Conifer Translational Genomics Network
Coordinated Agricultural Project

Genomics in Tree Breeding and Forest Ecosystem Management

Module 15 – Genomic Selection

Ross Whetten – North Carolina State University

Historical review: Genomic selection

In 1998, Chris Haley and Peter Visscher had a vision …

- “Relationship information derived from marker information will replace the standard relationship matrix; thus, the average relationship coefficients that this represents will be replaced by actual relationships.”
- “In fact, future technological developments should make [QTL mapping] unnecessary and make possible high resolution maps of the whole genome, even, perhaps, to the level of the DNA sequence.”

Historical review: Genomic selection

More insights from Haley & Visscher, 1998

- “With high density marker information giving essentially complete information on relationships, different genomic regions could be given the weight appropriate to the variation controlled.”
- “Thus, selection for a limited number of detectable QTL can be complemented by genomic selection aimed at the residual genetic variation that is spread over the remainder of the genome.”
- “This information will not replace information that is already being collected because it will be a long time before phenotype can be predicted solely from DNA sequence. Thus, collection of good performance information remains crucial for the foreseeable future.”

Historical review: Genomic selection

The next step: finding a way to make it work…

- The possibility of high-density genetic marker data required new approaches to analyzing relationships between traits and markers, because the number of marker genotypes per individual would be much greater than the number of individuals with phenotypic data available for analysis
- A seminal paper by Meuwissen et al (2001) proposed two different approaches and tested them using simulated data
  - *A simplifying assumption about the variance explained by each marker*
  - *A more complex Bayesian approach to estimate individual variances*

Using genomic selection

The basic procedure

- Collect a “training population” of individuals, for which both genotypes and phenotypes are available. The number of individuals in the training population, the number of generations, and the density of marker loci at which individuals are genotyped are all positively correlated with predictive power of the resulting model
- Use the training population to create a statistical model with predictive power. Cross-validation is a suitable method of testing predictive power of different models
- Use the statistical model to predict genetic values of individuals in a “prediction population”, for which genotypes are available but phenotypes are not available. Close relationships between the training and the prediction population are favorable for genomic selection

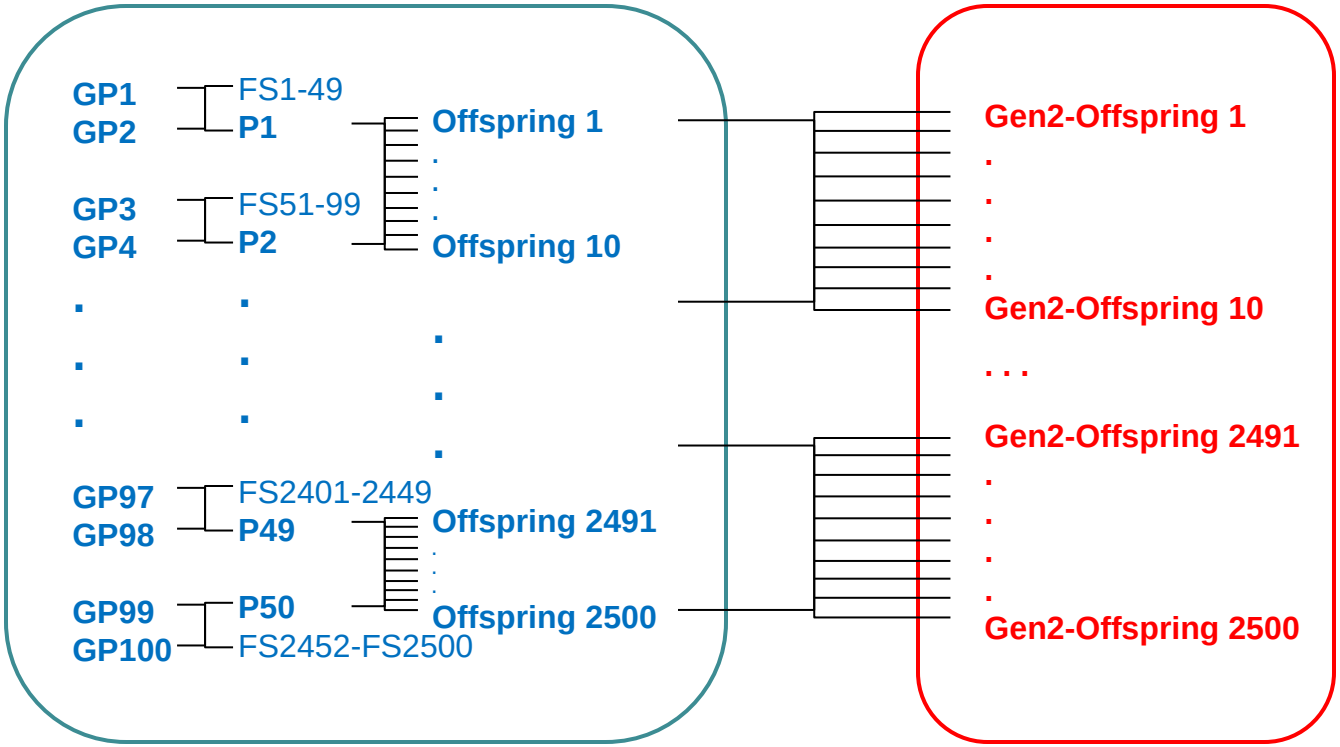
Testing models by cross-validation

Example cross-validation strategies

- Divide the training population into ten random subsets; carry out ten cycles of testing where each cycle consists of training the model on nine of the subsets and use the tenth as a validation population
- Sample at random, without replacement, a fraction (typically 50% to 90%) of the individuals for use as a training population and use the remaining individuals as the validation population. Repeat this sampling and validation as many times as desired
- If family relationships are likely to be an important part of the predictive power of the model, sampling can be done within family rather than at random, so that each training and validation set shares family relationships of the same sort expected between the training and prediction populations

Example population structure

Training and prediction populations in a breeding program



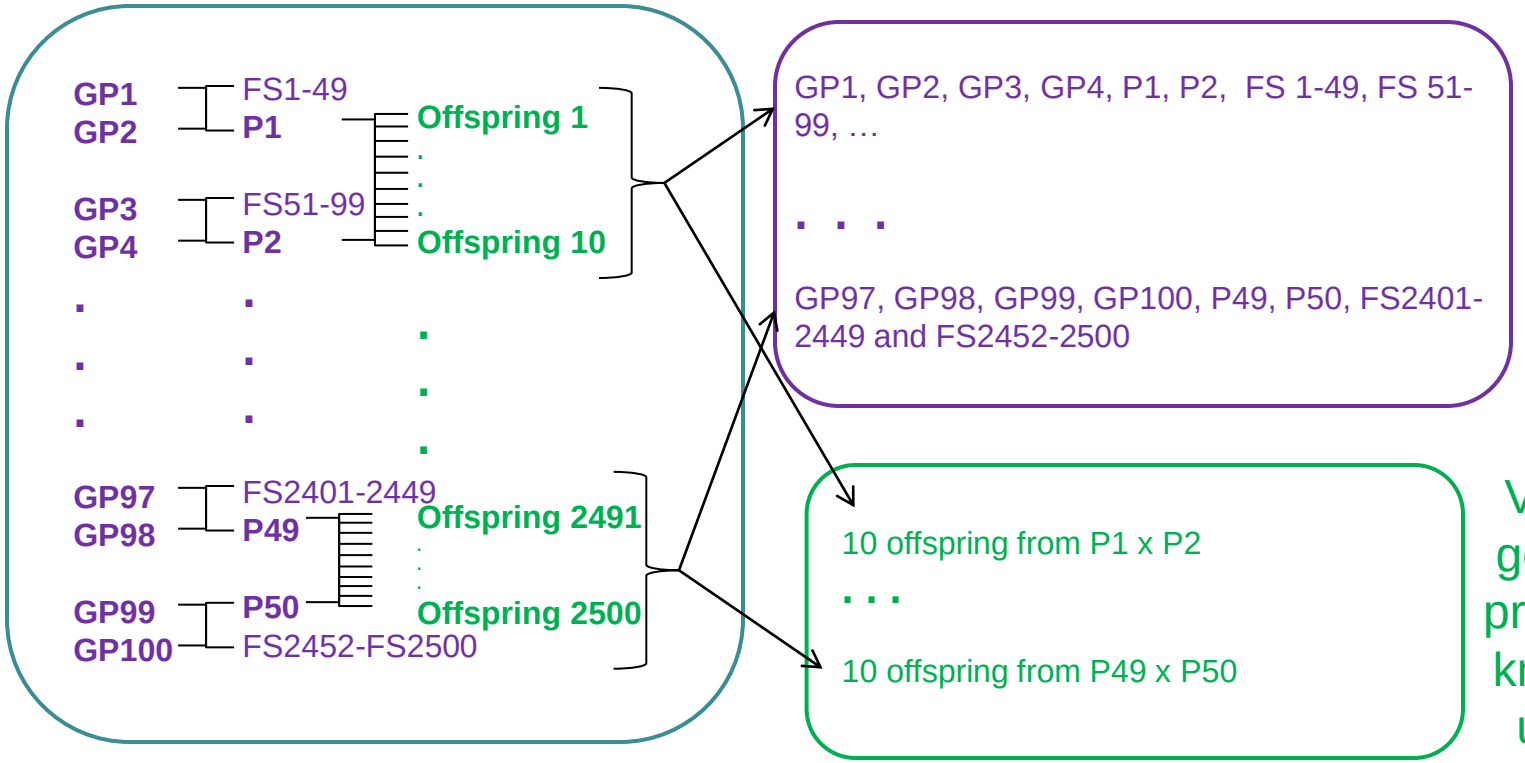
Training: genotypes and phenotypes known

Prediction: genotypes known, phenotypes unknown

Figure credit: Ross Whetten, North Carolina State University

Example cross-validation structure

Train model with grandparents, parents, and full-siblings of parents; validate model with offspring generation



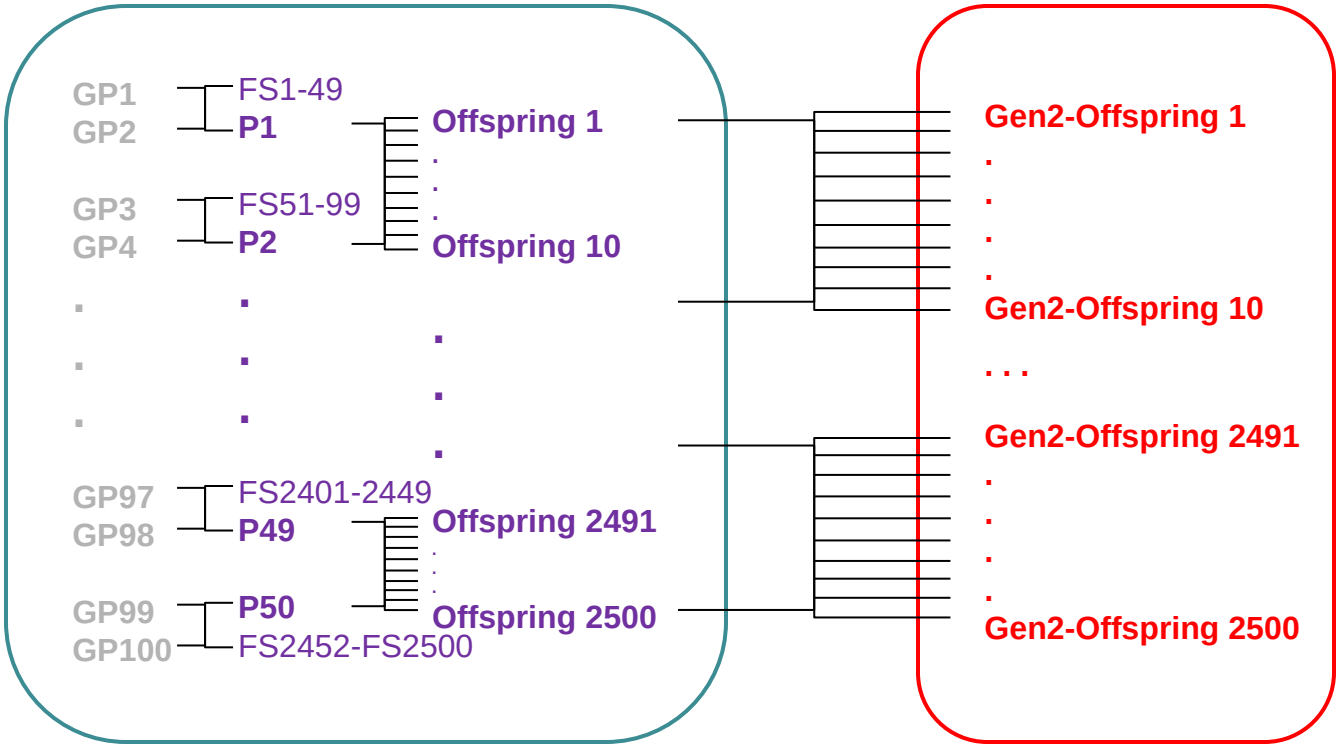
Training subset: genotypes and phenotypes used to train a statistical model: 100 GP, 2500 P

Validation subset: genotypes used to predict phenotypes, known phenotypes used to evaluate quality of model

Figure credit: Ross Whetten, North Carolina State University

Example population structure

Keep similar relationships between training and prediction populations as between training and cross-validation sets



Training: genotypes and phenotypes known

Prediction: genotypes known, phenotypes unknown

Figure credit: Ross Whetten, North Carolina State University

First approach: Simplify by assuming each marker explains equal variance

- The classic infinitesimal model of quantitative genetics is that quantitative traits are controlled by many genes, each of equal and small effect, and that those genes are distributed around the genome
- Assuming that all markers explain an equal fraction of the total phenotypic variance means that only the total genetic variance needs to be estimated
- Predicting genetic value of offspring under this model amounts to determining the proportion of alleles each individual in the prediction population has inherited from ancestors of known genetic value, and then summing the values of those alleles

Second approach: Use Bayesian methods to estimate marker variances

- QTL studies often detect a modest number of loci that affect a complex trait, and each locus may explain a different fraction of the total genetic variation in the trait
- Allowing each marker to have an independent variance, and assuming that many of those variances are zero or close to zero, is consistent with an oligogenic model of inheritance
- Predicting genetic value of offspring under this model requires that markers be in linkage disequilibrium with trait loci, so that the value assigned to each marker is reflective of the genetic value of nearby trait loci

Linear mixed models of genetic variation

… assume equal variance in phenotype is associated with each marker, and produce **B**est **L**inear **U**nbiased **P**redictions

- BLUP is a statistical method that reduces the difference between an estimate of the value of an individual and the mean of the relevant population, by a factor based on the ratio of residual (or error) variance to genetic variance for the trait of interest
- The relevant population is made up of individuals who are related to the individual in question. BLUP analysis uses a ‘relationship matrix’ to take relationships among individuals into consideration during prediction of breeding values
- This approach is consistent with the simplifying assumption that each marker or interval between markers accounts for an equal share of the total genetic variance

Bayesian approaches

…to predicting genetic value based on genotype data

- Bayesian methods are a way of incorporating prior knowledge or expectations with experimental data to refine an estimate of the probability of a particular hypothesis (Shoemaker et al, 1999)
- These methods often use a Markov Chain Monte Carlo (MCMC) process to estimate parameters for which exact values cannot be computed. A ‘Markov Chain’ is a series of ‘states’, or sets of values of experimental parameters, in which the state at a given step is dependent on the state at one or more previous steps
- MCMC methods explore the parameter space of possible parameter values to determine which values are most likely given the observed data. The Gibbs sampler and the Metropolis-Hastings algorithm are two among many ways to do this

Training a genomic selection model

Meuwissen et al (2001) tested three approaches

- BLUP – intervals between markers are assumed to have a common variance. The error variance is assumed to be known (true for the simulated dataset used in the publication)
- BayesA – intervals between markers can have different variances, which are estimated by a MCMC process using a scaled inverted-chi-square distribution as the prior, and Gibbs sampling to generate a posterior distribution from which the variances and effects are estimated
- BayesB – adds another parameter, π, as the probability that the variance of a particular interval is zero. A Metropolis-Hastings sampler is used to generate the posterior distribution

Requirements of statistical models

… for successful application of genomic selection

- The infinitesimal model requires only that we be able to estimate the proportion of alleles shared among related individuals
- The number of markers required to do this depends on the structure of the training and prediction populations, and on the degree of relatedness within and among populations; a few hundred markers may suffice if the populations are closely related
- The oligogenic model also provides predictive power based on the proportion of shared alleles among individuals, but will have more power if markers are in LD with trait loci
- The number of markers required to assure every trait locus is in LD with at least one marker will differ for different species

Different approaches …

… may be best for different target traits

- Phenotypes that involve the product of a single biochemical pathway may be more likely to fit the assumptions of an oligogenic model of genetic control – DGAT1 in cattle as an example (Kühn et al, 2004)
- Phenotypes that are the result of combined action of many biochemical pathways may be more likely to fit the assumptions of the infinitesimal model of genetic control
- Using a statistical model that closely matches the true pattern of genetic control will give the best results, but …
- We cannot ever be certain of the true pattern of genetic control; at best, we can estimate it from the data

Different approaches …

… may be best for different population structures

- Dairy cattle (particularly Holsteins) are highly inbred, with an effective population size worldwide of about 100 (on the female side) to 20 (on the male side). LD extends for about 100,000 bp
- A genotyping array for 50,000 SNPs is commercially available, at a cost of a few hundred US dollars per sample. This array provides high-density marker genotypes in LD with virtually all loci in the bovine genome
- Forest trees typically have effective population sizes of thousands to hundreds of thousands and LD often decays over distances of a few thousand base pairs
- Some tree species have high-density SNP arrays; other do not

LD is dependent on allele frequencies

… and on the linkage phase of alleles

- Assume a scenario of two loci, **A** and **B**, each with only a single allele in the starting population. The ancestral alleles will be denoted *A* and *B*
- Imagine now that mutations occur independently (i.e. at different times) to create new alleles at both loci. The new alleles will be denoted *a* and *b*. Assume *a* arises first
- The mutation of *A* to *a* can occur only on a chromosome bearing the *B* allele, but the mutation of *B* to *b* can occur on a chromosome bearing either *A* or *a*. The two loci are so close that they are never separated by recombination
- How is LD affected by allele frequencies in these two scenarios?

LD is dependent on allele frequencies

Modeling LD in two scenarios

$$r^2 = \frac{(p_{AB} - pApB)^2}{p_A p_B (1 - pA)(1 - pB)}$$

Scenario 1: *b* arises in the *A* background. p_{AB} is the frequency of the *Ab* haplotype, which is equal to the frequency of the *b* allele. The *ab* haplotype never occurs, and p_A and p_B in the equation are the frequencies of the *A* and *b* alleles

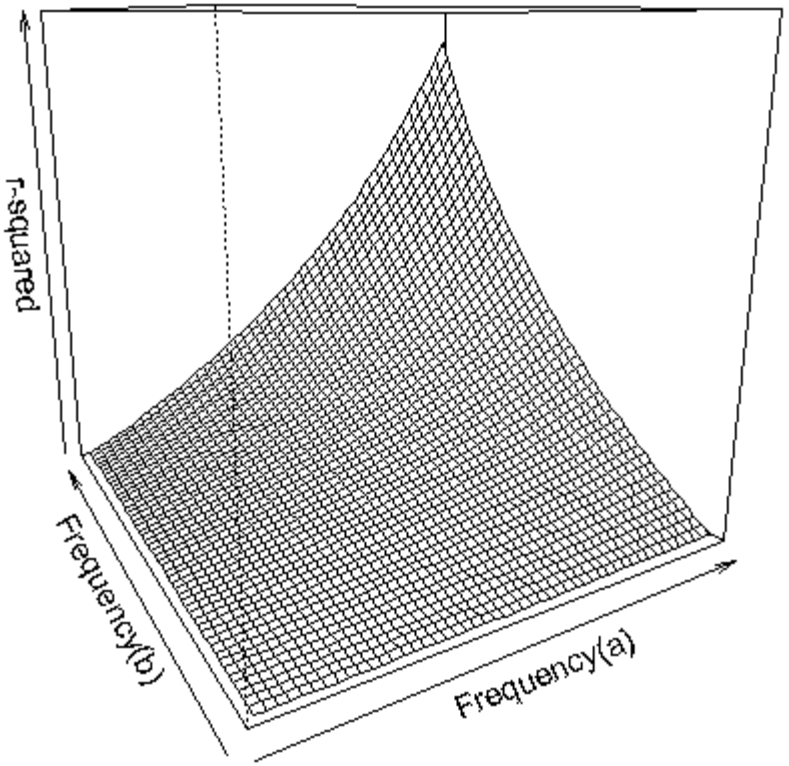
Scenario 2: *b* arises in the *a* background. p_{AB} is the frequency of the *ab* haplotype, which is equal to the frequency of the *b* allele. The *Ab* haplotype never occurs, and p_A and p_B in the equation are the frequencies of the *a* and *b* alleles

(Eberle et al., 2006)

Scenario 1: *b* arises in *A* background

Minor alleles have low LD with major alleles at neighboring loci

… except when all allele frequencies approach 0.5



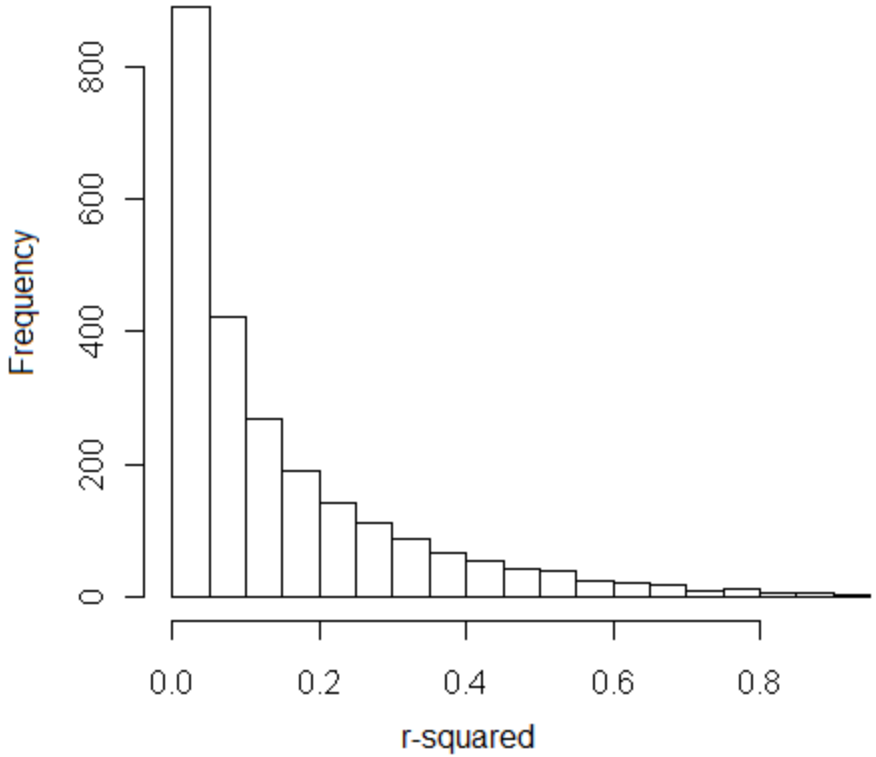
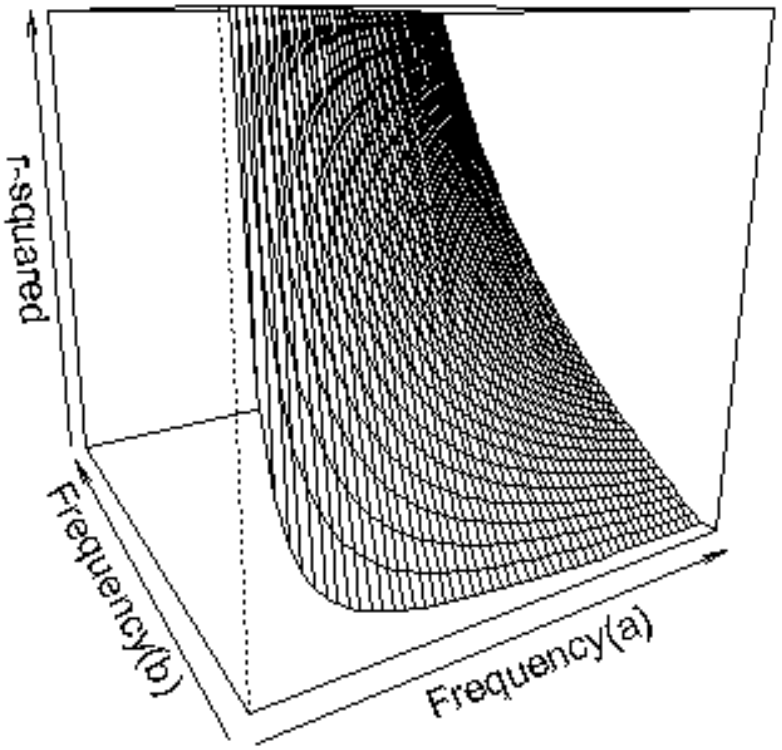


Figure credit: Ross Whetten, North Carolina State University

Scenario 2: *b* arises in *a* background

Minor alleles have high LD if allele frequencies are equal

… but the decline in LD is rapid as allele frequencies diverge from equality



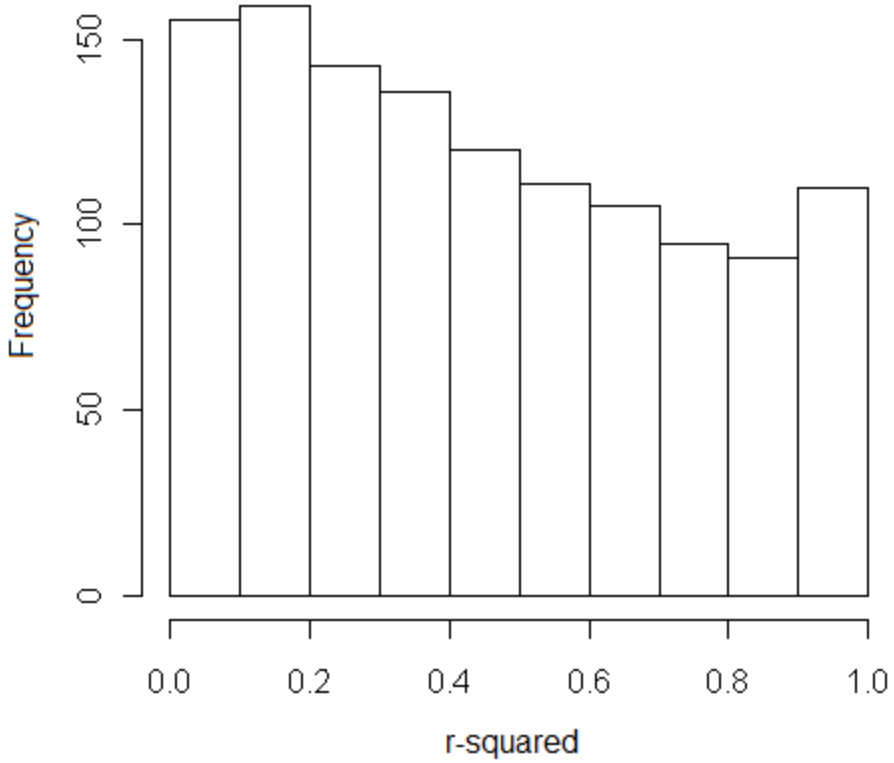


Figure credit: Ross Whetten, North Carolina State University

Allele frequency in populations

Breeding population size determines allele frequencies

- Minor allele frequency (MAF) is a population-specific parameter, and can change depending on which population is examined
- Many SNPs will have different MAF within a structured mating design with a small number of parents, than in the wild population or the breeding population as a whole
- Linkage equilibrium or disequilibrium relationships can also change within structured mating designs, due to sampling effects that limit the diversity of haplotypes in the smaller population
- Selection can also change haplotype frequencies and LD relationships, so advanced-generation breeding populations can differ from the initial population in important ways

LD in small populations

Sampling effects become stronger as N_e decreases

- Any sample of individuals is likely to include some individuals that carry genetic variants that are relatively rare in the population as a whole. The sub-population of progeny descended from those founders will show more genetic variation due to those variants than will progeny from the whole population, because the causal variants are more frequent in the sub-population than in the whole population
- The smaller the sample of founders, the greater the difference is likely to be between genetic variability due to those variants in the sub-population and variability due to those variants in the whole population
- SNP loci with minor alleles on the same chromosome homologues as the causal variants are more likely to have predictive power in the sub-population than in the whole population

LD vs. identity-by-descent

Different views of the same phenomenon

- Identical alleles at the same locus in two different individuals can be either identical-by-descent (IBD) or identical-by-state (IBS)
- The distinction is that IBD alleles are descended from a known common ancestor, while the last common ancestral source of IBS alleles is unknown
- A recent publication reported new methods for estimating IBS status at non-genotyped loci based on the similarity in state of genotyped loci (Powell et al, 2010)
- Other authors have presented simulation results suggesting that LD can be experimentally manipulated by reducing effective population size and increasing marker density

Multi-family mating design

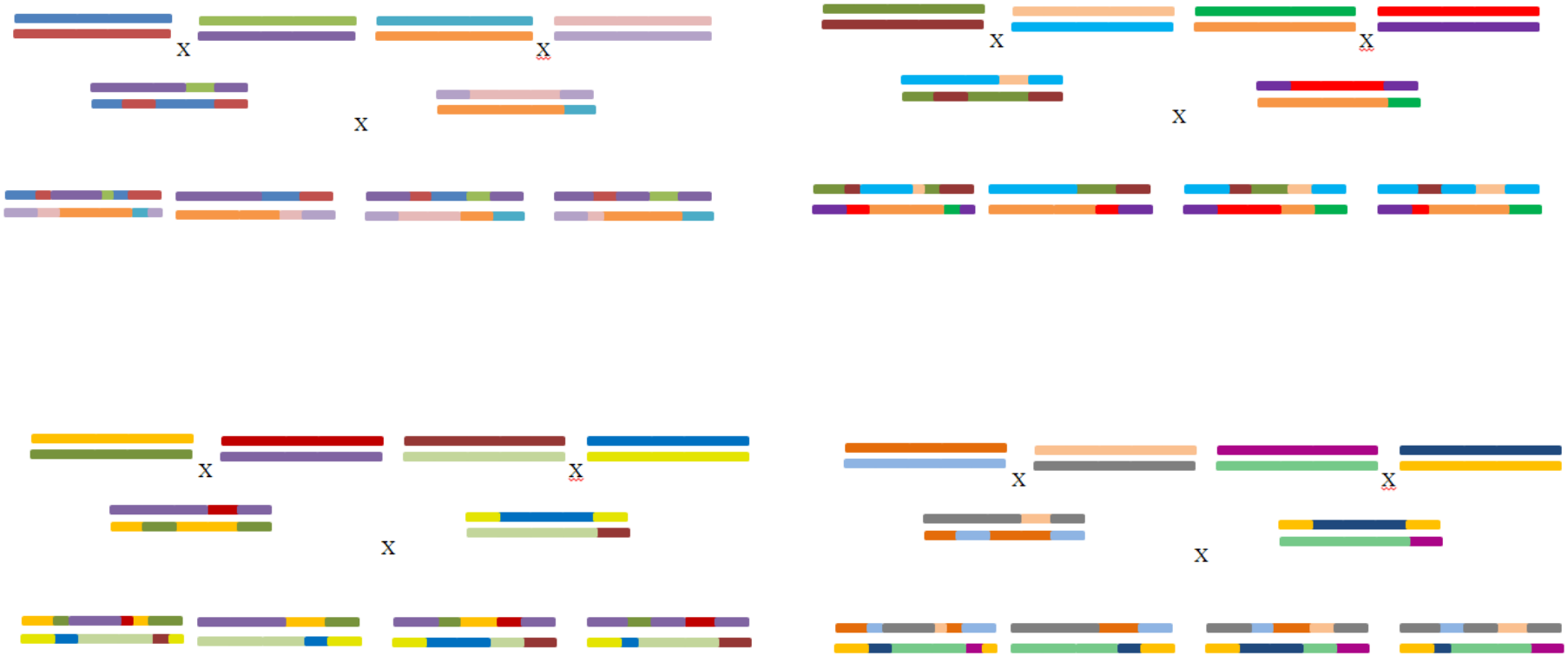


Figure credit: Ross Whetten, North Carolina State University

Multiple parents, many generations …

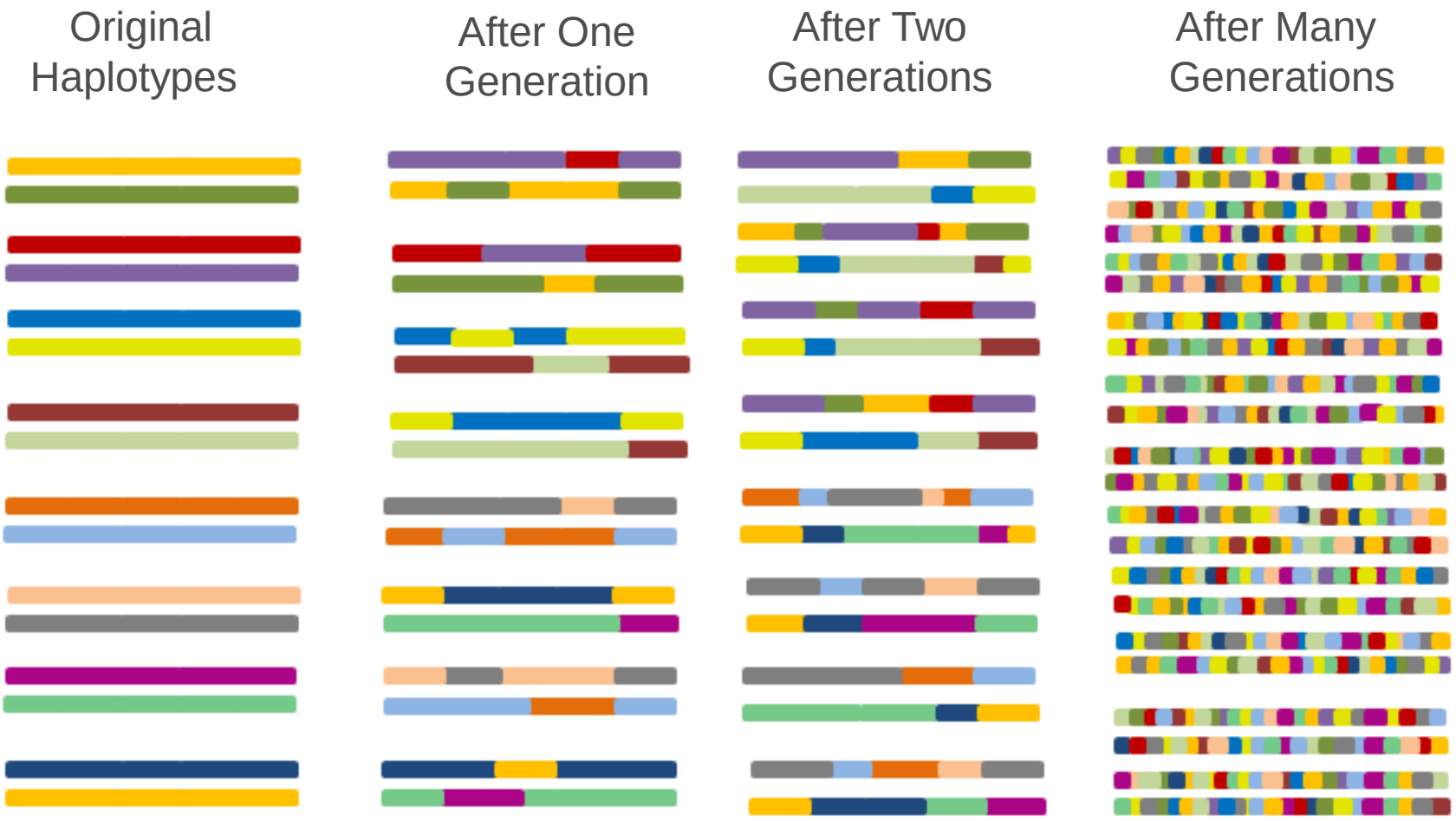


Figure credit: Ross Whetten, North Carolina State University

Results from dairy cattle breeding

Results are based on genotypes at 38,416 SNP loci

- Five yield traits, five fitness traits, 16 conformation traits, net merit
- 3576 bulls in training population, 1759 in prediction population
- R^2 values between estimated breeding values and observed breeding values were 0.05 to 0.38 greater using genomic data in non-linear models than using pedigree information alone
- Genomic data used in linear models gave R^2 values that averaged 0.01 lower than those based on non-linear models

  (VanRaden et al., 2009)

Tree breeding applications

What is required?

- A cost-effective genotyping platform that detects thousands to hundreds of thousands of loci in thousands of individuals
- At least 1000 individuals with genotypes and phenotypes for a training population and some genotyped relatives in a prediction population
- Adequate levels of LD or identity-by-descent to provide power, either due to small effective population size or to high population-wide LD

Tests of GS for trees are underway

Populations and genotyping platforms exist in multiple species

- Eucalyptus inter-specific hybrid populations have high levels of LD and a high-throughput assay for about 5000 Diversity Array (DArT) markers is available for genotyping
- Pine populations with limited effective population size (N_e ~ 20) and over 2000 phenotyped individuals will soon be available, and high-throughput genotyping methods using DNA sequencing are being adapted for pine
- Poplar genotyping arrays with over 200,000 SNPs are available and structured mating designs and progeny trials exist within breeding programs

What does the future hold?

Complete genome sequence information contains information on all trait loci and is the ultimate basis for GS

- Reference genome sequences already exist for poplar and eucalyptus and will be available in a few years for pine
- Re-sequencing to discover all variants in the genomes of a limited number of parents will become increasingly cost-effective over the next few years as costs continue to decrease
- Computational methods for imputing whole-genome sequence information to thousands of progeny who have been genotyped with a low-density SNP array are under development

(Meuwissen and Goddard, 2010)

Why use whole-genome sequences?

Greater predictive accuracy and more reliable models

- Simulations by Meuwissen and Goddard (2010) showed that densities of SNPs expected of whole-genome sequence data provided 40% more accurate predictions of breeding value than SNP densities from currently-available genotyping platforms
- The value of Bayesian or non-linear analytical methods may increase as SNP densities increase, at least for some populations and some traits
- If the causative loci that actually underlie genetic variation are included in models, the accuracy of the model should remain high for several generations before re-training is required

Summary

Genomic selection aims to predict value, not identify genes

- The predictive accuracy of genomic selection is a function of training population size, marker density, trait heritability, and genetic architecture of the trait
- The first two variables are under the breeder's control; the second two are not – careful choice of populations and marker platforms is important!
- Costs of DNA sequencing are dropping exponentially, so the number of markers available per unit cost will increase
- Re-training of statistical models will be needed every generation until marker density increases; eventually it may be needed only every five to ten generations when genome sequences are used

References cited

Chapman, N. H. and E. A. Thompson. 2002. The effect of population history on the lengths of ancestral chromosome segments. Genetics 162: 449-458.

Eberle, M.A., M. J. Rieder, L. Kruglyak, and D. A. Nickerson. 2006. Allele frequency-matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. Public Library of Science Genetics 2: e142. (Available online at: http://dx.doi.org/10.1371/journal.pgen.0020142) (verified 29 Feb 2012).

Grattapaglia, D and M. D. V. Resende. 2011. Genomic selection in forest tree breeding. Tree Genetics & Genomes 7:241 – 255 .

Haley, C.S. and P. M. Visscher. 1998. Strategies to utilize marker-quantitative trait loci associations. Journal of Dairy Science 81(Supp2): 85-97. (Available online at: http://dx.doi.org/10.3168/jds.S0022-0302(98)70157-2) (verified 29 Feb 2012).

Kühn, C., G. Thaller, A. Winter, O. R. Bininda-Emonds, B. Kaupe, G. Erhardt, J. Bennewitz, M. Schwerin, and R. Fries. 2004. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle Genetics 167: 1873–1881. (Available online at: http://dx.doi.org/10.1534/genetics.103.022749) (verified 29 Feb 2012).

Meuwissen , T.H.E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Meuwissen, T.H.E. and M. E. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing . Genetics 185: 623-631. (Available online at: http://dx.doi.org/10.1534/genetics.110.116590) (verified 29 Feb 2012).

Powell, J. E., P. M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11: 800-805. (Available online at: http://dx.doi.org/10.1038/nrg2865) (verified 29 Feb 2012).

Shoemaker, J. S., I. S. Painter, and B. S. Weir. 1999. Bayesian statistics in genetics: a guide for the uninitiated. Trends in Genetics 15: 354-358. (Available online at: http://dx.doi.org/10.1016/S0168-9525(99)01751-5) (verified 29 Feb 2012).

Stein, L. D. 2010. The case for cloud computing in genome informatics. Genome Biology 11: 207. (Available online at: http://dx.doi.org/10.1186/gb-2010-11-5-207) (verified 29 Feb 2012).

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92: 16–24. (Available online at: http://dx.doi.org/10.3168/jds.2008-1514) (verified 29 Feb 2012).

Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UCDAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture