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

...GTTCTGTCATATATATATATATATATATATATAT----CGTACTT...

...GTTCTGTCATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATAT---CGTACTT...

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

• simple
  ...CACACACACACACACACACACACACACACAC... 

• compound
  ...CACACACACA\textcolor{green}{TGTGTGTGTGTG}... 

• interrupted
  ...CACACA\textcolor{red}{TTCACACACATTTCACA}...
Repetitive sequences

• dinucleotides
  • AT repeat most common in plants
  • every 30-50 kb
  • number of repeats up to 30

• trinucleotides
  • occurs 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*)

...GTTCTGTCATATATATATATATATATATATATATATGCATCTTA...
Distribution in the genome

- distributed throughout the whole genome (BUT – reflects variability of the studied loci, i.e., restricted 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
  - repeats of one base – i.e., \((T)_{12}\)
  - *flanking regions* – less variable – possible to design consensual primers
Polymorphism detection

...GTTCTGTCATATATATATATATATCGTACTT...

total DNA

PCR – pair of specific primers
Microsatellite primers

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

- do exists for the target species (published)
  - (i.e., in *Molecular Ecology Resources*, formerly *ME Notes*)
    (see also database at http://tomato.bio.trinity.edu)
  - summary – http://botany.natur.cuni.cz/dna (Primery/Mikrosatelitové primery)
  - mined from onekp.com project (Matasci et al. 2014, Hodel et al. 2016)

- search the GeneBank – 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

1. **Total DNA**
2. **Restriction**
   - Fragments 300-700 bp
3. **Insertion to Plasmid**
4. **Transformation**
5. **Library Enrichment**
6. **Primer Design**
   - F: CTGT
   - R: CGCTT
   - F: GCCTGC
   - R: GCTATTTG
   - F: GCTTAC
   - R: GTACTTG
7. **Sequencing of Positive Clones**
8. **Detection of Positive Clones**
9. **Functionality and Polymorphism Tests**
Primer development – NGS


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 or 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)
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

6 ATATATATATAT

5 ATATATATATAT

5 ATATATATATAT.. 5 ATATATATATAT.. ATATATATATAT.. 5

ATATATATATAT.. ATATATATATAT.. ATATATATATATAT.. 5

ATATATATATATATAT

identical by descent (IBD)

identical in state (IIS)
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)
  - alleles 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
- 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

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 – $\rho_{ST}$ (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 $\rightarrow$ 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*

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


Gene flow – indirect estimation

Cynara cardunculus – 5 loci

SSRs

(A)

AFLP

(B)

Portis et al. 2005

<table>
<thead>
<tr>
<th>Locus</th>
<th>$F_{IS}$</th>
<th>$F_{TR}$</th>
<th>$F_{ST}$</th>
<th>$R_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDAT-01</td>
<td>-0.119</td>
<td>-0.016</td>
<td>0.093</td>
<td>0.078</td>
</tr>
<tr>
<td>CLIB-02</td>
<td>-0.089</td>
<td>0.139</td>
<td>0.201</td>
<td>0.178</td>
</tr>
<tr>
<td>CMAL-06</td>
<td>-0.076</td>
<td>0.068</td>
<td>0.133</td>
<td>0.182</td>
</tr>
<tr>
<td>CMAL-24</td>
<td>-0.036</td>
<td>0.136</td>
<td>0.166</td>
<td>0.210</td>
</tr>
<tr>
<td>CMAL-108</td>
<td>-0.014</td>
<td>0.071</td>
<td>0.083</td>
<td>0.185</td>
</tr>
<tr>
<td>Overall loci</td>
<td>-0.064</td>
<td>0.086*</td>
<td>0.141*</td>
<td>0.168*</td>
</tr>
</tbody>
</table>

* $P < 0.0001.$
Phylogeography

testing alternative migration hypotheses

- ABC – approximate Bayesian computation

*Alnus glutinosa* (Mandák et al. 2015)
Phylogeny inference

Self-(in)compatibility

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

Hybridization

<table>
<thead>
<tr>
<th>Species</th>
<th>176</th>
<th>176</th>
<th>278</th>
<th>278</th>
<th>176</th>
<th>190</th>
<th>269</th>
<th>269</th>
<th>179</th>
<th>179</th>
<th>93</th>
<th>93</th>
<th>278</th>
<th>278</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. latifolia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. angustifolia</td>
<td>210</td>
<td>210</td>
<td>286</td>
<td>286</td>
<td>196</td>
<td>196</td>
<td>287</td>
<td>287</td>
<td>193</td>
<td>193</td>
<td>101</td>
<td>101</td>
<td>280</td>
<td>280</td>
</tr>
<tr>
<td>T. x glauca</td>
<td>180</td>
<td>210</td>
<td>278</td>
<td>286</td>
<td>190</td>
<td>196</td>
<td>269</td>
<td>287</td>
<td>179</td>
<td>193</td>
<td>93</td>
<td>101</td>
<td>278</td>
<td>280</td>
</tr>
<tr>
<td>advanced hybrid</td>
<td>176</td>
<td>210</td>
<td>278</td>
<td>286</td>
<td>190</td>
<td>196</td>
<td>287</td>
<td>287</td>
<td>179</td>
<td>193</td>
<td>93</td>
<td>101</td>
<td>278</td>
<td>280</td>
</tr>
</tbody>
</table>

Snow et al. 2010
Population study

Systematic study

Literature


