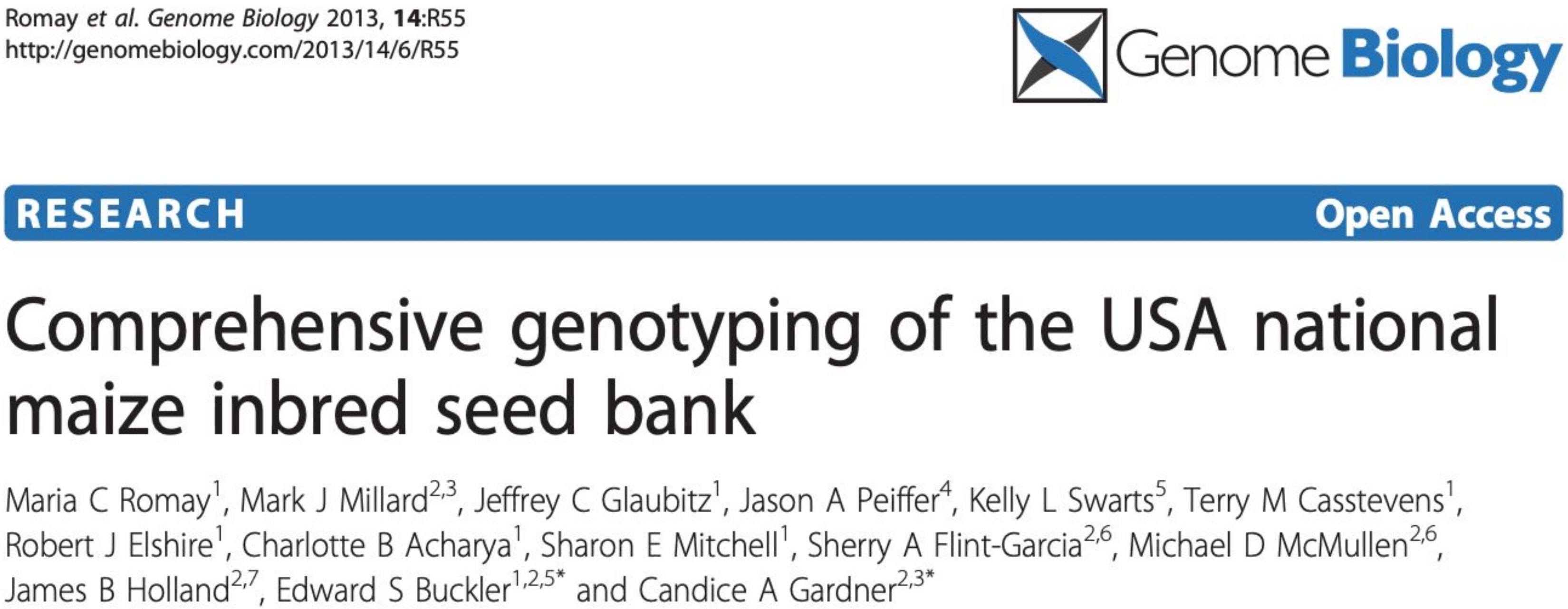
BCB/EEOB 546X
Computational Skills for Biological Data

Final Project Presentation

CAPTIVE

Daniel Kohlhase
Alejandro Ledesma
Cassie Winn
Anderson Verzegnazzi



Romay *et al. Genome Biology* 2013, **14**:R55
http://genomebiology.com/2013/14/6/R55

Genome Biology

RESEARCH Open Access

Comprehensive genotyping of the USA national maize inbred seed bank

Maria C Romay[1], Mark J Millard[2,3], Jeffrey C Glaubitz[1], Jason A Peiffer[4], Kelly L Swarts[5], Terry M Casstevens[1], Robert J Elshire[1], Charlotte B Acharya[1], Sharon E Mitchell[1], Sherry A Flint-Garcia[2,6], Michael D McMullen[2,6], James B Holland[2,7], Edward S Buckler[1,2,5*] and Candice A Gardner[2,3*]

Cornell University, Ithaca, NY
Department of Agriculture (USDA)
Iowa State University, Ames , IA
North Carolina State University, Raleigh, NC
University of Missouri, Columbia, MO

Cited by 284

Article accesses: 3,705

Biological Relevance

- Maize (Zea mays L.) is one of the most important crops in the world (human food, animal feed, and raw material for some industrial processes)
- Germplasm banks are huge sources of diversity
  - Diversity is important for association mapping
  - Only a modest amount of the available diversity is present in the commercial germplasm
- Takes time and money to evaluate and sequence

Paper Relevance

- **Genotyping by sequencing (GBS)** - Elshire et al., 2011, Institute for Genomic Diversity, Cornell University, Ithaca, New York
    - Procedure that is simple, quick, extremely specific and high reproducible
    - Provides a large number of markers across the genome at low cost per sample
    - Methylation-sensitive REs, avoiding repetitive regions of genomes and lower copy regions targeted with two to three fold higher efficiency

Paper Relevance

- **Genotyping by sequencing (GBS)**
  - High-quality reference genome for the maize inbred B73 to align the position of the SNPs
  - GBS enables characterization of germplasm collections on a genome-wide scale
  - Expands the number of individuals and markers under study
  - Increases the chances of discovering more uncommon or rare alleles

↓

GWAS and GS

Objectives

1. Compare GBS sequencing technology with other available options
2. Explore the potential of GBS to help with curation and use of germplasm
3. Evaluate genetic diversity and population structure both across the genome and between groups of germplasm
4. Investigate the history of recombination and LD through the different breeding groups
5. Explore the potential of the collection as a resource to study the genetic architecture of quantitative traits.

Description of Data

Genotypic data (GBS)

- 2,815 maize inbred accessions (USDA-ARS NCRPIS collection) some sequenced multiple times -> 4,351 samples
- 620,279 SNP markers across the Genome that are polymorphic among samples
- Data separated by chromosome in .txt files
- File sizes: 433 - 975 Mb
- Raw data file: 5.7 Gb

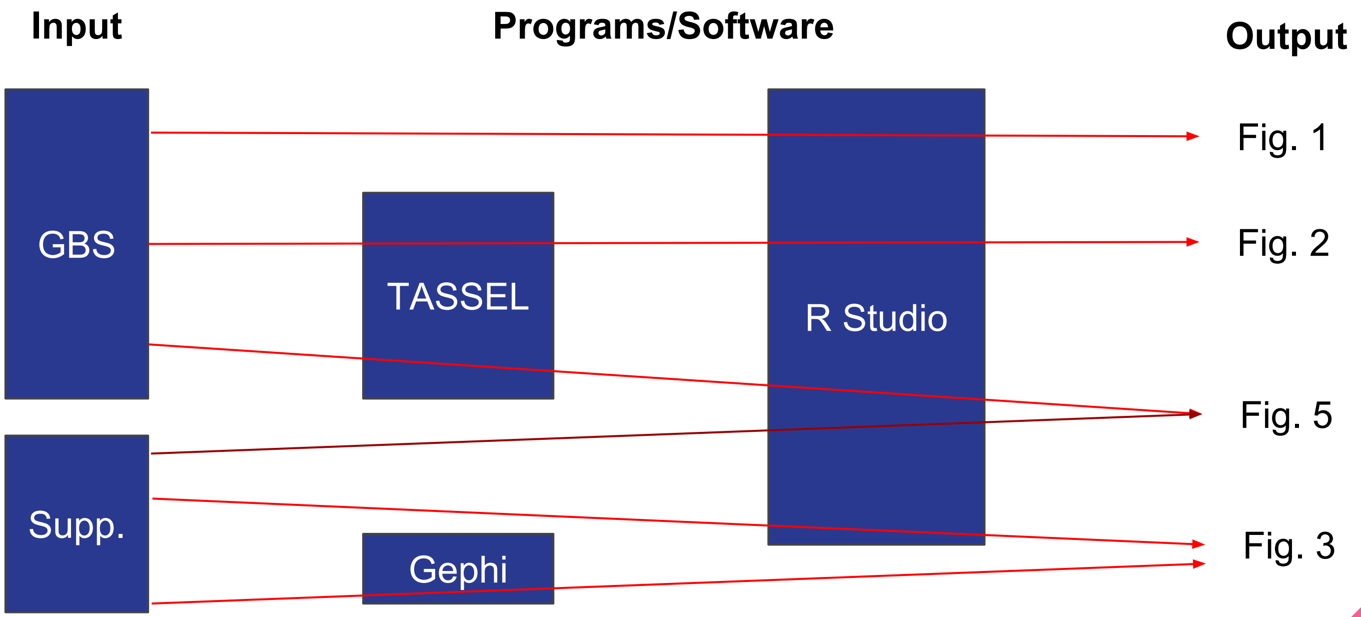
Supplemental data

- IBS data for each line and the 10 lines most closely related (generated using PLINK)
- Excel file with list of subgroups for each line (tropical, stiff-stalk, etc.)

Phenotypic data

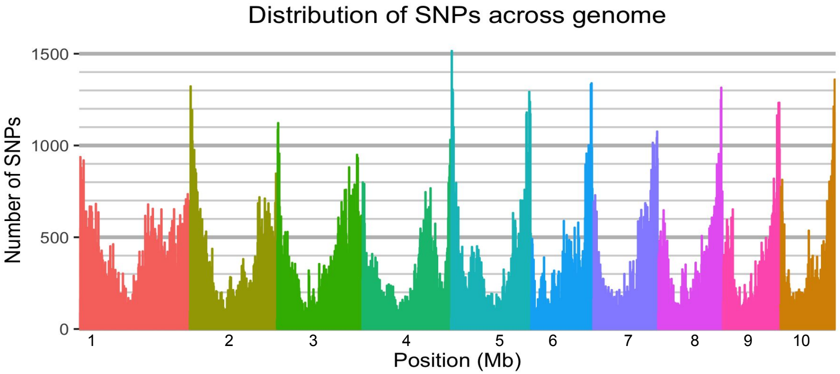
- 2,649 maize inbred accessions
  - .txt file 96.4 kb
  - BLUPs for kernel color, starch, GDD to silking

Description of Workflow



Results & Comparison - SNP Distribution

R Studio



Romay et al. 2013

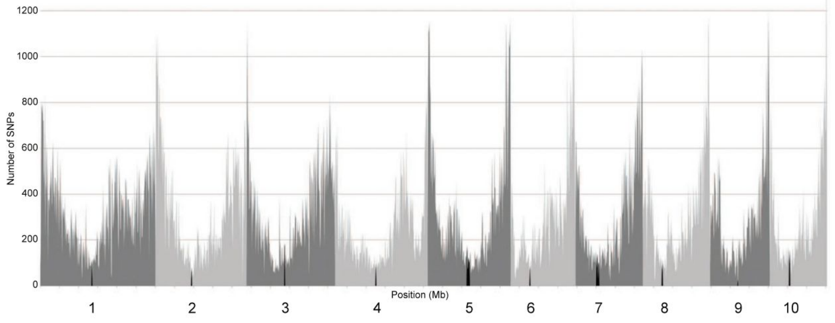
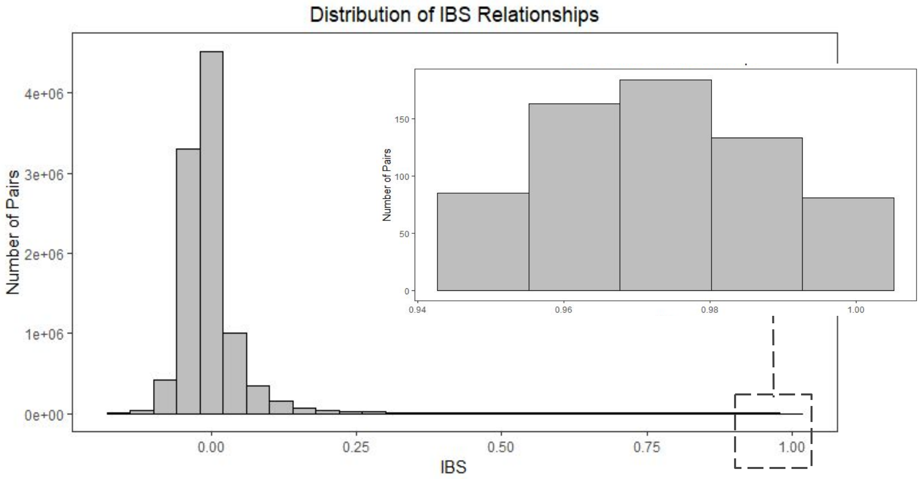


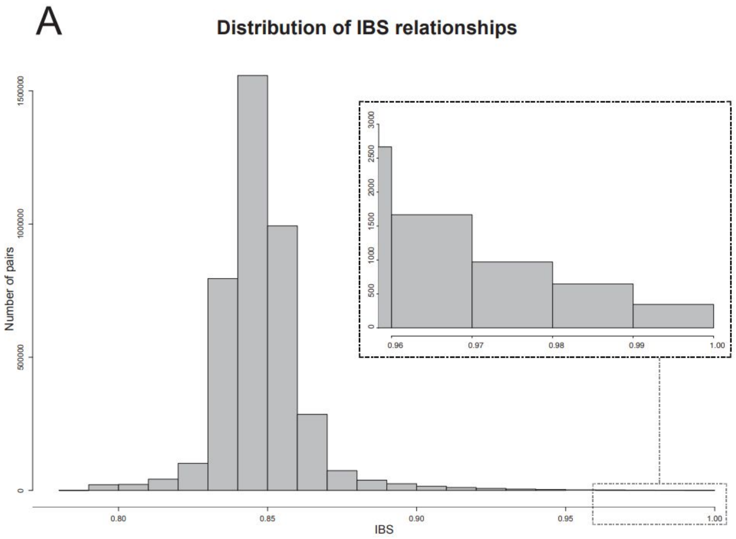
Figure 1 distribution of single-nucleotide polymorphisms (SNPs) across the genome. Distribution of the number of SNPs found in 1 Mb windows across the 10 maize chromosomes. Centromere positions are shown in black.

Results & Comparison - IBS Relationships

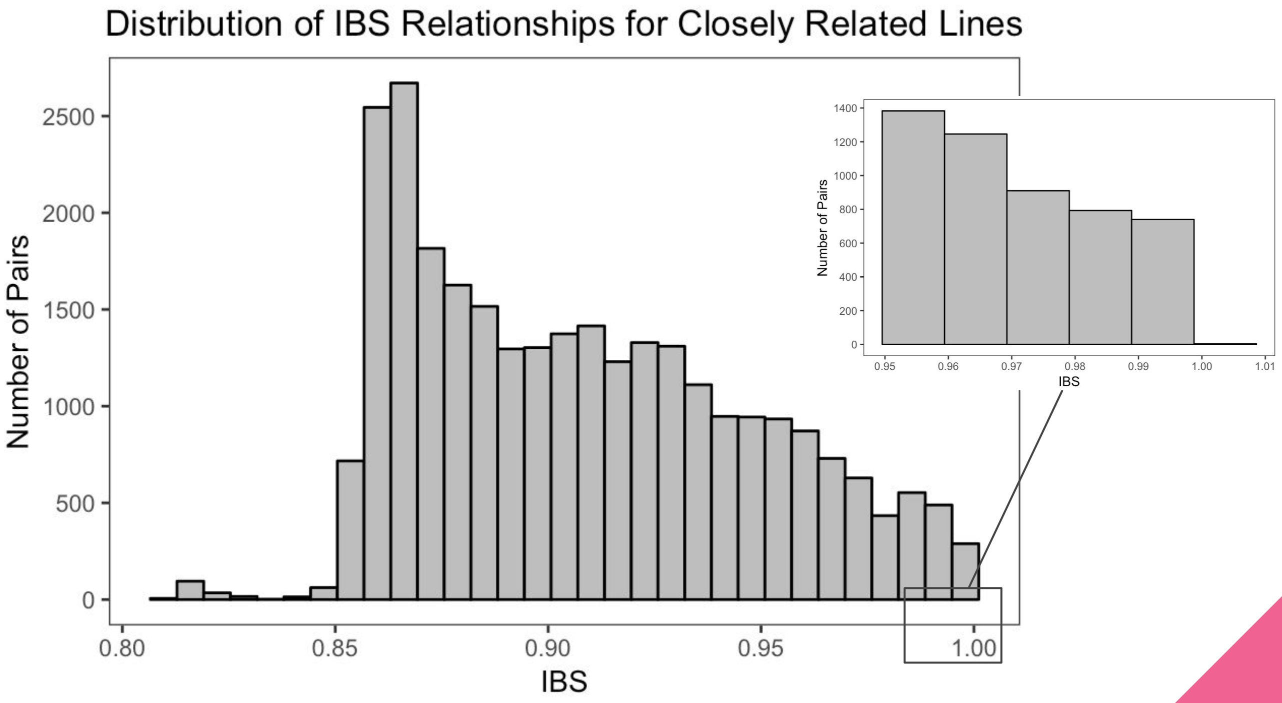
TASSEL + R Studio



Romay *et al.* 2013 - PLINK

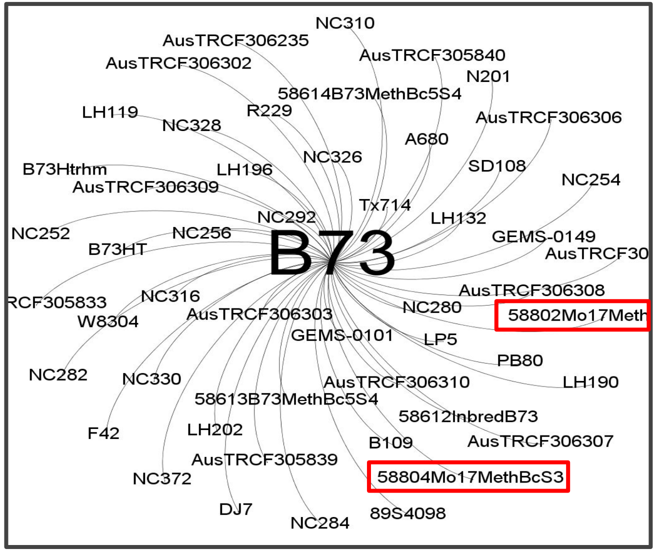


Results & Comparison - IBS Relationships



Results & Comparison - B73 Network Diagram

Gephi-0.9.2



Romay et al. 2013 (Gephi-0.8)

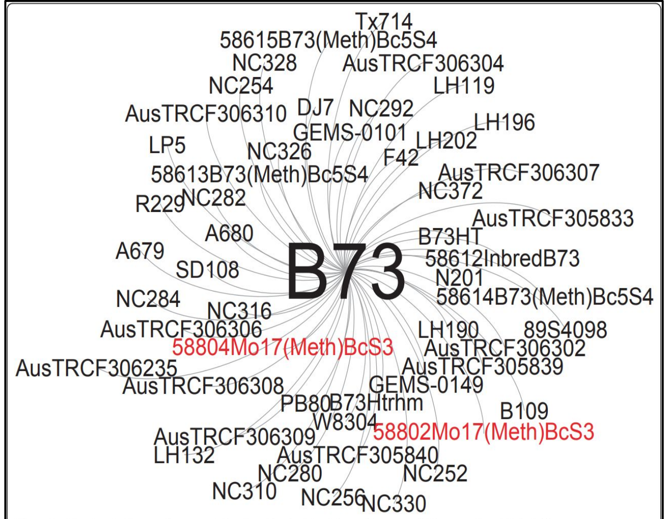
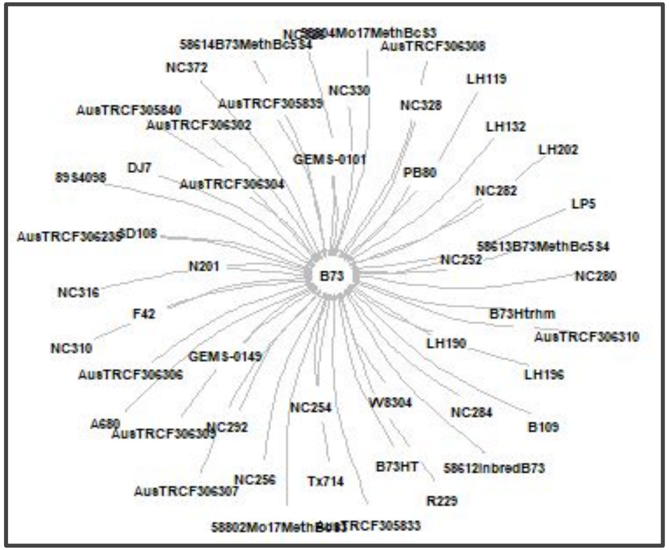


Figure 3 ***B73*** **network diagram.** Network relationships of maize inbred lines with values of IBS greater than 0.97 for *B73*.

- B73 and the 51 closest related inbred lines based on IBS > 97 %
- Identify accessions misclassified
- Select best sources for multiplication/distribution
- Eliminate duplications
- Select core collection

Results & Comparison - B73 Network Diagram

R Studio



Romay et al. 2013 (Gephi-0.8)

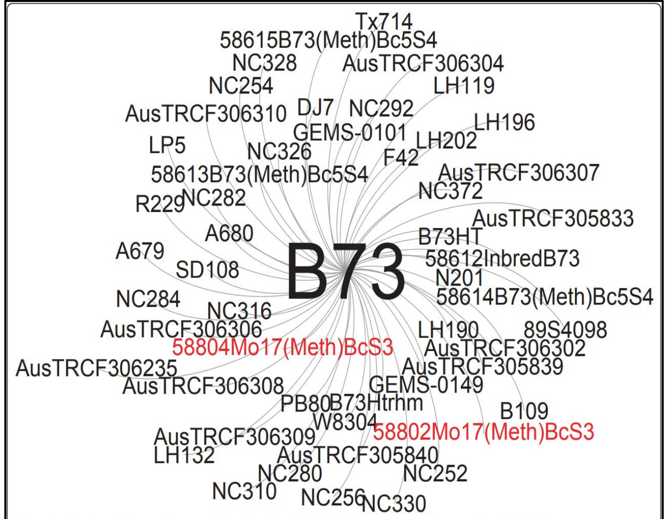
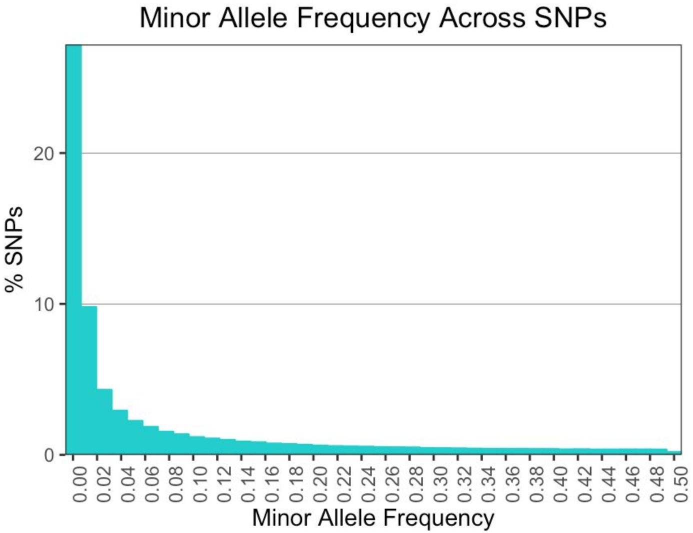


Figure 3 ***B73* network diagram.** Network relationships of maize inbred lines with values of IBS greater than 0.97 for *B73*.

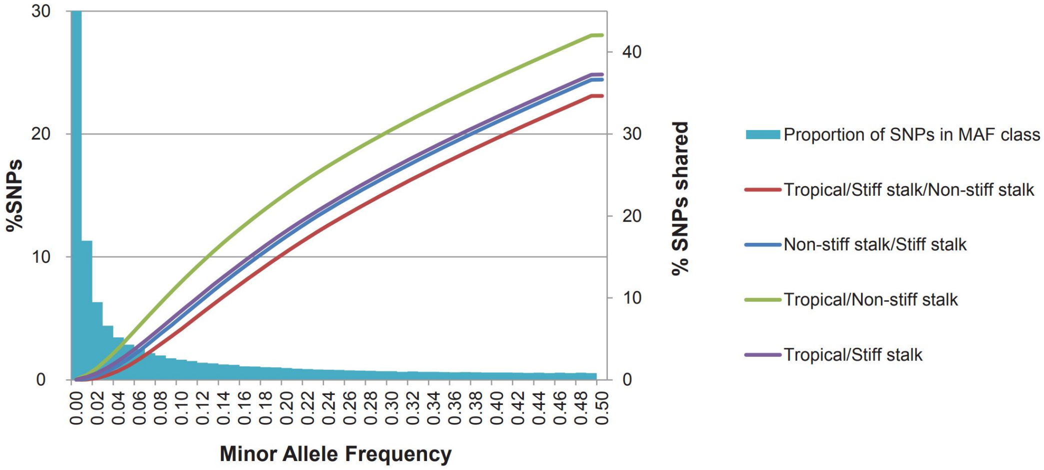
- R studio package “igraph” and documentation (Ognyanova, 2018)
- Additional file 2 was provided distribution IBS relationships and the 10 closest neighbors for each inbred line.
- We extracted the lines with > 97% IBS using Excel functions.

MAF and % of SNPs shared between subpopulations

TASSEL + R Studio



Romay *et al.* 2013



GWAS

TASSEL

Romay et al. 2013

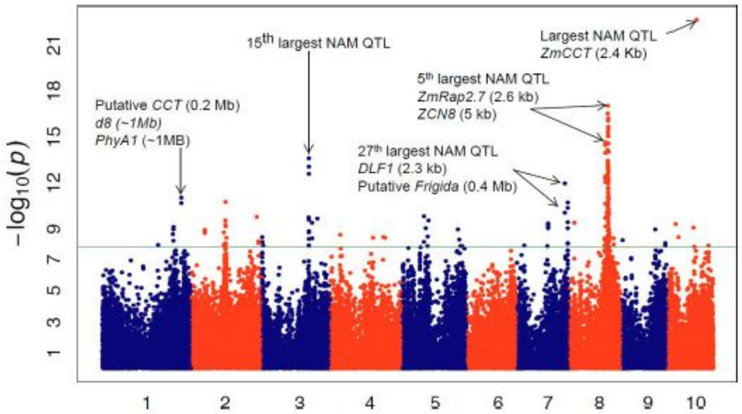


Figure 11 Genome-wide association study (GWAS) for growing degree days to silking. GWAS for growing degree days to 50% silking on 2,279 maize inbred lines. NAM, nested association mapping; QTL, quantitative trait loci.

- PLINK issues
  - Learning new program
  - Input formats
- TASSEL issues
  - Phenotypic data
  - Relationship File

Challenges & Conclusion

- Close!
- Multiple platforms/software
  - Collaboration opportunities
  - Learning Opportunities
  - Version Control
  - More Detailed Methods Description?
- File Format
  - Raw vs input
  - Filtering Ambiguity

Thank you!