1. Introduction – overview of the methods

Tomáš Fér

tomas.fer@natur.cuni.cz
What are molecular markers?

- Information about an organism obtained from analysis of its molecules – proteins, DNA

- Marker – information unit – targeted or randomly chosen part of the total information

- Markers tell about genetic similarity (kinship) of individuals, populations or species
Marker examples

- information about enzyme molecule (e.g., its charge and mobility)

- sequence of nucleotides in DNA chain or specific (variable) position (SNP)

- length of DNA fragment
Aspects of molecular data

• give an information about individual genotype
• information not dependent on environment conditions (no plasticity)
• assumption about selective neutrality – i.e., no influence on individual fitness
  • non-coding DNA regions
  • identical functionality of all isozymes
• qualitative information – fragment presence, allele, nucleotide
• unique information about organism – clone identification
• subject to modification during generative reproduction – recombination
Deposition of genetic information

- **nucleus**
  - different ploidy level
  - recombination
  - biparental transfer

- **plastids**
  - 1 circular molecule
  - without recombination
  - uniparental transfer

- **mitochondrion**
  - structurally complicated
  - common restructuring
  - uniparental transfer
## Genome characteristics

<table>
<thead>
<tr>
<th></th>
<th>nDNA animals</th>
<th>plants</th>
<th>cpDNA</th>
<th>mtDNA animals</th>
<th>plants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>heritability</strong></td>
<td>biparental</td>
<td>biparental</td>
<td>angiosperms maternal, conifers paternal</td>
<td>maternal</td>
<td>maternal</td>
</tr>
<tr>
<td><strong>structure</strong></td>
<td>linear</td>
<td>linear</td>
<td>circular</td>
<td>circular</td>
<td>circular, complex</td>
</tr>
<tr>
<td><strong>size (kb)</strong></td>
<td>$4.9\times10^4 - 7.0\times10^8$</td>
<td>$5.0\times10^4 - 3.0\times10^8$</td>
<td>$71 - 214$</td>
<td>$15 - 20$</td>
<td>$200 - 2400$</td>
</tr>
<tr>
<td><strong>substitution rate</strong></td>
<td>$3.5\times10^{-9}$</td>
<td>$4.1 - 5.7\times10^{-9}$</td>
<td>$0.86 - 1.20\times10^{-9}$</td>
<td>$56\times10^{-9}$</td>
<td>$0.36 - 0.50\times10^{-9}$</td>
</tr>
<tr>
<td><strong>substitution rate relative to plant mtDNA</strong></td>
<td>8.1</td>
<td>11.4</td>
<td>2.4</td>
<td>130.2</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>foreign sequences</strong></td>
<td>common</td>
<td>common</td>
<td>rare</td>
<td>rare</td>
<td>common</td>
</tr>
<tr>
<td><strong>structural mutations</strong></td>
<td>common</td>
<td>common</td>
<td>rare</td>
<td>rare</td>
<td>common</td>
</tr>
<tr>
<td><strong>recombination</strong></td>
<td>yes</td>
<td>yes</td>
<td>intramolecular</td>
<td>no</td>
<td>inter- and intramolecular</td>
</tr>
</tbody>
</table>

Lowe et al. 2004
Molecular markers overview

1. proteins – isozymes

2. DNA markers

- RFLP (Restriction Fragment Length Polymorphism)
- PCR based – analysis of DNA fragments
  - order of nucleotides – DNA sequences
  - „whole genome“ analysis – fragment length polymorphism
    - RAPD (Random Amplified Polymorphic DNA)
    - AFLP (Amplified Fragment Length Polymorphism)
    - ISSRs (Inter Simple Sequence Repeats)
  - information from specific genome regions
    - PCR-RFLP (Polymerase Chain Reaction – RFLP)
    - microsatellites (Simple Sequence Repeats – SSRs)
    - SSCP (Single Strain Conformation Polymorphism)
- whole genome markers – SNP, whole genome sequencing
  - RADseq
  - HybSeq
Isoenzymes (isozymes, allozymes)

- proteins catalysing basic biochemical reactions
- extracted from living tissues – preferably leaves
- individual molecules (=alleles) electrophoretically separated according to differences in their mobility (electric charge)
- visualisation by „color reactions“

- limited variation
+ codominant marker

1st locus
1 allele

2nd locus
5 alleles
Isoenzyme gel example
RFLP

Restriction Fragment Length Polymorphism

- DNA is specifically cleaved to fragments by restriction endonuclease
- electrophoresis – length separation
- large amount of fragments – specific part is visualized by hybridization with labelled probe (Southern blotting) – e.g., visualization of cpDNA only
- variability – insertion/deletion or mutation in restriction site

+ highly reproducible pattern
- labelled probes required
RFLP gels examples

http://www.ufpe.br/biolmol/Tec-mol-biol/RFLP-real.JPG
PCR-RFLP

• amplification of specific DNA region using PCR with two specific primers
• use of consensual primers (e.g., cpDNA)
  • applicable to almost all plant species
• restriction of amplified region by different restriction enzymes
• electrophoresis of fragments

+ universal method
+ codominat marker when ITS is used
– lower variability when cpDNA is used
tRFLP

terminal - Restriction Fragment Length Polymorphism

Extract DNA

PCR 16S rRNA genes with labelled primer

Cut DNA with Restriction Enzymes

Gel Electrophoresis of terminal restriction fragments

http://hpl.umces.edu/faculty/bcrump/TRFLP.pdf
tRFLP

terminal - Restriction Fragment Length Polymorphism

Siu N. et al. (2014). Open J. Ecol. 4: 434-455
RAPD
Random Amplified Polymorphic DNA

- fragments are generated using PCR with one arbitrary primer (decanucleotide)
- electrophoretic separation of fragments according to their length
- polymorphism is caused by
  - mutation in the place where primers anneal (*priming site*)
  - insertion/deletion in the amplified DNA

+ simple method
- dominant marker
- results difficult to reproduce
RAPD gel example
AFLP
Amplified Fragment Length Polymorphism

• combination of RFLP and subsequent PCR of selected fragments
• DNA restriction by two different enzymes
• selective amplification of a subset of fragments
• fluorescence visualization of fragments on the gel (using automated sequencer)

+ highly polymorphic, reflects variability of „whole“ genome
+ high reproducibility, reliability
- dominant marker – homo- a heterozygotes cannot be distinguished
AFLP — Amplified Fragment Length Polymorphism

1. restriction

2. ligation

3. amplification

Mueller & Wolfenbarger (1999), TREE
AFLP gel example
Microsatellites
SSRs – simple sequence repeats

- tandem repeats of several nucleotides

AGGC\textcolor{red}{TATATATATAGGCA} 1
AGGC\textcolor{red}{TATATATA}--\textcolor{green}{GGCA} 2

- alleles – differ by number of repeats

+ codominant marker, highly variable
+ relationships among alleles can be assessed
- necessity to develop primers for study species
Example of microsatellite analysis

http://www.qub.ac.uk/bb-old/prodohl/gel/Images/multiplex_set_1b.gif
Microsatellites - homozygotes and heterozygotes (Nuphar lutea)
ISSRs – **Inter Simple Sequence Repeats**

- length variation of regions between microsatellite loci
- primer – microsatellite sequence

+ variable marker, simple
- dominant marker

J. Košnar, unpubl.
SSCP – Single Strand Conformation Polymorphism

- method for finding unknown point mutation
- PCR amplification of target region
- denaturation – electrophoresis of ssDNA
- mutation changes tertiary structure (conformation) of the chain and thus its mobility in gel
SSCP gel example
DNA sequencing

- determination of the sequence of nucleotides in DNA chain
- Sanger sequencing – use of automated sequencers – fluorescence base labelling
- specific primers for PCR amplification of the target region
Sequencing

1. coding genes – *conservative*
   - systematics at the level of families, genera (*rbcL*)
2. spacers, introns – *variable regions*
   - systematics at the level of genera, species and below
     (*trnL-trnF, atpB-rbcL, ITS*)

- **chloroplast genes**
  - *rbcL*
  - *atpB*
  - *matK* ...

- **nuclear genes**
  - *ITS*
  - *18S rDNA*
  - *26S rDNA* ...
Chloroplast DNA

Khoa et al., 2014
Next generation sequencing – NGS
massively parallel sequencing, high-throughput sequencing

- parallel sequencing of millions of fragments
- bioinformatics to deal with huge amount of information

Diverse approaches/applications
- shotgun sequencing (genome, transcriptome – RNAseq)
- sequence capture (enrichment) – HybSeq
- restriction-based reduction – RADseq
- amplicon sequencing
- metasequencing ...

Several platforms
- Roche (454) – pyrosequencing
- Illumina (Solexa)
- SOLiD (ABI)
- Pacific Biosciences
- Ion Torrent
- ...
RAD-seq sequencing
Restriction-site-associated DNA sequencing

SNP
single-nucleotide polymorphism

• variability in DNA at one particular site (base mutation)
  ...
ACTGGAG\textcolor{red}{T}CGACTG...
...
ACTGGAGA\textcolor{red}{C}GACTG...

• mostly biallelic – two variants (e.g., A/T)
• codominant marker
• known SNPs × new variants detection
• detection (see http://en.wikipedia.org/wiki/SNP_genotyping)
  • sequencing
  • PCR-RFLP
  • SSCP, TGGE, DGGE (conformation based)
  • hybridization (allele-specific probes)
  • MALDI-TOF mass spectrophotometry (Sequenom MassARRAY...)
  • microarrays
  • NGS (e.g., RADseq)
Hyb-Seq

- solution phase hybridization
- enrichment of target sequences

Bi et al. (2012) BMC Genomics
### Types of molecular markers

#### important differences

- **variability**
  - high × low

<table>
<thead>
<tr>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>microsatellites</td>
<td>allozymes</td>
</tr>
<tr>
<td>AFLP</td>
<td>chloroplast markers</td>
</tr>
<tr>
<td>SNPs</td>
<td></td>
</tr>
</tbody>
</table>
# Types of molecular markers

## Important differences

- **Variability**
  - high × low

- **Heritability**
  - dominant × codominant

<table>
<thead>
<tr>
<th>Dominant</th>
<th>Codominant</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFLP</td>
<td>allozymes</td>
</tr>
<tr>
<td>RAPD</td>
<td>microsatellites</td>
</tr>
<tr>
<td>RFLP</td>
<td>(PCR-RFLP of e.g. ITS)</td>
</tr>
<tr>
<td></td>
<td>SNPs</td>
</tr>
</tbody>
</table>
Types of molecular markers
important differences

• variability
  high × low

• heritability
  dominant × codominant

• recombination
  yes × no

<table>
<thead>
<tr>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>nuclear markers (diploid, polyploid)</td>
<td>organelar (cp, mt) (haploid)</td>
</tr>
</tbody>
</table>
Types of molecular markers
important differences

• variability
  - high × low

• heritability
  - dominant × codominant

• recombination
  - yes × no

• transfer to next gener.
  - biparental × uniparental

<table>
<thead>
<tr>
<th>biparental</th>
<th>uniparental</th>
</tr>
</thead>
<tbody>
<tr>
<td>nuclear markers</td>
<td>angiosperms – cp and mt DNA maternally (seeds)</td>
</tr>
<tr>
<td>(seeds and pollen)</td>
<td>gymnosperms – cp DNA paternally (pollen), mtDNA maternally</td>
</tr>
</tbody>
</table>
Types of molecular markers

important differences

- variability: high × low
- heritability: dominant × codominant
- recombination: yes × no
- transfer to next gener.: biparental × uniparental
- mutation rate: high × low

<table>
<thead>
<tr>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>microsatellites</td>
<td>allozymes</td>
</tr>
<tr>
<td>introns, spacers</td>
<td>exons (esp. 1st and 2nd positions)</td>
</tr>
</tbody>
</table>
Which factors influence markers?

DNA markers (non-coding)
- mutations
- selection
- gene flow
- genetic drift

Allozyme markers
- variability increase
- variability decrease

Population level

Individuals
## Utility of markers in different types of studies

<table>
<thead>
<tr>
<th></th>
<th>RFPL and PCR-RFLP</th>
<th>Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allozymes</td>
<td>nDNA</td>
</tr>
<tr>
<td>Genetic diversity</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Population differentiation</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Gene flow</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Polyploidy</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Hybridization</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Phylogeny</td>
<td>(+)</td>
<td>+</td>
</tr>
<tr>
<td>Individual genotyping</td>
<td>(+)</td>
<td>++</td>
</tr>
<tr>
<td>Phylogeography</td>
<td>(+)</td>
<td>?</td>
</tr>
</tbody>
</table>

### Scale

- +++: excellent
- ++: good
- +: OK
- (+): has been used
- -: unlikely to be useful or useless
- ?: uncertain or not used

Based on Lowe et al. 2004
Literature

# Molecular markers in botany

1. **(11.10.)** molecular markers – characteristics, differences, technique overview
2. **(18.10.)** molecular markers – overview of applications and questions
3. **(25.10.)** isozymes – electrophoresis, evaluation of codominant data, population genetics
4. **(1.11.)** DNA – structure, PCR techniques, applications, dominant markers (AFLP, RAPD, ISSRs...), data evaluation
5. **(8.11.)** restriction techniques (RFLP, PCR-RFLP), cpDNA, phylogeography
6. **(15.11.)** microsatellites – nuclear, chloroplast, isolation, data evaluation, applications
7. **(22.11.)** Sanger sequencing – cpDNA, genes and non-coding regions
8. **(29.11.)** sequencing II – nuclear DNA, nrDNA, ITS, low-copy genes
9. **(6.12.)** NGS (next-generation sequencing) – principles and applications...
10. **(13.12.)** RADseq – SNP analysis, population genomics
11. **(20.12.)** HybSeq – enrichment methods, phylogenomics
12. **(3.1.2018)** student presentations, exam...