LaByRInth

An Improved Algorithm for Low-Coverage Biallelic Genetic Imputation

Jason Vander Woude
Outline

1. Project History
2. Genetics
3. Imputation
   a. LB-Impute
   b. LaByRInth
      i. Modeling
4. Future Work
Kansas State Agronomy

1. Correlate physical features with genetics in wheat and wheatgrass
   a. Plant height
   b. Number of seeds per head
2. Fit equations to data to model distributions
3. LaByRInth imputation
Chromosomes

1. Chromosomes encode genetic information

2. Chromosome is a sequence of bonded bases
   a. A/T and C/G
   b. 5’ and 3’
   c. 5’ATGACACTGTGACA3’ uniquely identifies
Heterozygous and Homozygous

1. Wheat has 42 chromosomes

2. Chromosomes come in pairs (homologs)
   a. Homologs serve same genetic purpose
   b. Base pairs can be completely different
   c. Each parent contributes one chromosome to each homologous pair
Genetic Imputation

1. Expensive to collect all genetic information

2. Patterns are expected based on known breeding

3. Imputation is used to fill in the gaps
   a. Build a mathematical model of the expected process
   b. Use known genetic sites to infer unknown sites
How well does imputation actually work?
Genetics of a Biallelic Homologous Chromosome Pair

- Reference base in both homologs
- Alternate base in both homologs
- 1 homolog with Ref. and 1 with Alt.
- 524 positions in the chromosomes
- 100 different plants
LB-Impute
Genetics vs LB-Impute
**LB-Impute**

1. Leaves large sections of the chromosome un-imputed
2. Designed for F2 populations
LaByRInth

1. Low-coverage Biallelic R-package Imputation
2. Initially supposed to be re-write of LB-Impute (Java to R)
3. Found many areas for improvement
   a. Project took unexpected direction
   b. A few weeks became more than a year
4. Open source
Modeling Strategies

1. Option 1: Use a model that ignores some biology (varying levels)
   - Often able to exactly “solve” the model
Modeling Strategies

1. Option 1: Use a model that ignores some biology (varying levels)
   ○ Often able to exactly “solve” the model

2. Option 2: Use a model that accurately captures biology
   ○ May not be able to “solve” the model exactly
Modeling Strategies

1. Option 1: Use a model that ignores some biology (varying levels)
   ○ Often able to exactly “solve” the model

2. Option 2: Use a model that accurately captures biology
   ○ May not be able to “solve” the model exactly

3. An analogy: find the area of a circle
   ○ Use a polygon to approximate the area
     i. Area of polygon may be able to be exactly computed
   ○ Use formula $\pi r^2$
     i. $\pi$ cannot be represented exactly
LaByRInth Strategy

1. Have not found a good way to do option 2 (capture biology)

2. Two different ideas for option 1 (exact solution to model)
   a. Extend LB-Impute strategy to other generations
      i. Assumes we can segment the chromosome
   b. New method based on different biological assumptions
      i. Assume limited genetic change during reproduction
How well does LaByRInth work?
Genetics
Genetics vs LaByRInth
This Summer

1. Implement and test both concept methods
   a. Real data
   b. Simulated data
2. Write and submit paper
3. Package code and release
Thanks to my advisors,

Dr. Nathan Tintle (Dordt)
Dr. Jesse Poland (Kansas State)
Dr. Mike Janssen (Dordt)
Questions?
Fitting Gaussian Curves

Outline

↣ Project History

17-LDH-STN-SAG-1009
BIC = 33420

17-LDH-STN-SAG-1009
BIC = 34321

Density

Height Difference (cm)
Mean Constrained