New imputation strategies optimized for crop plants:
FILLIN (Fast, Inbred Line Library Imputation)
FSFHap (Full Sib Family Haplotype)

Kelly Swarts
PAG
Allele Mining
1/11/2014
Imputation is the projection of genotypes onto missing data

- Much of the new genotypic data is low coverage. e.g. Genotyping-by-Sequencing (GBS)
  - Cheaper
  - A problem for association studies and genomic predictions
    - Reduces power
    - Could reduce accuracy

- Projecting higher density to lower density data
Imputation Strategies

- Generate haplotypes
  - identity by state/descent segments and clustering
  - previously phased haplotypes (e.g., 1000 Genomes)

- Impute back onto target samples, usually some form of Hidden Markov Model (HMM)
Limitations of Current Implemented Algorithms

- Most designed for humans
  - Low structural variation (impute 100%)
  - Designed for obligate outcrossing species
- Slow
  - Unable to impute current ∼40k maize GBS lines
- Beagle v.4 (Browning and Browning 2013) current standard
New approaches optimized for crop plants

• FILL IN: Generalized approach for all populations
  – Optimized for large sparse genotypic matrix (GBS data)
  – Leverage inbred lines and segments to generate haplotypes

• FSFHap: Finely map recombination break points in full sib families
  – eg. Nested Association Mapping populations (NAM)
  – Fine mapping, understand patterns of recombination across genome
Viterbi for modeling transition between haplotypes

- First-order Hidden Markov Model (Rabiner LR 1989)

  True states defined as IBD parental genotypes, observations are SNP calls in target genotypes

- Viterbi finds most likely path through the state space

- Transition probabilities set by expectation-maximization (EM)-Baum-Welch algorithm, allowed to vary by distance between markers

- Emission probabilities defined by rates of sequencing error
Generate block haplotypes (42k taxa)

Impute back onto sample using haplotypes by block

- Impute one nearest neighbor haplotype
- Impute with two best haplotypes using Viterbi HMM

Impute 64 site subsets of blocks

- One haplotype
- Using Viterbi HMM
- Use two, resolve hets

DO NOT IMPUTE
FSFHap for full sib families

- FSFHap similar, optimized for full sib families
- Haplotypes from high coverage genotyped, known parents
- Viterbi HMM to determine recombination break points
Accuracy calculated by masking known genotypes

- Mask genotypes at high read depth
  - $P(\text{called Homo}|\text{Het})$
  - for $n=7$, 0.992

- Calculate accuracy by $R^2$ between known and imputed genotypes

- Beagle v.4 run on default settings

- Haplotypes not consistent for all algorithms (all use native)
Test datasets

- Three datasets to compare and understand differences between Beagle v.4 and FILL IN

HETEROZYGOSITY

DIVERsITY

Temperate Inbreds

Diversity Panel

Diverse Landraces
Temperate Inbreds

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Allele Frequency</th>
<th>Beagle v.4</th>
<th>FILL IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Genotype</td>
<td>Multiple $R^2$</td>
<td>Multiple $R^2$</td>
<td>Multiple $R^2$</td>
</tr>
<tr>
<td>Minor</td>
<td>0.0456</td>
<td>0.942</td>
<td>0.981</td>
</tr>
<tr>
<td>Predicted Het</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Het</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inbred, ex-PVP and Iowa breeding programs
Inbred Diversity Panel

### Known Genotype

- **Allele Frequency**
  - Beagle v.4
  - Multiple $R^2$: 0.883
- **Predicted Het**
  - FILL IN
  - Multiple $R^2$: 0.988

#### Frequency

- Taxon Coverage: 0.0, 0.4, 0.8
  - Frequency: 0, 40, 80, 140

*Inbred, selected for diversity*
Diverse Landraces

**Known Genotype**

- **Allele Frequency**
  - Multiple $R^2$: 0.0637

- **Beagle v.4**
  - Multiple $R^2$: 0.662

- **FILL IN**
  - Multiple $R^2$: 0.488

**Heterozygous, 96 from across Americas, 180 SW US, 96 Spanish**
FILLIN is better at imputing inbreds
FILLIN is MAF-independent for inbreds and biparental families.
FILLIN accurate even with no external information for related lines

FILLIN when external information excluded from haplotype generation

- **RILS**
- **Diverse Inbreds**
- **Landraces**

Sample Size

![Graph showing allele imputation proportion and $R^2$ with sample size](image)
Consensus imputation improves accuracy

<table>
<thead>
<tr>
<th></th>
<th>Diverse Landraces</th>
<th>Diverse Inbreds</th>
<th>Temperate Inbreds</th>
</tr>
</thead>
<tbody>
<tr>
<td>FILL IN</td>
<td>0.488</td>
<td>0.988</td>
<td>0.981</td>
</tr>
<tr>
<td>Beagle</td>
<td>0.662</td>
<td>0.883</td>
<td>0.942</td>
</tr>
<tr>
<td>Both agree</td>
<td>0.713</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>Beagle, when FILL IN does not impute</td>
<td>0.626</td>
<td>0.767</td>
<td>0.833</td>
</tr>
</tbody>
</table>
FILL IN is faster than Beagle v.4
24 full sib families, one common parent
Should you use it?

- Implemented in TASSEL 5 (very soon!)

Good for very large, inbred, low coverage datasets (eg. GBS for crop plants)

- If significant amounts of biologically missing data, FILL IN/FSFHap is less likely to overimpute.

- For full-sib families, use FSFHap if confident in parental genotypes

- For heterozygous germplasm, Beagle v.4 is better
Imputation Group

- Buckler Lab
  - Ed Buckler
  - Alberto Romero
  - Dong An
  - Peter Bradbury (FSPHap and Viterbi HMM)

- CIMMYT
  - HuiHui Li
  - Charles Chen