Usage Cases of GBS

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

Cornell CBSU Workshop
September 13-14, 2012
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

B73

ApeKI site (GCWGC)

Loss of cut site

Mo17

(→) 64-base sequence tag

< 450 bp
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)

Illumina 55K

Ram Sharma
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)
Linkage mapping of QTL in a biparental cross

- In maize, we use the reference genome to order markers.
- With *ApeKl*, 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 TASSEL.
  - Others?
GBS analysis of Teosinte/W22 BC$_2$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</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 BC$_2$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>
Parallel domestication of the *Shattering1* genes in cereals

Zhongwei Lin¹, Xianran Li¹, Laura M Shannon², Cheng-Ting Yeh³,⁴, Ming L Wang⁵, Guihua Bai¹,⁶, Zhao Peng⁷, Jiarui Li⁷, Harold N Trick⁷, Thomas E Clemente⁸, John Doebley², Patrick S Schnable³,⁴, Mitchell R Tuinstra⁹, Tesfaye T Tesso¹, Frank White⁷ & Jianming Yu¹

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
• In contrast, 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 maize inbreds

The best hit for kernel color lies within $Y1$
GWAS of a more complex trait directly hits known flowering time genes

However, even with ~660k SNPs we almost missed Ghd7 (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 maize inbreds

-\log_{10}(p)

1 3 5 7 9 12 15 18 21

1 2 3 4 5 6 7 8 9 10

Vgt1 (62Kb)

Ghd7 (5Kb)

1 SNP

80 SNPs

Zhiwu Zhang

Alex Lipka
Genomic Selection & GWAS

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

• In contrast, 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.
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:
*liguleless1* and *liguleless2* explain the two “biggest” leaf angle QTL

**Upper leaf angle**

- Associations with positive effect
- Associations with negative effect
- Linkage QTL peak

**Chromosomes**

Tian, Bradbury, et al. 2011 *Nature Genetics*
ligule
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

ApeKI site (GCWGC)

Loss of cut site

(→) 64-base sequence tag

< 450 bp
Genetically mapping individual GBS alleles

SNPs (e.g., from Illumina 55K chip)

RILs (e.g., from IBM)

B73  Mo17  Heterozygote
Genetically mapping individual GBS alleles

SNPs (e.g., from Illumina 55K chip)

RILs (e.g., from IBM)

↓ does GBS tag map here?

B73  Mo17   Heterozygote
Genetically mapping individual GBS alleles

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

does GBS tag map here?

B73  Mo17  Heterozygote
Genetically mapping individual GBS alleles

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

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

Binomial Test for linkage

- prob. success: segregation ratio of the SNP being tested (~0.5)
- n trials: n RILs with GBS tag (10)
- n successes: n co-occurrences with presumed parental allele at SNP being tested (co-segregation)
- p-value: 0.00098 (<10⁻³)

does GBS tag map here?
Genetically mapping individual GBS alleles

These 10 SNPs all tie ($p = 9.8 \times 10^{-4}$)

Binomial Test for linkage

- prob. success: segregation ratio of the SNP being tested ($\sim 0.5$)
- n trials: n RILs with GBS tag (10)
- n successes: n co-occurrences with presumed parental allele at SNP being tested (co-segregation)
- p-value: 0.00098 ($<10^{-3}$)

SNPs (e.g., from Illumina 55K chip)

RILs (e.g., from IBM)

(→) 64-base sequence tag (GBS coverage $\sim 0.4x$)

B73 Mo17 Heterozygote
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>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...

- 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>
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
   - \( \approx 5\% \) of B73 sequence is not in the B73 reference genome
Mapping Chr0 and *de novo* contigs via GBS

• The sequences of Chr0 contigs are known
  ▪ so we know which *ApeKI* GBS tags are present
• *De novo* contigs constructed from 454 whole genome sequencing
  ▪ by collaborators at CSHL (*Ware et al.*)
  ▪ can predict *ApeKI* GBS tags from these
• Created a pipeline to genetically map novel contigs using linkage populations
• Used IBM GBS data for proof of concept
Adjacent tags on a Chr0 or de novo contig can be merged into haplotypes

ApeKl site (GCWGC)

B73

contig

→ 64-base sequence tag

< 450 bp

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

Novel? (not included in B73 RefGen_v2)
Genetically mapping GBS haplotypes improves resolution

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

RILs (e.g., from IBM)

These 3 SNPs all tie ($p = 1.5 \times 10^{-5}$)
Genetically mapping contigs via GBS

<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>novel: 8</td>
</tr>
<tr>
<td>B73 454 (k96)</td>
<td>3,964,387</td>
<td>CSHL</td>
<td>non-novel: 36,041</td>
</tr>
<tr>
<td>FLCrDNA</td>
<td>61,477</td>
<td>CSHL</td>
<td>novel: 407</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>non-novel: 10,776</td>
</tr>
</tbody>
</table>

\[ p < 10^{-7} \]
### Genetically mapping contigs via GBS

<table>
<thead>
<tr>
<th>Contigs</th>
<th>Total #</th>
<th>Source</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></td>
</tr>
<tr>
<td>B73 454 (k96)</td>
<td>3,964,387</td>
<td>CSHL</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>10,776</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Contigs</th>
<th>Total #</th>
<th>Source</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>---</td>
</tr>
<tr>
<td>B73 454 (k96)</td>
<td>3,964,387</td>
<td>CSHL</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>10,776</td>
</tr>
</tbody>
</table>

\[ p < 10^{-7} \]
Some regions of reference genome are missing large chunks

Telomere of Chr4 is a prime target for future improvement

Novel 454k96 contigs genetically mapped on Chr4
Conclusions – Improving the genome

• GBS data from a mapping population where one of the parents is the reference genome can help improve that reference genome

• Can help place:
  ▪ unanchored contigs (chromosome 0)
  ▪ contigs/BACS that have been misplaced (wrong chromosome)
  ▪ novel contigs from *de novo* sequencing (missing from the reference)

• These improvements incorporated into B73 RefGenV3

• Can uncover major structural differences between lines
This coming year – Improving the genome

• Add in GBS data from NAM for much higher resolution
  ▪ Recently constructed an imputed GBS framework map of NAM
  ▪ Anchor as many novel genes & contigs as possible

• Use GBS SNP calls in NAM plus >20,000 additional maize lines and map tags by LD (association mapping)
  ▪ Further improve genetic mapping resolution?
  ▪ Preliminary results: Median resolution = 50Kb
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