Shifting to HDF5

• Hierarchical Data Format – supports very large data sets and complex data structures.
• Widely used in climate and astromonomy communities
• TBT – files can approach 2 Tb in size
• Compressed HDF5 can be 40 times smaller
• Access times looks very good
• Working to fuse TOPM, TBT, and Keyfile into one HDF5 repository

Why can GBS be complicated? Tools for filtering, error correction and imputation.

Edward Buckler
USDA-ARS
Cornell University
http://www.maizegenetics.net
Maize has more molecular diversity than humans and apes combined

Silent Diversity (Zhao PNAS 2000; Tenallion et al, PNAS 2001)

1.34%

0.09%

1.42%

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
Maize genetic variation has been evolving for 5 million years

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>5mya</td>
<td>Modern Variation Begins Evolving</td>
</tr>
<tr>
<td>4mya</td>
<td>Sister Genus Diverges</td>
</tr>
<tr>
<td>3mya</td>
<td>Zea species begin diverging</td>
</tr>
<tr>
<td>2mya</td>
<td>Maize domesticated</td>
</tr>
<tr>
<td>1mya</td>
<td>Modern Variation Begins</td>
</tr>
</tbody>
</table>

What are our expectations with GBS?
High Diversity Ensures High Return on Sequencing

- Proportion of informative markers
  - Highly repetitive – 15% not easily informative
  - Half the genome is not shared between two maize line
    - Potentially all of these are informative with a large enough database
  - Low copy shared proportion (1% diversity)
    - Bi-parental information = (1-0.01)^64bp = 48% informative
    - Association information = (1-0.05)^64bp = 97% informative

Expectation of marker distribution

Biparental population

- Biallelic, 17%
- Too Repeative, 15%
- Non-polymorphic, 18%

- Presence / Absence, 50%

Across the species

- Multiallelic, 34%
- Too Repeative, 15%
- Non-polymorphic, 1%

- Presence / Absence, 50%
Sequencing Error

Illumina Basic Error Rate is ~1%

- Error rates are associated with distance from start of sequence
  - Bad – GBS puts these all at the same position
  - Good – Reverse reads can correct
  - Good – Error are consistent and modelable
Reads with errors

- Perfect sequences:
  \(0.99^{64} = 52.5\%\) of the 64bp sequences are perfect
  47.5 are NOT perfect

The errors are autocorrelated so the proportion of perfect sequence is a little higher, and those with 2 or more is also higher.

Do we see these errors?
- Assume 10,000 lines genotyped at 0.5X coverage

<table>
<thead>
<tr>
<th>Base</th>
<th>Type</th>
<th>Read # (no SNP)</th>
<th>Read # (w/ SNP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Major</td>
<td>4950</td>
<td>4900</td>
</tr>
<tr>
<td>C</td>
<td>Minor</td>
<td>17</td>
<td>67 (50 real)</td>
</tr>
<tr>
<td>G</td>
<td>Error</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>T</td>
<td>Error</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>
Do Errors Matter?

- Yes – Imputation, Haplotype reconstruction
- Maybe – GWAS for low frequency SNPs
- No – GS, genetic distance, mapping on biparental populations

Expectations of Real SNPs

- Vast majority are biallelic
- Homozygosity is predicted by inbreeding coefficient
- Allele frequency is constrained in structured populations
- In linkage disequilibrium with neighboring SNPs
Filters in TagsToSNPByAlignmentMTPlugin

- Only calls bi-allelic (hard coded now)
  - Two most common alleles used
- Inbreeding coefficient (-mnF)
  - If have inbred samples definitely use, very powerful for errors and paralogues
- Minimum minor allele frequency (-mnMAF)
  - Very important if do not have other tools for filtering (bi-parental populations or LD)
  - Set for >=1% if no other filter method present
MergeDuplicateSNPsPlugin

- When restriction sites are less than 128bp apart, we may read SNP from both directions (strands)
- ~13% of all sites
- Fusing increases coverage
- Fixes errors
- -misMat = set maximum mismatch rate
- -callHets = mismatch set to hets or not

GBSHapMapFiltersPlugin

- Basic filters for coverage of sites, taxa inbreeding coefficient, and LD
- -mnTCov = minimum taxa coverage (e.g. 0.05)
- -mnSCov = minimum site coverage, proportion of taxa with call (e.g. 0.10)
- -mnMAF = minimum minor allele frequency (e.g. 0.01)
GBSHapMapFiltersPlugin

- \textbf{mnF} = minimum inbreeding coefficient (e.g. 0.9) – \textit{Don’t use with outcrossers}
- \textbf{hLD} = require that sites are in high local LD, currently parameters are hard coded, so difficult to tune without using the code.
  - Tests a sliding window of 100 surrounding sites, and looks for a Bonferonni corrected \( P<0.01 \)
  - Useful but can be slow option.
  - More work needed here.

\textbf{Biparental populations}

Limited range of alleles, expected allele frequencies, high LD
Maize RIL population expectations

- Allele frequency 0% or 50%
- Nearby sites should be in very high LD ($r^2 > 50\%$)
- Most sites can be tested if multiple populations are available

Bi-parental populations allow identification of error, and non-Mendelian segregation
Bi-parental populations allow identification of error, and non-Mendelian segregation

Median error rate is 0.004, but there is a long tail of some high error sites
BiParentalErrorCorrectionPlugin

- `-popM = REGEX population identification (e.g. “Z[0-9]{3}”)`
- `-popF = population File (not implemented) instead of popM option`
- `-mxE = maximum error rate (e.g. 0.01); calculated from non-segregating populations`

BiParentalErrorCorrectionPlugin

- `-mnD = distortion from expectation (e.g. 2.0); the test uses both the binomial distribution and this distortion to classify segregation.`
- `-mnPLD = minimum linkage disequilibrium r^2= 0.5; this is calculated within each population, and then the median across segregating populations is used`
MergIdenticalTaxaPlugin

• Fuse taxa with the same name. Useful for checks and duplicated runs. Also useful in determining error rates
• -xHets = exclude heterozygotes calls (e.g. true)
• -hetFreq= frequency between hets and homozygous calls (e.g. 0.76)

Product of Filtering

• After filters, in maize we find 0.0018 error rate
  – AA<>aa = < 0.0018
  – AA<>Aa = 0.8 at low coverage
• SNPs in wrong location <~1%. Lower in other species.
Clean Up and Imputation

HapMap

MergeDuplicateSNPsPlugin
Merge reads from opposite sides

GBSHapMapFiltersPlugin
Site Coverage, Taxa Coverage, Inbreeding Coefficient, LD

BiParentalErrorCorrectionPlugin
Error rate estimation, LD filters

MergeIdenticalTaxaPlugin
Error rate estimation, LD filters

HapMap

Inbreds Partially SOLVED

Heterozygous Partially SOLVED

Imputation

Imputation & Phasing

GWAS

Kinship Distance Phylogeny LD GS

Process File (data structure)

Missing Data

Two major sources:

- **Sampling**
  - Low coverage often used in big genomes with inbred lines
  - Differential coverage caused by fragment size biases

- **Biological**
  - Region on genome not shared between lines
  - Cut site polymorphisms

*We want to impute the missing sampling but not the biological*
Standard Imputation

Lots of algorithms: FastPhase, NPUTE, BEAGLE, etc.

These are appropriate for high coverage loci, inbreds, and regions where biological missing is a rare condition

Some can be slow for sample sizes that we have.

FastImputationBitFixedWindow

- Imputation approach focused on speed and large sets of taxa with some closely related individuals.
- Nearest neighbor approach, fixed window sizes
- Strengths: Very accurate <1% error, much faster than other algorithms 100X
- Weakness: Not good a recombination junctions, heterozgyosity
- Code in TASSEL – not plugin, but available
Hidden Markov Model
TASSEL GBS Imputation

• Developed by Peter Bradbury
• Aimed a GBS and biparental populations
• Hidden Markov Model
• Very accurate at determining boundaries
• Works well on Maize NAM inbred lines, and probably others.
  • AA <> BB error rate– 0.00005
  • AB > AA – 0.0278
• Most problem appears in faulty populations
• Available as TASSEL 4.0 plugin

Only 50% of the maize genome is shared between two varieties

Maize

Plant 1
Plant 2
Plant 3

50%

Humans

Person 1
Person 2
Person 3

99%

Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005
Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010
Mapping all the alleles (TagCallerAgainstAnchor)

- Most maize alleles have no position on the reference map
- Map allele presence (TagsByTaxa) versus a anchor SNP map (HapMap)
- 8.7M alleles were mapped in <24 hours using 100 CPU cluster

Physical and genetic mapping of 8.7 million GBS alleles

- Only 29% of alleles are simple - physical and genetic agree
- 55% of alleles are easily genetically mappable
- Many complex alleles are rarer, so 71% of alleles are genetic and/or physically interpretable.
- With more samples and better error models perhaps 90% will be useable
Using the Presence/Absence Variants

- In species like maize, this is the majority of the data
- Less subject to sequencing error
- Need imputation methods to differentiate between missing from sampling and biologically missing

Future

- Need better integration of Whole Genome Sequence data with pipeline
  - Add information on premature cut sites or mutated cut sites
- Use paired-end read information
- Full incorporation of presence/absence variants
- Increase range of imputation tools and phasing for structure populations
- Quantitative genotype tools for polyploids/GS