Population Diversity in Legumes from Genomic to Intragenic Scales

National Center for Genome Resources
Santa Fe, New Mexico, USA

Jimmy Woodward
jew@ncgr.org
Medicago HapMap Project

- Collaboration with Nevin Young, UMN
- Define haplotype structure of *Medicago truncatula*
- Gather associated phenotype data
- Examine in detail genomic role of legume–rhizobium symbiosis (model for nitrogen fixation)
Linkage disequilibrium in Medicago

- Haplotypes: 14 – 36 Kbp
- Glycine soja: 8 – 14 Kbp
- Arabidopsis: ~8Kbp

\[ R^2 = 0.367 \]
Sequencing & Analysis Plan

- Sequence 32 lines deeply
  - 2x90 short insert PE
  - 2x54 long insert PE
  - ~30X coverage
- Sequence 350-400 additional lines
  - 2x90 short insert PE
  - ~5X coverage
- Variant detection
  - SNP
  - Indel
  - Structural variant
- Haplotype definition
- Population genetic studies
- Comparisons to symbiotic phenotypes
Germplasm selection

- Deep 30:
  - Core 16 from INRA-Montpellier
  - Jemalong A17 (reference line)
  - 3 related subspecies, 1 diploid alfalfa
  - Other strategically important lines

- Shallow ~350:
  - Core 192 from INRA-Montpellier
  - 64 from USDA/GRIN
  - Accessions from Spain, Morocco, Tunisia, Syria
Progress to date

- >200 Gbp sequence generated
- 28 of 32 core lines finished
- SNP analysis underway for core lines
In a 5-Mbp well-finished region of Chr 5, SNP density is 1 SNP / 23 Kbp
Initial SNP results

A17

DZA315-16

F83005-5

R108

0 SNPs
≤ 2 SNPs/kb
≤ 4 SNPs/kb
≤ 6 SNPs/kb
≤ 8 SNPs/kb
≤ 10 SNPs/kb
> 10 SNPs/kb
A17 SNPs versus SNPs found in at least one of the other three lines

DZA315-16
F83005-5
R108

3,492,706
27,163
30,581

A17
Two way comparisons with A17

<table>
<thead>
<tr>
<th></th>
<th>HM005</th>
<th>HM029</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,224,648</td>
<td>2,113,172</td>
</tr>
<tr>
<td>HM005</td>
<td>34,734</td>
<td>39,005</td>
</tr>
<tr>
<td>HM101</td>
<td>23,010</td>
<td>18,739</td>
</tr>
<tr>
<td>HM006</td>
<td>1,123,838</td>
<td></td>
</tr>
<tr>
<td>HM101</td>
<td>34,630</td>
<td></td>
</tr>
<tr>
<td>HM029</td>
<td></td>
<td>18,739</td>
</tr>
</tbody>
</table>
HM005 v HM006 v HM029

HM005

HM006

HM029

1,702,949

566,771

340,952

233,320

106,615

89,027

1,702,949

89,027
Two way comparisons
Deep sequencing of ~30 lines will uncover most of the common polymorphisms
Shallow sequencing will enable haplotype definition
Release of SNPs from deep 30 late 2009
Sequencing should be complete by mid-2010
Analysis through 2012
Cowpea PCR amplicon resequencing

- Collaboration with Doug Cook, UCD
- Much smaller scale
- Previous work:
  - 1440 orthologous loci identified in Mt, Gm, Lj – “COS” (Conserved Orthologous Sequences)
  - PCR primers designed in conserved exons flanking introns
  - Amplicons Sanger-sequenced in mapping parent lines for Vigna unguiculata, Cicer arietinum, Cajanus cajan, Phaseolus vulgaris, Lens culinaris, Arachis spp.
Sequencing plan

- gDNA from 757 cowpea lines pooled
- 1440 PCR reactions using pooled gDNA
- PCR products purified, normalized, and pooled
- Pooled PCR products sequenced on 1 lane of 2x90 Solexa
- >2.5 Gbp generated
- Reads aligned to Sanger consensus sequences
Theory vs. Reality

- Expected coverage - ~1 Mbp of unique sequence – 2,500 Mbp should give ~3x coverage of 750 genotypes
- Initial alignment allowed 12 mismatches and up to 5 nonunique alignments; 15% of reads aligned
- 10% of Sanger sequences had no reads aligning
First-Pass alignment results

- >16,000 SNPs/indels called against Sanger reference
- 16.6 SNPs/Kbp
- 457 variants with >2 alleles
- Reads also aligned against soybean genome: >35,000 variants called
Causes of poor alignment

- Incomplete Sanger reference
- Off-target amplification
- Large indel/MNP polymorphism
Alternative Alignment/Assembly method

- Bin Solexa reads by BLAST alignment against Sanger sequences
- Assemble binned reads
- Iteratively extend assemblies
- Identify large variants
- Develop reference integrating large variants
- Re-align reads for SNP/indel identification
Summary

- Solexa sequencing of large amplicon pools is tractable but nontrivial
- No haplotype information is possible
- Proper normalization of gDNA and amplicon pools is crucial
- Haplotype information may be possible via multiplexed hybridization-capture sequencing
Acknowledgements

Medicago
HapMap

Cowpea COS
Resequencing

NCGR:
Greg May
Joann Mudge
John Crow
Andrew Farmer
Ryan Kim
Leonda Clendennen

Minnesota:
Nevin Young
Peter Tiffin
Michael Sadowsky

UC Davis:
Doug Cook
George Breuning
Varma Penmetsa
Noelia Carrasquilla-Garcia

INRA – Montpellier:
Thomas Bataillon
Jean Marie Prosperl
Joelle Ronfort