Accuracy of Eight Genomic Selection Models in the Prediction of Salt Tolerance in Alfalfa

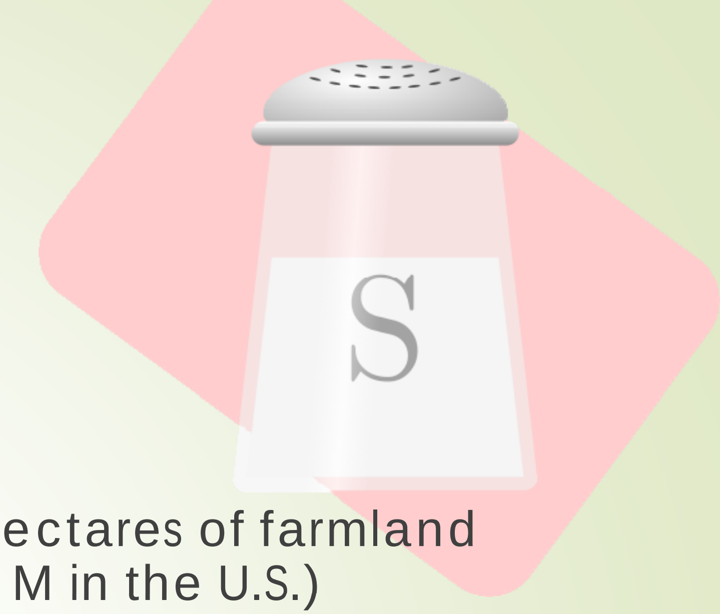
Charles Hawkins, Ph.D.

United States Department of Agriculture Agricultural Research Service, Long-Xi Yu group

Alfalfa

- Perennial, cool-season forage legume
- Used for hay, silage, pasture
- In 2016, 58M tons were produced in the U.S., including 2.2M in WA and 2.2M in UT (USDA NASS)

Salt



- A 1989 study: Worldwide, 351.5 M hectares of farmland were afflicted with high salinity (6.2 M in the U.S.)
- Primarily sodium salts, but also calcium, magnesium, potassium, iron, boron, sulphate, carbonate, and bicarbonate salts
- Saline soil is bad for crop productivity
- High soil salt draws water out of plants, subjecting them to osmotic stress similar to drought
- Salts taken up by the plant can also cause direct toxicity to plant tissues
- Global losses to salt in 2014 were estimated at $27 Bn
- Irrigation can increase field salinity, esp if drainage is poor
- Saline fields require additional irrigation water to flush out salt (leaching fraction)

Alfalfa and Salt

- Soil salinity is measured by soil electrical conductivity (EC)
- Typically measured in deciSiemens per meter (dS/m)
- Alfalfa is classified as moderately salt-sensitive
- 50% yield loss at 8.8 dS/m

Conventional Breeding

- Phenotypic Selection
  - Evaluate traits of each generation, select based on evaluations
  - Accurate but slow
- Pedigree-Based Selection (BLUP)
  - Generate an estimated breeding value (EBV) based on pedigree
  - Less accurate, but faster
  - All (full) siblings receive the same EBV

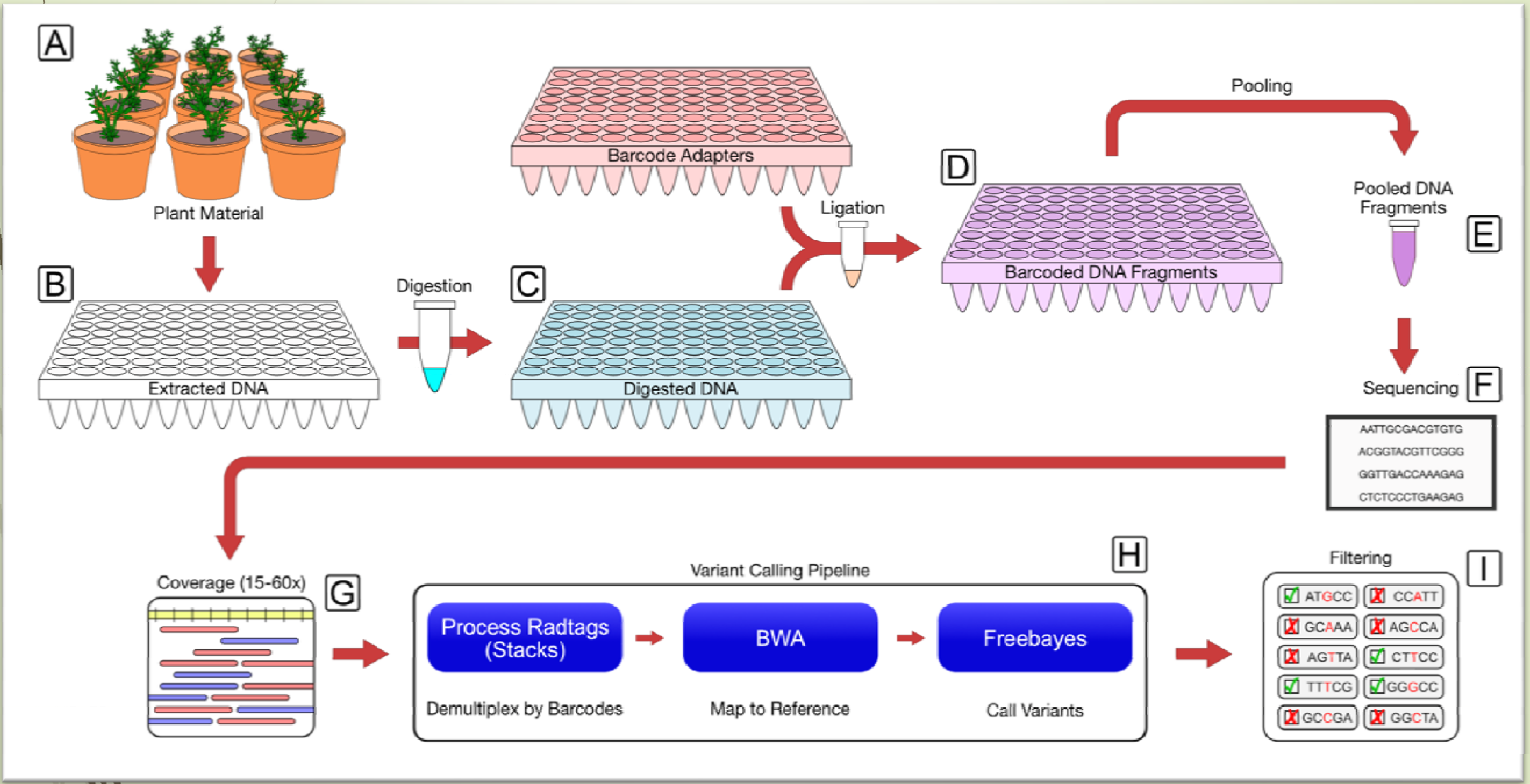
Marker-based Breeding

- Markers: Any locus that varies within your population that you know about and can test for
- Discover markers and evaluate trait(s) of interest
- Determine association between markers and traits
- With this, subsequent generations can be selected based on markers
  - Testing for markers is quick, can be done on young plants
- Questions:
  - What type of marker?
  - How are associations determined?

GBS

- Restriction digest genomic DNA, sequence ends of restriction fragments
- Reduced Representation – get sequence for about $^{1}/_{7}$ of the genome in total
- Effects from unsequenced regions can be captured via linkage
- GBS can generate 10,000+ SNP / MNP markers
- Less costly than whole-genome sequencing (WGS)

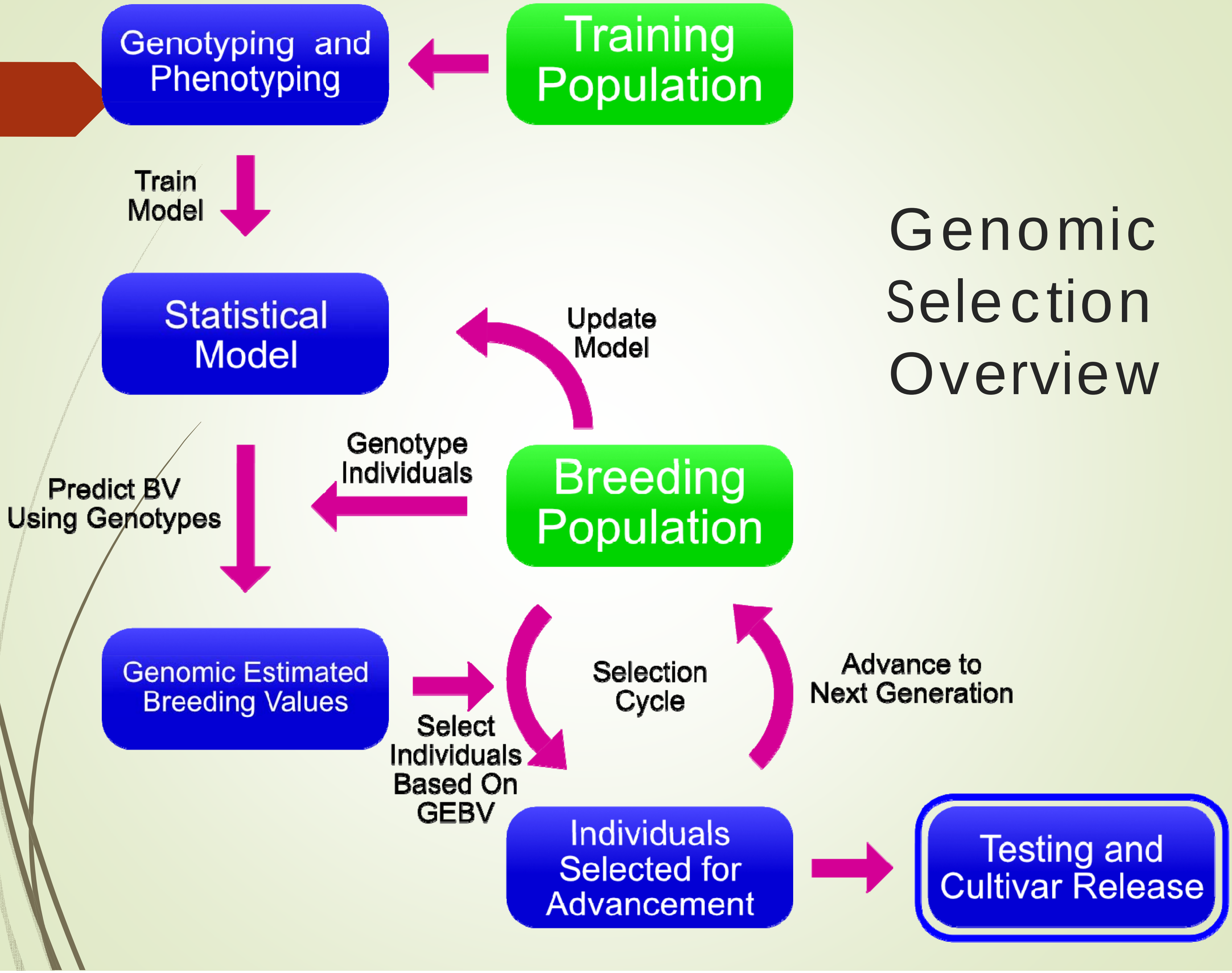
GBS Pipeline



Discovery of marker-trait associations

- Conventional Marker-Assisted Selection (MAS)
  - Probe each marker for significant association with trait
  - Identify the top few markers (usually 5-10)
  - Select for plants that have more of the “good” marker variants, drive good markers to fixation
- Genomic Selection (GS)
  - Train a statistical or machine-learning model using all the markers
  - For plants under selection, generate a Genomic Estimated Breeding Value (GEBV) using the trained model, then select based on GEBVs
  - Already being used for cattle breeding

Genomic Selection Overview



Training Population

Genotyping and Phenotyping

Train Model

Statistical Model

Update Model

Genotype Individuals

Breeding Population

Predict BV Using Genotypes

Genomic Estimated Breeding Values

Select Individuals Based On GEBV

Selection Cycle

Advance to Next Generation

Individuals Selected for Advancement

Testing and Cultivar Release

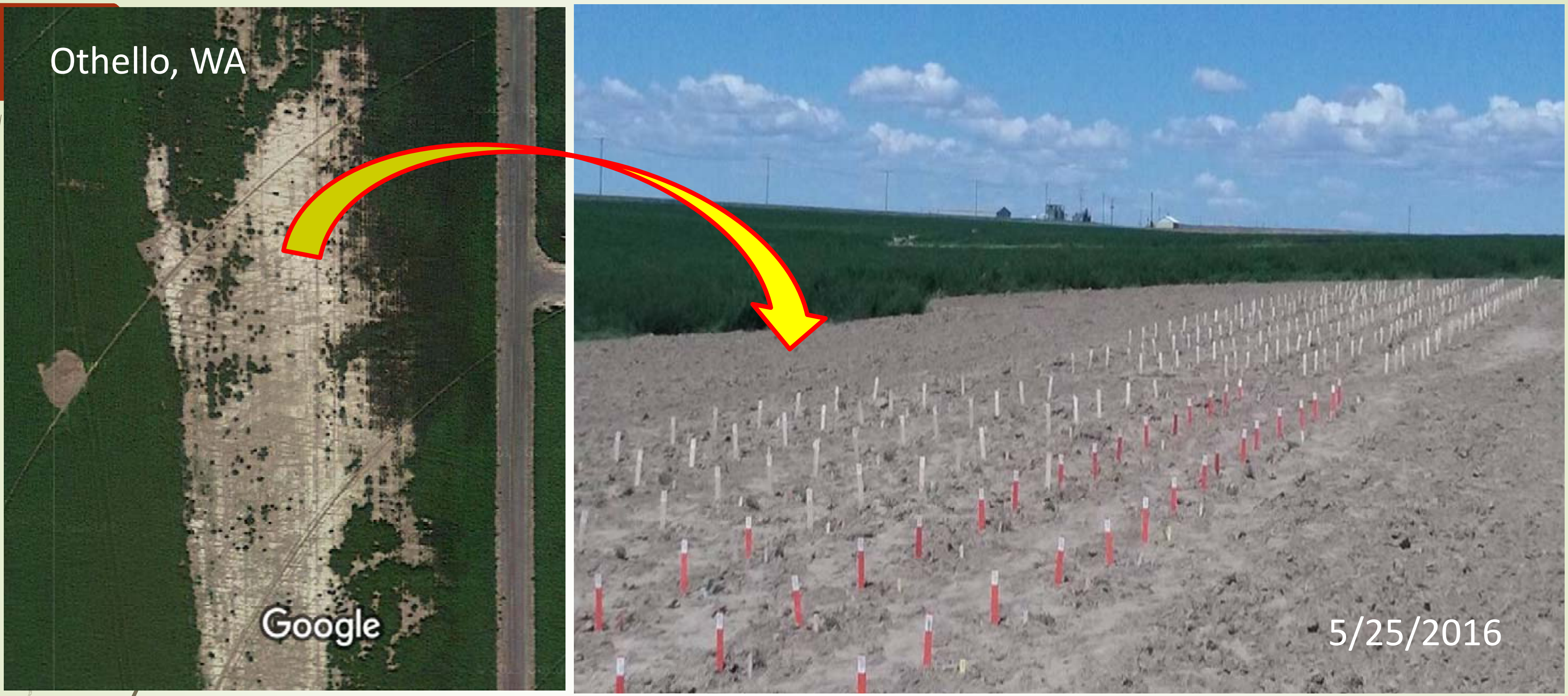
Cross-Validation

- How do we know our model will make good predictions before starting the breeding cycle?
- Cross-Validation
- Randomly assign plants to be part of the "training set" or the "validation set"
- Train the model based on the training set, see how well it predicts the traits of the validation set. Then pick a new training set and repeat. Accuracy is the average correlation between predicted and measured trait values over 800 replicates
- Cross-validation also helps us choose between models and set the parameters of the model we've chosen

Our Project

- Breed alfalfa for improved salt tolerance using Genomic Selection
- Starting material is 280 plants of already-improved alfalfa from Logan, UT
  - Previously bred for salt survival via three cycles of phenotypic selection, one cycle for survival and forage production
- Traits of Interest: Health measures under salt stress in a field and greenhouse, yield under salt stress in a field
  - Field: Single plants grown in Castle Dale, UT; health scores
  - Field: 3 replicates, one plant per plot, in Othello, WA; yield
  - Greehnouse; various growth metrics

Field test for alfalfa salt tolerance is in progress in the Othello farm of WSU


Othello, WA

5/25/2016

6/23/2016

6/23/2016

6/23/2016

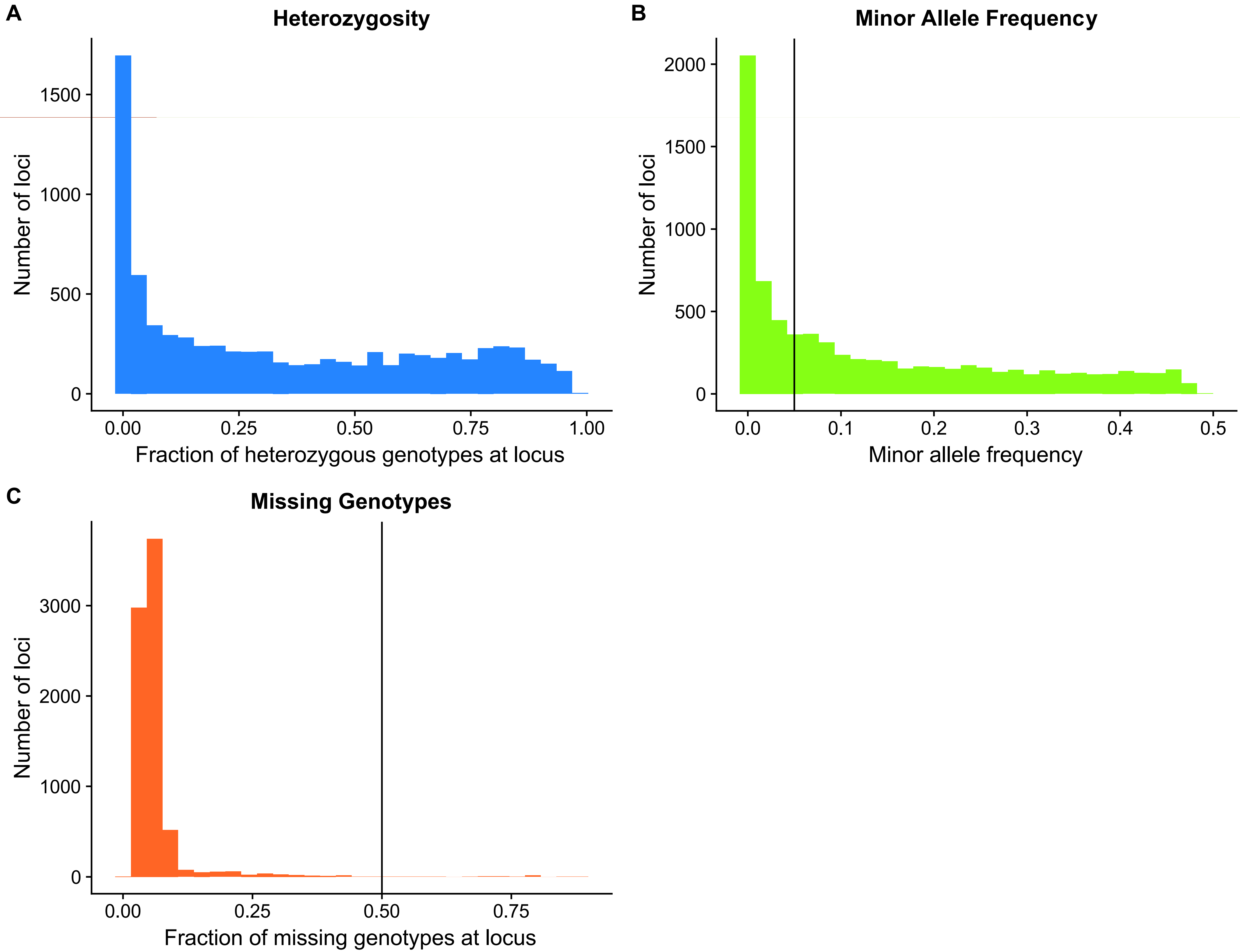
August, 2016

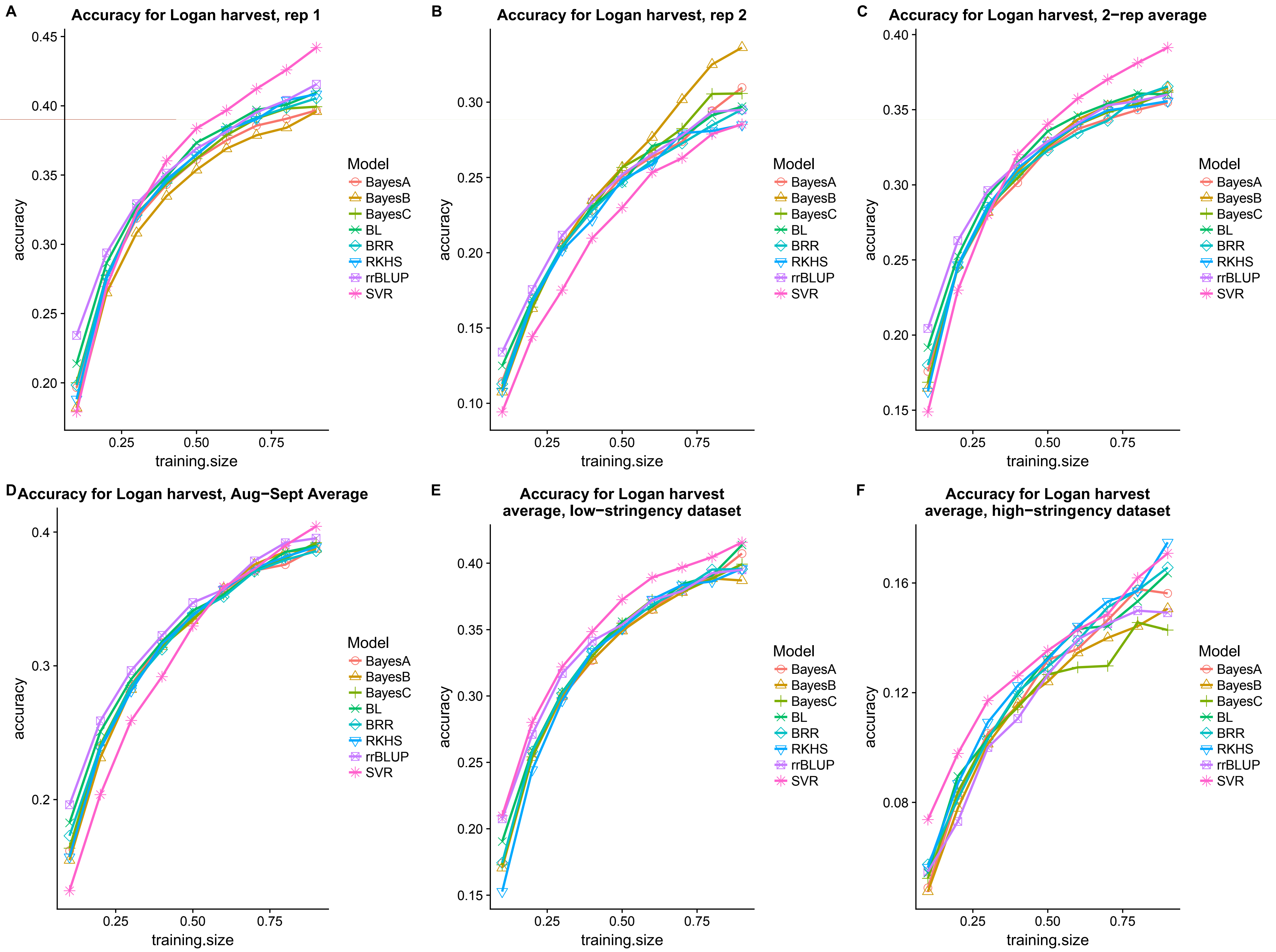
Marker Filtering

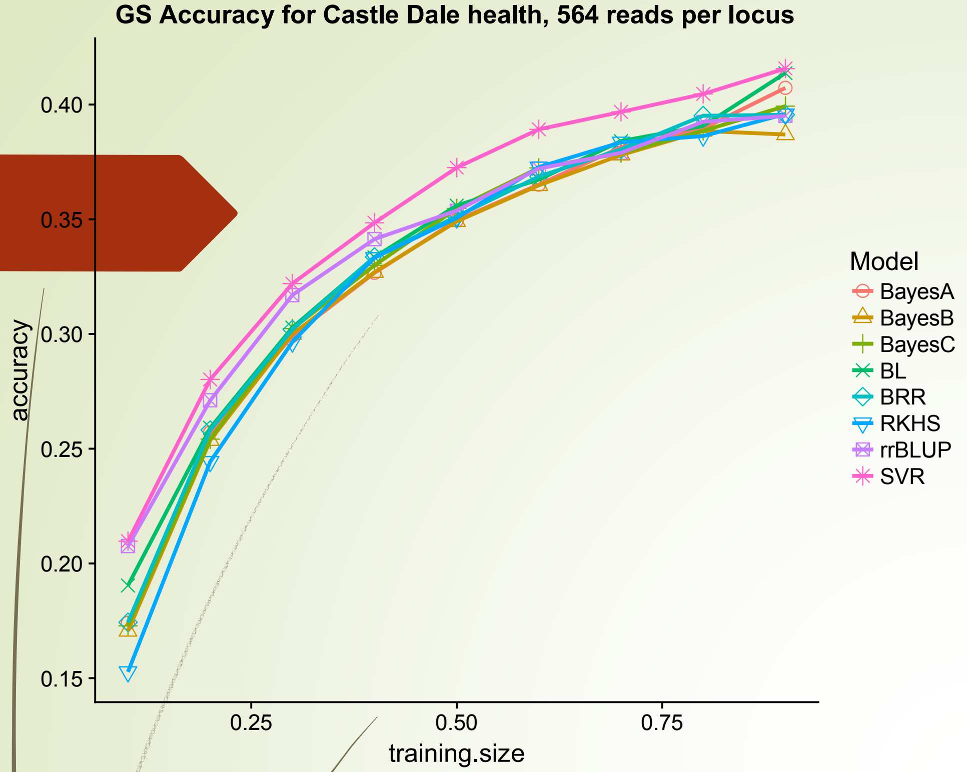
- To be called, a locus must have a read depth of 1410 reads (avg 5 reads per sample)
- To be used, markers had to pass the following tests:
  - Quality score > 20
  - No more than 50% of plants unknown for that marker
  - Less-frequent marker variant must be in at least 5% of plants
- To be used, plants must have no more than 50% of markers unknown

Genotyping Results

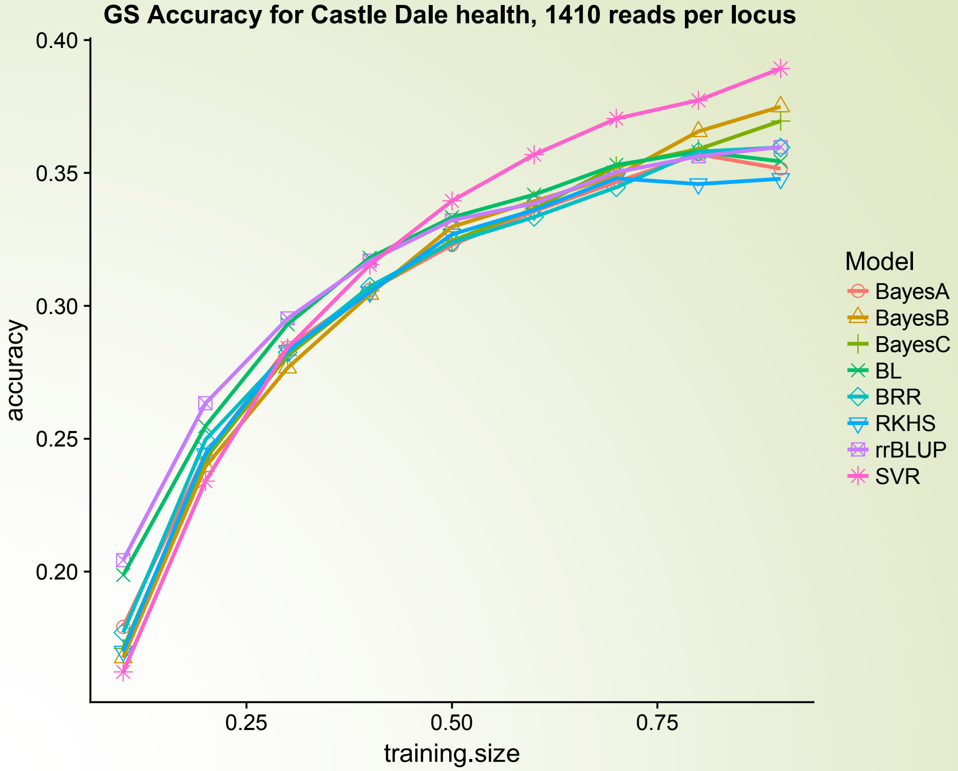
- Genotyping-by-sequencing done on an Illumina HiSeq 2000, 100bp single-ended reads
- 240,444,007 sequencing reads obtained
- 31,948,048 could be located within the genome (mapped)
- 7,679 markers obtained, 4,315 passed filtering
- No plants were excluded for having too many missing marker genotypes
- Tested 8 models: Ridge regression, Bayesian ridge regression (BRR), Bayesian Lasso (BL), BayesA, BayesB, BayesC, reproducing kernel Hilbert Space (RKHS), support vector regression (SVR)


A
Heterozygosity
Number of loci
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
Fraction of heterozygous genotypes at locus
B
Minor Allele Frequency
Number of loci
2000
1500
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
Minor allele frequency
C
Missing Genotypes
Number of loci
3000
2000
1000
0
0.00
0.25
0.50
0.75
Fraction of missing genotypes at locus

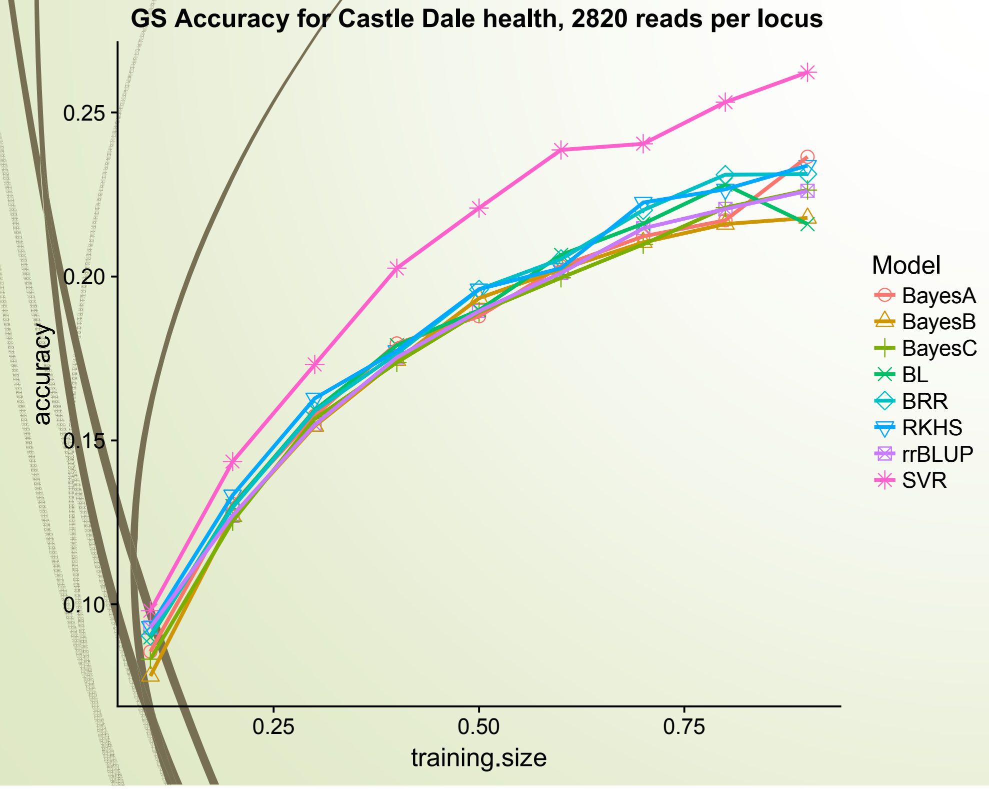

A
Accuracy for Logan harvest, rep 1
B
Accuracy for Logan harvest, rep 2
C
Accuracy for Logan harvest, 2−rep average
D
Accuracy for Logan harvest, Aug−Sept Average
E
Accuracy for Logan harvest average, low−stringency dataset
F
Accuracy for Logan harvest average, high−stringency dataset
accuracy
training.size
0.25
0.50
0.75
Model
BayesA
BayesB
BayesC
BL
BRR
RKHS
rrBLUP
SVR
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.4
0.3
0.2
0.16
0.12
0.08



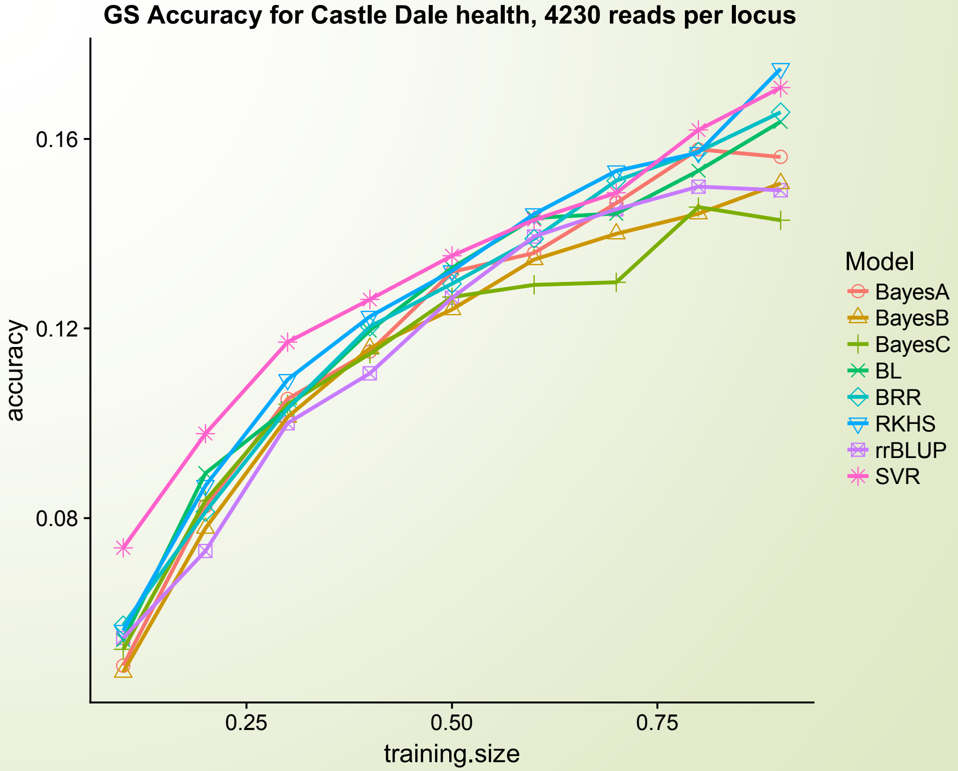
GS Accuracy for Castle Dale health, 564 reads per locus



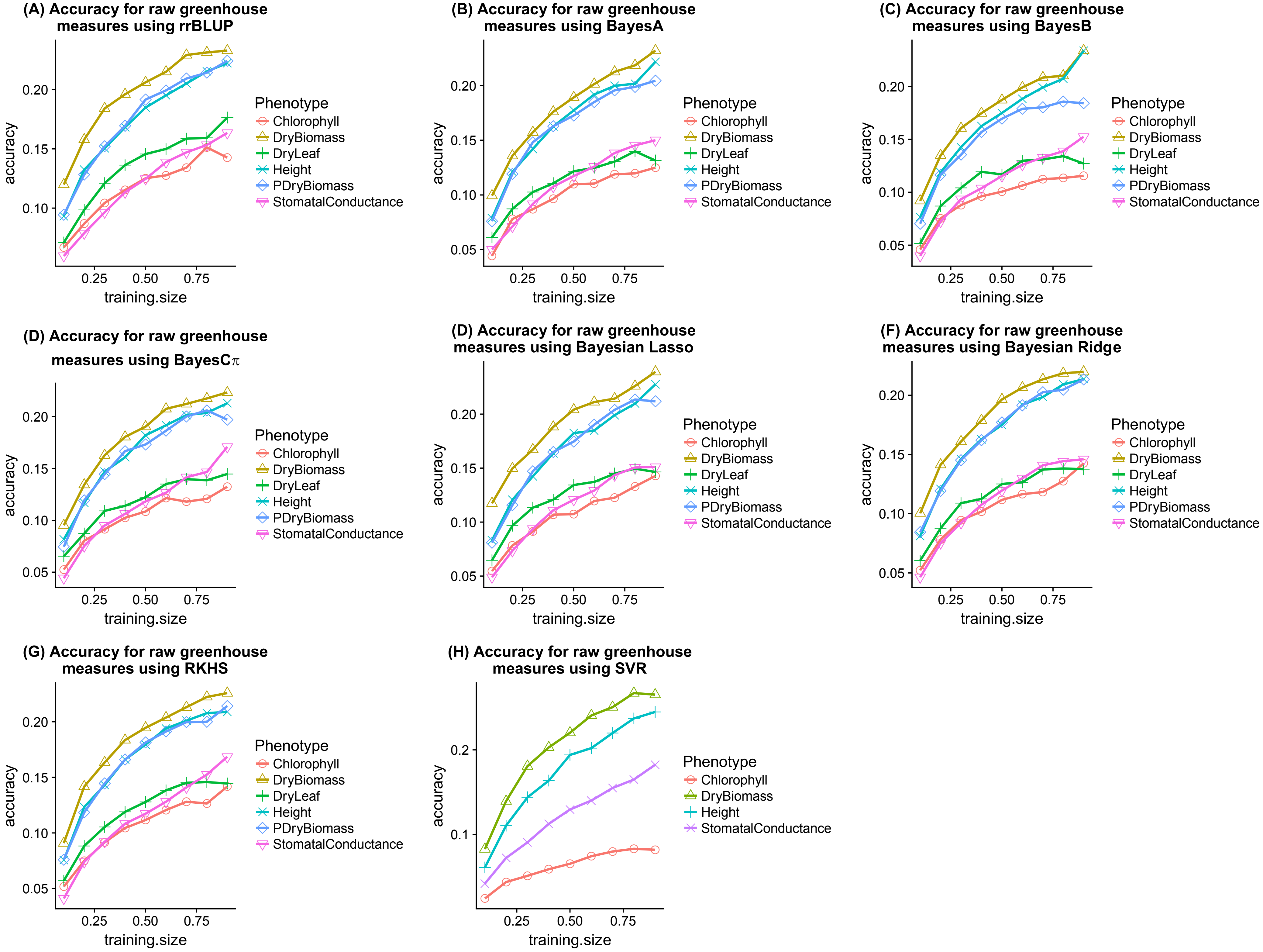
GS Accuracy for Castle Dale health, 1410 reads per locus

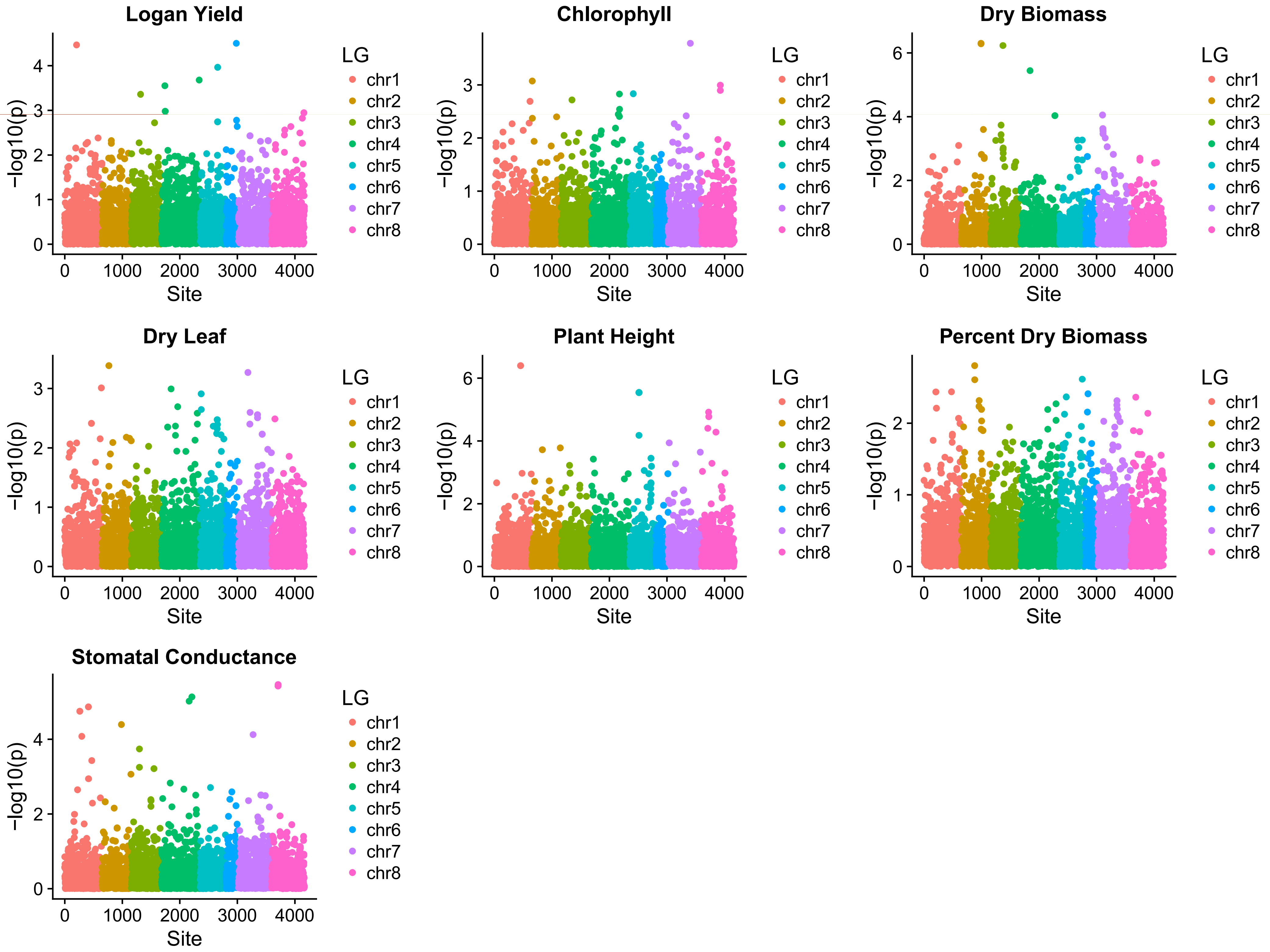


GS Accuracy for Castle Dale health, 2820 reads per locus



GS Accuracy for Castle Dale health, 4230 reads per locus


(A) Accuracy for raw greenhouse measures using rrBLUP
(B) Accuracy for raw greenhouse measures using BayesA
(C) Accuracy for raw greenhouse measures using BayesB
(D) Accuracy for raw greenhouse measures using BayesCπ
(D) Accuracy for raw greenhouse measures using Bayesian Lasso
(F) Accuracy for raw greenhouse measures using Bayesian Ridge
(G) Accuracy for raw greenhouse measures using RKHS
(H) Accuracy for raw greenhouse measures using SVR
accuracy
training.size
0.05
0.10
0.15
0.20
0.1
0.2
0.25
0.50
0.75
Phenotype
Chlorophyll
DryBiomass
DryLeaf
Height
PDryBiomass
StomatalConductance


Logan Yield
Chlorophyll
Dry Biomass
Dry Leaf
Plant Height
Percent Dry Biomass
Stomatal Conductance
−log10(p)
Site
0
1000
2000
3000
4000
0
1
2
3
4
6
LG
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8

Notable Results

- Highest accuracy is 43%, for SVR under rep 1 of the Othello dataset
  - Minimum for GS to outperform other techniques is around 30%
- Loci below an average of 5 reads per sample are not informative and add only noise
- Loci in the 5-15 reads/sample range may still be have predictive value

Acknowledgements

- Advisor, Dr. Long-Xi Yu
- Dr. Mike Peel, collaborator
- Xiang-Ping Liu, collaborator
- Lab Techs, Martha Rivera and Bill Boge
- Farm workers, Jose-Louis and Jesse
- Funding sources, USDA-NIFA and the USDA-ARS base fund