. 1994*
  - modification of one repeat with probability *p*
  - modification of more than one repeat with probability *1-p*

Null alleles

- loss of PCR product due to mutation in *priming site*
- i.e., heterozygosity underestimation – some heterozygotes scored as homozygotes
- identification using a pedigree study – allele not inherited
- frequency is higher when heterologous primers are used (cross-amplification from related species)
- frequency could be estimated based on H-W disequilibrium (i.e., software Cervus)

SSRs and SNPs comparison

SSRs

- every 2-30 kbp
- mutation rate 10^{-3} to 10^{-4}
- high allelic richness
- more private alleles
- higher degree of homoplasy
- limited number of loci

SNPs

- more numerous in the genome (every 100-300 bp)
- mutation rate 10^{-9}
- mainly bi-allelic
- fewer private alleles
- less prone to homoplasy
- many more loci

advantages of SSRs over SNPs

- little ascertainment bias (i.e., systematic deviation from theoretical expectations due to , i.e., nonrandom sampling)
- higher success rate of cross-amplification
- accuracy is easy to assess in pedigree analyses (due to many alleles per locus)

drawbacks of SSRs over SNPs

- large sample sizes needed for accurate estimation of allelic frequencies
- rapid mutation could complicate parentage reconstruction
- poor indicators of long-term population history due to backward mutations
- might not accurately reflect the underlying genomic diversity
- complicated screening (capillary gel electrophoresis)
- need to include common controls among studies

Guichoux E. et al. (2011): Current trends in microsatellite genotyping. Mol. Ecol. Res. 11: 591-611.
Hodel R.G.J. et al. (2016): The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. Appl. Pl. Sci. 4:1600025.

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 coefficients for microsatellites
  - R_{ST} – analogue of F_{ST} (Slatkin 1995)
    - SMM included (*stepwise mutation model* – based on variance in allele lengths)
    - 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)
  - RSTcalc (Goodman 1997)

Application of microsatellites

- parentage analysis
  - parent identification of seeds (seedlings) in populations
  - outcrossing rate
- clone identification
- population-genetic studies
  - inbreeding, H-W equilibrium testing
  - gene flow, migration
  - population history, effective population size changes…
- phylogeography
- systematics
  - problematical application – allele homology?
  - only at the level of closely related species
  - necessary to use many loci (to cover the „whole genome“ variation)
  - cpDNA SSRs
- hybridization
  - possible to distinguish F1 and advanced (F2, B1) hybrids

Parentage analysis

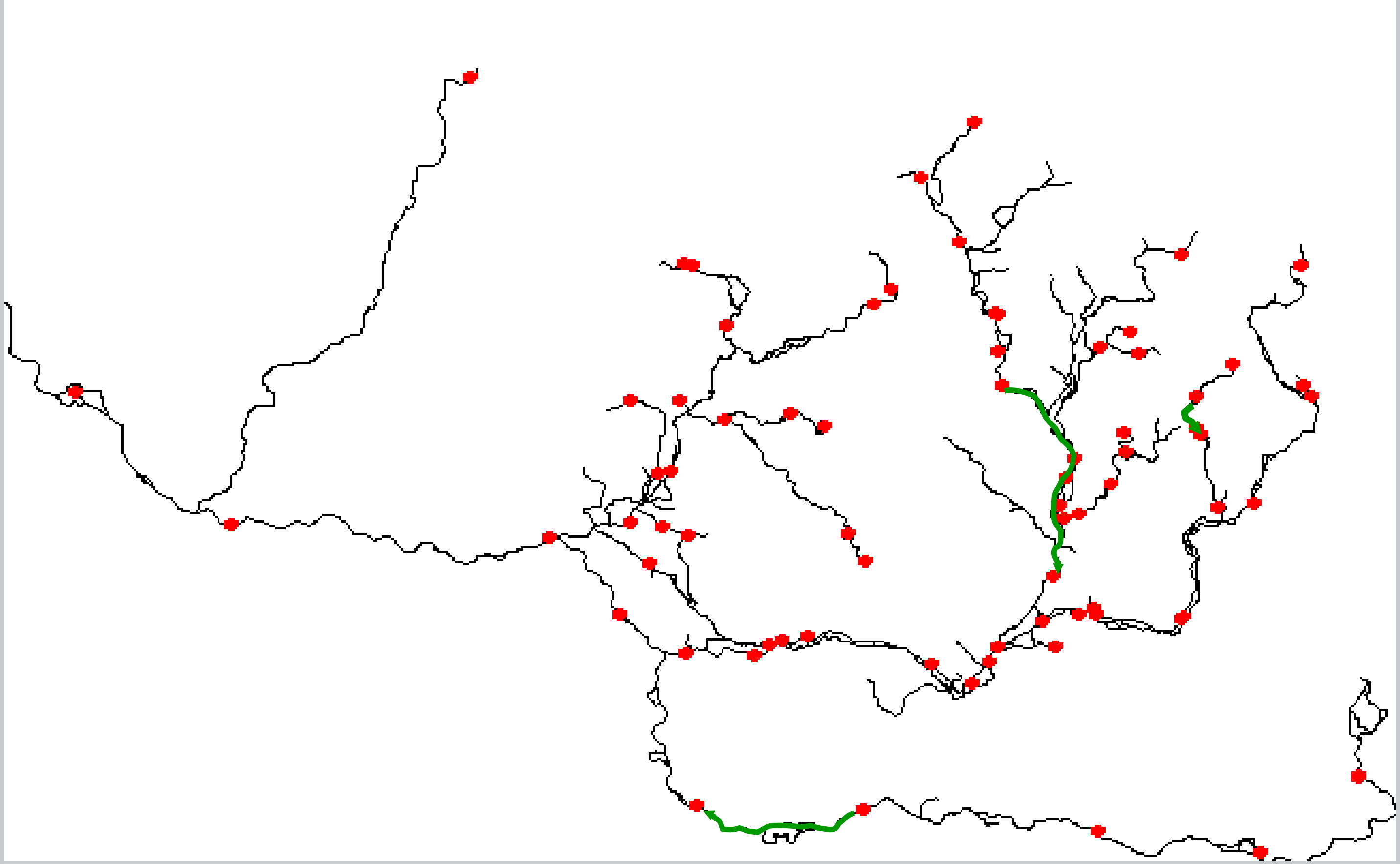
- direct estimate of distance and frequency of dispersal
  - seeds – distances between seeds and their parents
  - pollen – distances between parent pairs
- fitness of particular genotypes in population
  - participation of „individuals-fathers“ at pollination and fertilization
- outcrossing rate
  - % of seeds originated by allogamy
- assumptions
  - genotypes of all potential parents available (relatively low amount of individuals)
  - variable marker – microsatellites, AFLP

Methods of *parentage analysis*

- *exclusion analysis*
  - incompatibility between parental and progeny genotypes → rejection of hypothesis
  - i.e., rejection of all parents but one or two
  - problems – scoring errors, null alleles, mutations
- *categorical allocation*
  - calculation of LOD score (*logarithm of the likelihood ratio*)
  - parents have the highest LOD score
  - advantage – less sensitive to errors and mutations
- software – i.e., CERVUS (Marshall et al. 1998)

Clone identification

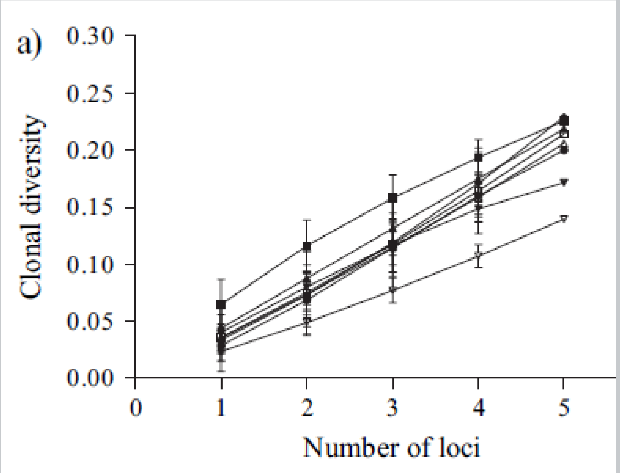
- clone = the same multilocus genotype (i.e., same alleles at all loci)

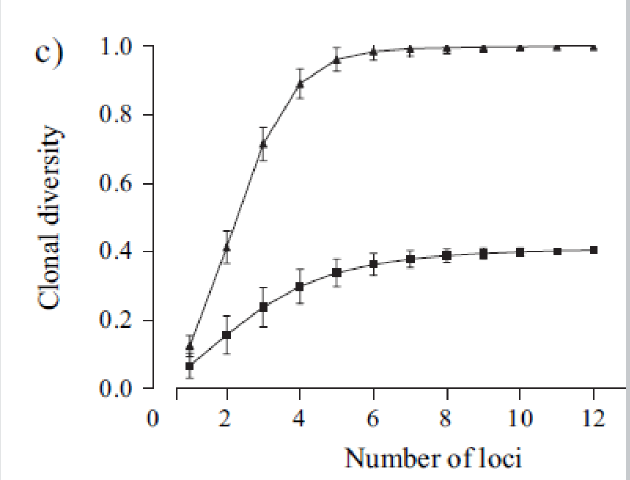


Phragmites australis in the river Labe (Fér & Hroudová 2009)

Clone identification

- take care of discrimination possibility of markers
- *marker power*





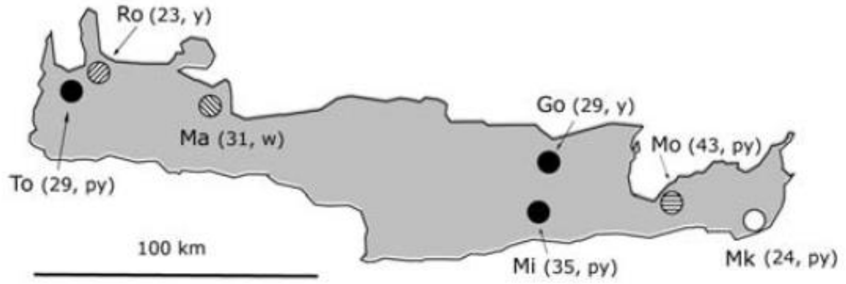
insufficient variability

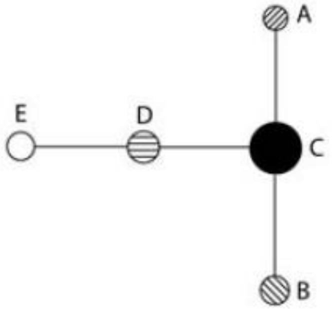
sufficient variability

- MLG (*multilocus genotype*)
  - if found more than ones – P_{sex} calculation, i.e., probability that this MLG could originate just by chance during different generative event – software GenClone, MLGSIM

Arnaud-Haond et al. (2005): Assessing genetic diversity in clonal organisms: Low diversity or low resolution? Combining power and cost efficiency in selecting markers. *Journal of Heredity* 96:434-440
Arnaud-Haond et al. (2007): Standardizing methods to address clonality in population studies. *Molecular Ecology* 16: 5115–5139

Gene flow – indirect estimation





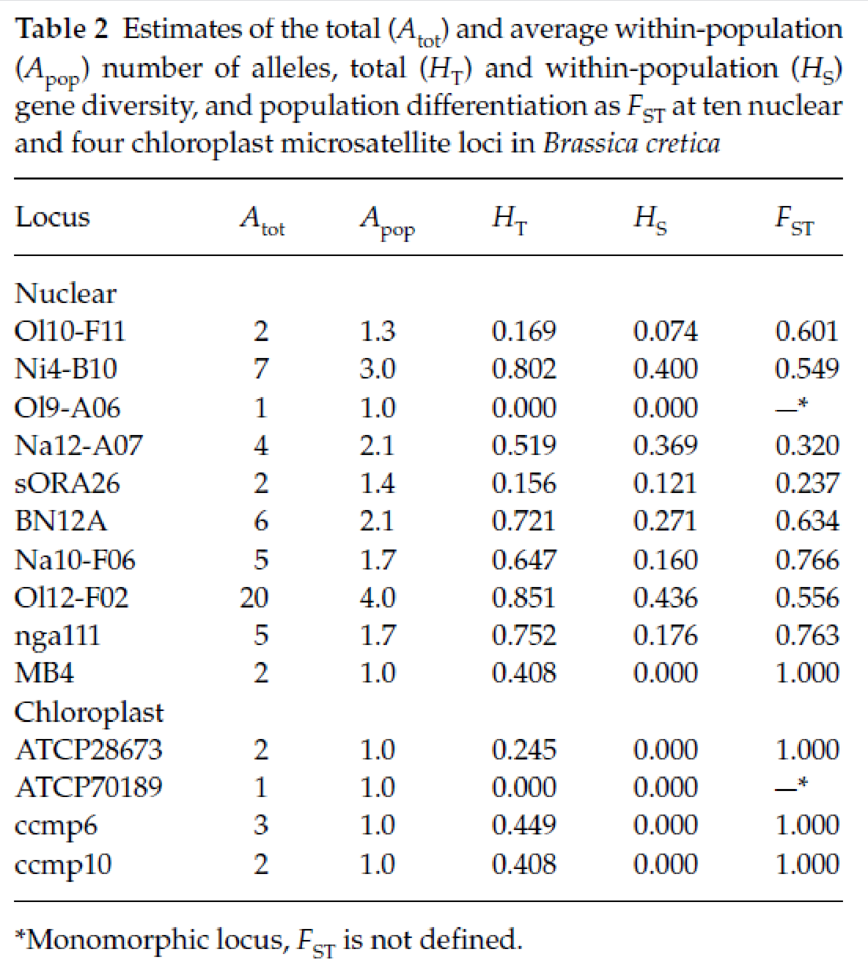


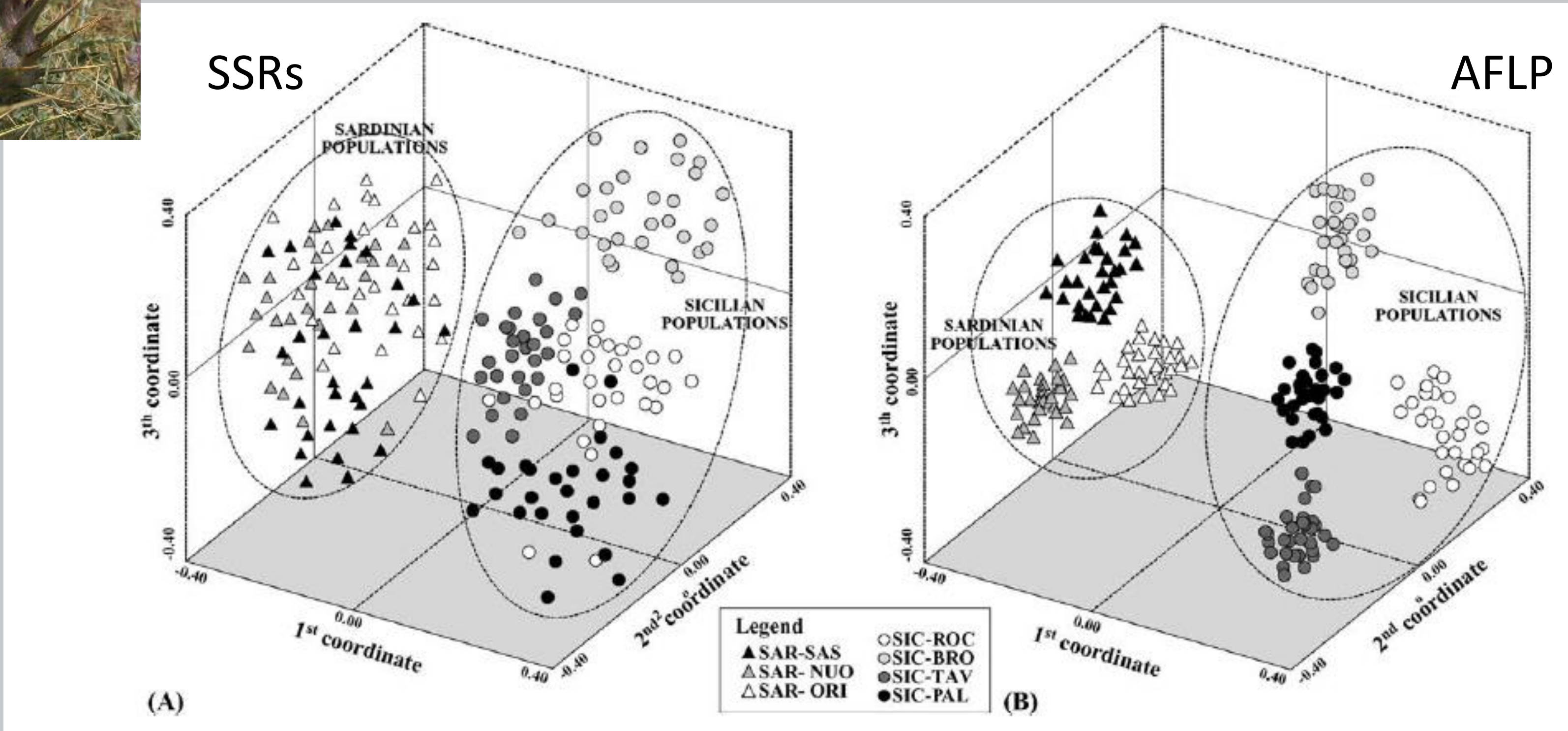
Table 2 Estimates of the total (A_{tot}) and average within-population (A_{pop}) number of alleles, total (H_T) and within-population (H_S) gene diversity, and population differentiation as F_{ST} at ten nuclear and four chloroplast microsatellite loci in *Brassica cretica*

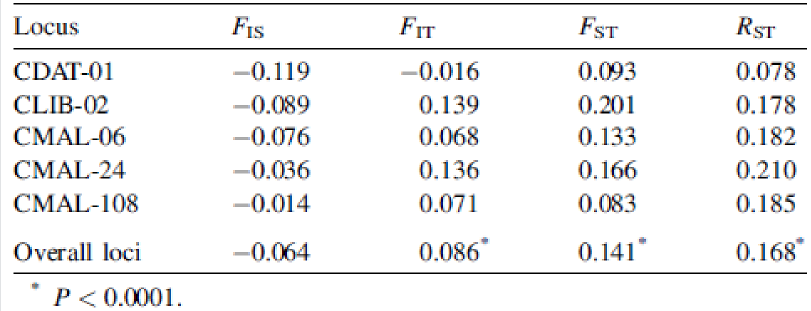
Locus	A_{tot}	A_{pop}	H_T	H_S	F_{ST}
Nuclear					
Ol10-F11	2	1.3	0.169	0.074	0.601
Ni4-B10	7	3.0	0.802	0.400	0.549
Ol9-A06	1	1.0	0.000	0.000	—*
Na12-A07	4	2.1	0.519	0.369	0.320
sORA26	2	1.4	0.156	0.121	0.237
BN12A	6	2.1	0.721	0.271	0.634
Na10-F06	5	1.7	0.647	0.160	0.766
Ol12-F02	20	4.0	0.851	0.436	0.556
nga111	5	1.7	0.752	0.176	0.763
MB4	2	1.0	0.408	0.000	1.000
Chloroplast					
ATCP28673	2	1.0	0.245	0.000	1.000
ATCP70189	1	1.0	0.000	0.000	—*
ccmp6	3	1.0	0.449	0.000	1.000
ccmp10	2	1.0	0.408	0.000	1.000

*Monomorphic locus, F_{ST} is not defined.

Edh K. et al. (2007): Nuclear and chloroplast microsatellites reveal extreme population differentiation and limited gene flow in the Aegean endemic *Brassica cretica* (Brassicaceae). Mol. Ecol. 16, 4972-4983.

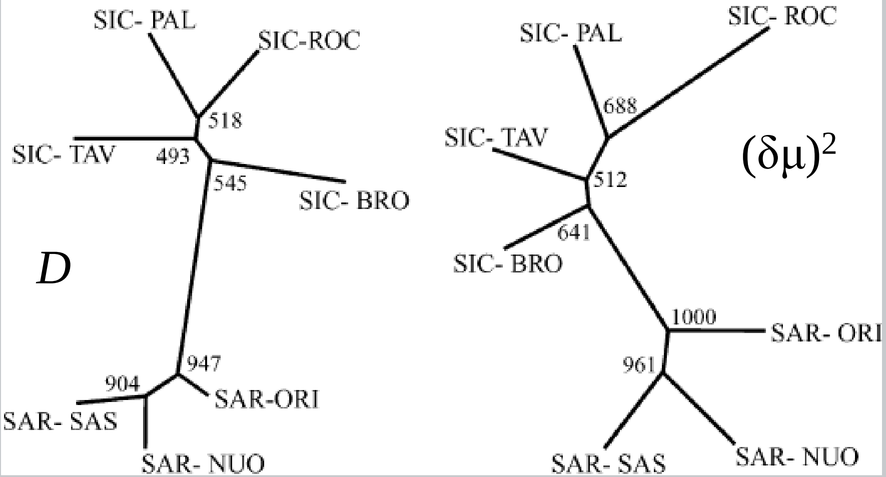
Cynara cardunculus – 5 loci





Locus	F_{IS}	F_{IT}	F_{ST}	R_{ST}
CDAT-01	−0.119	−0.016	0.093	0.078
CLIB-02	−0.089	0.139	0.201	0.178
CMAL-06	−0.076	0.068	0.133	0.182
CMAL-24	−0.036	0.136	0.166	0.210
CMAL-108	−0.014	0.071	0.083	0.185
Overall loci	−0.064	0.086*	0.141*	0.168*

* $P < 0.0001$.

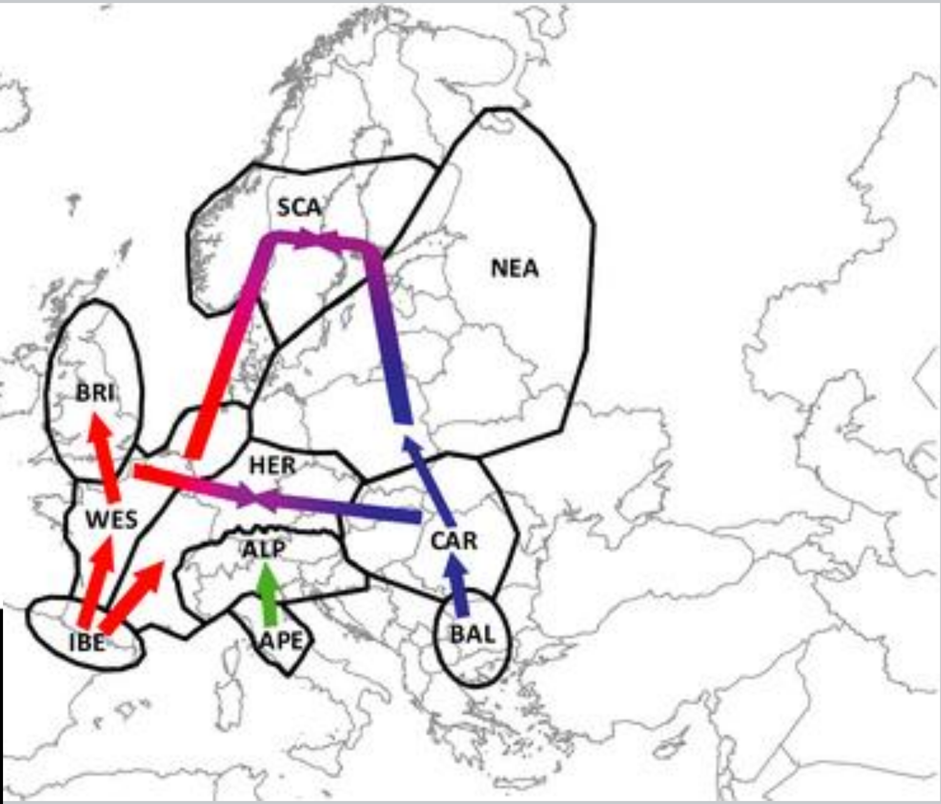


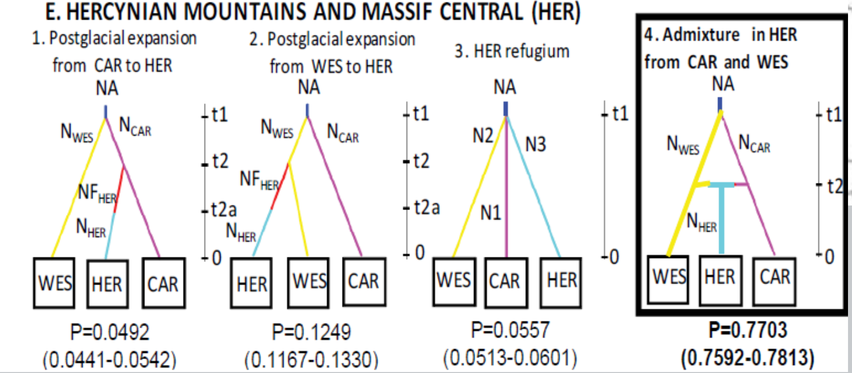
Portis et al. 2005

Phylogeography

testing alternative migration hypotheses

- ABC – approximate Bayesian computation





Alnus glutinosa (Mandák et al. 2015)

Phylogeny inference

Hordeum
cpDNA microsatellites

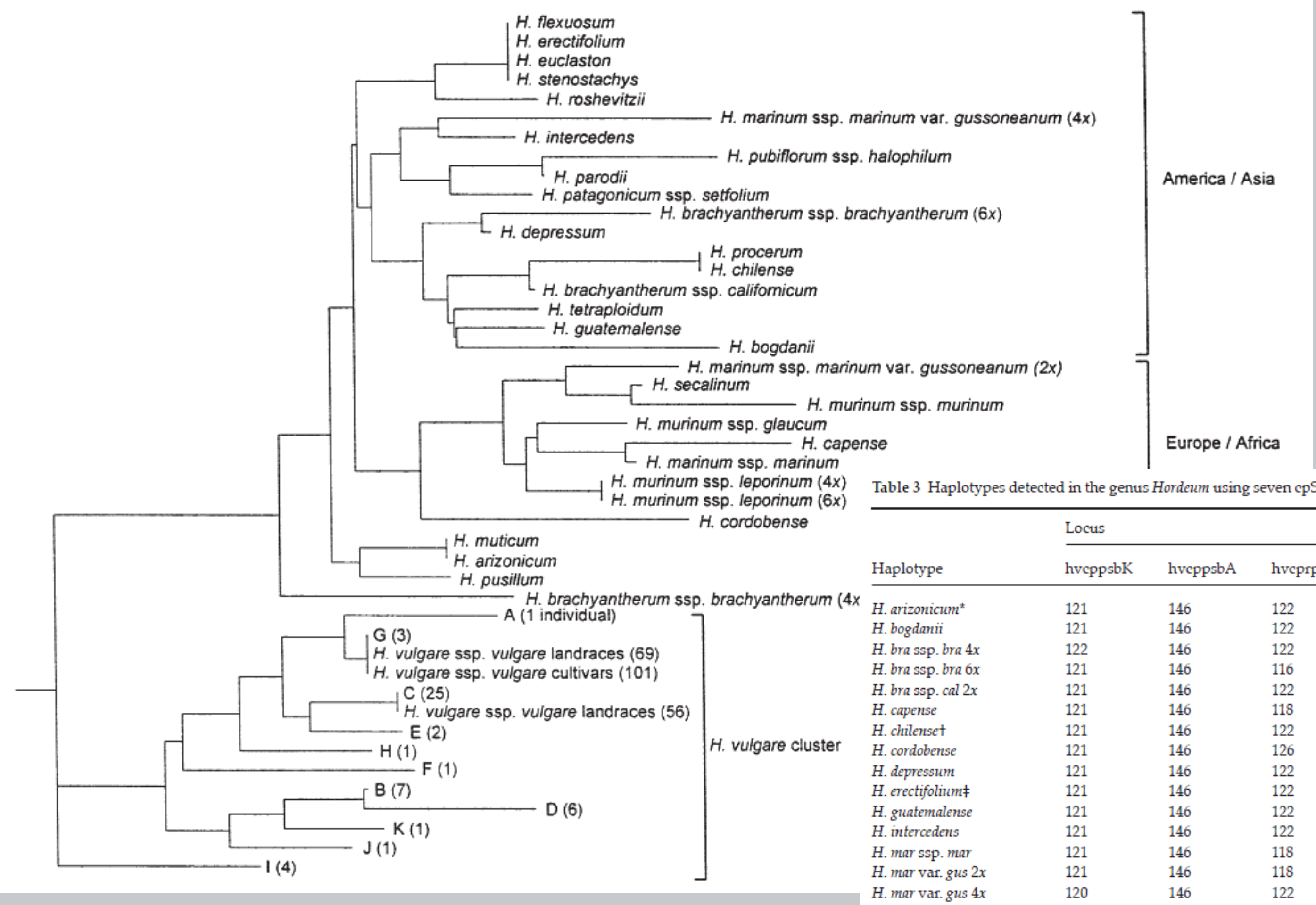

H. flexuosum
H. erectifolium
H. euclaston
H. stenostachys
H. roshevitzii
H. marinum ssp. marinum var. gussoneanum (4x)
H. intercedens
H. pubiflorum ssp. halophilum
H. parodii
H. patagonicum ssp. setfolium
H. brachyantherum ssp. brachyantherum (6x)
H. depressum
H. procerum
H. chilense
H. brachyantherum ssp. californicum
H. tetraploidum
H. guatemalense
H. bogdanii
America / Asia
H. marinum ssp. marinum var. gussoneanum (2x)
H. secalinum
H. murinum ssp. murinum
H. murinum ssp. glaucum
H. capense
H. marinum ssp. marinum
Europe / Africa
H. murinum ssp. leporinum (4x)
H. murinum ssp. leporinum (6x)
H. cordobense
H. muticum
H. arizonicum
H. pusillum
H. brachyantherum ssp. brachyantherum (4x
A (1 individual)
G (3)
H. vulgare ssp. vulgare landraces (69)
H. vulgare ssp. vulgare cultivars (101)
C (25)
H. vulgare ssp. vulgare landraces (56)
E (2)
H (1)
F (1)
B (7)
D (6)
K (1)
J (1)
I (4)
H. vulgare cluster

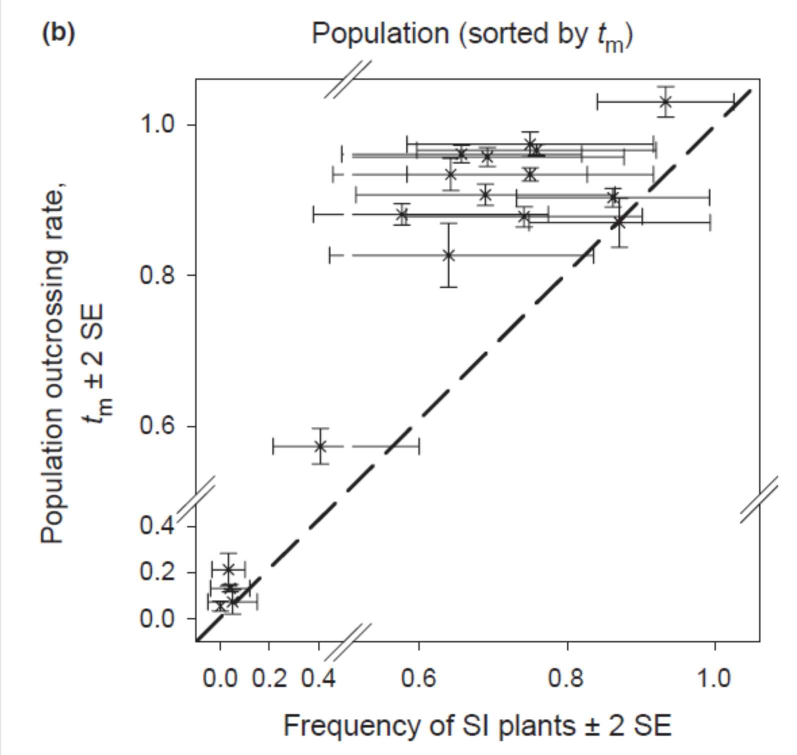
Table 3 Haplotypes detected in the genus *Hordeum* using seven cpSSRs

	Locus						
Haplotype	hvcppsbK	hvcppsbA	hvcprpoA	hvcprps12	hvcptrnS1	hvcptrnS2	hvcptrnLF
*H. arizonicum**	121	146	122	148	128	115	101
H. bogdanii	121	146	122	148	135	102	99
H. bra ssp. *bra* 4x	122	146	122	148	128	112	101
H. bra ssp. *bra* 6x	121	146	116	148	128	116	99
H. bra ssp. *cal* 2x	121	146	122	148	128	114	99
H. capense	121	146	118	148	135	113	101
H. chilense†	121	146	122	152	128	114	99
H. cordobense	121	146	126	148	128	114	103
H. depressum	121	146	122	148	128	116	99
H. erectifolium‡	121	146	122	148	128	114	100
H. guatemalense	121	146	122	148	128	113	99
H. intercedens	121	146	122	148	128	116	101
H. mar ssp. *mar*	121	146	118	148	128	113	101
H. mar var. *gus* 2x	121	146	118	148	128	112	102
H. mar var. *gus* 4x	120	146	122	148	128	116	97
H. mur ssp. *mur*	120	146	118	148	128	112	101
H. mur ssp. *gla*	121	146	118	148	128	115	101
H. mur ssp. *lep* 4x§	121	146	118	148	128	114	101
H. parodii	121	146	122	148	128	115	102
H. pat ssp. *set*	121	146	122	148	128	116	102
H. procerum	121	146	122	152	128	114	99
H. pub ssp. *hal*	121	145	122	148	128	115	102
H. pusillum	121	146	122	148	128	113	101
H. roshevitzii	121	146	122	148	128	103	100
H. secalinum	121	146	118	148	128	112	101
H. stenostachys	121	146	122	148	128	114	100
H. tetraploidum	121	146	122	148	128	115	99

Provan J. et al. (1999): Polymorphic chloroplast simple sequence repeats for systematic and population studies in the genus *Hordeum*. Molecular Ecology 8, 505-511.

Self-(in)compatibility

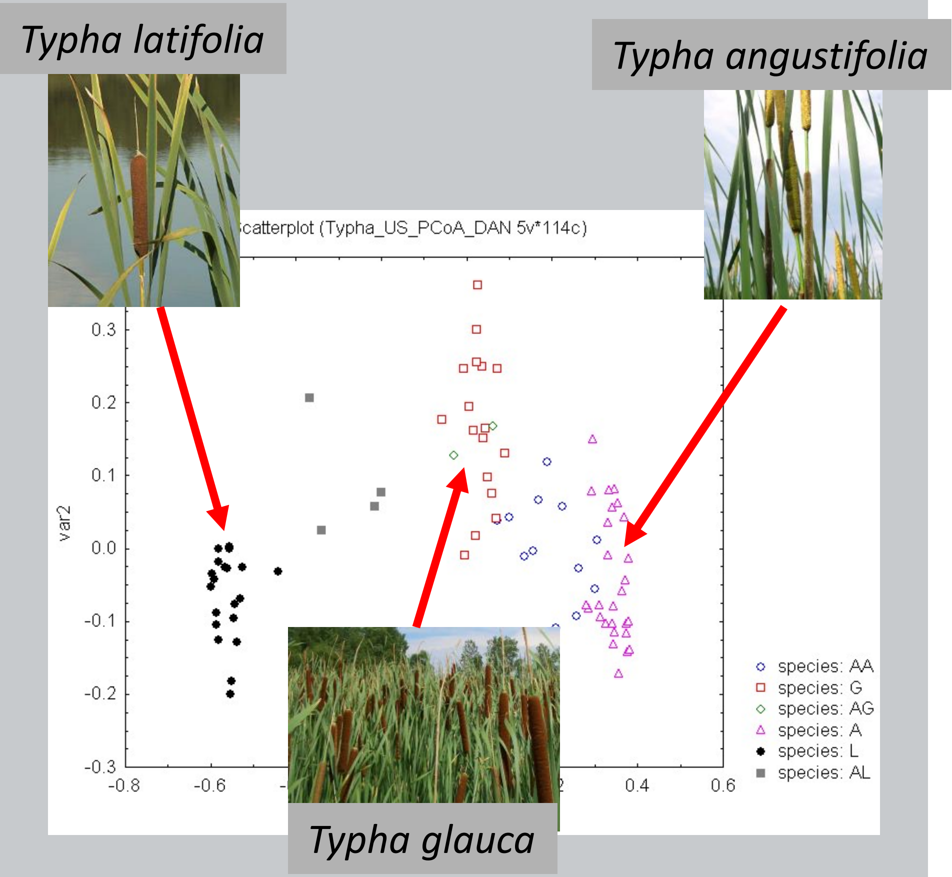
- % of seeds originated by allogamy, i.e. in parentage analysis is first and second parent the same
- *outcrossing rate*



Willi Y. & Määttänen K. (2010): Evolutionary dynamics of mating system shifts in *Arabidopsis lyrata*. Journal of Evolutionary Biology 23: 2123–2131.

Hybridization

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

Typha angustifolia

Typha glauca

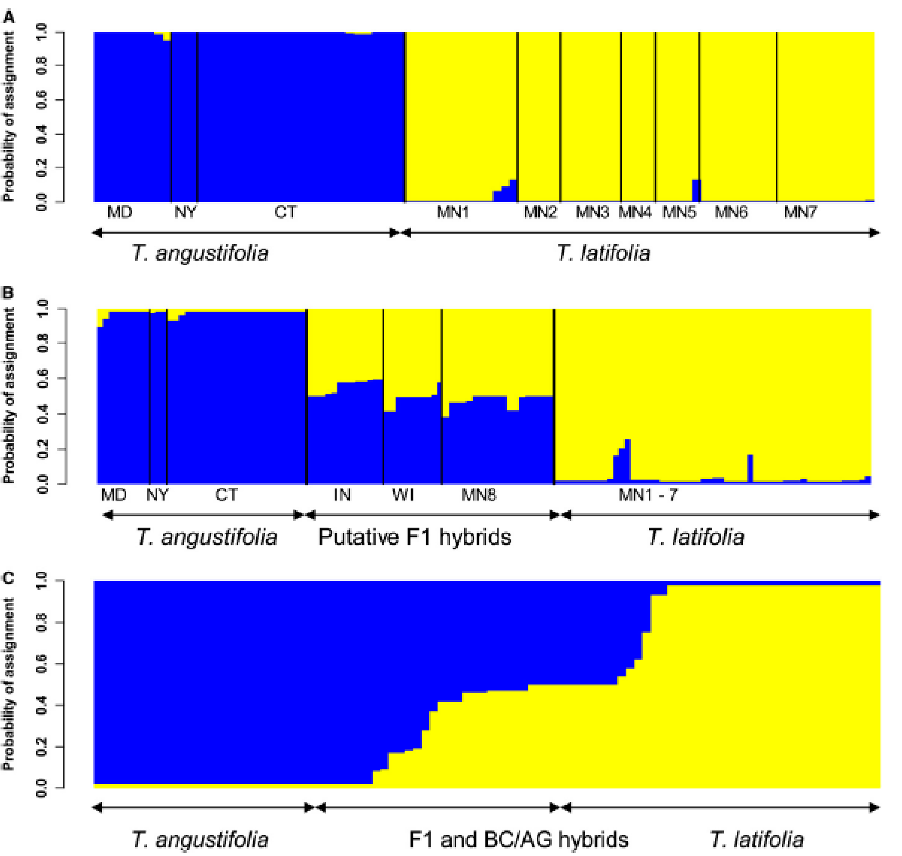


Fig. 1. Barplots of STRUCTURE analyses for (A) RAPD-identified *T. angustifolia* and *T. latifolia* from Collection Group 1; (B) RAPD-identified *T. angustifolia*, *T. latifolia*, and putative F_1 hybrids from additional Collection Group 1 sites in Indiana, Wisconsin, and Minnesota; and (C) mixed populations in Michigan (Collection Group 2) that included advanced-generation/backcrossed individuals (AG/BC). Each bar represents a single individual. Locations sampled are listed in Tables 1 and 2. Six SSR loci were used in these STRUCTURE analyses (TA 21 was not included).

Snow et al. 2010

Population study

Kameyama Y. et al. (2001): Patterns and levels of gene flow in *Rhododendron metternichii* var. *hondoense* revealed by microsatellite analysis. *Molecular Ecology* 10:205-216

Systematic study

Provan J. et al. (1999): Polymorphic chloroplast simple sequence repeat primers for systematic and population studies in the genus *Hordeum*. *Molecular Ecology* 8:505-511

Literature

Vieira M.L.C. et al. (2016): *Microsatellite markers: what they mean and why they are so useful*. Genetics and Molecular Biology 39: 312-328

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

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

Kantartzi S.K. (2013): *Microsatellites. Methods and Protocols*. Springer

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

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

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.

Clark L.V. & Jasieniuk M (2011): *POLYSAT: an R package for polyploid microsatellite analysis*. Molecular Ecology Resources 11: 562-566.