Genotyping By Sequencing (GBS)
Method Overview

RJ Elshire, JC Glaubitz, Q Sun, JV Harriman
ES Buckler, and SE Mitchell

http://www.maizegenetics.net/
Topics Presented

• Background/Goals
• GBS lab protocol
• Illumina sequencing review
• GBS adapter system
• How GBS differs from RAD
• Modifying GBS for different species
• Examples
• Future plans
Background

Genotyping by sequencing (GBS) in any large genome species requires reduction of genome complexity.

I. Target enrichment

• Long range PCR of specific genes or genomic subsets

• Molecular inversion probes

• Sequence capture approaches hybridization-based (microarrays)

II. Restriction Enzymes (REs)

*Technically less challenging*

• Methylation sensitive REs filter out repetitive genomic fraction
QTL are often located in non-coding regions

*Vgt1, Tb, B* regulatory regions 60-150kb from gene

Map large numbers of genome-wide markers

Exon capture
Our goal is to create a public genotyping/informatics platform based on next-generation sequencing.
Open Source

• Method available for anyone to use / modify.
• Analysis pipeline details and code are public.
• Promote dataset compatibility.
• Method published in *PLoS ONE* to promote accessibility.
• Genotype calls publically available.
Overview of Genotyping by Sequencing (GBS)

- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers
GBS is a simple, highly multiplexed system for constructing libraries for next-gen sequencing

- Reduced sample handling
- Few PCR & purification steps
- No DNA size fractionation
- Efficient barcoding system
- Simultaneous marker discovery & genotyping
- Scales very well
GBS 96-plex Protocol
(http://www.maizegenetics.net/gbs-overview)

1. Plate DNA & adapter pair
2. Digest DNA with RE
3. Ligate adapters
GBS Adapters and Enzymes

Barcode Adapter

Barcode (4-8 bp)

Illumina Sequencing Primer 1

“Sticky Ends”

Common Adapter

Illumina Sequencing Primer 2

Restriction Enzymes

ApeKI 5’ G CWGC 3’
PstI CTGCA G
EcoT22I ATGCA T
GBS 96-plex Protocol
(http://www.maizegenetics.net/gbs-overview)

1. Plate DNA & adapter pair
2. Digest DNA with RE
3. Ligate adapters
4. Pool 96 samples
   Clean-up
5. PCR
Pooled Digestion/Ligation Reactions (n=94)

Pooled Digestion/Ligation Reactions (n=94) → PCR → GBS “Library”

PRC primers:

O1

O2

O1

P1

O2

P2

P1

P2

Insert


GBS 96-plex Protocol
(http://www.maizegenetics.net/gbs-overview)

1. Plate DNA & adapter pair

2. Digest DNA with RE
3. Ligate adapters
4. Pool 96 DNA aliquots

Clean-up

5. PCR

Clean-up

6. Evaluate fragment sizes
Perform Titration to Minimize Adapter Dimers Before Sequencing

NOTE: Done once with a small number of samples.
Adapter dimers constitute only 0.05% of raw sequence reads

Optimal adapter amount
Small Fragments are Enriched in GBS Libraries

- B73 RefGen v1
- IBM (B73 X Mo17) RILs

ApeKI fragment size (bp)

Total Fragments

- 0% - 5%
- 5% - 10%
- 10% - 15%
- 15% - 20%
- 20% - 25%
- 25% - 30%
- 30% - 35%
- 35% - 40%

Fragment size distribution for different libraries.
Illumina Sequencing by Synthesis
Illumina NGS Platform

Flowcell
8 channels

Solid Phase Oligos
Cluster Formation Amplifies Sequencing Signal

1. Denatured “Library”
2. Bridge Amplification
3. Flow cell with bound oligos
4. Linearization
Sequencing by Synthesis

HiSeq 2000 Flowcell

P1 primer

TT G T G C T G C A

Sequencing by Synthesis
"In Phase"

"Out of Phase"
GBS captures barcode and insert DNA sequence in single read.
Variable Length GBS Barcodes Solves Sequence Phasing Issues

• First 12 nt used to calculate phasing.
• Algorithm assumes random nt distribution.
• Incorrect phasing causes incorrect base calls.

• Good design and modulating the RE cut site position with variable length barcodes produces even nt distribution.
Invariant GBS barcodes cause loss of signal intensity.
Successful GBS sequencing run.
Most significant GBS technical issue?

DNA quality
GBS does not use standard “Y” adapters
Same-ended Fragments Do Not Form Clusters

Denatured “library”

Flow Cell with Bound Oligos

Cluster Formation & “Linearization”

Bridge Amplification

Cleaved from surface

No P1 binding site
GBS vs. RAD

- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers
Digest

Ligate adapters

Pool

Random shear

Size select

Ligate Y adapters

PCR

RAD

GBS

Reference

Davey et al. 2011
Modifying GBS

Considerations for using GBS with new species and / or different enzymes.
Why Modify the GBS Protocol?

- More markers
- Fewer markers (deeper sequence coverage per locus)
- Increase multiplexing
- More genome appropriate (avoid more repetitive DNA classes)
- Other novel applications (i.e., bisulfite sequencing).
Genomic DNA Digest with methylation-sensitive RE

Digest with methylation-sensitive RE

Genomic DNA

PCR
Sampling large genomes with methylation-sensitive restriction enzymes.

Methylated DNA - 5-base cutter

Unmethylated DNA - 6-base cutter

GBS Library

Sequenced Fragments
Genome Sampling Strategies Vary by Species

Dependent on Factors that Affect Diversity:

• Mating System
  (Outcrosser, inbreeder, clonal?)

• Ploidy
  (Haploid, diploid, auto- or allopolyploid?)

• Geographical Distribution
  (Island population, cosmopolitan?)
Other Factors

• **Genome size**
  – The size of the genome has some bearing on the size of the fragment pool.

• **Genome composition**
  – The composition of the genome can affect the frequency and distribution of the cut sites.
### GBS SNP calls in *Sorghum bicolor*

<table>
<thead>
<tr>
<th>Alleles</th>
<th>Chromosome</th>
<th>SNP Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>S10_29954</td>
<td>C/A</td>
<td>10 A N A C A N N N C C N N N N N N N N C N</td>
</tr>
<tr>
<td>S10_29956</td>
<td>A/C</td>
<td>10 C N C A C N N A A N N N N N N N N A N</td>
</tr>
<tr>
<td>S10_29965</td>
<td>T/C</td>
<td>10 N C N C T C N N T T N N N N N N N N T N</td>
</tr>
<tr>
<td>S10_84178</td>
<td>C/G</td>
<td>10 N N C N G N N N N N N N C N N N N N N</td>
</tr>
<tr>
<td>S10_84512</td>
<td>T/C</td>
<td>10 T T N N T N T N N N N N C N N N N N N</td>
</tr>
<tr>
<td>S10_84515</td>
<td>C/T</td>
<td>10 C C N N C N C N N N N T N N N N N N N N</td>
</tr>
<tr>
<td>S10_80356</td>
<td>G/A</td>
<td>10 G A G A G N G G N N G G G G R N A N R G</td>
</tr>
<tr>
<td>S10_80356</td>
<td>G/A</td>
<td>10 G N N A G A G N N G G G N N A N G G G</td>
</tr>
<tr>
<td>S10_115500</td>
<td>T/C</td>
<td>10 T C T C T C T N N T T N N N N C N N T T</td>
</tr>
<tr>
<td>S10_113205</td>
<td>T/C</td>
<td>10 T C N N N N N N N N N N N N N N N N N N</td>
</tr>
<tr>
<td>S10_158101</td>
<td>G/A</td>
<td>10 G N N A G N N N N N N N G A A N G G N</td>
</tr>
<tr>
<td>S10_158101</td>
<td>G/A</td>
<td>10 G A N A G A N N N G N G N N N N N N G G N</td>
</tr>
<tr>
<td>S10_156374</td>
<td>G/A</td>
<td>10 G A N A N N G N G N N N N N N N N G G N</td>
</tr>
<tr>
<td>S10_156379</td>
<td>T/G</td>
<td>10 G T N T G N N N G G N G T T T T N G N</td>
</tr>
<tr>
<td>S10_160391</td>
<td>C/T</td>
<td>10 T C N C T N T N T N N T T C C C N T N N</td>
</tr>
<tr>
<td>S10_180445</td>
<td>G/A</td>
<td>10 G A G N G A G G G G N G A A N G G G</td>
</tr>
<tr>
<td>S10_200048</td>
<td>C/G</td>
<td>10 N G C G C G C C C C C C C C C C G G N C N</td>
</tr>
<tr>
<td>S10_218810</td>
<td>C/T</td>
<td>10 C C N N C N C N N C C N C N C C N C C C</td>
</tr>
<tr>
<td>S10_234030</td>
<td>T/G</td>
<td>10 T N T N T G T N T T T T T N N G T T N T</td>
</tr>
<tr>
<td>S10_234030</td>
<td>T/G</td>
<td>10 T G T G T G N T T T T T N N G G T T T T</td>
</tr>
<tr>
<td>S10_288307</td>
<td>G/A</td>
<td>10 N A N A N N G G G N G N N N A A N N N N N</td>
</tr>
<tr>
<td>S10_253729</td>
<td>G/T</td>
<td>10 G T G T G N N G G G N N N N N T G G G</td>
</tr>
<tr>
<td>S10_253730</td>
<td>C/T</td>
<td>10 C T C T C N N C C C C C C N N N N N N T T C C C</td>
</tr>
<tr>
<td>S10_261151</td>
<td>T/C</td>
<td>10 T C T C T C N N T T N N T N N T N N T N N</td>
</tr>
<tr>
<td>S10_268092</td>
<td>C/A</td>
<td>10 N N N N A N A N A N N N N N N N N N N N N</td>
</tr>
<tr>
<td>S10_268093</td>
<td>G/A</td>
<td>10 N N N N A N A N A N N N N N N N N N N N N</td>
</tr>
<tr>
<td>S10_278414</td>
<td>C/T</td>
<td>10 N N N N C N N N N C T C N N C T N N C N C</td>
</tr>
<tr>
<td>S10_282172</td>
<td>T/C</td>
<td>10 T N T N N N N N N N N N N N N N N N N T T</td>
</tr>
<tr>
<td>S10_287523</td>
<td>G/T</td>
<td>10 N N N N G G G G G N N N N N N N N N N N N</td>
</tr>
<tr>
<td>S10_290220</td>
<td>A/T</td>
<td>10 A T N N A N A N A A A A A A A N T T A A A</td>
</tr>
<tr>
<td>S10_290220</td>
<td>A/T</td>
<td>10 A T A T N N A A N A A N A A N N T T A A A</td>
</tr>
<tr>
<td>S10_294189</td>
<td>C/T</td>
<td>10 N N N N C N N N C C C N N N N C T N N C N C</td>
</tr>
<tr>
<td>S10_294189</td>
<td>C/T</td>
<td>10 C T C T C N N C N N C C C N C T T T N N C N C</td>
</tr>
<tr>
<td>S10_312580</td>
<td>A/C</td>
<td>10 A C A N A N A A A A A A A A A A C C N A A A</td>
</tr>
<tr>
<td>S10_347038</td>
<td>T/C</td>
<td>10 T C T C T C T T N T T T N N T N C N T T T T</td>
</tr>
</tbody>
</table>
Missing Data Strategies

- **Impute**
- **Technical Options**
  - Reduce the multiplexing level
  - Sequence the same library multiple times
- **Molecular Options**
  - Choose less frequently cutting enzymes
GBS Libraries in New Species

“Life is like a bowl of chocolates”.
Deer Mouse

ApeKI

PstI
Shrub Willow

**ApeKI**

**PstI**

**EcoT22I**
GBS Adapter Design

a) Barcode adapter

DNA insert

Barcode sequence

Common adapter

5' -ACACTTTTCCCTACACGAGCTCTTCCGATCT

3' -TGAGAAAGGGAGCGCTCGAGAAGGCTAGA

5' -ACACTTTTCCCTACACGAGCTCTTCCGATCT

b) PCR primer 1

Complement binds to flowcell oligo 1

Binds to 3' strand of barcode adapter

5' -ATGATACGGCGACCCAGATCTACACTTCTTCTACAGCGCTCTTCCGATCT

5' -ACACTTTTCCCTACACGAGCTCTTCCGATCT

Paired-end sequencing primer 1

c) PCR primer 2

Binds to 3' strand of common adapter

Complement binds to flowcell oligo 2

TCTAGCCTTCTCAGTCTCTCGAGAGATCTTCAGGCTCTGTCGCTAGAGCATACGCAAGAGACGAAAC

TCTAGCCTTCTCAGTCTCTCGAGAGATCTTCAGGCTCTGTCGCTAGAGCATACGCAAGAGACGAAAC

Paired-end sequencing primer 2
Barcode Design Considerations

- Barcode sets are enzyme specific
  - Must not recreate the enzyme recognition site
  - Must have complementary overhangs
- Sets must be of variable length
- Bases must be well balanced at each position
- Must different enough from each other to avoid confusion if there is a sequencing error.
  - At least 3 bp differences among barcodes.
- Must not nest within other barcodes
- No mononucleotide runs of 3 or more bases

http://www.deenabio.com/services/gbs-adapters
Future Priorities?

Reduce cost per sample
• Increase multiplexing levels (ApeKI 384-plex)

Robotics and adapters acquired
First libraries next week
Bioinformatics / Core Lab Platform

- Developing a complete pipeline from sample submission to marker database for end users.
- Analysis pipeline for species with a reference genome.
- Analysis pipeline for species without a reference genome.
- Document our experiences with various sample types and troubleshooting.