GBS Usage Cases: Non-model Organisms

Katie E. Hyma, PhD
Computational Biology Service Unit
Cornell University
The Best Advice I can Give You:

STOP
The Best Advice I can Give You:

Does my Experimental Design make Sense?

Do I have a Plan for Analysis?

Do my Results make Biological Sense??
Why can GBS be complicated?

- Sequencing Error
- SNP calling
- Missing data
- Large data sets
- Population type
- Biological Reality
Is GBS more complicated for non-model organisms?

- I don’t have a reference genome
- My reference genome is in a pre-assembly state, or the physical ordering is inaccurate
- My species is outcrossing
- I want to use GBS data for population genetics
Roadmap

- SNP calling particulars and file formats – reference pipeline
- Outcrossing species - Vitis example
- No reference genome
- Working with a less than perfect reference genome
- Population Genetics
Sequence based SNP calling: What’s in a SNP?

<table>
<thead>
<tr>
<th>Total Reads</th>
<th># minor allele reads to call heterozygote (1% error rate)</th>
<th># minor allele reads to call heterozygote (0.33% error rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Sequence based SNP calling:
What’s in a SNP?

- Total read depth: 7
- Major allele (“G”) count: 4
- Minor allele (“C”) count: 3
- SNP call: “S” (heterozygous G/C)
Sequence based SNP calling: What’s in a SNP?

- Total read depth: 7
- Major allele ("G") count: 6
- Minor allele ("A") count: 1
- SNP call: "G" (invariant)
Sequence based SNP calling: What’s in a SNP?

- Total read depth: 3
- Major allele ("G") count: 3
- Minor allele ("C") count: 0
- SNP call: "G" (homozygote)
Sequence based SNP calling: What’s in a SNP?

- Total read depth: 3
- Major allele ("G") count: 2
- Minor allele ("A") count: 1
- SNP call: “R” (A/G heterozygote)
Sequence based SNP calling: What’s in a SNP?

Depth of Coverage

Whole Genome Shotgun

GBS

Sites

Tags

Depth

Depth
Heterozygote undercalling in *Vitis* spp.

Without filters Aa > AA error rate (heterozygote undercalling) in *Vitis* is **30-50%**

50,951 SNPs
Mean depth 5.7x
Heterozygote Undercalling Solutions

1. Focus on the high coverage genotypes (through VCF plugin)
   Example: Vitisgen project

2. Use genetics (ex. Maize project)

- Each project is different, and different solutions are appropriate for each situation
SNP calling in *Vitis* (heterozygous, outcrossing)

- Read Depth of Each allele is captured for SNP calling in *Vitis* using the Variant Call format (VCF)
- 0/0:3,0,0:3:88:0,0,108
- Plugins available in TASSEL 3 to output VCF format
  - tbt2vcfPlugin
  - MergeDuplicateSNP_vcf_Plugin
  - Does NOT work with HDF5 TBT/TOPM
- VCF integration in TASSEL 4 in progress

Qi Sun  
Jon Zhang

Jeff Glaubitz  
Terry Casstevens
Sequenced based SNP calling: What’s in a SNP?

Likelihood score calculation. No allele frequency based prior probability was used due to mixed population structures.

\[
L(1/1) = P\left(n_1,n_2,n_3,n_4|1/1\right) \\
= \frac{n!}{n_1!n_2!n_3!n_4!} \left(1 - \frac{3e}{4}\right)^{n_1} \left(\frac{e}{4}\right)^{n_2+n_3+n_4}
\]

and

\[
L(1/2) = P\left(n_1,n_2,n_3,n_4|1/2\right) \\
= \frac{n!}{n_1!n_2!n_3!n_4!} \left(0.5 - \frac{e}{4}\right)^{n_1+n_2} \left(\frac{e}{4}\right)^{n_3+n_4}
\]

Hohenlohe PA, Cresko WA et al. PLOS Genetics 2010 Feb 26;6(2):e1000862.
Heterozygote undercalling in *Vitis* spp.
Working With VCF Files

- VCFtools

- VCFtools output formats:
  - plink *** can be used as TASSEL input
  - 012
  - IMPUTE
  - Idhat
  - BEAGLE-GL

- VCFtools is installed on CBSU BioHPC Computing Lab
  - [http://cbsu.tc.cornell.edu/lab/lab.aspx](http://cbsu.tc.cornell.edu/lab/lab.aspx)
How to Get VCF Output

**Tassel 3.0**

1. QseqToTagCountPlugin/FastqToTagCountPlugin
2. MergeMultipleTagCountPlugin
3. TagCountToFastqPlugin
4. Align
5. SAMConverterPlugin
6. QseqToTBTPlugin/FastqToTBTPlugin (use \(-y\) to output in TBTByte format)
7. MergeTagsByTaxaFilesPlugin
8. tbt2vcfPlugin
9. MergeDuplicateSNP_vcf_Plugin

- running each plugin without any options will output a list of available options on the console
- there is no plugin to merge taxa for VCF files. Merge samples by name at the TBT stage if desired using the flag \(-x\) with the MergeTagsByTaxaFilesPlugin.
Outcrosses

- Use pseudo-testcross markers (Aa x aa)
- Leverage deep sequencing of parental genotyping to phase markers
- Filter by Mendelian errors and/or Linkage Disequilibrium
- Methods development in *Vitis* for genetic mapping with GBS data
Dense maps from GBS data – in development for *Vitis*

Chromosome 1 markers from male parent in a VitisGen family (phased) and ordered by physical position

Major allele frequency distribution
Dense maps from GBS data – in development (*Vitis*)

- ~26,000 markers for a single family
- $0.4 < AF < 0.6$
- Separate maps for male and female parents
- No parental genotypes
- Up to 50% missing data per SNP locus

Chromosome 1 markers from male parent in a VitisGen family (phased) and ordered by physical position
No Reference Genome

- SNPs can be called with the UNEAK pipeline if there is no reference genome. (Fei Lu presented this morning)

- Be careful with your sampling strategy
  - Optimized for single species populations
  - Allele frequency matters
No Reference Genome

- Take care when analyzing highly divergent species/populations together - artificially inflated amount of fixed differences observed when divergence > 1.5%

Ex. The same locus may be identified as two separate loci

Ex. Some loci that are monomorphic within species but diverged > 1bp/tag between species will not be detected
No Reference Genome

- Allele frequency matters for SNP calling in the UNEAK pipeline – and is influenced by your population/sampling strategy.
- Undersampled diversity (and true rare alleles) will look like error.
Less than Perfect Reference

- SNPs can be called with the reference pipeline even on genomes that are not fully assembled
  - Hint: concatenate contigs into pseudo-chromosomes, but pad with “Ns” in between contigs to avoid aligning to junctions – for 64 bp tags pad with > 64 “Ns”
Less than Perfect Reference

- Assembly errors can affect alignment and SNP calling.
- Common with short read assemblies
- Reference allele does not match GBS alleles

Reference genome assembly

Biological Reality

Paralogous regions
Population Genetics and GBS

- Know what you need for your analyses before planning your project

- GBS is a powerful reduced representation approach
  - molecular methods
  - bioinformatics pipelines

- The GBS bioinformatics pipelines were developed and optimized for analyzing large mapping populations, not for small population genetics studies
Population Genetics and GBS

- Filter – expect to filter up to 90% of SNPs when imputation is not practical

- Pedigree independent filtering:
  - Read depth/genotype quality
  - Percentage of absent calls (missingness)
  - Minor allele frequency
  - For inbred species, inbreeding coefficient filters out most paralogous sites

- Likely to require customized bioinformatics
- Be ready for post-processing of data into other formats! Have a plan!
The Best Advice I can Give You:

STOP
The Best Advice I can Give You:

Does my Experimental Design make Sense?

Do I have a Plan for Analysis?

Do my Results make Biological Sense??