Conifer Translational Genomics Network
Coordinated Agricultural Project

Genomics in Tree Breeding and Forest Ecosystem Management

Module 13 – Marker Informed Breeding (MIB) – Association Discovery and Evaluation

Nicholas Wheeler – Oregon State University

Marker informed breeding is many things

- Marker informed program management (MIPM – see Module 12)
  - *Fingerprinting, paternity analyses, characterizing population genetic variation*
- Marker assisted selection (MAS)
  - *Central dogma of molecular breeding involves the utilization of molecular marker fingerprints to improve selection efficiency in plant breeding programs (Eathington et al., 2007)*

Current status of MAS in tree improvement

- MAS in forest trees is mostly in the discovery/research phase
- Why the slow adoption of such a promising technology?
  - *Highly heterozygous trees and large diverse populations*
  - *Out-crossed species in linkage equilibrium*
  - *Poor understanding of the genetic architecture of traits*
  - *Lack of simply inherited traits*
  - *Very modest proportion of the genome characterized*
  - *Few scientists working in this area*
  - *Little industrial investment*
  - *High cost of program development*

Three approaches to MAS (classified by mapping precision)

FIGURE 5

Classification of three different types of marker-trait associations relevant to *Eucalyptus* MAS

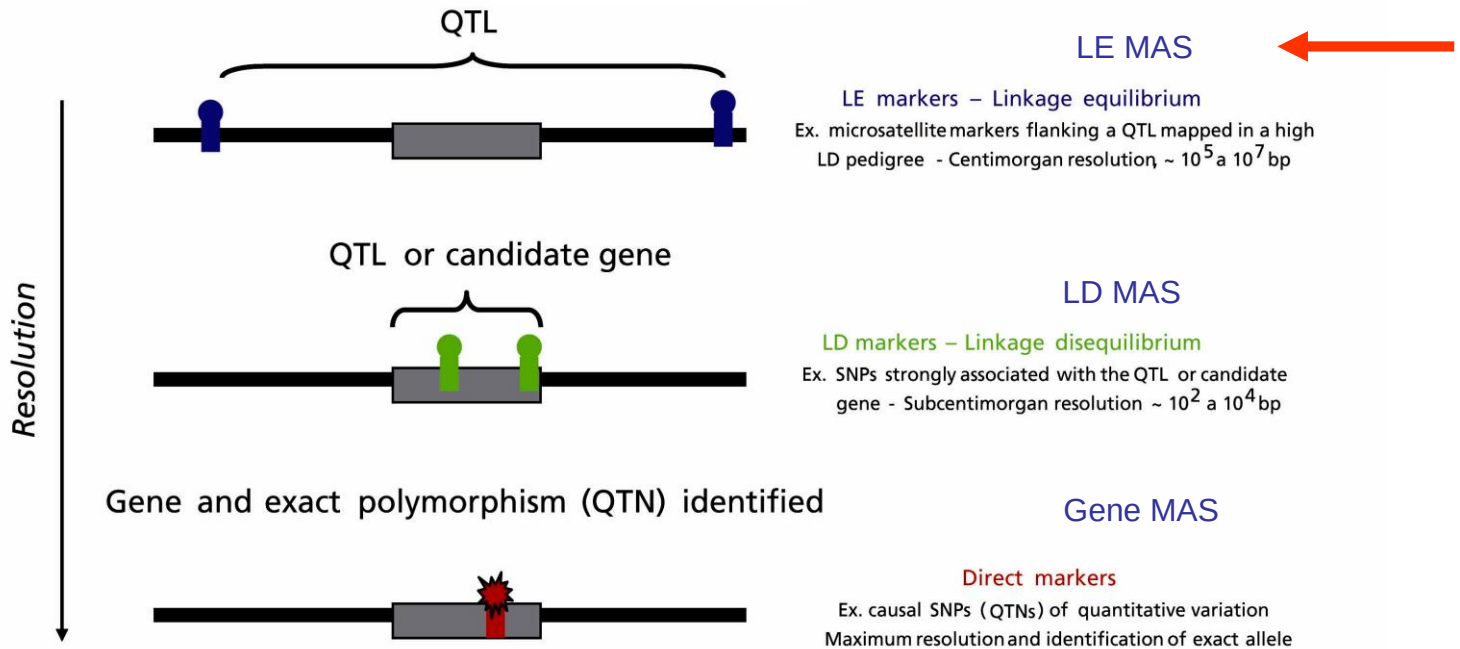


Figure Credit: Modified from Grattapaglia, 2007.

TABLE 1

Loblolly pine mapping populations and phenotypic traits for QTL analyses of physical and chemical wood properties

	Pedigree		
	Detection	*Verification*	*Unrelated*
Grandparents	$G_1 \times G_2$ $G_3 \times G_4$	$G_1 \times G_2$ $G_3 \times G_4$	$G_5 \times G_6$ $G_7 \times G_8$
Parents	$P_1 \times P_2$	$P_1 \times P_2$	$P_5 \times P_6$
Progeny	172	457	445
Trait and rings analyzed			
Wood-specific gravity (*ewsg, lwsg*)[a]	Rings 2–11	Rings 4–6	Rings 4–6
Percentage of late wood (*%lw*)	Rings 2–11	Rings 4–6	Rings 4–6
Microfibril angle (*emfa, lmfa*)	Rings 3, 5, 7	Ring 6	Ring 6
Cell wall chemistry (*ecwc, lcwc*)[b]	Ring 5	Ring 6	Not assayed

[a] *wsg* is a measure of the total amount of cell wall substance and within an annual ring has three main determinants: *wsg* of earlywood (xylem cells having thin walls and large lumens: *ewsg*), *wsg* of latewood (xylem cells with thicker walls and smaller lumens: *lwsg*), and the percentage of latewood (*%lw*).

[b] Mass peaks collected from the pyrolysis molecular beam mass spectrophotometer were associated with the amounts of α-cellulose, galactan, mannan, xylan, and lignin, collectively termed *cwc* traits.

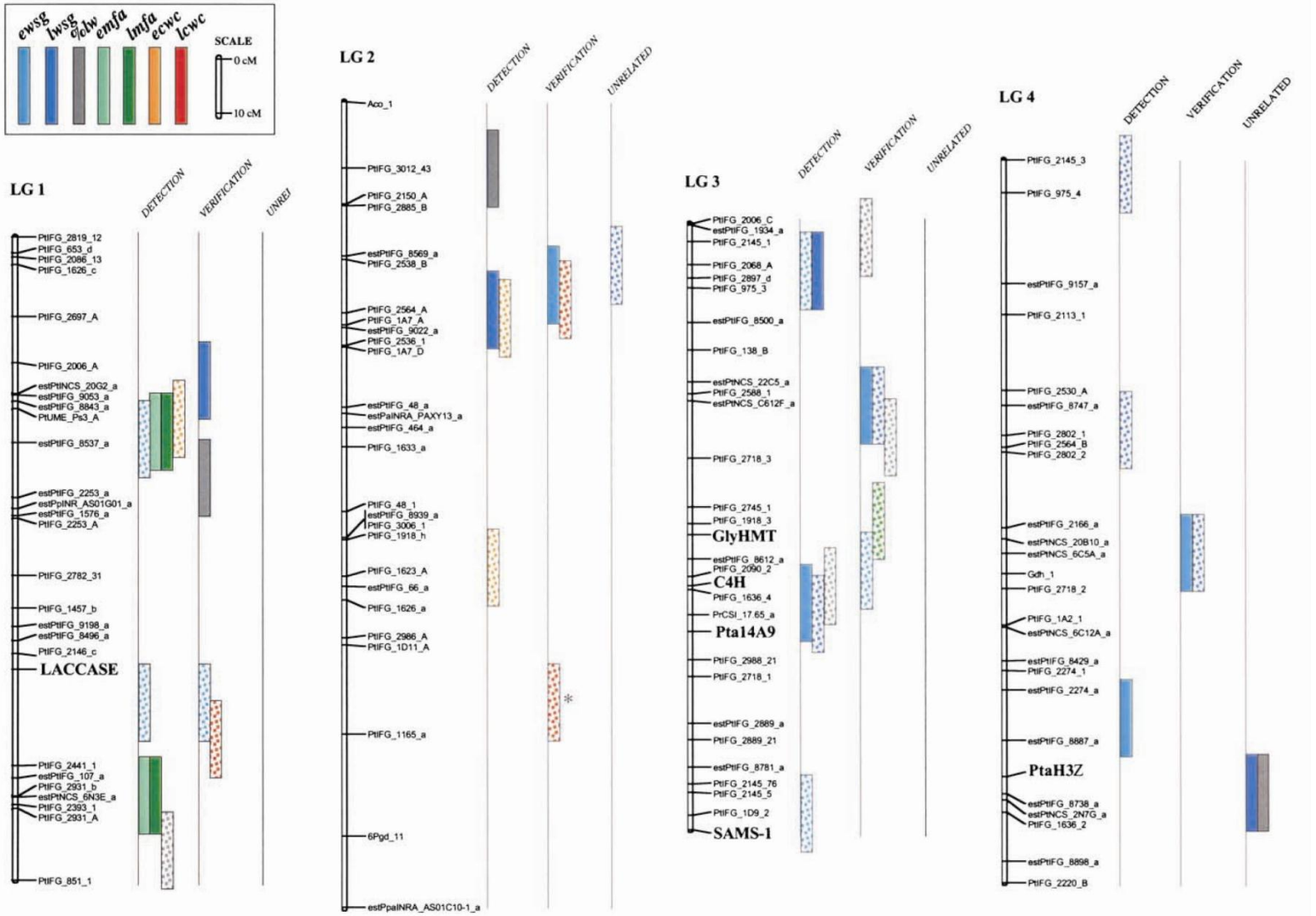


Figure Credit: Modified from Brown et al., 2003

TABLE 2

Summary of QTL verified by repeated detection in loblolly pine

LG	Trait	Interval[a]	*P* value[b]		PVE[c]			Candidate genes
			Verification	*Unrelated*	*Detection*	*Verification*	*Unrelated*	
1	*ewsg*	2146_c:2441_1	0.004*		7.2	3.0		laccase
2	*lwsg*	2150_A:1A7_A		0.005*	5.4		2.3	
3	*ewsg*	2090_2:17.65_a	0.006*		6.6	3.2		C4H, GlyHMT, Pta14A9
5	*ewsg*	15.01_a:2220_A	0.003*		6.0	2.9		
5	*ewsg*	2963_3:2090_4	0.004**[d]		5.6	3.5[d]		AGP6
5	*%lw*	2933_1:15.01_a	0.002**[d]	0.006*	7.2	3.1[d]	2.0	
6	*ewsg*	2802_3:8972_a	0.0008**		6.6	3.5		PAL-2
6	*%lw*	2874_1:8702_a	0.007*[e]		11.0[e]	3.1[e]		
		8702_a:2009_a						CCoAOMT
6	*ecwc*	2874_1:8702_a	0.006*[d]		6.4	4.4[d]		
7	*lwsg*	1916_2:2361_2		0.0105*	5.9		2.4	C3H, 4CL, PtaAGP4
7	*%lw*	1916_2:2361_2		0.0006**	6.1		2.4	C3H, 4CL, PtaAGP4
8	*%lw*	719_A:1916_4	0.004*		5.8	1.8		
10	*%lw*	2145_2:1635_A	0.005*		8.7	2.5		
12	*ewsg*	8542_a:3012_2	0.015*		5.4	2.3		

[a] Marker interval on the consensus genetic map of loblolly pine. Markers that are underlined are common to the genetic maps of the populations compared. Markers not underlined denote interval boundaries inferred from homologous flanking markers.

[b] * and ** represent chromosome-wide significance at $P < 0.05$ and 0.01, respectively, except for marker-trait associations detected by the two-QTL model (see footnote [d] below).

[c] Percentage of the phenotypic variance explained by a QTL.

[d] A QTL detected only by the two-QTL model in the *verification* population with significance levels as in Sewell *et al.* (2000): *, $0.01 > P > 0.005$; **, $P < 0.005$. PVE refers in this case to that explained by the pair of QTL detected.

[e] QTL detected by the two-QTL model in both the *detection* and *verification* populations.

Three approaches to MAS (classified by mapping precision)

FIGURE 5

Classification of three different types of marker-trait associations relevant to *Eucalyptus* MAS

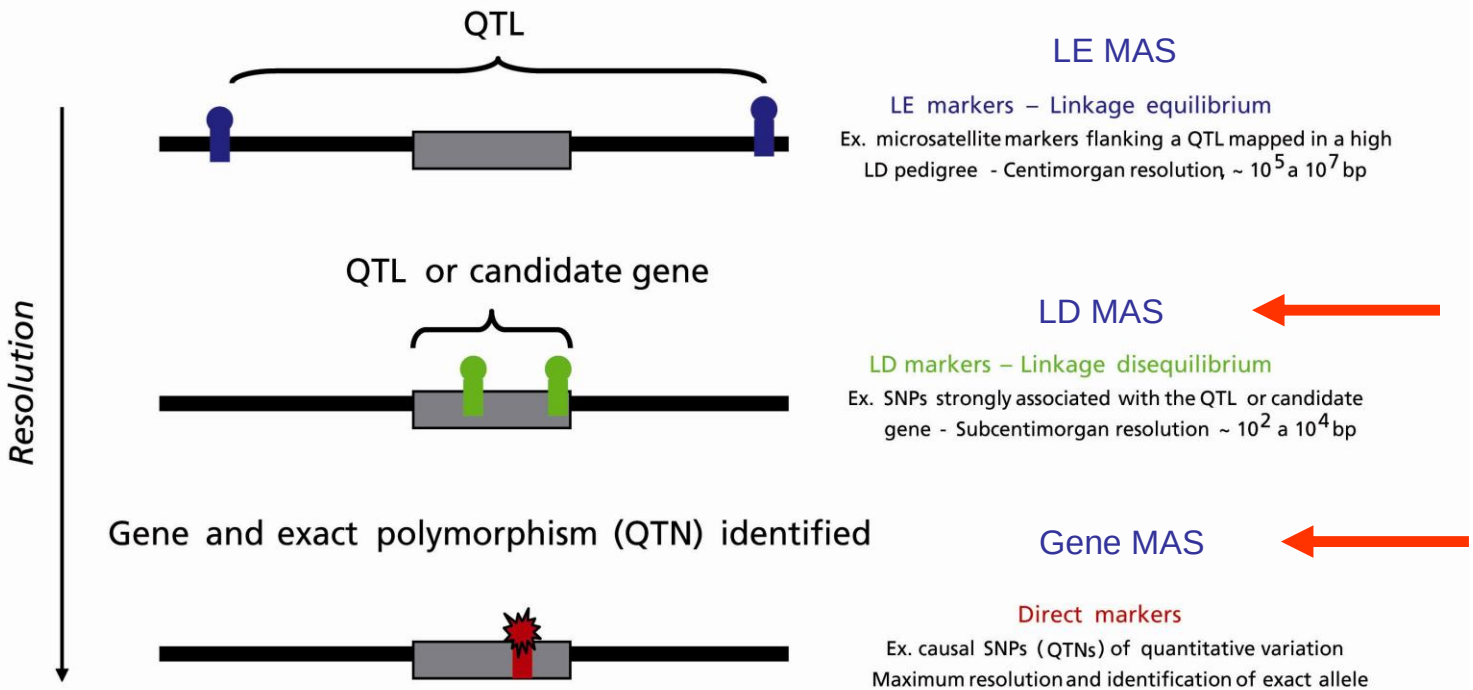


Figure Credit: Modified from Grattapaglia, 2007

LD MAS / Gene MAS

- We will treat LD MAS and Gene MAS (gene assisted selection) together here, though the distinction might be relevant under some circumstances. We will collectively refer to the discovery approaches for identifying LD QTL-trait associations as association mapping or association genetics
- The fundamental distinction between association mapping and LE QTL mapping is that the latter relies on genetic linkage following one or two generations of crossing, while the former utilizes historical, population-level LD
- This has enormous implications for practical application in forest tree improvement programs

A loblolly pine association population

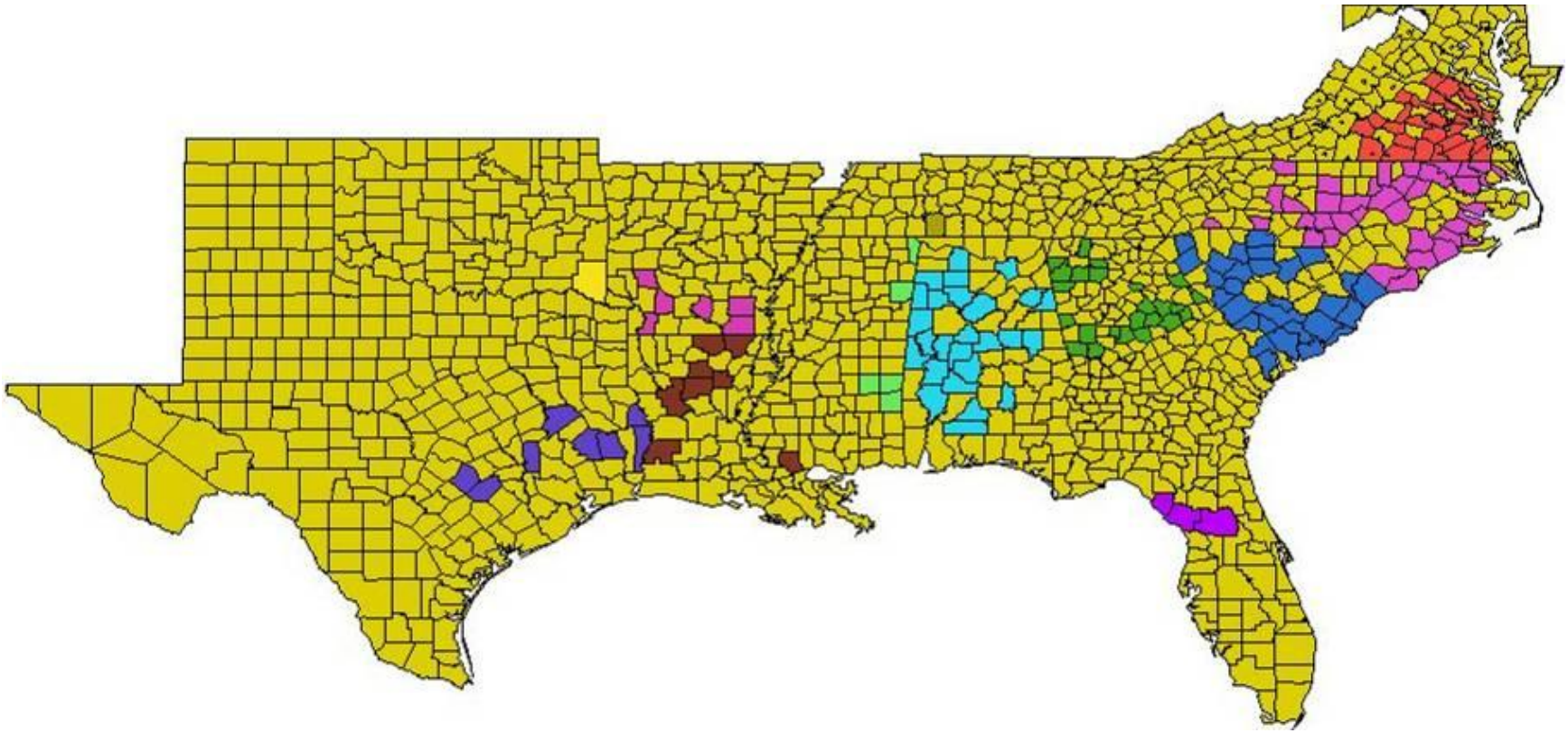
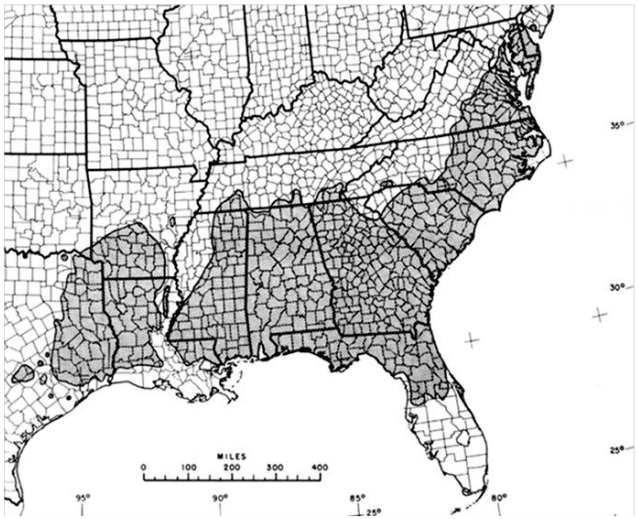
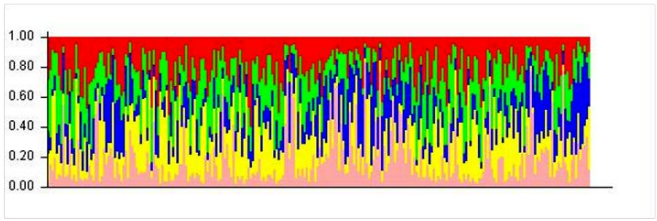


Figure Credit: Barry Goldfarb, North Carolina State University

Pinus taeda L

Continuous range, no clear population genetic structure

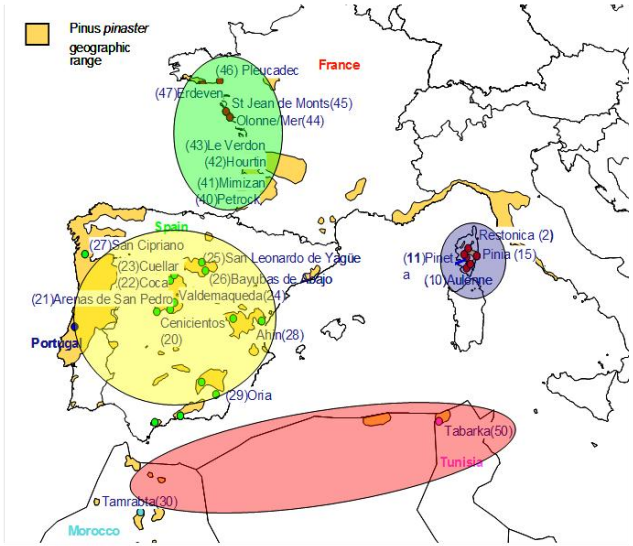


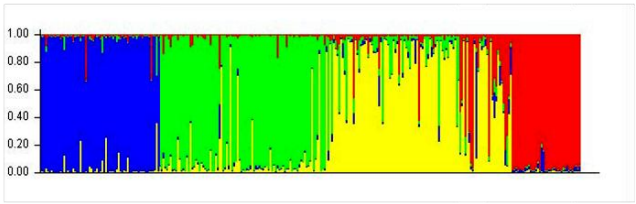


ADEPT project

Pinus pinaster Ait.

Fragmented range, significant population structure





TREESNIPS project
(also *P. sylvestris*, *Picea abies* and oaks)

Figure Credit: David Neale, University of California, Davis

Associations between SNPs and wood properties

- Gonzalez-Martinez et al. 2007.
  - *This was the first multi-gene association genetic study in forest trees to be reported*
  - *It demonstrated feasibility of candidate gene strategies for dissecting complex traits*

- Study details
  - *Genetic associations were tested between 58 SNPs from 20 candidate genes and wood properties (specific gravity, % latewood, microfibril angle, and wood chemistry – cellulose, lignin content) on over 400 clonally replicated individuals*
  - *Population structure assessed (22 nuclear SSR) and kinship removed*

Many associations were identified

- Many significant associations were identified between wood traits and genes known to be associated with lignin and cellulose biosynthesis
- Many SNPs gave consistent associations with the same trait measured at different ages
- Some SNPs were consistent with co-location of candidates and QTL

	EWSG	LWSG	%LW	MFA
Juvenile Wood	4	5	4	5
Transition Wood	7	7	4	9
Mature Wood	6	1	3	5
All Age	6	3	3	0
PCA	11	2	2	6
Total	34	18	16	25

EWSG – earlywood specific gravity; LWSG – latewood specific gravity; %LW – percent latewood; MFA – microfibril angle

Significant association with candidate genes for wood quality traits after correction for multiple testing using the positive FDR method (Q-values)

Trait	Wood-age type	Gene	SNP	N	Marker effect			FDR
					F	*P*	R^2	*Q-value*
ewsg	*transition*	*sams-2*	M44	403	6.7595	0.0013	0.0327	0.0630
	all age	*cad*	M28	409	7.7480	0.0005	0.0347	**0.0228**
	PCA	*cad*	M28	366	6.5945	0.0015	0.0351	0.0742
lw	*transition*	*lp3-1*	Q5	431	7.9007	0.0004	0.0357	**0.0248**
ewmfa	*transition*	*α-tubulin*	M10	374	8.3766	0.0040	0.0221	**0.0062**
	PCA	*α-tubulin*	M10	370	13.508	0.0003	0.0355	**0.0078**

Table Credit: Table used with permission of the Genetics Society of America from “Association Genetics in Pinus Taeda L I. Wood properties”, Gonzalez-Martinez et al. Genetics 175: 399-409, 2007; permission conveyed through Copyright Clearance Center, Inc.

FIGURE 2. Genotypic effects (box plots) of SNPs that showed significant genetic association (after correction for multiple testing) with earlywood specific gravity (cad SNP M28 and sams-2 SNP M44) and percentage of latewood (lp3-1 SNP Q5 and 4cl SNP M7 in the east of the Mississippi Valley range)

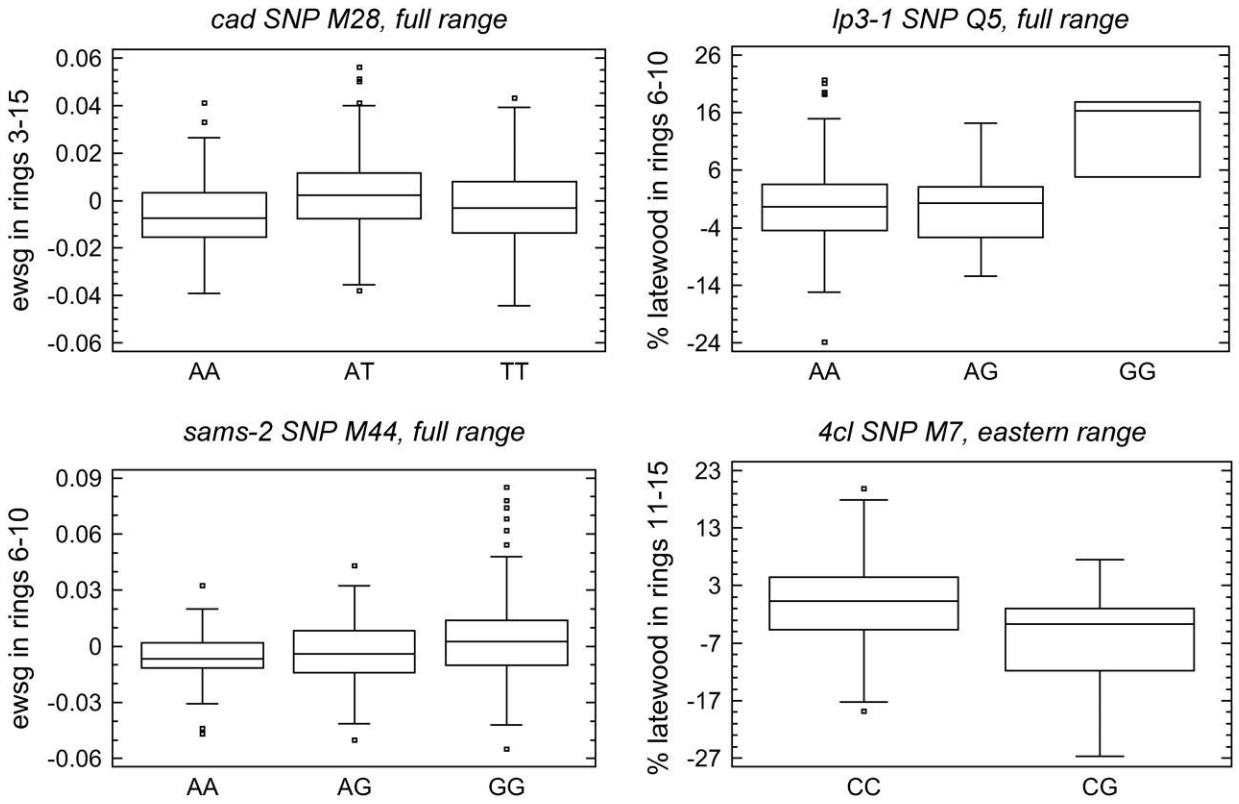
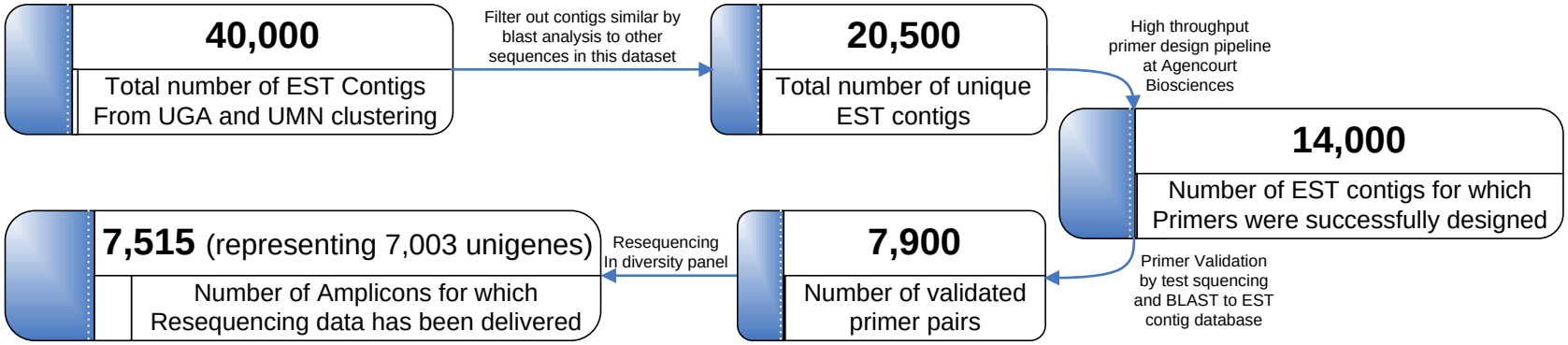


Figure Credit: Figure used with permission of the Genetics Society of America from “Association Genetics in Pinus Taeda L I. Wood properties”, Gonzalez-Martinez et al. Genetics 175: 399-409, 2007; permission conveyed through Copyright Clearance Center, Inc.

ADEPT2 re-sequencing status



~40,000 SNPs called
Average of ~6 SNPs per amplicon
Average amplicon size is 450bp

Figure Credit: Jennifer Lee, University of California, Davis

Finding genes associated with pitch canker resistance

	SNP	Best Hit
1	0_15227_01_159	ATP binding protein, lectin-like protein kinase
2	0_15382_01_99	geranylgeranyl transferase type I beta subu
3	0_2234_01_128	putative long-chain acyl-CoA synthetase
4	0_6323_01_240	DELLA protein
5	0_9288_01_370	***** No hits found ******
6	1_3327_01_113	***** No hits found ******
7	2_4484_02_622	plastid hexose transporter
8	2_6181_02_400	hexokinase
9	2_8946_02_435	Cucumber peeling cupredoxin
10	CL4336Contig1_01_180	n.a.

10 significant SNPS, accounting for 3.5% of phenotypic variation. Heritability ~0.28. Thus, total genetic variation explained is ~10 – 15%

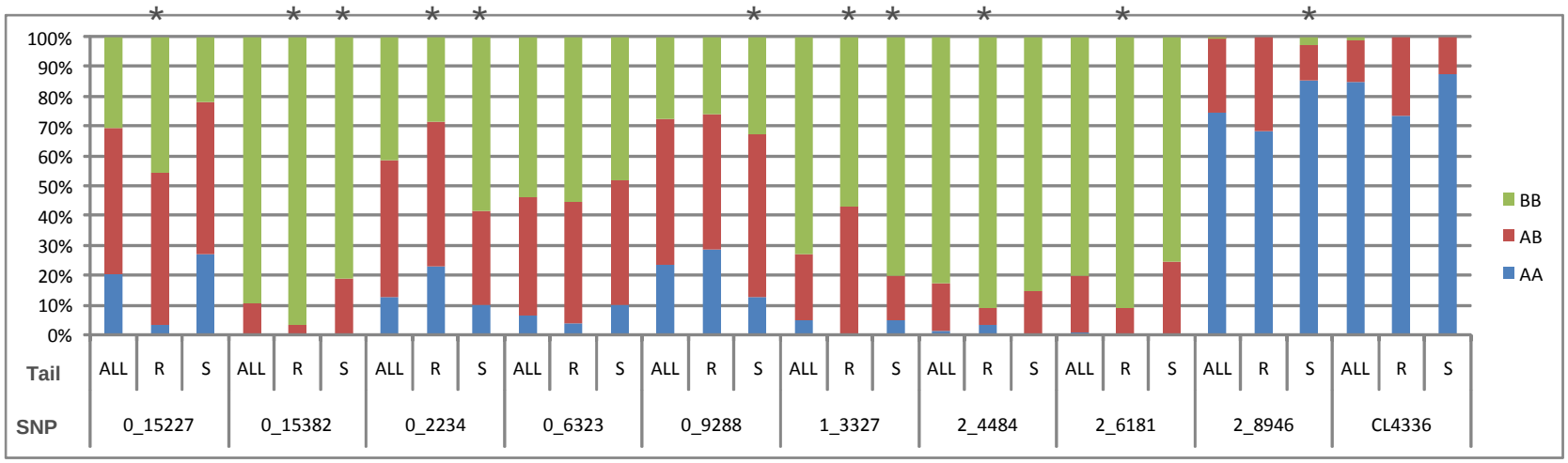


Figure Credit: John Davis, University of Florida

Genes associated with water use efficiency, nitrogen content, and height

Trait	Heritability (Broad)	# of SNP Associations	% Phenotype Explained	% Genotype Explained
WUE (Carbon 13)	0.50 ± 0.05	5-7†	7.1	~14.2
% Nitrogen	0.42 ± 0.06	5*	7.0	~16.7
Height	0.43 ± 0.05	1	<.01	-

† Gonzalez-Martinez et al., 2008, two SNPs, 7% phenotypic variance explained

* One of these five loci explained most of the variation

What have we learned from these studies?

- Both approaches – candidate gene and whole genome screens – appear to be effective for dissecting complex traits in experimental populations of trees
- Desirable alleles can be identified for breeding and conservation and their breeding values and mode of action can be estimated (effect of allelic substitution)
- Correcting for multiple testing greatly reduces the number of associations statistically confirmed
- To the extent that testing was performed over two or more populations, there was an encouraging level of validation
- The size of effects described is consistent with the proportion of the genome studied

The future of MAS in forestry

- LD MAS and Gene MAS show promise of satisfying elements of the vision breeders have for MAS but questions remain
  - *What level of gain might we expect from the addition of MAS?*
  - *Would it be economical to do so?*
  - *Can we verify effects? Are there G X E interactions? Do our studies allow for dissection of epistatic effects?*
  - *How and when would we apply association in the tree improvement cycle?*

Contrasting traditional selection approaches with MAS

Method	Reliability	Years per Stage‡	Total Cycle Time	%Gain*	%Gain per Year
Seedling Progeny Testing†	0.03	5+3+1+6	15	7%	0.5%
Clonal Progeny Testing†	0.70	5+3+1+2+6	17	122%	7%
Marker-based Selection	0.05	5+3+1	9	12%	1%
Marker-based Selection	0.15	5+3+1	9	35%	4%
Marker-based Selection	0.25	5+3+1	9	58%	6%
Marker-based Selection	0.35	5+3+1	9	81%	9%

‡ Values refer to years for selection, grafting and breeding, raising seedlings, vegetative propagation, and field testing, in that order. * Gain based on selection intensity of 2.33 for seedling testing and 1.75 for clonal testing.

Table Credit: Patrick Cumbi, North Carolina State University

Contrasting traditional approaches with MAS

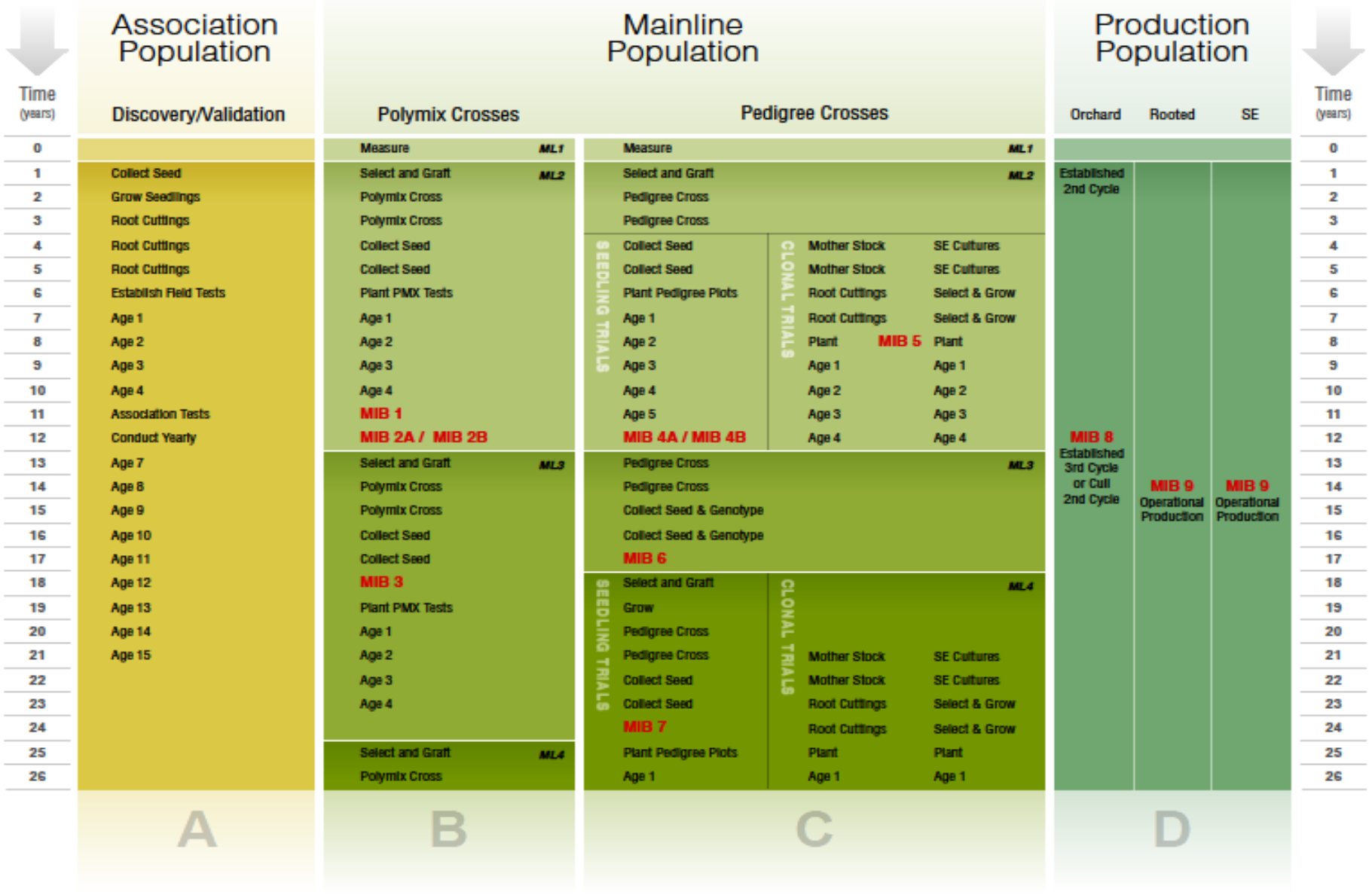
Selection Method	Reliability	# of Genotypes	Field Testing Cost ($)	Vegetative propagation Cost ($)	Cost to Genotype 25 Markers ($)	Total Selection Cost ($)	% Gain*	Cost per % Gain ($)
Seedling-Progeny†	0.03	16,000	$80,000	-	-	$80,000	7%	$11,455
Clonal - Progeny†	0.70	4,000	$80,000	$80,000	-	$160,000	122%	$1,310
Marker-based	0.05	16,000	-	-	$96,042	$96,042	12%	$8,251
Marker-based	0.15	16,000	-	-	$96,042	$96,042	35%	$2,750
Marker-based	0.25	16,000	-	-	$96,042	$96,042	58%	$1,650
Marker-based	0.35	16,000	-	-	$96,042	$96,042	81%	$1,179

Table Credit: Patrick Cumbi, North Carolina State University

What is needed to make association viable?

Can we verify effects? Are there G X E interactions? Do our studies allow for dissection of epistatic effects? Answers to these questions remain largely unknown and getting them will require considerably more work.

- Appropriate populations
- Repeated trials
  - *Time*
  - *Space*
  - *Genetic background*
- Whole genome scan (all or nearly all genes represented by 1 or more SNPs, genotyped for relevant populations)
- Dedicated scientists
- Dedicated long-term funding from industrial partners

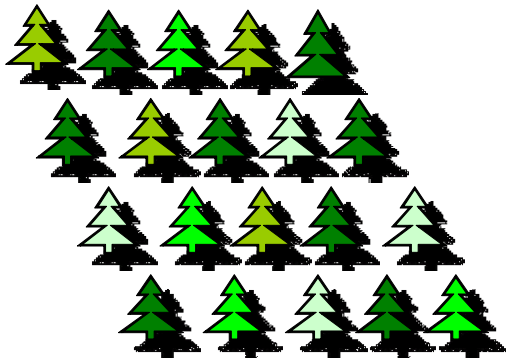


www.pinegenome.org/ctgn

Figure Credit: Nicholas Wheeler, Oregon State University.

CTGN CAP

Using LD MAS plus phenotypes for forward selection (picking superior individuals for next generation breeding)



SNP genotypes of parents and potential genotypes of progeny at 3 loci controlling MFA (micro fibril angle)

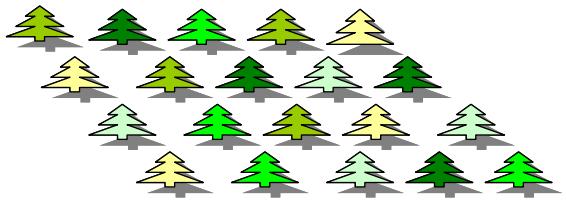
Colors represent controlled crosses (full-sib families)

Locus 1	Locus 2	Locus 3	Breeding Value
A/A x A/T	C/G X G/G	A/T X A/T	
A / A 1 A / T 1	C / G 1 G / G 1	A / A 1 A / T 2 T / T 1	FamAve: + 7.0 BestInd: + 14 WorstInd: 0
A/T X A/T	C/C X C/G	T/T X A/T	
A / A 1 A / T 2 T / T 1	C / C 1 C / G 1	A / T 1 T / T 1	FamAve: + 5.5 BestInd: + 11 WorstInd: 0
A/G X A/A	C/G X C/G	A/T X A/T	
A / A 1 G / A 1	C / C 1 C / G 2 G / G 1	A / A 1 A / T 2 T / T 1	FamAve: + 10 BestInd: + 19 WorstInd: + 2
A = + 2 T = 0 G = +5	C = + 2 G = 0	A = + 4 T = -1	

Figure Credit: Nicholas Wheeler, Oregon State University.

MIB 4B: Marker assisted breeding

- Using LD MAS to identify superior individuals to intermate with the intent of pyramiding favorable alleles at multiple loci
  - *Colors represent full-sib families with varying numbers of favorable disease resistance alleles*



Field trial – Age 4 to 8

Phenotypes of individual trees for disease resistance at 3 loci
S = Susceptible R = Resistant
R is dominant
(Assumes all R phenotypes are actually heterozygotes and SNPs exist that can distinguish all alleles)*

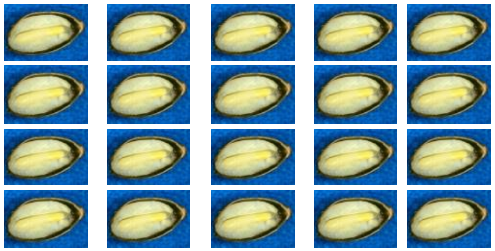
	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	sss	ssR	sRs	sss	Rss
Row 2	Rss	sss	ssR	RsR	ssR
Row 3	RsR	sRs	sss	Rss	RsR
Row 4	Rss	sRs	RsR	ssR	sRs

* Actual genotype for C1R3 is [s/R, s/s, s/R]

Figure Credit: Nicholas Wheeler, Oregon State University.

MIB 5 and/or 7: Early culling - individual tree selection to populate genetic tests

- Using LD MAS to identify superior individuals to move forward into seedling and or clonal trials (two-stage forward selection)



Phenotypes of individual seedling / seed for disease resistance at 3 loci
s = Susceptible R = Resistant
R is dominant
(Assumes all R phenotypes are actually heterozygotes and SNPs exist that can distinguish all alleles)*

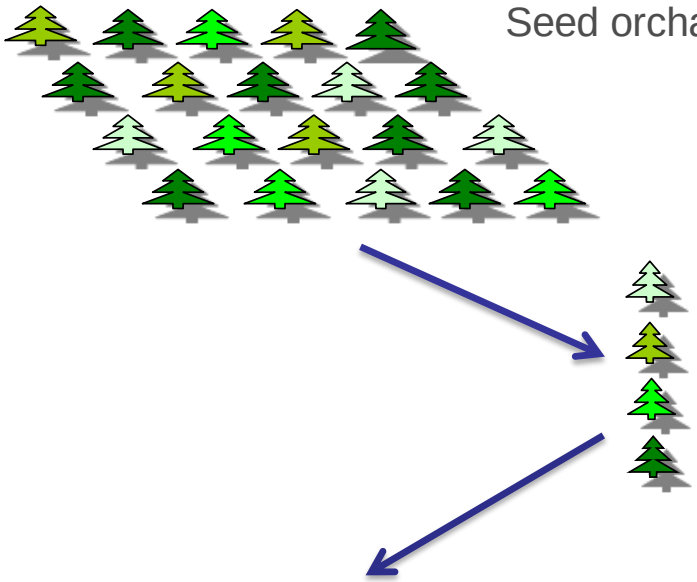
	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	sss	ssR	sRs	sss	sss
Row 2	sRR	Rss	sss	Rss	sss
Row 3	sRs	sRs	RRR	sRS	ssR
Row 4	sss	sRs	sss	sss	sss

* Actual genotype for C1R3 is [s/s, s/Rs/s]

Figure Credit: Nicholas Wheeler, Oregon State University.

MIB 8: Cull existing seed orchard

Using LD MAS Plus Phenotypic Selection to Cull Seed Orchards (Backward Index Selection)
Colors represent female parent (Half-sib family)



SNP Genotypes at 3 loci controlling MFA (microfibril angle)

Locus 1	Locus 2	Locus 3	Breeding Value
A / T	C / C	T / T	+ 4
A / A	C / G	A / T	+ 9
A / A	C / C	A / A	+ 16
A / T	C / G	A / A	+ 12
A = + 2 T = 0	C = + 2 G = 0	A = + 4 T = -1	

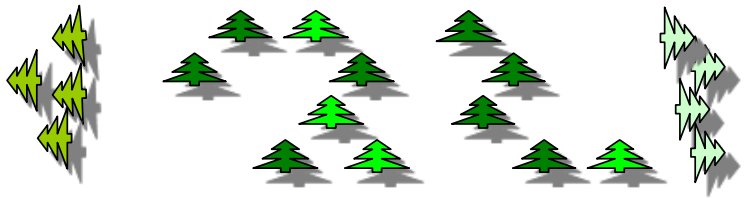


Figure Credit: Nicholas Wheeler, Oregon State University.

Association genetics: A review

- A comprehensive association genetics study should provide the following (White et al. 2007)
  - *An estimate of the number of loci controlling quantitative traits of interest*
  - *An estimate of the proportion of phenotypic variation explained/locus*
  - *An estimate of the effects of allelic substitution*
  - *The identity and putative function of each significantly associated gene*
  - *The SNP allele and haplotype frequencies in the population*
  - *The mechanism of gene action at each locus (additive, dominant)*
  - *The genetic markers that are either the causative mutation (QTN) or are in complete or nearly complete LD with the QTN*
  - *Verified associations in multiple populations (breeding populations)*

Genomic selection: An alternative to association genetics

Markers are used to infer kinship

- In genomic selection, markers are used to indicate the extent to which a progeny may be related to a favorable parent. That is, what proportion of the parent's genome is represented in the progeny?
- Requires as many or more SNP markers as association, but does not require association trials themselves (populations). Work is done directly within elite lineages

References cited

- Brown, G. R., D. L. Bassoni, G. P. Gill, J. R. Fontana, N. C. Wheeler, R. A. Megraw, M. F. Davis, M. M. Sewell, G. A. Tuskan, and D. B. Neale. 2003. Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda L)* III. QTL verification and candidate gene mapping. Genetics 164: 1537-1546.
- Eathington, S. R., T. M. Crosbie, M. D. Edwards, R. Reiter, and J. K. Bull. 2007. Molecular markers in a commercial breeding program. Crop Science 47: s154-s163. (Available online at: http://dx.doi.org/10.2135/cropsci2007.04.0015IPBS) (verified 2 June 2011).
- Gonzalez-Martinez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson, and D. B. Neale. 2007. Association Genetics in Pinus Taeda L I. Wood properties. Genetics 175: 399-409. (Available online at: http://dx.doi.org/10.1534/genetics.106.061127) (verified 2 June 2011).
- Gonzalez-Martinez, S. C., D. Huber, E. Ersoz, J. M. Davis, and D. B. Neale. 2008. Association genetics in Pinus taeda L. II. Carbon isotope discrimination. Heredity 101: 19-26. (Available online at: http://dx.doi.org/10.1038/hdy.2008.21) (verified 2 June 2011).
- Grattapaglia, D. 2007. Marker–assisted selection in Eucalyptus. p. 251-281. *In*. E. P. Guimaraes, J. Ruane, B. D. Scherf, A. Sonnino, and J. D. Dargie (ed.) Marker assisted selection: Current status and future perspectives in crops, livestock, forestry and fish. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.
- White, T. L., W. T. Adams, and D. B. Neale. 2007. Forest Genetics. CAB International, Oxfordshire, U.K. (Available online at: http://bookshop.cabi.org/?page=2633&pid=2043&site=191) (verified 2 June 2011).

Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UCDAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture