Genomic tools in breeding

Ross Whetten, Professor

NC State University, Raleigh NC

Pine Plantation Research and Decision Support Tool Rollout

May 16-17, 2017 Athens, GA

PINEMAP

USDA United States Department of Agriculture National Institute of Food and Agriculture

Outline

Overview of molecular and statistical models of genetic variation

Strategies for building and testing genetic models

Experimental results with 56 families of loblolly pine

Finding genetic variants associated with adaptation

PINEMAP

Overview of pine genome

12 chromosomes, each present in two copies in most cells

25,000 to 50,000 different genes plus a lot of DNA with no known function; most genes are 500 to 5000 bases long

Total DNA per diploid cell is about 45 billion bases (A, C, G, or T)

In principle, each base position could exist in one or another alternate states or “alleles”, e.g. A vs C.

PINEMAP

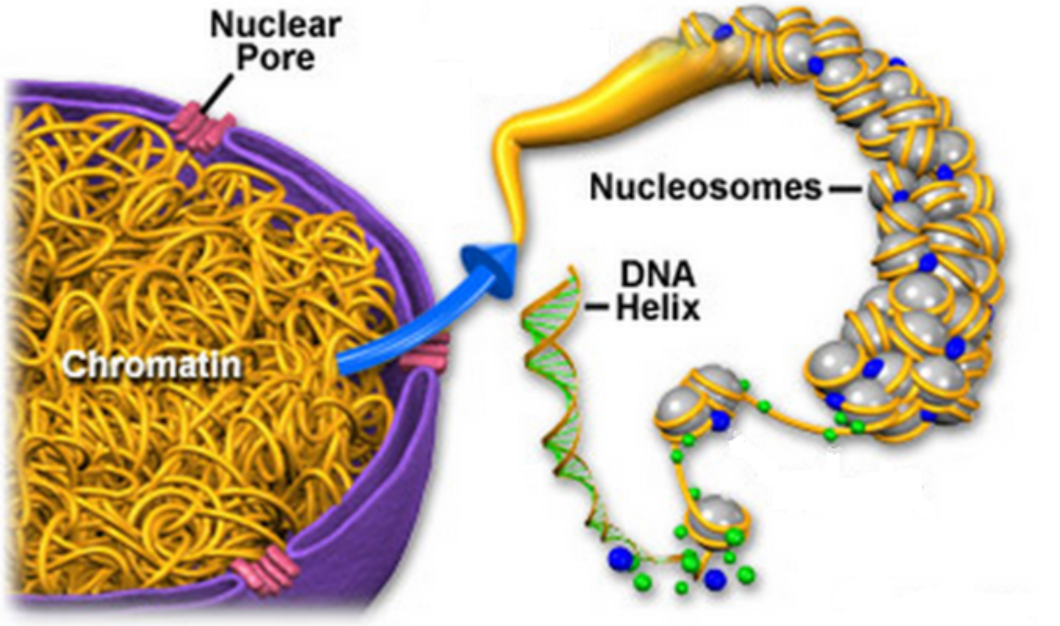
DNA is tightly packed

Imagine nucleus of cell is the size of a basketball

24 chromosomes, each a single DNA strand 7 miles long

Some DNA regions are “accessible”, others are wound around proteins and condensed

Genes tend to be in accessible regions



PINEMAP

Genetic variation – molecular view

Two ways to classify genetic variants

- By effect: “causal” have effects on traits; “neutral” have no effect on any trait
- By location: “structural” in genes that produce protein or RNA; “regulatory” in regions that control how genes are expressed; “inter-genic” are outside of genes or regulatory regions

Effect and location are partly independent

- neutral variants can occur in structural, regulatory, or intergenic regions
- causal variants typically define the boundaries of structural or regulatory regions

PINEMAP

Genetic variation – statistical view

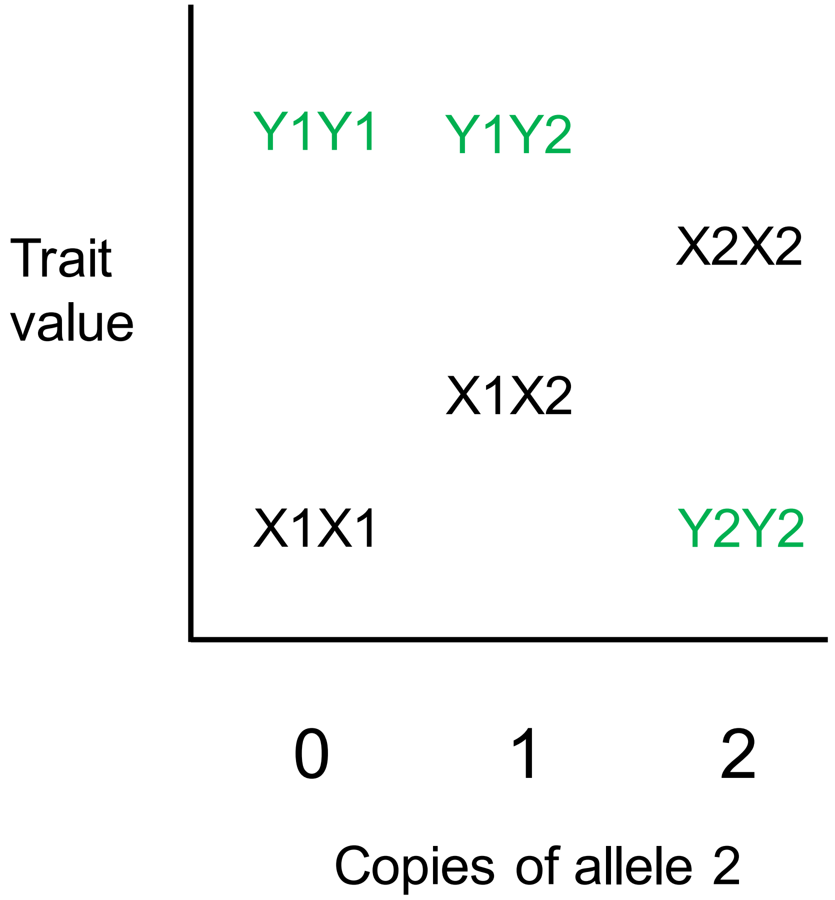
Genetic variation is additive or non-additive

- Additive: passed from parent to offspring in predictable ways, most useful for breeders
- Non-additive – less predictable, more difficult to measure or use in breeding
  - Dominance due to interactions between different versions of the same gene
  - Epistasis due to interactions between different genes

PINEMAP

Statistical approach

- Allelic effects can be
  - Additive (X)
  - Dominant (Y)
  - Epistatic (if trait value of X2X2 individual depends on genotype at Y)



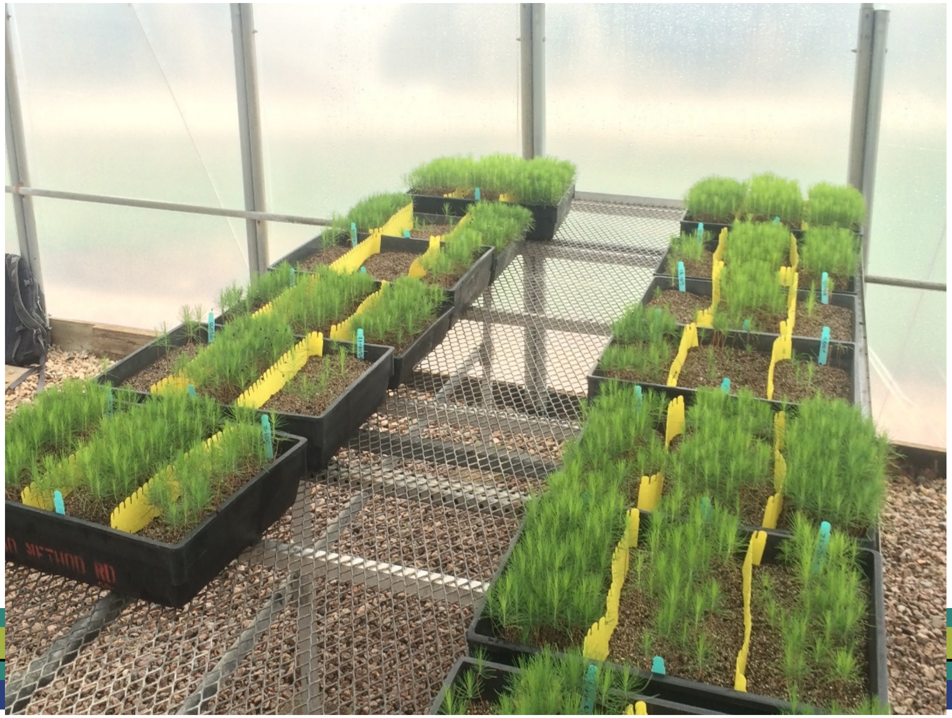
Finding functional genetic variation

Genes (DNA) are expressed to produce RNA

- Isolating RNA from a specific tissue gives access to genes expressed in that tissue
- If we are interested in all tissues, one option is to work with seedlings – they contain all cell types except reproductive organs
- Converting RNA to DNA and sequencing the DNA allows discovery of genetic variation both in sequences of expressed genes, and in the relative levels of gene expression

Plant material for RNA isolation

- 3-month old greenhouse-grown OP seedlings – whole plants, with potting mix washed off roots before freezing in liquid N_2
- Goal is to measure “family mean” gene expression levels
- SNP genotypes called from alignment to reference sequences



PINEMAP

Finding functional genetic variation

DNA sequences are compared to "reference transcriptome"; 86,000 previously-described RNA sequences from pine

- About 40,000 different RNA transcripts are detected; levels of some are highly variable while others are not
- About 49,000 SNPs are detected; 22,000 have no missing data across the 56 families of interest
- Many possible explanatory variables, relatively few observations to be explained – use cross-validation to reduce risk of bias

PINEMAP

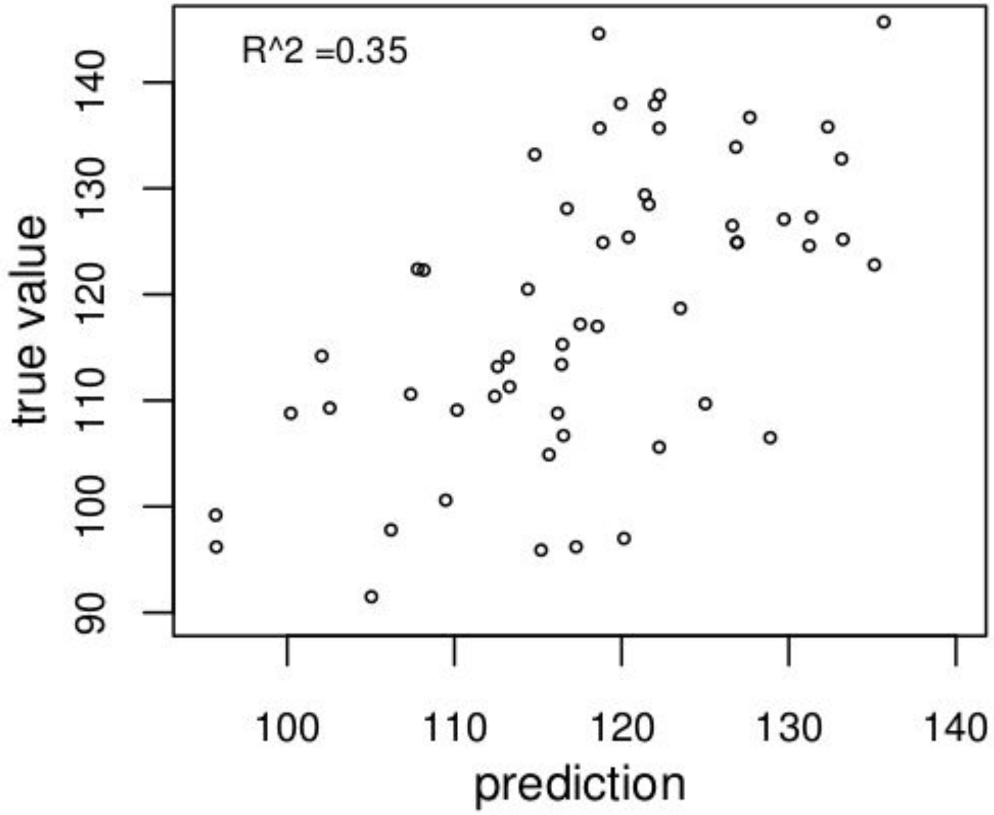
Finding functional genetic variation

Cross-validation: use some of the data to create a statistical model that predicts trait value from SNPs or expression levels; then test the accuracy of the model on other data not used to create the model

- Leave one family out for prediction
  - Use remaining 55 families in nested leave-one-out:
    do 55 cycles of regression of 54 trait values on each predictor
  - Choose predictors that are significantly associated with trait value in all 55 regressions
- Use the selected set of predictors to estimate the trait value of the 56th family (not used in any of the regressions)
- Repeat for all 56 families, so each has an estimated value based on the others

PINEMAP

Predicting breeding values based on gene expression levels



- Linear regression of gene expression levels against known breeding values of a subset of families identifies a subset of genes
- A covariance matrix based on those genes is then used to predict the breeding value of a single family
- This approach explains about one-third of the variation in known breeding values

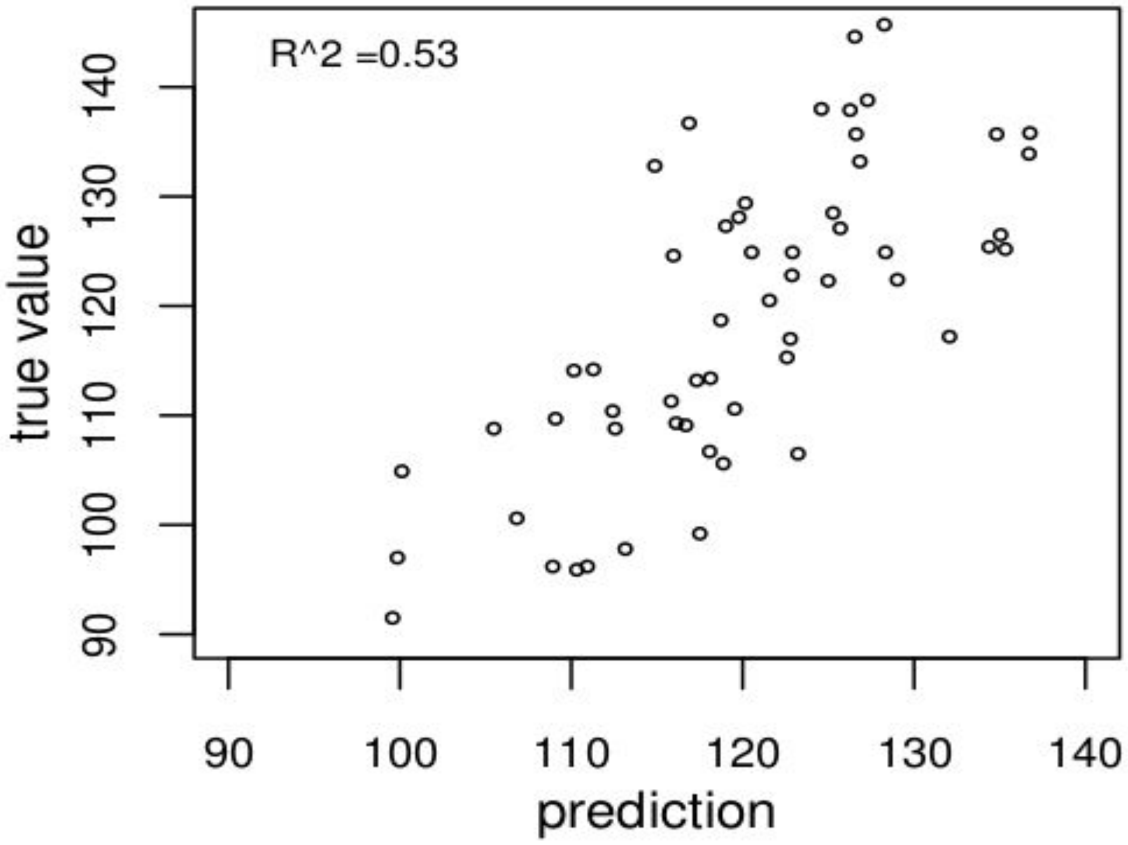
PINEMAP

Predicting breeding values based on SNP genotypes

- About 49,000 putative SNPs were identified among the 56 families; about 22,000 had no missing data
- Use regression of SNP genotypes against known breeding values for a subset of families, as before, to select which SNP loci to use in covariance matrix
- Covariance matrix based on those SNPs describes relationships among families at the subset of the genome that is associated with breeding value.
- Plot all predicted breeding values against known breeding values to summarize results across multiple cycles of testing

PINEMAP

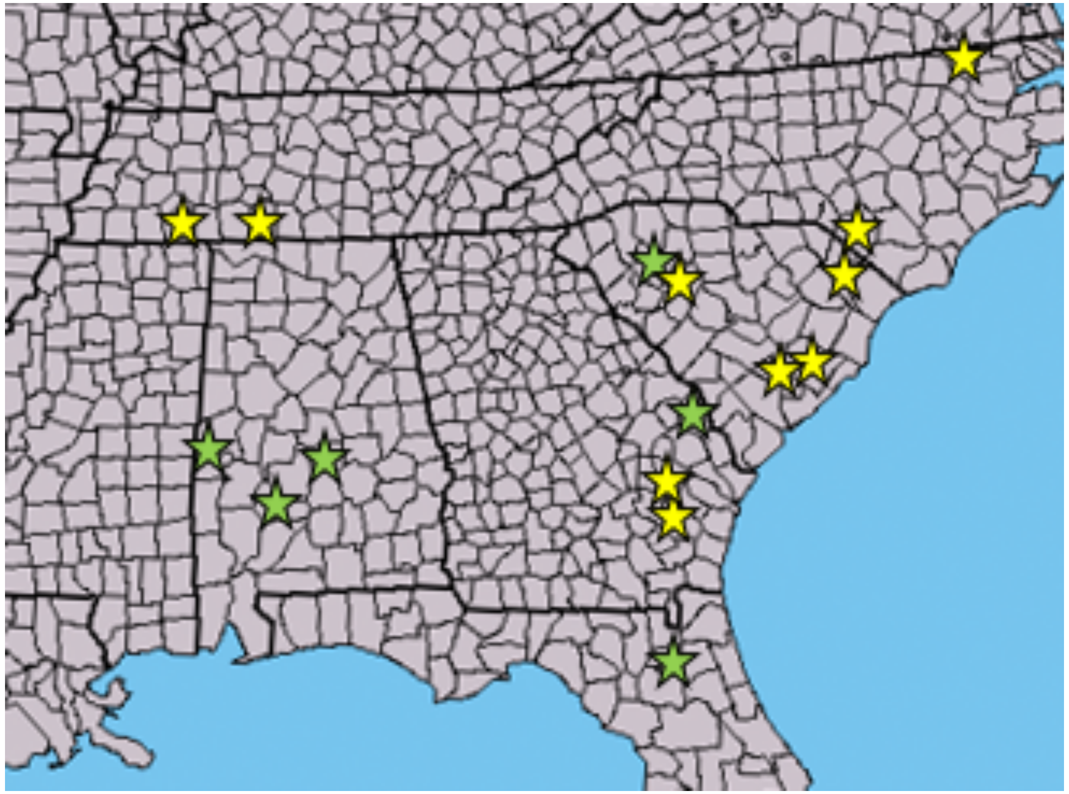
Predicting breeding values based on SNP genotypes



About half the variation in known breeding value can be predicted based on SNPs

Finding genetic variants underlying adaptation

- Use long-term field trials of the same genetic material planted in many locations
- Identify SNP markers in structural or regulatory regions of the genome
- Test for association of SNPs with height growth as proxy for adaptation



PINEMAP

Summary

- Gene expression levels and SNPs in expressed genes have some predictive power within a set of 56 coastal families
- This may prove useful for reducing progeny testing effort, but is unlikely to allow prediction of performance outside the region of origin
- Genetic variation in adaptability exists, and can probably be identified, but will require analysis of trees planted outside the region of origin

PINEMAP