Conifer Translational Genomics Network
Coordinated Agricultural Project

Genomics in Tree Breeding and Forest Ecosystem Management

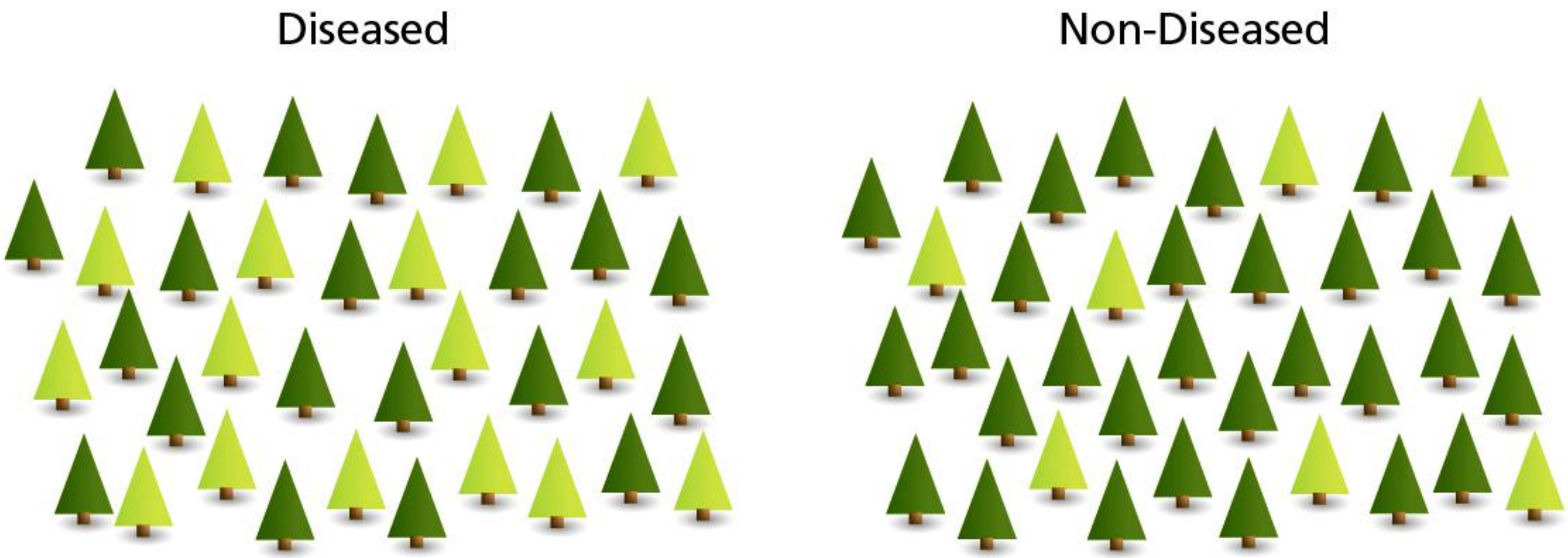
Module 11 – Association Genetics

Nicholas Wheeler & David Harry – Oregon State University

What is association genetics?

- Association genetics is the process of identifying alleles that are disproportionately represented among individuals with different phenotypes. It is a population-based survey used to identify relationships between genetic markers and phenotypic traits
  - *Two approaches for grouping individuals*
    - *By phenotype (e.g. healthy vs. disease)*
    - *By marker genotype (similar to approach used in QTL studies)*
  - *Two approaches for selecting markers for evaluation*
    - *Candidate gene*
    - *Whole genome*

Association genetics: conceptual example



Genotype	Diseased	Non-diseased	Total
▲ BR-S	17	7	24
▲ BR-R	20	30	50
	37	37	

$$\chi^2_{.05} = 5.377$$

$$p < 0.025$$

Figure Credit: Nicholas Wheeler, Oregon State University

Comparing the approaches

Criteria	Family-based QTL Mapping	Population-based Association Mapping
Number of markers	Relatively few (50 – 100's)	Many (100's – 1000's)
Populations	Few parents or grandparents with many offspring (>500)	Many individuals with unknown or mixed relationships. If pedigreed, family sizes are typically small (10's) relative to sampled population (>500)
QTL analysis	Easy or complex. Sophisticated tools minimize ghost QTL and increase mapping precision	Easy or complex. Sophisticated tools reduce risk of false positives
Detection depends on	QTL segregation in offspring, and marker-trait linkage within-family(s)	QTL segregation in population, and marker-trait LD in mapping population
Mapping precision	Poor (0.1 to 15 cM). QTL regions may contain many positional candidate genes	Can be excellent (10's to 1000's kb). Depends on population LD
Variation detected	Subset (only the portion segregating in sampled pedigrees)	Larger subset. Theoretically all variation segregating in targeted regions of genome
Extrapolation to other families or populations	Poor. (Other families not segregating QTL, changes in marker phase, etc)	Good to excellent. (Although not all QTL will be segregate in all population / pedigree subsamples)

Essential elements of association genetics

- Appropriate populations
  - *Detection*
  - *Verification*
- Good phenotypic data
- Good genotypic data
  - *Markers (SNPs): Number determined by experimental approach*
  - *Quality of SNP calls*
  - *Missing data*
- Appropriate analytical approach to detect significant associations

Flowchart of a gene association study

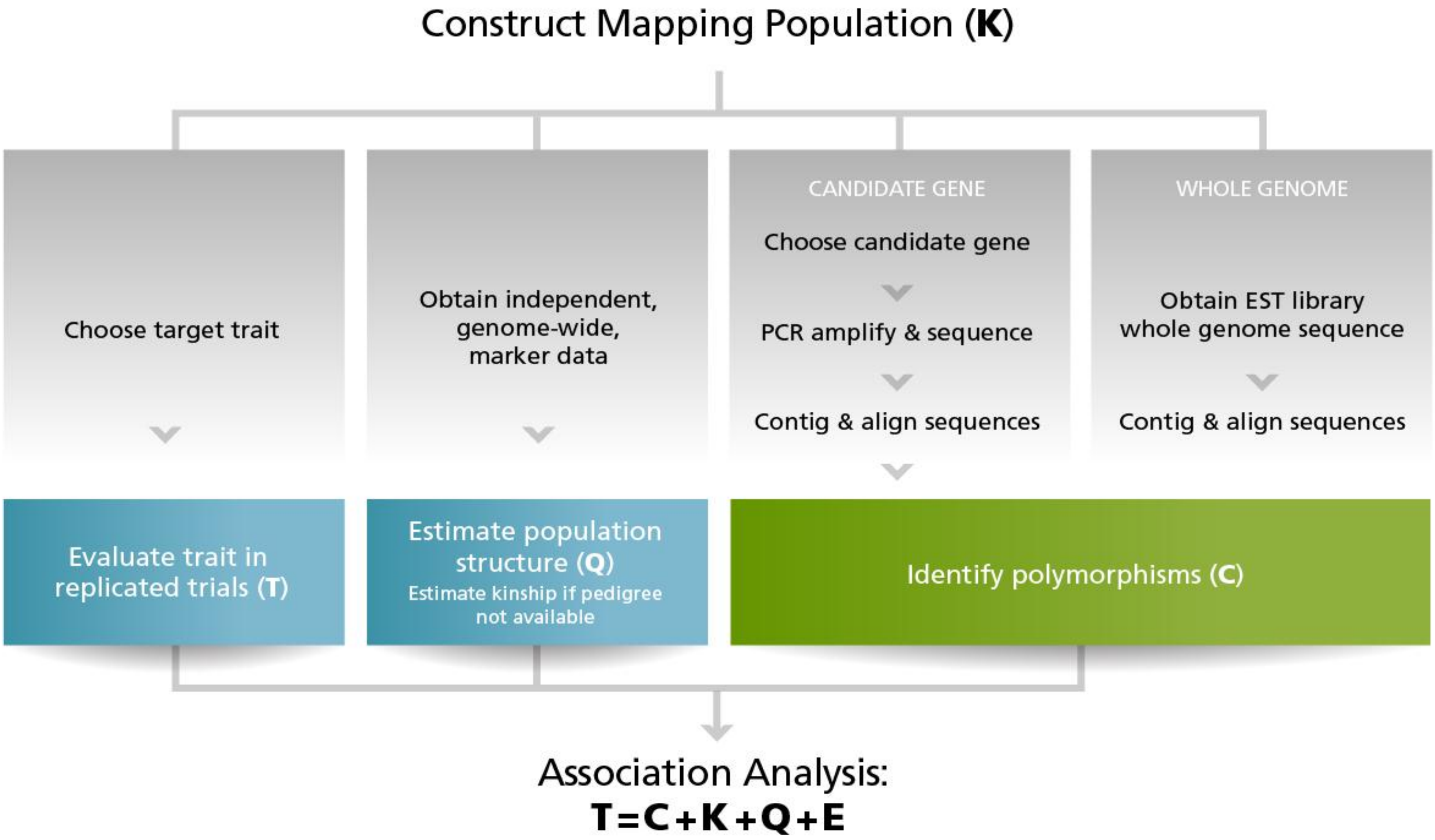
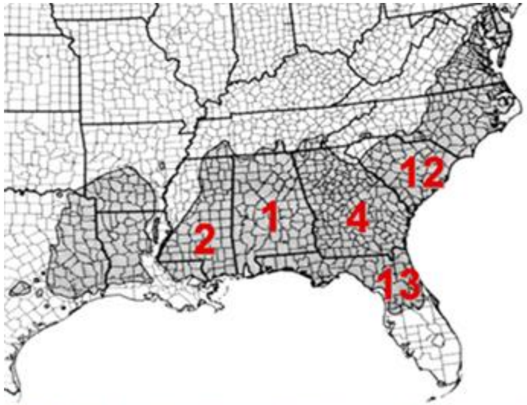
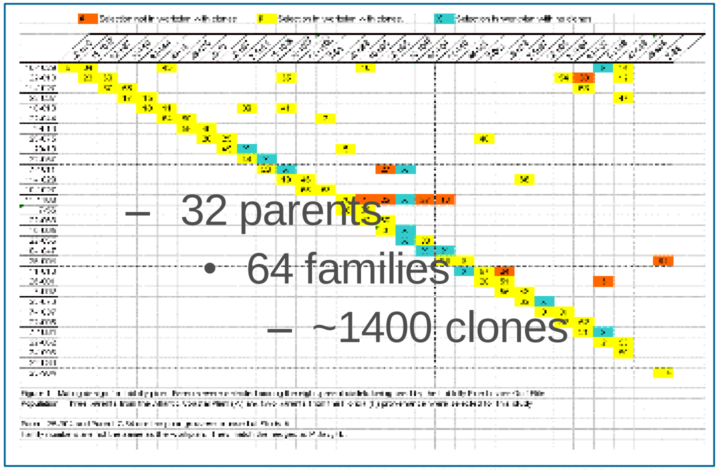

Construct Mapping Population (**K**)
Choose target trait
Obtain independent, genome-wide, marker data
CANDIDATE GENE
Choose candidate gene
PCR amplify & sequence
Contig & align sequences
WHOLE GENOME
Obtain EST library whole genome sequence
Contig & align sequences
Evaluate trait in replicated trials (**T**)
Estimate population structure (**Q**)
Estimate kinship if pedigree not available
Identify polymorphisms (**C**)
Association Analysis:
T=C+K+Q+E

Figure Credit: Modified from Flint-Garcia et al., 2005

An association mapping population with known kinship

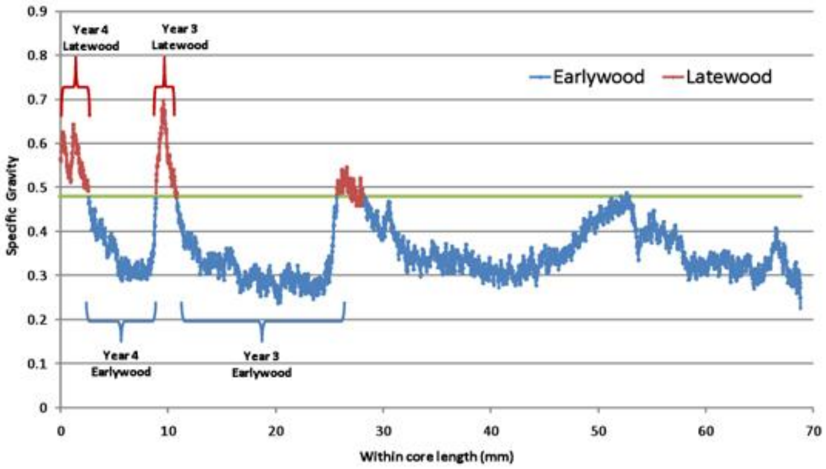


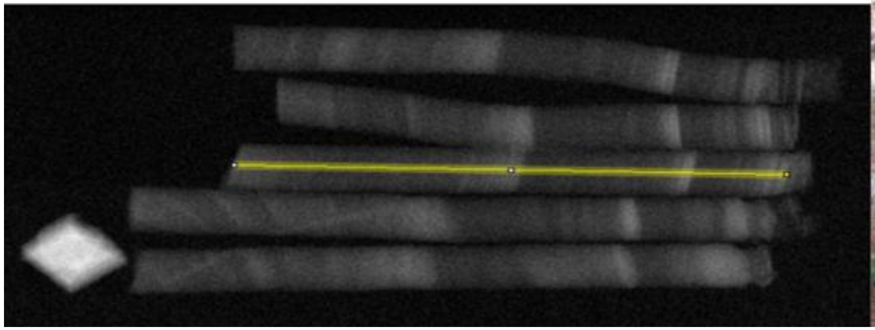


- 32 parents
  - 64 families
    - ~1400 clones

Figure Credits: Cooperative Forest Genetics Research Program, University of Florida

Phenotyping: Precision, accuracy, and more





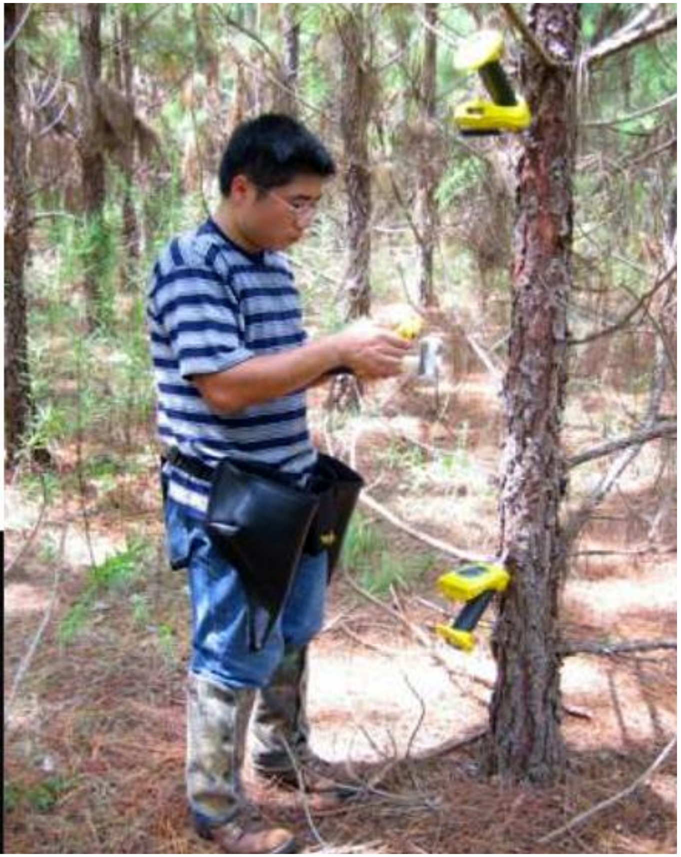


Figure Credits: Gary Peter, University of Florida

Genotyping: Potential genomic targets

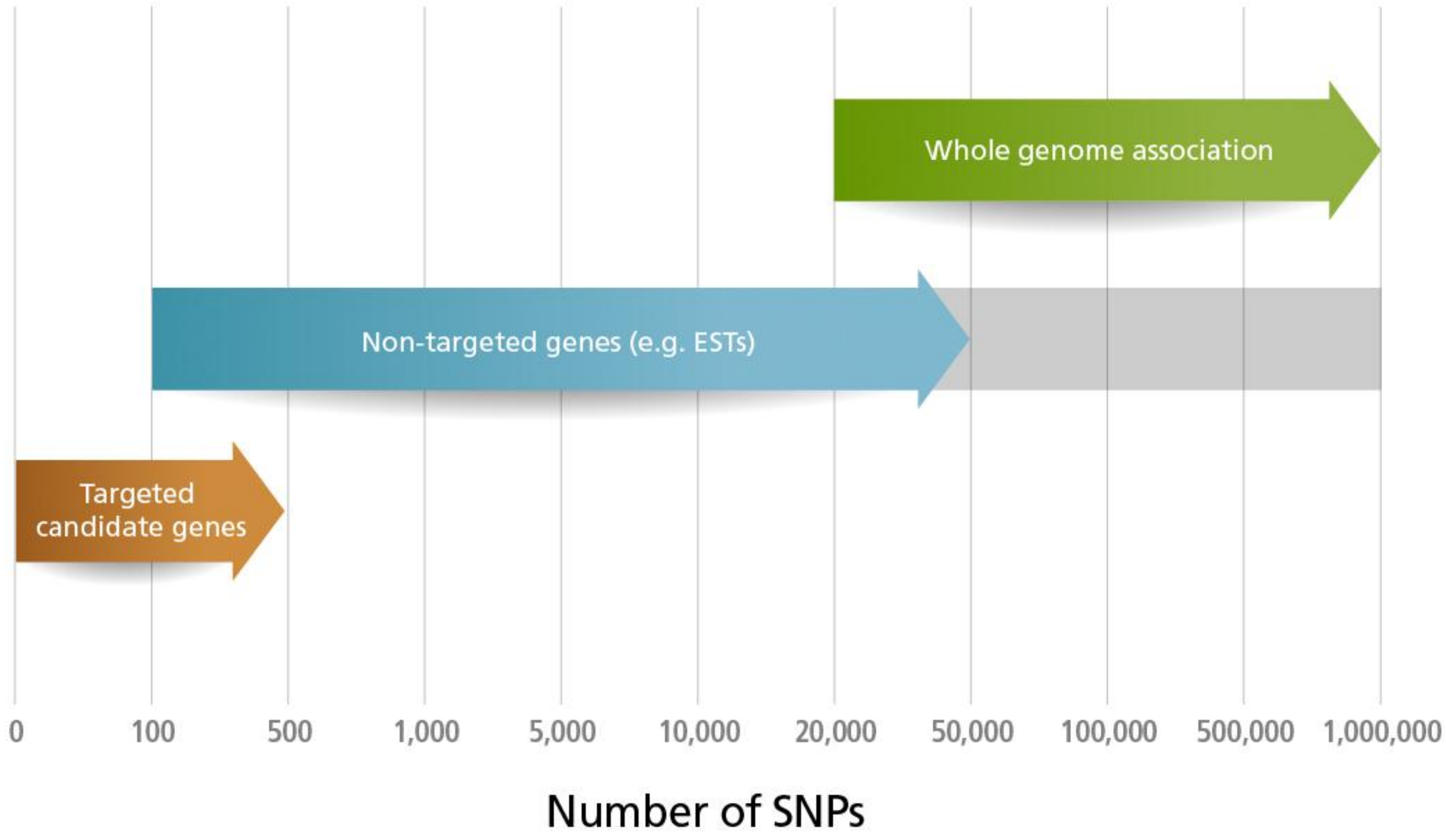
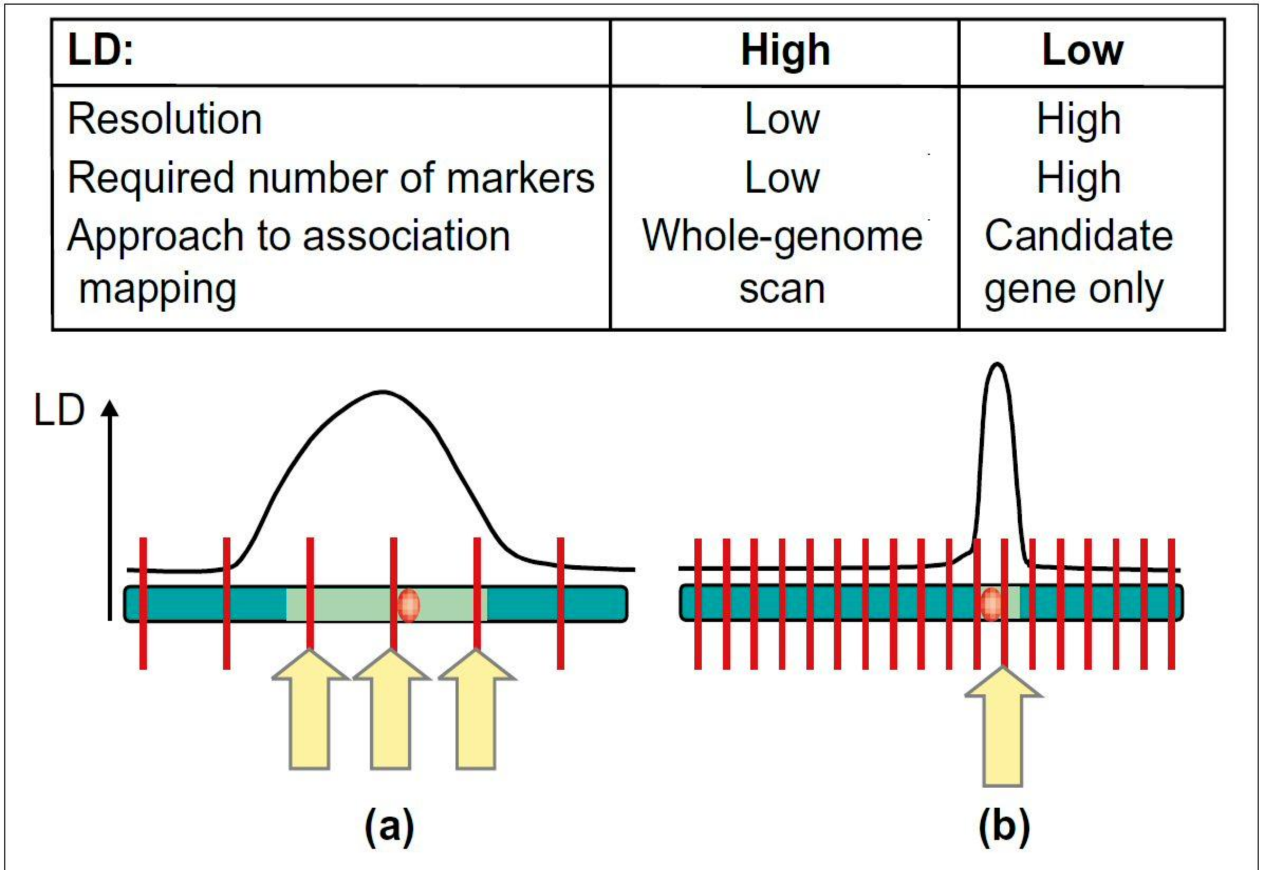


Figure Credit: Nicholas Wheeler and David Harry, Oregon State University

Whole genome or candidate gene? Let’s look again at how this works



LD:	High	Low
Resolution	Low	High
Required number of markers	Low	High
Approach to association mapping	Whole-genome scan	Candidate gene only

Figure Credit: Reprinted from Current Opinion in Plant Biology, Vol 5, Rafalski, Applications of single nucleotide polymorphisms in crop genetics, pages 94-100, 2002, with permission from Elsevier

Local distribution of SNPs and genes

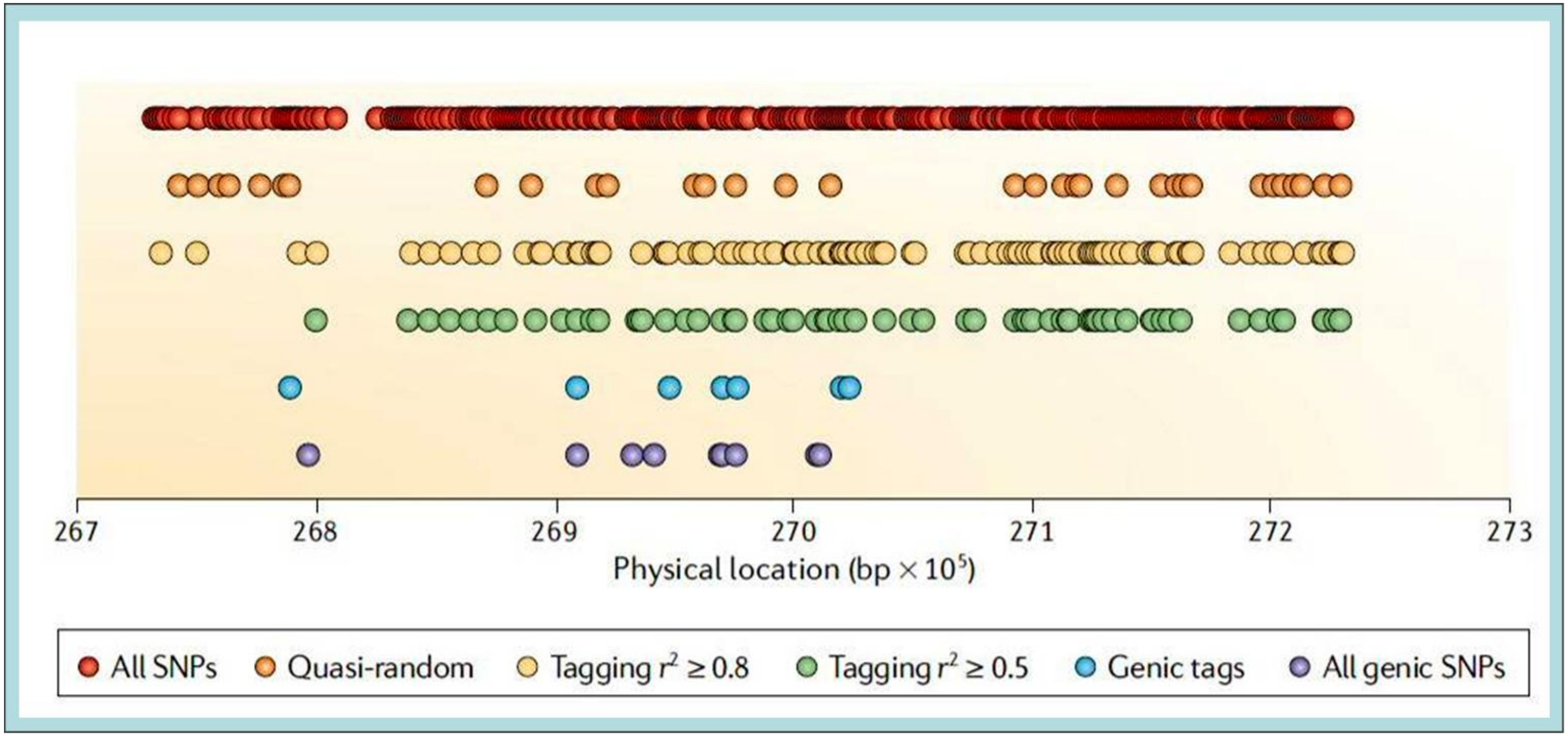


Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Jorgenson and Witte, 2006

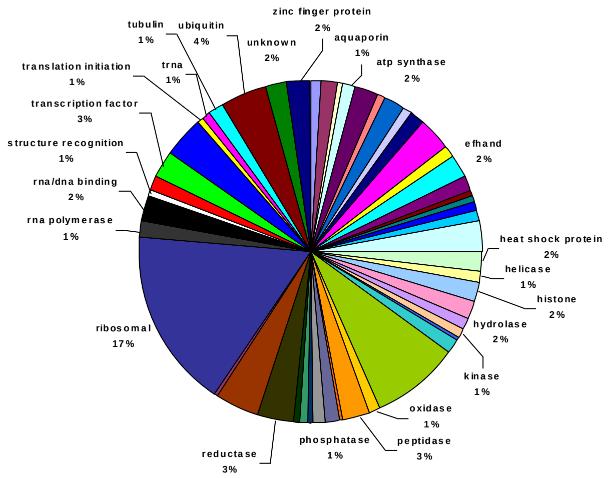
Candidate genes for novel (your) species

- Availability of candidate genes
  - *Positional candidates*
  - *Functional studies*
  - *Model organisms*
  - *Genes identified in other forest trees*

Candidate genes for association studies

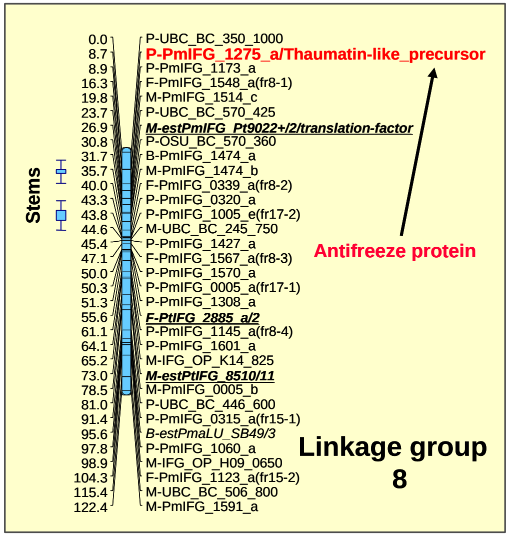
Functional candidates

- By homology to genes in other species
- By direct evidence in forest trees



Positional candidates

- QTL analyses in pedigrees



Expression candidates

- Microarray analyses
- Proteomics
- Metabolomics

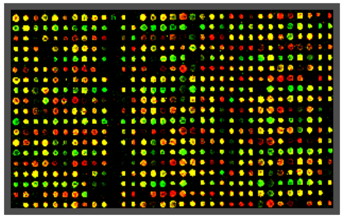


Figure Credits: Kostya Krutovsky, Texas A&M University

Potential genotyping pitfalls

- Quality of genotype data
  - *Contract labs, automated base calls*
- Minor allele frequency
  - *Use minimum threshold, e.g. MAF ≥ 0.05 or MAF ≥ 0.10*
  - *Rare alleles can cause spurious associations due to small samples (recall that D' is unstable with rare alleles)*
- Missing data !!!
  - *Alternative methods for imputing missing data*

Statistical tests for marker/trait associations

- SNP by trait association testing is, at its core, a simple test of correlation/regression between traits
- In reality such cases rarely exist and more sophisticated approaches are required. These may take the form of mixed models that account for potential covariates and other sources of variance
- The principle covariates of concern are population structure and kinship or relatedness, both of which may result in LD between a marker and a QTN that is not predictive for the population as a whole

Causes of population structure

- Geography
  - *Adaptation to local conditions (selection)*
- Non-random mating
  - *Isolation / bottlenecks (drift)*
  - *Assortative mating*
  - *Geographic isolation*
- Population admixture (migration)
- Co-ancestry

Case-control and population structure

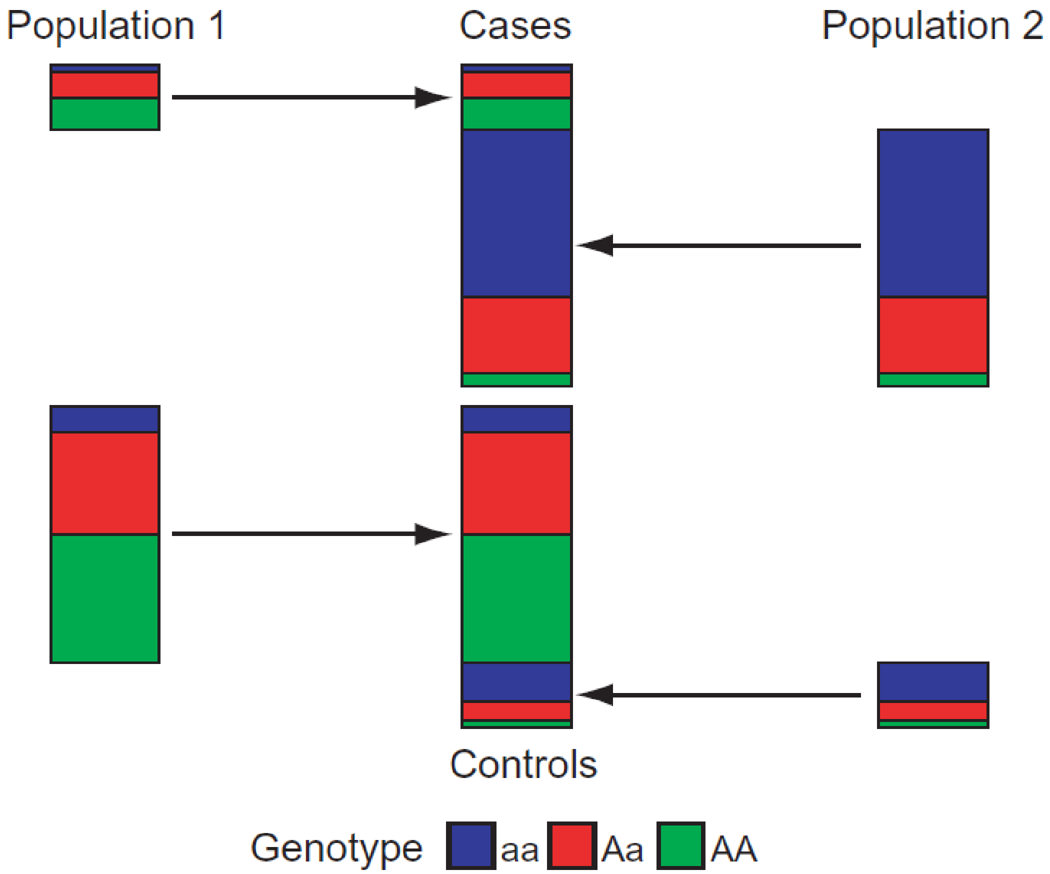


Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics Marchini et al., 2004.

Accommodating population structure

- Avoid the problem by avoiding admixted populations or working with populations of very well defined co-ancestry
- Use statistical tools to make appropriate adjustments

Detecting and accounting for population structure

- Family based methods

- Population based methods
  - *Genomic control (GC)*
  - *Structured association (SA)*
  - *Multivariate*

- Mixed model analyses (test for association)

Family based approaches

- Avoid unknown population structure by following marker-trait inheritance in families (known parent-offspring relationships)
- Common approaches include
  - *Transmission disequilibrium test (TDT) for binary traits*
  - *Quantitative transmission disequilibrium test (QTDT) for quantitative traits*
  - *Both methods build upon Mendelian inheritance of markers within families*
- Test procedure
  - *Group individuals by phenotype*
  - *Look for markers with significant allele frequency differences between groups*
- For a binary trait such as disease, use families with affected offspring
- Constraints
  - *Family structures must be known (e.g. pedigree)*
  - *Limited samples*

Population based: Genomic control

- Because of shared ancestry, population structure should translate into an increased level of genetic similarity distributed throughout the genome of related individuals

- By way of contrast, the expectation for a causal association would be a gene specific effect

- Genomic control (GC) process
  - *Neutral markers (e.g. 10-100 SSRs) are used to estimate the overall level of genetic similarity within a sampled population*
  - *In turn, this proportional increase in similarity is used as an inflation factor, sometimes called λ, used to adjust significance probabilities (p-values)*
  - *For example, $p\text{-value}_{(adj)} = p\text{-value}_{(unadj)} /(1+ \lambda)$*
  - *Typical values of λ are in the range of ~0.02-0.10*

Structured association

- The general idea behind structured association (SA) is that cryptic population history (or admixture) causes increased genetic similarity within groups
- The challenge is to determine how many groups (K) are represented, and then to quantify group affinities for each individual
- Correction factors are applied separately to each individual, based upon the inferred group affinities
- SA is computationally demanding

Multivariate methods

- Multivariate methods build upon co-variances among marker genotypes
- Multivariate methods such as PCA offer several advantages over SA
- Downstream analysis of SA and PCA data are similar

Mixed model approaches

- Mixed models test for association by taking into account factors such as kinship and population structure, provided by other means
- Provides good control of both type 1 (false positive associations) and type 2 (false negative associations) errors

Tassel mixed model : $y_i = \overset{1}{X}\overset{2}{\beta} + \overset{3}{S}\overset{4}{\alpha} + \overset{5}{Q}\overset{6}{v} + \overset{7}{Z}\overset{8}{u} + e$

	Location ID		SNP ID	Population ID		Genotype ID			
Trait	L1	L2	SNP1	P1	P2	G1	G2	G3	G4

$$\begin{pmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8 \end{pmatrix} \begin{matrix} =\\ =\\ =\\ =\\ =\\ =\\ =\\ = \end{matrix} \begin{pmatrix} 1 & 0\\ 1 & 0\\ 1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\\ 0 & 1\\ 0 & 1 \end{pmatrix} * \begin{pmatrix} b_1\\ b_2 \end{pmatrix} + \begin{pmatrix} 1\\ 1\\ 1\\ 0\\ 0\\ 0\\ 1\\ 1 \end{pmatrix} * \begin{pmatrix} a_1 \end{pmatrix} + \begin{pmatrix} 1 & 0\\ 0 & 1\\ 0 & 1\\ 0 & 1\\ 1 & 0\\ 0 & 1\\ 1 & 0\\ 0 & 1 \end{pmatrix} * \begin{pmatrix} v_1\\ v_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} u_1\\ u_2\\ u_3\\ u_4 \end{pmatrix} + \begin{pmatrix} e_1\\ e_2\\ e_3\\ e_4\\ e_5\\ e_6\\ e_7\\ e_8 \end{pmatrix}$$

$$y_i = X\beta + S\alpha + Qv + Zu + e_i$$

$$y_3 = b_1 + a_1 + v_2 + u_3 + e_3$$

🟨 = Measured trait

🟧 = Fixed effects (BLUE = Best Linear Unbiased Esitimates)

🟦 = Random effects (BLUP = Best Linear Unbiased Predictions)

Figure Credit: Fikret Isik, North Carolina State University

Significant associations for diabetes distributed across the human genome

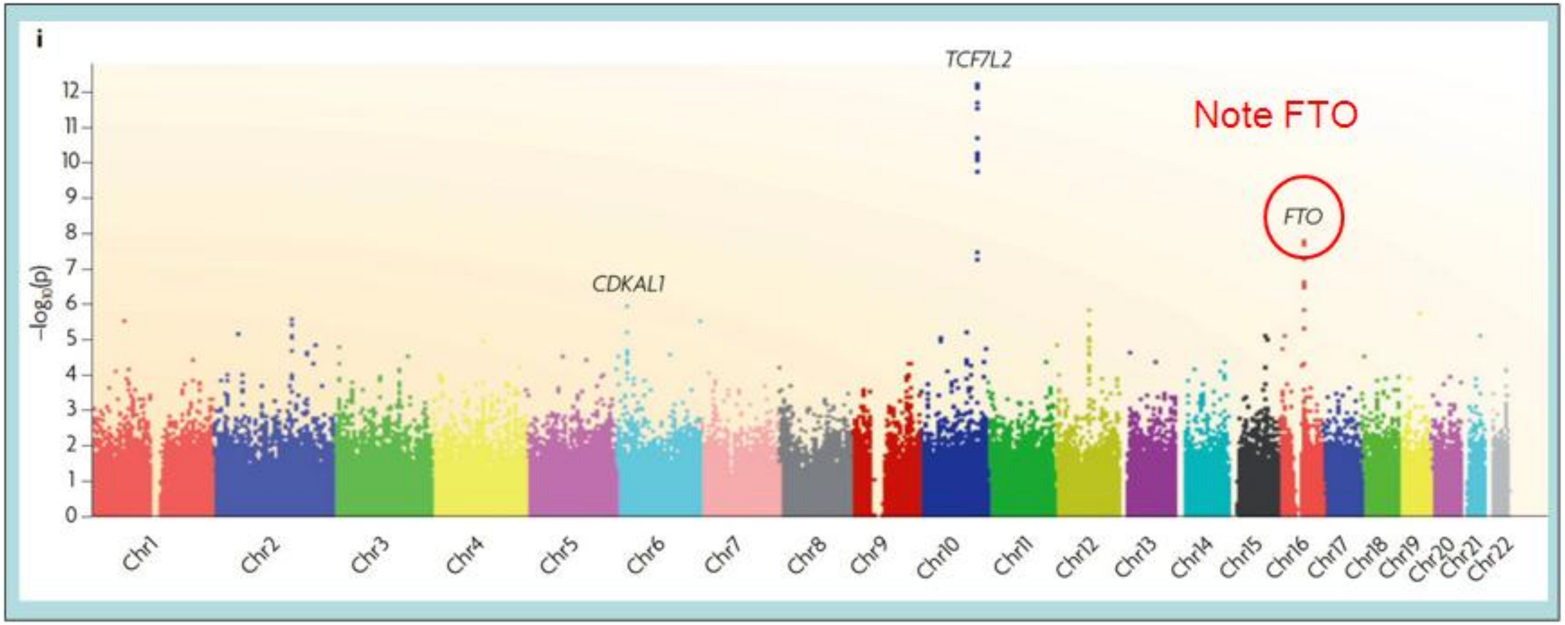


Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics McCarthy et al., 2008.

Association genetics: Concluding comments

- Advantages
  - *Populations*
  - *Mapping precision*
  - *Scope of inference*
- Drawbacks
  - *Resources required*
  - *Confounding effects*
  - *Repeatability*

References cited

- Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdos, and E. S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635. (Available online at: http://dx.doi.org/10.1093/bioinformatics/btm308) (verified 2 June 2011).

- Flint-Garcia, S. A., A. C. Thuillet, J. M. Yu, G. Pressoir, S. M. Romero, S. E. Mitchell, J. Doebley, S. Kresovich, M. M. Goodman, and E. S. Buckler. 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant Journal 44: 1054-1064. (Available online at: http://dx.doi.org/10.1111/j.1365-313X.2005.02591.x) (verified 2 June 2011).

- Jorgenson, E. and J. Witte. 2006. A gene-centric approach to genome-wide association studies. Nature Reviews Genetics 7: 885-891. (Available online at: http://dx.doi.org/10.1038/nrg1962) (verified 2 June 2011).

- Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly. 2004. The effects of human population structure on large genetic association studies. Nature Genetics 36: 512-517. (Available online at: http://dx.doi.org/10.1038/ng1337) (verified 2 June 2011).

- McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J.P.A. Iaonnidis, and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9: 356-369. (Available online at: http://dx.doi.org/10.1038/nrg2344) (verified 2 June 2011).

References cited

- Neale, D. B., and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends in Plant Science 9: 325-330. (Available online at: http://dx.doi.org/10.1016/j.tplants.2004.05.006) (verified 2 June 2011).

- Pearson, T. A., and T. A. Manolio. 2008. How to interpret a genome-wide association study. Journal of the American Medical Association 299: 1335-1344. (Available online at: http://dx.doi.org/10.1001/jama.299.11.1335) (verified 2 June 2011).

- Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5: 94-100.

- Wheeler, N. C., K. D. Jermstad, K. Krutovsky, S. N. Aitken, G. T. Howe, J. Krakowski, and D. B. Neale. 2005. Mapping of quantitative trait loci controlling adaptive traits in coastal Doublas-fir. IV. Cold-hardiness QTL verification and candidate gene mapping. Molecular Breeding 15: 145-156. (Available online at: http://dx.doi.org/10.1007/s11032-004-3978-9) (verified 2 June 2011).

- Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203-208. (Available online at: http://dx.doi.org/10.1038/ng1702) (verified 2 June 2011).

Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UC DAVIS

USDA
United States
Department of
Agriculture

National Institute
of Food and
Agriculture