DNA Pooling as a Low-cost Method to Detect Important Genomic Regions for Difficult Traits in Beef Cattle

Introduction

- Main limiting factor affecting power in the design of GWAS is cost of genotyping and phenotyping
- Genotyping costs range from $100 to $250 per sample using arrays
  - Implies genotyping cost of millions for achieving sufficient power with complex traits
- Reduce cost through DNA pooling?

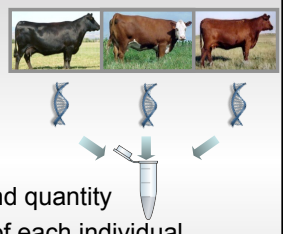
DNA pooling

- Pooling is not a new idea
- Proposed/utilized to reduce genotyping costs for several types of genomic studies (e.g. QTL; Dekkers, 2000)
- More recently for analysis of bead data from Illumina® chips
  - Macgregor et al., 2007 (human)
  - Huang et al., 2010 (bovine)

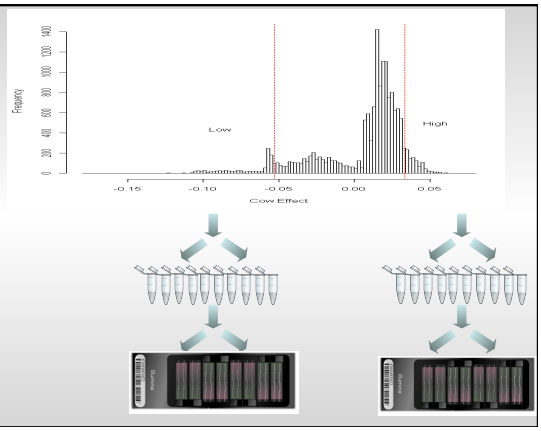
DNA pooling

- Create pools from extremes of phenotypic distributions into pools
  - Estimate allele frequencies between extremes
  - Same concept as case/control
- Larger pools sizes reduce influence of single animals through sampling/mixing
  - Pool size is a balance of phenotyping, array, sample collection and DNA extraction costs

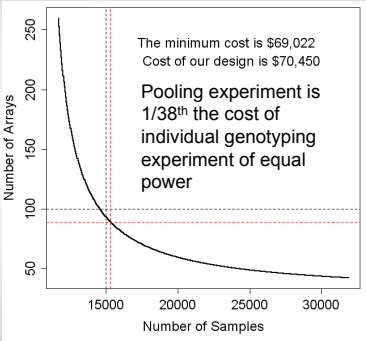
Steps in creating DNA pools



- Extract DNA
  - Tissue
  - Blood
  - FTA cards
- Determine quality and quantity
- Pool equal amount of each individual
- Properly mix pool to ensure consistency of individuals in pool



DNA pooling optimization



Pooling projects at USMARC

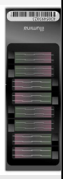
- Pilot – Hereford/Angus bulls
- Propensity to bloat
  - Collaboration with New Zealand
  - Case-control pools, selection study
- Fertility
  - Collaboration with several large US ranches
  - Heifer pregnancy/success after two exposures
- Disease resistance
  - USMARC treatment database
  - Lung lesions on random animals at slaughter
  - Case-control pools

Estimating allele frequencies

- Frequencies are a function of red and green intensities on the chip (Illumina)
- Calculate a pooling allele frequency

PAF = (Red + k) / (Red + Green)

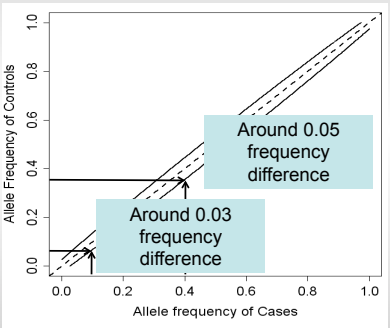
k is a constant to adjust effects of individual chip/ array location (derived across all SNP)



Variance component analysis

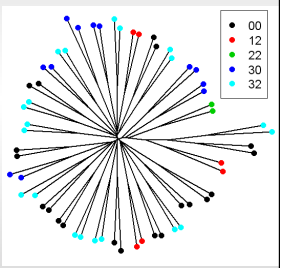
- Using bead level data from chips
  - Calculate pooling allele freq (PAF) from red/green
  - Normalize PAF per Macgregor et al. (2007)
- Sources of variance at multiple levels
  - Technical variance (beads within array, array)
    - Separate estimate for each SNP
  - Pool level
    - Binomial sampling and pool construction (CV^2)
    - Pool at tissue or DNA level

Frequency differences

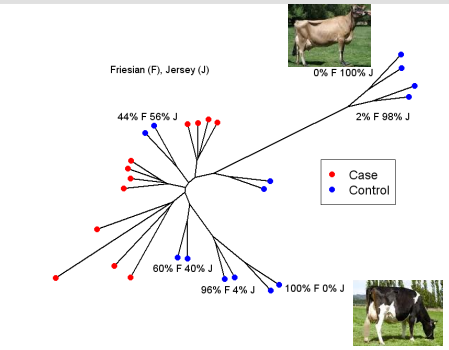


Want case/control to be similar

- Distance diagram based on correlations of PAF among pools
- This example is from a lung lesion pilot project
- Distances consistent between controls and other pools



Cases and unselected controls not always well matched



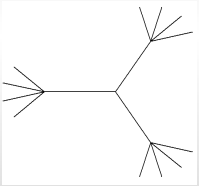
Handling stratification differences

- Adjusting for population hidden population stratification
- Compute correlation among pools using PAF for all SNP
- Spectral decomposition to obtain eigenvalues and eigenvectors
- Use eigenvectors as covariates that correspond to eigenvalues accounting for most of the variance to adjust for hidden population stratification
- Individual SNP have minimal effect on the covariates so analysis Is not circular

Correlation matrix for example

1	.5	.5	.5	0	0	0	0	0	0	0	0
.5	1	.5	.5	0	0	0	0	0	0	0	0
.5	.5	1	.5	0	0	0	0	0	0	0	0
.5	.5	.5	1	0	0	0	0	0	0	0	0
0	0	0	0	1	.5	.5	.5	0	0	0	0
0	0	0	0	.5	1	.5	.5	0	0	0	0
0	0	0	0	.5	.5	1	.5	0	0	0	0
0	0	0	0	.5	.5	.5	1	0	0	0	0
0	0	0	0	0	0	0	0	1	.5	.5	.5
0	0	0	0	0	0	0	0	.5	1	.5	.5
0	0	0	0	0	0	0	0	.5	.5	1	.5
0	0	0	0	0	0	0	0	.5	.5	.5	1

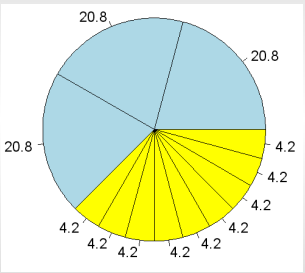
Phylogenetic tree for 3 breeds equally distant from one another



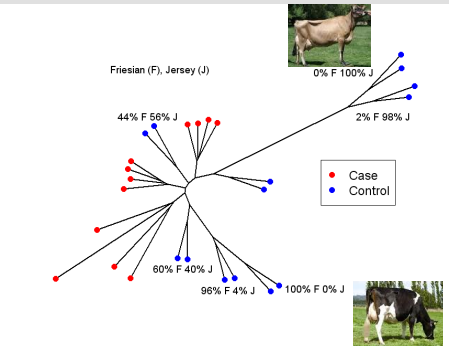
Covariates to adjust for population stratification

0	-.5	0
0	-.5	0
0	-.5	0
0	-.5	0
-.5	0	0
-.5	0	0
-.5	0	0
-.5	0	0
0	0	-.5
0	0	-.5
0	0	-.5
0	0	-.5

Variance accounted for by each Eigen Value

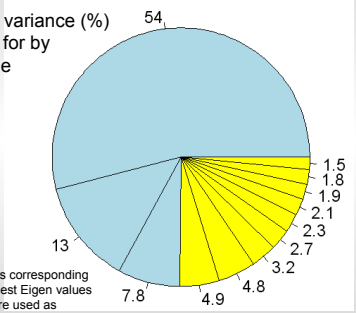


Back to New Zealand Bloat example



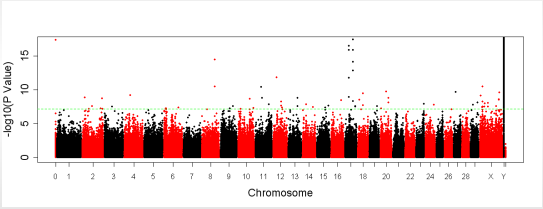
Accounting for Population stratification using Eigen Vectors (like EigenStrat)

Labels are variance (%) accounted for by Eigen Value



Eigen vectors corresponding to the 3 Largest Eigen values (in blue) were used as covariates

Manhattan Plot for New Zealand Industry Bloat Study



Conclusion

- We think pooling can be used to reduce the cost of GWAS
- The mixed model approach does a good job of accounting for technical and biological variation.
  - Should prevent over-interpretation of frequency differences caused by non-genetic sources of variation (sampling, beads, pool construction, hidden population stratification)

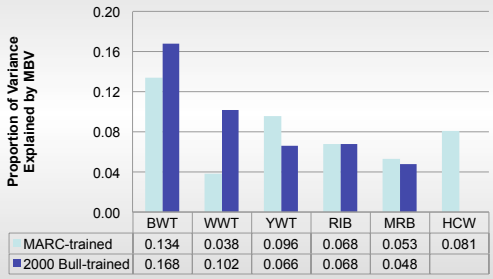
Acknowledgements

John Keele
Tim Smith
Tara McDaneld
Eduardo Casas
Warren Snelling
Steve Simcox
Kevin Tennill
Shuna Jones
Dee Griffin
Andy King
Tommy Wheeler
Sam Nejezcheleb
B.J. Johnson
Tammy Sorensen
USMARC Cattle Operations
Commercial Cooperators
John Pollak

2,000 bull project predictions

- Whole genome selection pilot project with in cooperation with breed associations
  - Over 2,000 industry bulls genotyped with Illumina BovineSNP50
- Formed prediction equations using USMARC GPE data as well as deregressed EPDs from the 2,000+ bulls

Proportion of variance explained



	BWT	WWT	YWT	RIB	MRB	HCW
MARC-trained	0.134	0.038	0.096	0.068	0.053	0.081
2000 Bull-trained	0.168	0.102	0.066	0.068	0.048	

2000 Bulls MBV

- Resulting MBV are being sent to breed associations today
  - Sent by Mark Thallman
- Contact us with any questions

Prediction Equations

- Equations are available at:
  https://www.ars.usda.gov/sp2UserFiles/Place/54380510/2000 Bull Prediction Equations.xlsx

  Link from www.marc.usda.gov ->
  Genetics & Breeding ->
  Documents

- Special thanks to Kristina Weber

Questions?