Usage Cases of GBS

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

Cornell CBSU Workshop
Sept 15-16, 2011
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
  - NAM-GWAS
- Genomic selection
- Phylogeny
  - Improving reference genome assembly
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Genomic selection
- Phylogeny
- Improving reference genome assembly
Marker Discovery

- Develop SNP assays from polymorphic tags at same location
- Develop PCR primers from adjacent tags & hope for large indels
Some potential applications of GBS Data

• Marker discovery
  ▪ convert to SNP or PCR assays of indels
• Linkage mapping of QTL in a biparental cross
• Fine-mapping QTL
• Bulked segregant analysis
• Genome Wide Association Studies (GWAS)
• NAM-GWAS
• Genomic selection
• Phylogeny
• Improving reference genome assembly
Some potential applications of GBS Data

• Marker discovery
  ▪ convert to SNP or PCR assays of indels
• Linkage mapping of QTL in a biparental cross
• **Fine-mapping QTL**
• Bulked segregant analysis
• Genome Wide Association Studies (GWAS)
• NAM-GWAS
• Genomic selection
• Phylogeny
• Improving reference genome assembly
Some potential applications of GBS Data

• Marker discovery
  ▪ convert to SNP or PCR assays of indels
• Linkage mapping of QTL in a biparental cross
• Fine-mapping QTL
• Bulked segregant analysis
• Genome Wide Association Studies (GWAS)
• NAM-GWAS
• Genomic selection
• Phylogeny
• Improving reference genome assembly
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Genomic selection
- Phylogeny
- Improving reference genome assembly
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Genomic selection
- Phylogeny
- Improving reference genome assembly
Some potential applications of GBS Data

• Marker discovery
  ▪ convert to SNP or PCR assays of indels
• Linkage mapping of QTL in a biparental cross
• Fine-mapping QTL
• Bulked segregant analysis
• Genome Wide Association Studies (GWAS)
• NAM-GWAS
• Genomic selection
• Phylogeny
• Improving reference genome assembly
Phylogeny of switchgrass association panel

Fei Lu, Postdoc, Buckler lab, unpublished
Some potential applications of GBS Data

• Marker discovery
  ▪ convert to SNP or PCR assays of indels
• Linkage mapping of QTL in a biparental cross
• Fine-mapping QTL
• Bulked segregant analysis
• Genome Wide Association Studies (GWAS)
  • NAM-GWAS
• Genomic selection
• Phylogeny
• Improving reference genome assembly
The NAM population was built for NAM-GWAS

NAM population:

25 DL:
- B97
- CML103
- CML228
- CML247
- CML277
- CML332
- CML333
- CML52
- CML69
- Hp301
- Ii144
- Ki11
- Ki3
- Ky21
- M162W
- M37W
- Mo18W
- Mo37W
- NC350
- NC358
- Oh43
- Oh7B
- P39
- Tx303
- Tz28

× B73

F1s:

SSD:

1
2

NAM

...
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)
  ▪ 50 million high quality SNPs
• Precise knowledge of crossover locations in NAM RILs allows us to more accurately project sequences of parents onto RILs:
Advantages of GWAS on NAM

- Population structure eliminated by design
- Project HapMap SNPs To 5000 RILs Accurately
- Can Compare Linkage & GWAS
- Wide range of incremental models can be evaluated
- Control for 90% Genetic Variance
NAM-GWAS works beautifully for simple traits

Cob color

Direct hit within p1 locus

Tian, Bradbury, et al 2011 Nature Genetics
NAM-GWAS works beautifully for simple traits

Kernel color  Within 200bp of the promoter of Y1
Liguleless1 and Liguleless2 explained the two “biggest” leaf angle QTL.
NAM – GWAS software is implemented on CBSU’s High Performance Computing Cluster

- Accessible through the web. (http://cbsuapps.tc.cornell.edu/namgwas.aspx)
- Calculation is carried out on the BioHPC computer cluster.
- GWAS analysis using the maize Hapmap v1 genotyping data.
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Genomic selection
- Phylogeny
- Improving reference genome assembly
Improving the maize B73 reference genome

• The B73 reference genome accurate for B73 but less so for other maize lines (e.g., Mo17)
• Even for B73, some regions of the genome are in the wrong place
• 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
• Some regions of the genome are missing
  ▪ ≈5% of B73 sequence is not in the B73 reference genome
Improving the maize B73 reference genome

• The B73 reference genome accurate for B73 but less so for other maize lines (e.g., Mo17)
• Even for B73, some regions of the genome are in the wrong place
• 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
• 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
Genetically Mapping Individual GBS Alleles

<table>
<thead>
<tr>
<th>SNPs (e.g., from Illumina 55K chip)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RILs (e.g., from IBM)</td>
</tr>
</tbody>
</table>

- **B73**
- **Mo17**
- **Heterozygote**
### Genetically Mapping Individual GBS Alleles

<table>
<thead>
<tr>
<th>SNP (e.g., from Illumina 55K chip)</th>
<th>RILs (e.g., from IBM)</th>
</tr>
</thead>
</table>

- **B73**
- **Mo17**
- **Heterozygote**
Genetically Mapping Individual GBS Alleles

SNPs (e.g., from Illumina 55K chip) →

RILs (e.g., from IBM)

B73  Mo17  Heterozygote

← SNP being tested
Genetically Mapping Individual GBS Alleles

SNPs (e.g., from Illumina 55K chip) (→) 64-base sequence tag (GBS coverage ~0.4x)

RILs (e.g., from IBM)

B73  Mo17  Heterozygote

SNP being tested
**Genetically Mapping Individual GBS Alleles**

<table>
<thead>
<tr>
<th>SNPs (e.g., from Illumina 55K chip)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial Test for linkage</td>
</tr>
<tr>
<td><strong>prob. success</strong>: segregation ratio of the SNP being tested (~0.5)</td>
</tr>
<tr>
<td><strong>n trials</strong>: n RILs with GBS tag (10)</td>
</tr>
<tr>
<td><strong>n successes</strong>: n co-occurrences with presumed parental allele at SNP being tested (co-segregation)</td>
</tr>
<tr>
<td><strong>p-value</strong>: 0.00098 (&lt;10⁻³)</td>
</tr>
</tbody>
</table>

(➔) 64-base sequence tag (GBS coverage ~0.4x)

**RILs (e.g., from IBM)**

<table>
<thead>
<tr>
<th>B73</th>
<th>Mo17</th>
<th>Heterozygote</th>
</tr>
</thead>
</table>

SNP being tested
Genetically Mapping Individual GBS Alleles

SNPs (e.g., from Illumina 55K chip) (→) 64-base sequence tag (GBS coverage ~0.4x)

RILs (e.g., from IBM)

These 10 SNPs all tie

B73    Mo17    Heterozygote
Genetically Mapping Individual GBS Alleles in IBM

<table>
<thead>
<tr>
<th>Min #Successes</th>
<th>max Recomb.</th>
<th>p -value</th>
<th>nGBS tags</th>
<th>nB73</th>
<th>nMo17</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>&lt;5%</td>
<td>&lt;10^{-3}</td>
<td>485,860</td>
<td>266,192</td>
<td>219,668</td>
</tr>
<tr>
<td>20</td>
<td>&lt;5%</td>
<td>&lt;10^{-6}</td>
<td>235,531</td>
<td>123,094</td>
<td>112,437</td>
</tr>
<tr>
<td>30</td>
<td>&lt;5%</td>
<td>&lt;10^{-7}</td>
<td>140,713</td>
<td>73,829</td>
<td>66,884</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...

...but far less so for other maize lines

• 9.3% of Mo17 tags genetically map to different chromosome than they align to
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>
Improving the maize B73 reference genome

• The B73 reference genome accurate for B73 but less so for other maize lines (e.g., Mo17)
• Even for B73, some regions of the genome are in the wrong place
• Some large (multiple BAC) contigs could not be anchored
  ▪ assigned to “chromosome 0”
  ▪ 30 chr 0 contigs in B73 RefGenV1
  ▪ 17 chr 0 contigs in B73 RefGenV2
• Some regions of the genome are missing
  ▪ ≈5% of B73 sequence is not in the B73 reference genome
Some contigs are on “Chr0” & some chunks are missing from the B73 assembly

• The sequence of Chr0 contigs is known
  ▪ so we know which ApeKI GBS tags are there
• De novo contigs are being constructed from WGS*
• Created a pipeline to genetically map novel contigs using linkage populations
• Used IBM GBS data for proof of concept
  ▪ Genetically anchored 20% of full length cDNAs (407 novel)
  ▪ 8 of 17 chromosome 0 contigs in B73 RefGenV2
  ▪ Small proportion (>70,000) of the de novo WGS contigs of B73 & Mo17 (as most of these 5 million contigs are very small)

* B73 (454), FLcDNA: Shiran Pasternak, Josh Stein, Andrew Olson, Doreen Ware (CSHL)
  Mo17 (Illumina GAII): Hainan Zhao, Jinsheng Lai (Chinese Agricultural Univ.)
Improving the reference genome

ApeKI site (GCWGC)

(→) 64-base sequence tag

< 450 bp

B73
Improving the reference genome

ApeKI site (GCWGC)
Improving the reference genome

ApeKI site (GCWGC)

B73

contig

(→) 64-base sequence tag

< 450 bp

de novo (e.g., from 454 or Illumina sequence)
Improving the reference genome

ApeKI site (GCWGC)

B73

(contig)

(< 450 bp)

(→) 64-base sequence tag

de novo (e.g., from 454 or Illumina sequence)

Novel? (not included in B73 RefGen_v2)
Improving the reference genome

SNPs (e.g., from Illumina 55K chip) → 64-base sequence tag (GBS coverage ~0.4x)

RILs (e.g., from IBM)

These 3 SNPs tie
Improving the reference genome

<table>
<thead>
<tr>
<th>Contigs</th>
<th>Total #</th>
<th>Source</th>
<th># contigs genetically mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr0</td>
<td>17</td>
<td>B73 RefGen_v2</td>
<td>8</td>
</tr>
<tr>
<td>B73 454 (k96)</td>
<td>3,964,387</td>
<td>CSHL</td>
<td>3,408</td>
</tr>
<tr>
<td>FLcDNA</td>
<td>61,477</td>
<td>CSHL</td>
<td>407</td>
</tr>
<tr>
<td>Mo17 Illumina GA II</td>
<td>1,051,007</td>
<td>CAU</td>
<td>5,988</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36,041</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10,776</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27,055</td>
</tr>
</tbody>
</table>
## Improving the reference genome

<table>
<thead>
<tr>
<th>Contigs</th>
<th>Total #</th>
<th>Source</th>
<th># contigs genetically mapped</th>
<th>novel</th>
<th>non-novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr0</td>
<td>17</td>
<td>B73 RefGen_v2</td>
<td>8</td>
<td>8</td>
<td>---</td>
</tr>
<tr>
<td>B73 454 (k96)</td>
<td>3,964,387</td>
<td>CSHL</td>
<td>3,408</td>
<td>3,408</td>
<td>36,041</td>
</tr>
<tr>
<td>FLcDNA</td>
<td>61,477</td>
<td>CSHL</td>
<td>407</td>
<td>407</td>
<td>10,776</td>
</tr>
<tr>
<td>Mo17 Illumina GA II</td>
<td>1,051,007</td>
<td>CAU</td>
<td>5,988</td>
<td>5,988</td>
<td>27,055</td>
</tr>
</tbody>
</table>

\[ p < 10^{-7} \]
>70% contigs genetically map to within 1 Mb of true position
>70% contigs genetically map to within 1 Mb of true position

- This is just with the IBM population
- Will vastly improve with NAM
Some regions of reference genome are missing large chunks

Telomere of Chr4 is a prime target for future improvement
This coming year – Improving the genome

• Add in GBS data from NAM for much higher resolution
  ▪ Anchor as many novel genes & contigs as possible

• Reorder contigs within BACs by LD (HapMapV2) & using GBS data from linkage populations (NAM & IBM)
  ▪ Also check order and orientation of BACs themselves
Some potential applications of GBS Data

- Marker discovery
  - convert to SNP or PCR assays of indels
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Genomic selection
- Phylogeny
- Improving reference genome assembly