Molecular markers in plant systematics and population biology

10. RADseq, population genomics

Tomáš Fér
tomas.fer@natur.cuni.cz
RADseq
Restriction site-associated DNA sequencing

- genome complexity reduced by DNA cutting by restriction endonuclease(s)
- only sequences associated with restriction sites are sequenced
- many modifications of the basic protocol
1. digestion with one enzyme
2. adapter/barcode ligation
3. pooling
4. mechanical shearing
5. size selection
6. adapter ligation
7. PCR amplification with adapter specific primers
8. sequencing

RADseq modifications

- sequencing of fragments adjacent to single restriction enzyme cut sites
  - original RADseq (Baird et al. 2008)
  - 2bRAD (Wang et al. 2012)

- sequencing of fragments flanked by two restriction enzyme cut sites
  - single enzyme, indirect size selection
    - GBS – genotyping by sequencing (Elshire et al. 2011)
    - SBG – sequence-based genotyping (Truong et al. 2012)
  - double enzyme, indirect size selection
    - CRoPS – complexity reduction of polymorphic sequences (Orsouw et al. 2007)
  - single enzyme, direct size selection
    - RRLs – reduced representation libraries (van Tassel et al. 2008)
    - MSG – multiplexed shotgun genotyping (Andolfatto et al. 2011)
    - ezRAD (Toonen et al. 2013)
  - double enzyme, direct size selection
    - ddRAD – double-digest RAD (Peterson et al. 2012)
Sequence next to RE cut site

- uses IIB restriction enzymes – cleave DNA upstream and downstream of the recognition site
- results in short fragments of uniform length
Sequence flanked by two RE cut sites

- common-cutter enzyme
- PCR size selection (shorter fragments preferentially amplified)

- common-cutter enzyme(s)
- proprietary kit for Illumina library preparation

- two enzymes
- size selection by automated gel cut
## Table 1 | Summary of trade-offs among five RADseq methods

<table>
<thead>
<tr>
<th></th>
<th>Original RAD</th>
<th>2bRAD</th>
<th>GBS</th>
<th>ddRAD</th>
<th>ezRAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options for tailoring number of loci</td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme or size selection window</td>
<td>Change restriction enzyme or size selection window</td>
</tr>
<tr>
<td>Number of loci per 1 Mb of genome size*</td>
<td>30–500</td>
<td>50–1,000</td>
<td>5–40</td>
<td>0.3–200</td>
<td>10–800</td>
</tr>
<tr>
<td>Length of loci</td>
<td>≤1kb if building contigs; otherwise ≤300 bp*</td>
<td>33–36 bp</td>
<td>&lt;300 bp*</td>
<td>≤300 bp*</td>
<td>≤300 bp*</td>
</tr>
<tr>
<td>Cost per barcoded or indexed sample</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Effort per barcoded or indexed sample</td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Use of proprietary kit</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Identification of PCR duplicates</td>
<td>With paired-end sequencing</td>
<td>No</td>
<td>With degenerate barcodes</td>
<td>With degenerate barcodes</td>
<td>No</td>
</tr>
<tr>
<td>Specialized equipment needed</td>
<td>Sonicator</td>
<td>None</td>
<td>None</td>
<td>Pippin Prep™</td>
<td>Pippin Prep™</td>
</tr>
<tr>
<td>Suitability for large or complex genomes†</td>
<td>Good</td>
<td>Poor</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Suitability for de novo locus identification (no reference genome)¶</td>
<td>Good</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Available from commercial companies</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

RADseq bioinformatics

• demultiplexing, trimming barcodes
• filtering reads
  • presence of expected restriction site
  • quality
• PCR duplicate removal

• reference genome existing
  • align reads to the genome
  • call SNPs – define genotypes/haplotypes

• reference genome missing
  • de-novo assembly of reads
  • call SNPs – define genotypes/haplotypes

• software – Stacks, pyRAD, AftrRAD, dDOCENT
RADseq data properties

- relatively short loci
- wide genomic distribution
- allelic dropout/null alleles
- large proportion of missing data
- orthology/paralogy – bioinformatic assessment
RADseq application

• phylogenomics
• population structure, phylogeography
• population genomics
• evolution of recently radiated groups
• hybridization, introgression
• ...
RADseq phylogenomics

Penstemon
- 75 species
- 13 sections

RADseq population structure

*Primula tibetica*
- 293 samples
- 61 populations
- 4 groups

Inferred demographic histories

Population genomics

- simultaneous study of numerous loci to better understand the roles of evolutionary processes (such as mutation, random genetic drift, gene flow and natural selection) that influence variation across genomes and populations
- **neutral loci** – will be similarly affected by demography and the evolutionary history of populations
- **loci under selection** (adaptive) – often behave differently and reveal ‘outlier’ patterns of variation
- identification of outlier loci (high or low $F_{ST}$ between populations)

Selection in *Arabidopsis thaliana*

- 180 lines from S and N Sweden
- massive variation in genome size due to 45S rDNA copy number variation
- massive global selective sweep (700-kb transposition)

RADseq in recently diversified group

- recently diversified group – closely related species
- phylogeny and detection of ancestral hybridization
- 40,000 loci

Testing for admixture

four-taxon D-statistic test

- two incongruent patterns of two biallelic SNPs (ABBA, BABA)
- these should be equally present under a scenario of ILS without gene flow
- excess of ABBA or BABA patterns is indicative of gene flow

testing admixture between *Diapensia purpurea* and *D. himalaica*

- 9 out of 18 tests detected significant signal
- congruent with reticulation in network

Comparison of RADseq and target enrichment

<table>
<thead>
<tr>
<th>Category</th>
<th>RAD-Seq</th>
<th>Sequence capture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker distribution and genomic context</td>
<td>Pro: Widely dispersed across genome</td>
<td>Pro: Can be tailored using new genomic information</td>
</tr>
<tr>
<td></td>
<td>Con: Anonymous, evolutionary processes largely unknown</td>
<td>Con: Purifying selection impacts allele frequencies</td>
</tr>
<tr>
<td>Practical considerations</td>
<td>Pro: Less expensive, faster</td>
<td>Pro: Works with low-quality and highly contaminated samples</td>
</tr>
<tr>
<td>Assembly and orthology identification</td>
<td>Pro: Deep coverage, high read overlap</td>
<td>Pro: Over-splitting less problematic</td>
</tr>
<tr>
<td>Variant-calling and genotyping</td>
<td>Pro: Fewer rare alleles may make errors easier to distinguish, phasing more straightforward</td>
<td>Pro: Fewer low-coverage rare alleles, no allele dropout</td>
</tr>
<tr>
<td>Information content</td>
<td>Pro: More overall information</td>
<td>Pro: More information per locus</td>
</tr>
<tr>
<td>Applications</td>
<td>Genome scans, rapid and inexpensive analyses, analyses using species in clades without genomic information, extremely shallow divergences and otherwise intractable relationships.</td>
<td>Comparisons across species, calibrating parameter estimates, targeting loci of known utility or interest, studies using poor-quality samples, studies requiring resolved gene trees, deeper phylogenetic studies.</td>
</tr>
</tbody>
</table>

Literature


