GBS Usage Cases: Examples from Maize

Jeff Glaubitz (jcg233@cornell.edu)
Senior Research Associate, Buckler Lab, Cornell University
Panzea Project Manager

Cornell IGD/CBSU Workshop
June 17-18, 2013
Some potential applications of GBS Data

- Marker discovery
- Phylogeny/Kinship
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Genomic selection
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Improving reference genome assembly
Marker Discovery

- GBS markers can be converted to SNPs or PCR assays of indels
- Develop SNP assays from polymorphic tags at same location
- Develop PCR primers from adjacent tags & hope for large indels
Phylogeny/Kinship

• Missing data not an issue for estimating pairwise genetic distance or kinship
  ▪ Each pair of individuals has large, “random” sample of markers in common

• Works really well even in non-model organisms
  ▪ Fei Lu’s previous talk on switchgrass

• Principle Coordinates Analysis better than Principle Components Analysis
  ▪ Uses distance matrix rather than every genotype
  ▪ Missing data not an issue for Prin. Coord. Analysis

• SNPs can be strongly affected by ascertainment bias
  ▪ Panel used to discover the SNPs can severely distort estimates of population genetic parameters (e.g., kinship, diversity)
  ▪ Industry SNPs on the Maize 55K SNP chip an extreme example
Less ascertainment bias than array-based SNPs?

Ram Sharma

Non stiff stalk, nss (106)
Stiff stalk, ss (28)
Tropical/sub tropical (66)
Popcorn (9)
Sweet corn (6)
Unclassified (67)
B73 (ss)
Mo17 (nss)
Less ascertainment bias than array-based SNPs?

Non stiff stalk, nss (106)
Stiff stalk, ss (28)
Tropical/sub tropical (66)
Popcorn (9)
Sweet corn (6)
Unclassified (67)
B73 (ss)
Mo17 (nss)
Less ascertainment bias than array-based SNPs?

Ram Sharma

Non stiff stalk, nss (106)
Stiff stalk, ss (28)
Tropical/sub tropical (66)
Popcorn (9)
Sweet corn (6)
Unclassified (67)
B73 (ss)
Mo17 (nss)
Less ascertainment bias than array-based SNPs?

Ram Sharma

Non stiff stalk, nss (106)
Stiff stalk, ss (28)
Tropical/sub tropical (66)
Popcorn (9)
Sweet corn (6)
Unclassified (67)
B73 (ss)
Mo17 (nss)
Linkage mapping of QTL in a biparental cross

• In maize, we use the reference genome to order markers
• With *ApeKI*, too many markers for traditional software (*MapMaker, JoinMap, R-QTL* etc.)
• Filter for a smaller set of markers with high coverage
• Use 6 base cutter (or combination of enzymes) for fewer markers with higher coverage
• *JoinMap* can handle at least 3,000 markers
• Newer software?
  ▪ *MSTMap* claims 10,000 – 100,000 markers
  ▪ Guided Evolutionary Strategy (*Mester et al. 2004, Comp Biol Chem* 28: 281) – currently being implemented in TASSLE
  ▪ Others?
HMM for calling hets/correcting errors in biparental RIL populations

\[ P(G \mid S) = \frac{P(S \mid G)P(G)}{P(S)} \propto P(S \mid G)P(G) \]

distance between nodes = \( \log[ P(\text{obs}_i \mid g_i)P(g_i \mid g_{i-1})] \)
HMM for calling hets/correcting errors in biparental RIL populations

Low % het RIL

<table>
<thead>
<tr>
<th>non-B73</th>
<th>het</th>
<th>B73</th>
</tr>
</thead>
</table>

actual calls

after HMM

Med % het RIL

<table>
<thead>
<tr>
<th>non-B73</th>
<th>het</th>
<th>B73</th>
</tr>
</thead>
</table>

High % het RIL

<table>
<thead>
<tr>
<th>non-B73</th>
<th>het</th>
<th>B73</th>
</tr>
</thead>
</table>

Physical Position on Chr1 (bp)
GBS analysis of Teosinte/W22 BC$_2$S$_3$

- 868 RILs in total
- Framework map with 492 SNPs
- 51,238 bi-allelic GBS markers (*ApeKl*)

<table>
<thead>
<tr>
<th>Machine</th>
<th># Samples</th>
<th>Mean # Barcoded Reads</th>
<th>CV</th>
<th>Non-Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAll-96-plex</td>
<td>680</td>
<td>357,624</td>
<td>48.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>GAll-48-plex</td>
<td>93</td>
<td>747,196</td>
<td>38.5%</td>
<td>31.4%</td>
</tr>
<tr>
<td>HiSeq</td>
<td>95</td>
<td>997,896</td>
<td>65.5%</td>
<td>33.6%</td>
</tr>
</tbody>
</table>
TeoW22 BC2S3 – Chr9 – 4122 bi-allelic GBS markers
Reproducibility – 60 RILs – Chr9
GBS analysis of Teosinte/W22 $\text{BC}_2\text{S}_3$

- 868 RILs in total
- Framework map with 492 SNPs
- 51,238 bi-allelic GBS markers ($ApeKI$)

<table>
<thead>
<tr>
<th>Machine</th>
<th># Samples</th>
<th>Mean # Barcoded Reads</th>
<th>CV</th>
<th>Non-Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAll-96-plex</td>
<td>680</td>
<td>357,624</td>
<td>48.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>GAll-48-plex</td>
<td>93</td>
<td>747,196</td>
<td>38.5%</td>
<td>31.4%</td>
</tr>
<tr>
<td>HiSeq (96-plex)</td>
<td>95</td>
<td>997,896</td>
<td>65.5%</td>
<td>33.6%</td>
</tr>
</tbody>
</table>
Maize $Sh1$ orthologs are located at seed shattering QTLs
Fine mapping QTL

- Need to saturate interval containing QTL with markers
- GBS a good source of markers
- Also need to collect recombinants in the interval
- Near-isogenic lines (NILs) helpful (Mendelize)
- Good reference genome
Fine Mapping of Domestication QTL in Maize

7,756 GBS SNPs along Chr?

Trait Y

136 RC-NILS

W22  Teo  Het  Trait Z
Genomic Selection & GWAS

- Complete data not required for genomic selection?
  - Closely linked markers in LD cover for each other
Accurate genomic prediction of height based on GBS data

2800 diverse maize breeding lines

R² = 0.82

USDA Ames Inbreds
Ridge-Regression BLUP

Jason Peiffer
Genomic Selection & GWAS

• Complete data not required for genomic selection?
  ▪ Closely linked markers in LD cover for each other

• Missing data are more problematic for GWAS
  ▪ imputation necessary, but might cause spurious results
  ▪ avoid false imputation of biologically missing regions
  ▪ area of active research
GWAS directly hits known Mendelian traits

GWAS of white vs. yellow kernels in 1,595 USDA Ames inbreds

The best hit for kernel color lies within $Y_1$
GWAS of a more complex trait directly hits known flowering time genes

Even with ~660k SNPs we almost missed *ZmCCT* (only 1 significant SNP)
GWAS of a more complex trait directly hits known flowering time genes

GWAS of growing degree days to silking in 2,279 inbreds

- $\log_{10}(p)$

1 3 5 7 9 12 15 18 21

ZmCCT (2.4 Kb) 1 SNP vs. ZmRap2.7 (2.6 Kb) 80 SNPs

Alex Lipka

Zhiwu Zhang
Genomic Selection & GWAS

• Complete data not required for genomic selection?
  ▪ Closely linked markers in LD cover for each other

• Missing data are more problematic for GWAS
  ▪ imputation necessary, but might cause spurious results
  ▪ avoid false imputation of biologically missing regions
  ▪ area of active research

• In NAM-GWAS, imputation is much less of an issue
  ▪ NAM = “Nested Association Mapping” population
The maize NAM population was built for NAM-GWAS
liguleless1 and liguleless2 explain the two “biggest” leaf angle QTL

Tian, Bradbury, et al 2011 Nature Genetics
We are using GBS to pinpoint the location of cross overs in the NAM RILs

• B73 is the reference genome: complete knowledge
• Remaining NAM parents whole genome sequenced via Illumina at 4x coverage (paired end random sheared)
  ▪ 26 million high quality SNPs
• Precise knowledge of crossover locations in NAM RILs allows us to more accurately project sequences of parents onto RILs:

![Genetic map diagram]

- B73
- MS71
- 1100 SNPs: 1100 SNPs from B73 to Z019E001
- GBS SNPs: Genetic loci from B73 to Z019E001
GBS data improves the resolution of joint linkage analysis in the NAM population

**Trait:** Days to silk (flowering time)

<table>
<thead>
<tr>
<th>Markers</th>
<th>Median QTL support interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,106 array SNPs</td>
<td>6.2 cM</td>
</tr>
<tr>
<td>171,479 GBS SNPs</td>
<td>2.6 cM</td>
</tr>
</tbody>
</table>

Preliminary NAM-GWAS results also show improved resolution
Recombination Rates for NAM from GBS Data

Peter Bradbury – USDA Scientist, Buckler lab, Cornell (unpublished)
The maize B73 reference genome: room for improvement?

1) The B73 reference genome accurate for B73 but less so for other maize lines (e.g., Mo17)

2) Even for B73, some regions of the genome are in the wrong place

3) Some large (multiple BAC) contigs could not be anchored
   - assigned to “chromosome 0”
   - 30 chr0 contigs in B73 RefGenV1
   - 17 chr0 contigs in B73 RefGenV2

4) Some regions of the genome are missing
   - ≈5% of B73 sequence is not in the B73 reference genome
The maize B73 reference genome: room for improvement?

1) The B73 reference genome accurate for B73 but less so for other maize lines (e.g., Mo17)

2) Even for B73, some regions of the genome are in the wrong place

3) Some large (multiple BAC) contigs could not be anchored
   - assigned to “chromosome 0”
   - 30 chr0 contigs in B73 RefGenV1
   - 17 chr0 contigs in B73 RefGenV2

4) Some regions of the genome are missing
   - ≈5% of B73 sequence is not in the B73 reference genome
Most tags can be mapped as individual alleles

- In a biparental cross such as maize IBM (B73 x Mo17)
- Provided that they are polymorphic between the parents

![Diagram showing genetic mapping in a biparental cross between B73 and Mo17, with an ApeK1 site (GCWGC) and a 64-base sequence tag.]
Genetically mapping individual GBS alleles in IBM

<table>
<thead>
<tr>
<th>Min # Successes</th>
<th>p-value</th>
<th>max Recomb.</th>
<th>Total # GBS tags mapped</th>
<th># B73 tags mapped</th>
<th># Mo17 tags mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>&lt;10^{-3}</td>
<td>&lt;5%</td>
<td>485,860</td>
<td>266,192</td>
<td>219,668</td>
</tr>
<tr>
<td>20</td>
<td>&lt;10^{-6}</td>
<td>&lt;5%</td>
<td>235,531</td>
<td>123,094</td>
<td>112,437</td>
</tr>
<tr>
<td>30</td>
<td>&lt;10^{-7}</td>
<td>&lt;5%</td>
<td>140,713</td>
<td>73,829</td>
<td>66,884</td>
</tr>
</tbody>
</table>
Genetically mapping individual GBS alleles in IBM

<table>
<thead>
<tr>
<th>Min # Successes</th>
<th>$p$-value</th>
<th>max Recomb.</th>
<th>Total # GBS tags mapped</th>
<th># B73 tags mapped</th>
<th># Mo17 tags mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$&lt;10^{-3}$</td>
<td>$&lt;5%$</td>
<td>485,860</td>
<td>266,192</td>
<td>219,668</td>
</tr>
<tr>
<td>20</td>
<td>$&lt;10^{-6}$</td>
<td>$&lt;5%$</td>
<td>235,531</td>
<td>123,094</td>
<td>112,437</td>
</tr>
<tr>
<td>30</td>
<td>$&lt;10^{-7}$</td>
<td>$&lt;5%$</td>
<td>140,713</td>
<td><strong>73,829</strong></td>
<td><strong>66,884</strong></td>
</tr>
</tbody>
</table>
B73 reference genome highly accurate for B73...

- 0.4% of B73 tags genetically map to different chromosome than they align to
B73 reference genome highly accurate for B73...

• 0.4% of B73 tags genetically map to different chromosome than they align to

...but far less so for other maize lines

• 9.3% of Mo17 tags genetically map to different chromosome than they align to
Only 50% of the maize genome is shared between two varieties

Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005
Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010
Some chunks of the B73 reference genome are in the wrong place

<table>
<thead>
<tr>
<th>Physical Chr</th>
<th>Start (Mb)</th>
<th>End (Mb)</th>
<th>Genetic Chr</th>
<th>Approx. Genetic Location (Mb)</th>
<th># Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>139.3</td>
<td>139.8</td>
<td>2</td>
<td>16.5–16.8</td>
<td>49</td>
</tr>
<tr>
<td>9</td>
<td>102.5</td>
<td>106.9</td>
<td>9</td>
<td>15–32</td>
<td>49</td>
</tr>
<tr>
<td>7</td>
<td>150.1</td>
<td>161.8</td>
<td>5</td>
<td>192–214</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>0.4</td>
<td>4</td>
<td>83–151</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>48.4</td>
<td>50</td>
<td>2</td>
<td>61–127</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>0.07</td>
<td>0.2</td>
<td>7</td>
<td>47–100</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>231.2</td>
<td>231.2</td>
<td>7</td>
<td>18–26</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>228.1</td>
<td>230.5</td>
<td>5</td>
<td>194–212</td>
<td>6</td>
</tr>
</tbody>
</table>
Some potential applications of GBS Data

- Marker discovery
- Phylogeny/Kinship
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Genomic selection
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Improving reference genome assembly