A targeted subgenomic approach for phylogenomics based on microfluidic PCR and high-throughput sequencing

Simon Uribe-Convers, Matt L. Settles and David C. Tank
University of Idaho

@uribe_convers  www.simonuribe.com
The era of Genomics
The era of Genomics

Sequence Capture  GBS/RADSeq

Whole Genome

Genome Skimming  Transcriptomes
The era of Genomics

Illumina platform

Genomic Library
Amplification
Sequencing
Targeted (sub-) genomics

Microfluidic PCR

- Using Fluidigm Access Array
- 48 x 48 (2304 PCRs)
- Ready for next-gen sequencing
Microfluidic PCR

- 4 primer reaction
- Dual barcodes and adapters are incorporated in the reaction
- No need for library preparation!

![Diagram](https://www.dddmag.com/sites/dddmag.com/files/legacyimages/Articles/2009_11/fluidigm.jpg)

- Primer: forward & reverse
- Conserved sequence
- Barcodes
- Sequencing adaptors
Primer design criteria

- Variable regions between 400-900bp
- Conserved flanking regions
- Every primer has the same annealing temperature (60°C)
Chloroplast data

- Six complete plastomes (via long PCR)
- Most variable regions in the chloroplast
- Designed 74 primer pairs
Chloroplast data

- 53 primer pairs were successfully validated
- 72% success rate
- The 48 most informative ones were chosen
  average variability 2.7% (0.8%-7.5%)
Nuclear data

- Low coverage genomic data
- Shotgun sequencing for four sample - three species
  HiSeq 2000 - 100bp paired-end reads
Nuclear data

- Compared our reads to public databases
  PPR gene family
  COSII

- Pipeline:
  BLAT
    Keeps reads and gene
  MAFFT
  IntronFinder from SolGenomics

Orthology, yes!
Target gene

Exon | 400-800 bp | Exon
F primer | | R primer

Diagram showing a target gene with exons flanked by primers.
Data Processing

Raw reads
- Trimming (optional)
  - different values for R1 and R2

- Merge reads
  - Min. 20 bp overlap
  - Red colors are joined reads
  - Grey colors are unpaired

- Very little missing data
- Split reads into samples by dual barcodes (demultiplexing)
- Split reads into amplicons by primers
  - Up to 2 primer mismatches
  - 4 last bp of primers must match to produce clean ends
Sample 1 - Region 1
Sample 1 - Region 1

Minimum 5 reads and 5% of all reads
Minimum 5 reads and 5% of all reads
**Neobartsia** - Orobanchaceae (Uribe-Convers et al. in prep; UIIdaho)

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>No. Primer Pairs</th>
<th>Validated Primer Pairs</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPR</td>
<td>44</td>
<td>26</td>
<td>59.09</td>
</tr>
<tr>
<td>COSII</td>
<td>130</td>
<td>25</td>
<td>19.23</td>
</tr>
<tr>
<td>ITS</td>
<td>4</td>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>ETS</td>
<td>4</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>Phototropin1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phototropin2</td>
<td>3</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>188</strong></td>
<td><strong>61</strong></td>
<td><strong>32.44</strong></td>
</tr>
</tbody>
</table>

576 samples

Nuclear: 21 PPR, 24 COSII, 1 ITS, 1 ETS, 1 Phototropin2

Chloroplast: 48 most variable regions

Total: ~50,000 bp
Castilleja - Orobanchaceae

96 samples
Nuclear: In primer design
Chloroplast: 48 most variable regions
Total: ~25,000 bp

Tank et al. in prep
Lachemilla - Rosaceae (Diego Morales-Briones et al., UIIdaho)

288 samples

Nuclear: 48 genes, Chloroplast: 48 most variable regions

Total: ~55,000 bp
Cucurbita - Cucurbitaceae (Heather-Rose Kates et al.; UFlorida)

22 species
Nuclear: 48 genes

Draba and Solanum - Solanaceae (Ingrid Jordon-Thaden et al.; Bucknell University)

Nuclear: Genes based on transcriptomes using MarkerMiner
Acknowledgments

Tank lab
  Diego Morales-Briones, Hannah Marx
  Sarah Jacobs, Maribeth Latvis

IBEST
  Sam Hunter, Dan New, Tamara Max