Genotype-by-sequence in oat

Nick Tinker, AAFC Ottawa
Acknowledgements

Coauthors:
Jesse Poland
Eric Jackson
Shiaoman Chao
Gerard Lazo
Becky Oliver

CORE:
Rick Jellen, Marty Carson, Howard Rines, Don Obert, Joe Lutz, Irene Shackelford, Abraham Korol, Aaron Beattie, Åsmund Bjørnstad, Mike Bonman, Jean-Luc Jannink, Mark Sorrells, Gina Brown-Guedira, Jennifer Mitchell Fetch, Steve Harrison, Catherine Howarth, Amir Ibrahim, Fred Kolb, Mike McMullen, Paul Murphy, Herb Ohm, Brian Rossnagel, Weikai Yan, Kelci Miclaus, Jordan Hiller, Jeff Maughan, Rachel Redman, Joe Anderson, Emir Islamicov ...
And others

Tinker Lab:
Charlene Wight
Kyle Gardner
Phil Couroux
Jiro Hattori
Biniam Hizbai
Benazir Marquez
Xiaomei Luo
Arsh Singh
Imagine ..... the complete oat genome!

← (this genome stretches across China) →

Imagine ..... every complete oat genome!

← (this genome stretches across China) →
State-of-the-art DNA sequencing

- Make 100’s of copies
- Put them through a shredder
- Try to put them back together

e.g. 150x coverage = 20 billion pieces x 100 bp = $100 K
Until then….

• Focus on differences among varieties
  • That’s what we care about the most

• Order differences by linkage (count recombinations)

• Associate with phenotype (also by linkage)
  • Mapping populations (2-parents, lots of kids)
  • ‘Natural’ populations (unknown family structure)
Single Nucleotide Polymorphism (SNP)

- The most common genetic difference

```
GTACCATGATCGCTAAGCTGACATGGCTTACGGCTTGAC
```

(A) ...........T...............G................
(B) ...........G...............G................
(C) ...........G...............A................
(D) ...........T...............A................
(E) ...........G...............G................

- SNP = SNP .... no matter how you find it!
  - “Old” non-sequence-based methods (AFLP, DArT)
  - Discover by sequence / assay by design
  - Discover and assay by sequencing (GBS)
SNP – discover by sequence, assay by design

Reference sequence

Sequence reads
By variety

Consistent within variety
CORE – Illumina SNP array

**cDNA** (Gene pieces)
- 20 varieties
- 9 million reads
- 18,000 templates

**DArT** (Genome pieces)
- 25 varieties
- 4 million reads
- 12,000 templates

...TGATCGCTA[G/T]CTGGCATGGCT......
- 80,000 predicted SNPs
- 4600 tested SNPs
- 2300 validated SNPs (Golden Gate)

- 6000 SNPS in progress (Infinium)
- (we estimate 4000 will work)
Genome Studio Software (example SNPs)
Genotype by sequence (GBS) - concept

- Discover and assay SNPs by direct sequencing
- Similar to SNP discovery for planned assay
  - But much larger numbers of and sequences
- Based on subset of genome (enzyme / amplify)

Variety A

----

Variety B
GBS details

1. Digest with *PstI* & *Mspl*

   ![Digest diagram](PstI_Mspl)

2. Ligate sequencing adapters + variety-specific barcode

   ![Ligation diagram](ligation)

3. Mix together (multiplex), amplify

   ![Mixing diagram](mixing)
5. Trim barcode, trim to 64 bases, keep track of variety

6. Identify all unique tags, count in each variety (Tag x Taxa)

7. Match tag pairs, call SNPs (across full data set)

Varieties: A, C, E

Varieties: B, D, H

Is a SNP

100,000

Varieties: A, C, D, H

Varieties: B, C, D, E

Not a SNP
GBS - caveats

• Missing data 10% to 70%
  – Depends on sequencing depth ("plexity")
  – Depends on how many SNPs you call
  – e.g. to get 95% complete, I could only call 2400 SNPs

• Bioinformatics is under development
  – Different methods give different SNP sets

• Massive storage and computing requirements

• Data sets too large for some software
  – More markers and more samples
  – Have not yet managed a consensus map
## DArT vs SNP vs GBS

<table>
<thead>
<tr>
<th></th>
<th>DArT</th>
<th>SNP</th>
<th>GBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay cost per sample</td>
<td>$50</td>
<td>$68</td>
<td>$20  *</td>
</tr>
<tr>
<td>Markers across all taxa</td>
<td>1500</td>
<td>4000  *</td>
<td>40 k → 100 k</td>
</tr>
<tr>
<td>Markers per population</td>
<td>300</td>
<td>800</td>
<td>4 k → 20 k</td>
</tr>
<tr>
<td>Missing data</td>
<td>&gt;5 %</td>
<td>&lt; 1%</td>
<td>10% → 50%</td>
</tr>
<tr>
<td>Co-dominant</td>
<td>0%</td>
<td>&gt; 25%</td>
<td>100% *</td>
</tr>
<tr>
<td>Genes / orthology</td>
<td>20%</td>
<td>100%</td>
<td>5%</td>
</tr>
<tr>
<td>Duplicate loci (map inconsistently)</td>
<td>&gt; 5%</td>
<td>&lt; 3 %</td>
<td>?</td>
</tr>
</tbody>
</table>
Are they useful?

- Already have more GBS data than anything else:
  - Not just more loci...... more varieties too
  - DArT: 350 diversity + 4 bi-parental populations
  - SNP: 108 diversity + 6 bp-pop (400)
  - GBS: 738 diversity + 8 bp-pop (700) + 16 iso-lines

- 10 x more likely to find \([\text{marker} - \text{QTL}]\) ?

- But missing data ...
  - Mapping difficulty ?
  - Association artefacts ?
  - MAS predictions ?
Simulated MAS using GBS

• Scenario:
  – 4 target loci, simulate with random GBS loci
  – Discover markers by association in odd # lines
  – Predict genotypes in even # lines

<table>
<thead>
<tr>
<th>Marker</th>
<th>f(A)</th>
<th>f(B)</th>
<th>f(H)</th>
<th>Chr</th>
</tr>
</thead>
<tbody>
<tr>
<td>avjp100014</td>
<td>568</td>
<td>782</td>
<td>20</td>
<td>1C</td>
</tr>
<tr>
<td>avjp100585</td>
<td>339</td>
<td>966</td>
<td>18</td>
<td>16A</td>
</tr>
<tr>
<td>avjp100261</td>
<td>150</td>
<td>1188</td>
<td>9</td>
<td>9D</td>
</tr>
<tr>
<td>avjp108144</td>
<td>1196</td>
<td>143</td>
<td>8</td>
<td>7C_17A</td>
</tr>
</tbody>
</table>
Predictive markers (target locus avjp100014)

204 “good” 141 “bad”
Predictive markers (target locus avjp100261)

62 “good”

287 “bad”
Simulated MAS using GBS

• Surprise!
  – Many loci in (near) perfect LD with a “QTL”
  – Bias in genome regions sampled by GBS ?
  – Too much good data ?

• Test predictions of array-based SNPs
  – Independently discovered
  – Most based on expressed genes
  – BUT.... SNP data only available for 108 varieties
Simulated QTLs based on Illumina SNPs

- 4 random Illumina cDNA SNPs (targets)
- Fit model with 54 odd # lines
- Test model with 54 even # lines

<table>
<thead>
<tr>
<th>Marker</th>
<th>f(A)</th>
<th>f(B)</th>
<th>Chr</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMI_ES01_c10033_104</td>
<td>56</td>
<td>52</td>
<td>1C</td>
</tr>
<tr>
<td>GMI_ES15_c10388_464</td>
<td>57</td>
<td>51</td>
<td>18D</td>
</tr>
<tr>
<td>GMI_ES01_c796_180</td>
<td>84</td>
<td>24</td>
<td>19A</td>
</tr>
<tr>
<td>GMI_ES15_c8064_341</td>
<td>25</td>
<td>83</td>
<td>3C</td>
</tr>
</tbody>
</table>
Predictive markers (locus GMI_ES01_c10033_104)

27 “good”  27 “bad”
Predictive markers (locus GMI_ES15_c8064_341)

41 “good” 13 “bad”
Case study with naked (hulless) oat

• Other genes (N2, N3, N4) or will we just find N1?
  – N1 was Mapped in Terra x Marion
  – Poorly placed by comparative mapping

• GBS data:
  – Diversity lines (*rare trait)
    • I only had phenotypes for 100 covered + 20 naked
  – 100 Terra x Marion progeny
  – 8 pairs of naked / covered iso-lines
    • OT253/Marion, from F6 heterozygotes
    • Developed by Solomon Kibite
Look at all pairwise linkages and LD’s

N1 ↔ GBS-tag

TxM Map
R < 0.05 = 28

Association
LD R2 > 50% = 18

16 N/C isolines
Alternate alleles fixed = 20

*avjp23455
avjp123459
avjp17890

20
1
3*
0
12
16
5
N1 predictions

• LD analysis based on partial data:
  – “training set” (the only ones I had data for at the time)
  – 20 naked / 100 covered

• Based on “avjp23455” we predicted:
  • 12 more naked lines (remaining 600 = covered)
  • All were correct!

<table>
<thead>
<tr>
<th>Training set</th>
<th>Predicated Naked Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>95Ab13050</td>
<td>98Ab7265</td>
</tr>
<tr>
<td>Boudrias</td>
<td>FL03184-FLID-B-S1</td>
</tr>
<tr>
<td>Bullion</td>
<td>FL04178-FLID-B-S-2</td>
</tr>
<tr>
<td>Gehl</td>
<td>HLA05AB1-34</td>
</tr>
<tr>
<td>IO1033</td>
<td>IL02-10836</td>
</tr>
<tr>
<td>IO1108</td>
<td>IL03-7936</td>
</tr>
<tr>
<td>IO114</td>
<td>LA02012-S-B-139-S2-B-S2-B-S2</td>
</tr>
<tr>
<td>IO1150</td>
<td>LA0210SBSBSSBSSB-S1</td>
</tr>
<tr>
<td>IO1191</td>
<td>LA03066SBS-S1</td>
</tr>
<tr>
<td>Navaro</td>
<td>Lennon</td>
</tr>
<tr>
<td></td>
<td>Nudist</td>
</tr>
<tr>
<td></td>
<td>Zuton</td>
</tr>
</tbody>
</table>
Conclusions

- GBS will be exceptionally efficient for tagging genes
  - Bi-parental AND association mapping
- Excellent potential for MAS
- SNP array or custom assays are better for
  - Genome analysis, comparative genomics
  - Critical genotyping, well-characterized targets
- Work required
  - Streamline informatics
  - Build GBS/SNP based consensus map
  - Evaluate consistency of map position
  - Evaluate genomic selection
American Oat Workers Conference
Ottawa Canada, 2014

Ottawa Marriott Hotel
100 Kent Street, Ottawa
Ontario K1P 5R7 Canada