Популяционная геномика

- definition of population genomics and ecological genomics
- genome-wide nucleotide sequence analysis and genome-wide association studies (GWAS)

were already briefly presented

- neutral theory of molecular evolution
- neutrality tests (DNASP software)
- search for genes under selection using genome wide scans (search for F_{ST}-outliers using LOSITAN software)

Genomics studies genetic variation in multiple genes or an entire genome at different levels

- within individuals
- between individuals

Comparative genomics

- **within populations**
- **between populations**
- **over the entire set of populations**

Population genomics

- **between different taxa (species, genera, families, etc.)**

Evolutionary genomics, Phylogenomics

- **within & between different complex illusive communities**

Metagenomics

- **association of genetic variation with environment**

Ecogenomics

What is population genomics?

- studies genome-wide variation and links genotypes to phenotypes and environment via linking genomic variation with phenotypic variation and environmental variables

- provides detailed data and description of genome-wide nucleotide and allelic variation for all types of molecular genetic markers (SNPs, SSR, etc.) in numerous selectively neutral loci (supposedly non-coding regions) as well as adaptive trait related candidate genes in multiple individuals and populations

- identifies genes under selection via **neutrality tests**, **outliers**, etc., using **genome wide scans**

- links genotypes to phenotypes via **genome wide association studies (GWAS)**

(**Spatial** or **landscape genomics** is one of applications of **population genomics**).

Population genomics

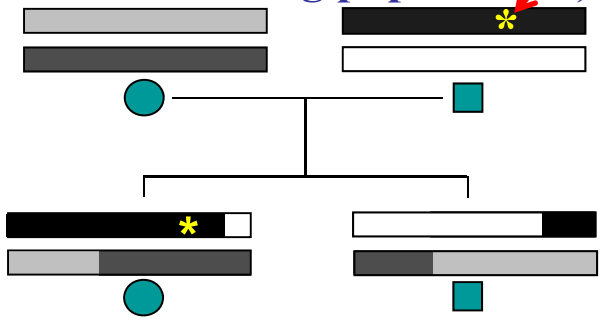
- **Genome-Wide Association Study** (GWA study, or **GWAS**), also known as Whole Genome Association Study (WGA study, or WGAS) or **Genome-Wide Association Mapping (GWAM)**
- TASSEL software for GWAS
- STRUCTURE software for inferring population structure
- LOSITAN software to search for F_{ST}-outliers

QTL or TDT mapping *versus* GWAM

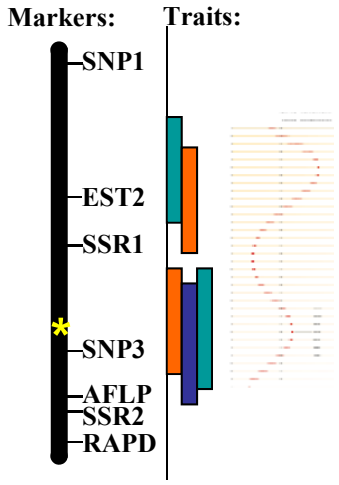
In both approaches association between marker and trait could be either due to the linkage to a causative gene (tighter in GWAM) or due to the causative gene itself

- **Quantitative Trait Loci (QTL) mapping in segregating progeny (exp. organisms), or**
- **Transmission Disequilibrium Test (TDT) in parent-offspring pedigrees (human & breeding populations)**

- Mapping progeny (experimental organisms)
- Mapping pedigree (human)
- LD is very high

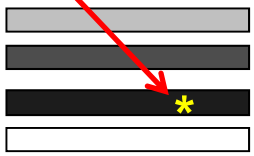


Linkage Group (LG)

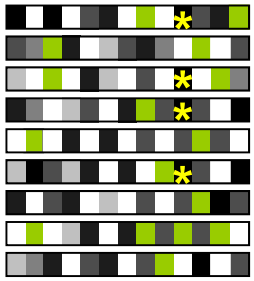


a favourable mutation

Genome-Wide Association Mapping (GWAM)



recombination in several generations



- LD is less than in QTL or TDT mapping

- Natural population (=multiple genetic backgrounds)
- **Problem**: selectively neutral population structure

Linkage Disequilibrium (LD) is a nonrandom association of alleles at linked loci

Use of candidate genes (functional markers) increases chances that association is due to causative gene

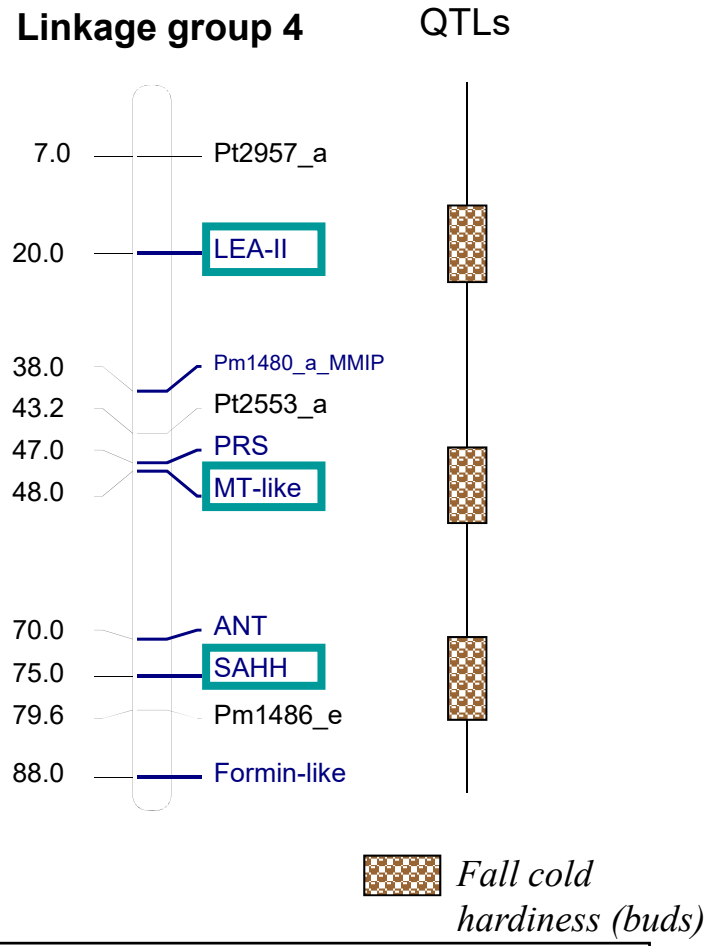
Collocation of adaptive trait related candidate genes with QTLs controlling adaptive traits:

Functional candidate genes:

LEA-II
(late embryogenesis abundant type II)
dehydrin-like protein
Induced by cold

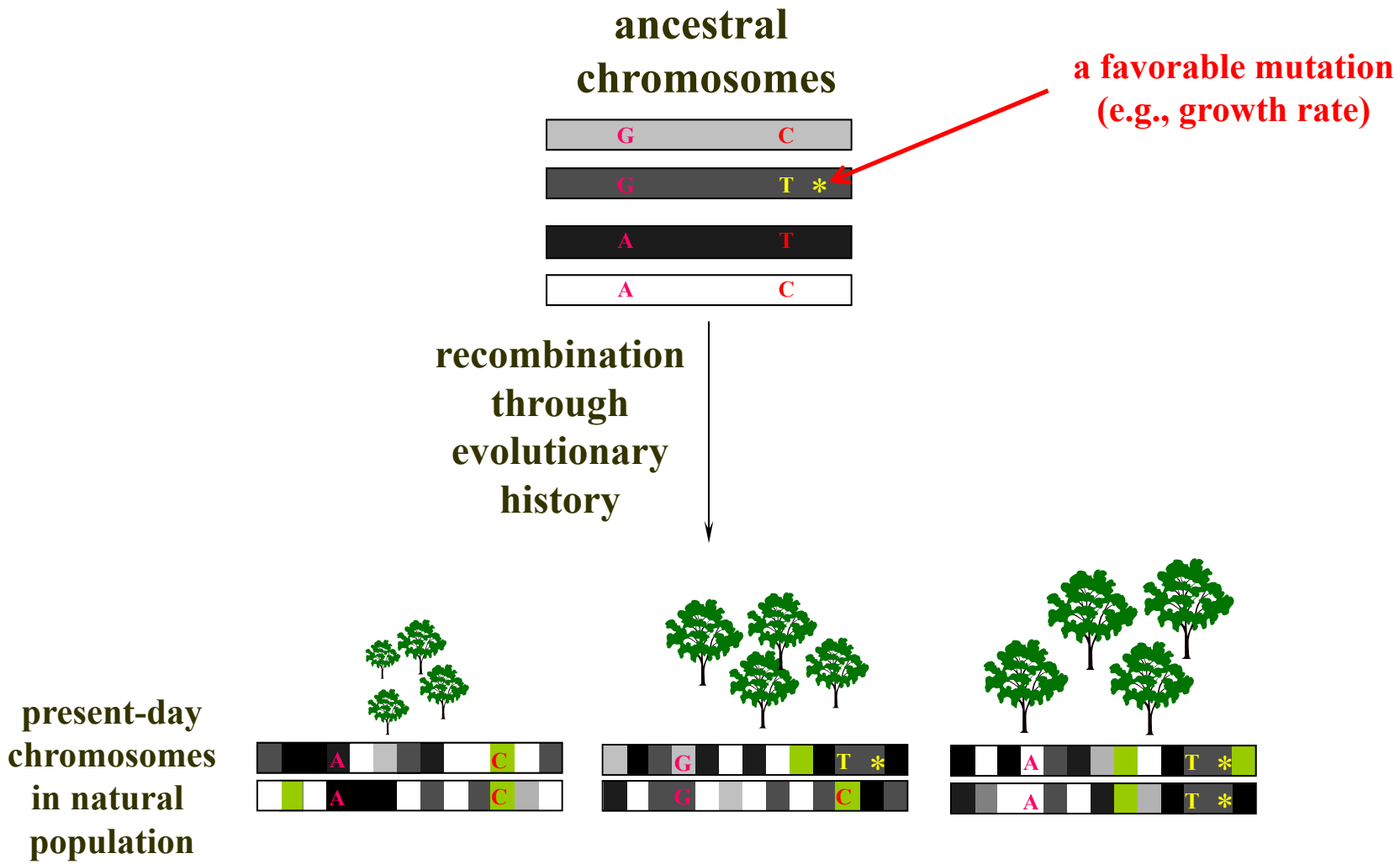
MT-like
(metallothionein-like protein)
stress-induced;
downregulated under water deficit

SAHH
(S-adenosyl-L-homocysteinas hydrolase)
upregulated under water deficit

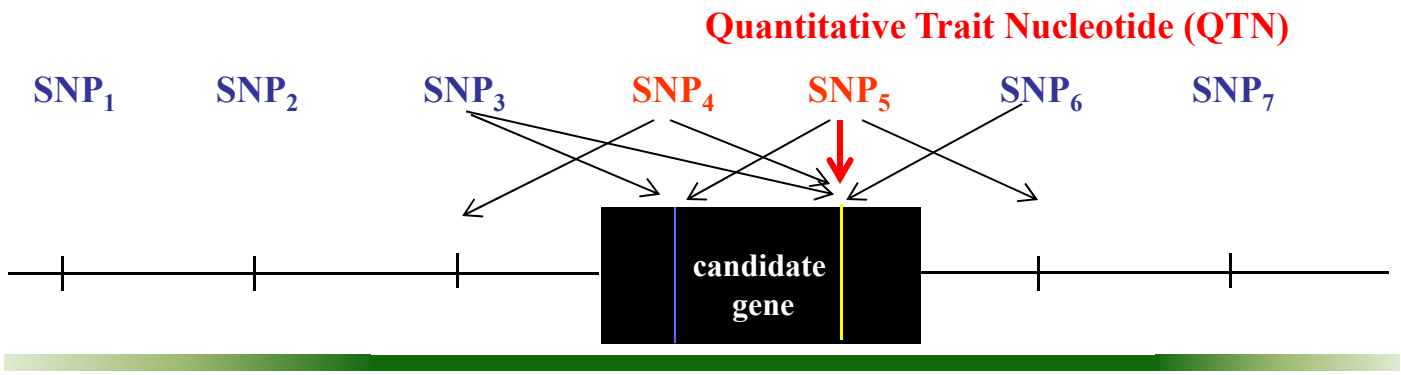


Wheeler N.C., Jermstad K.D., Krutovsky K.V. *et al.* 2005. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-Fir. IV. Cold-hardiness QTL verification and candidate gene mapping. *Molecular Breeding* 15: 145-156.

Genome-Wide Association Mapping using SNPs



Random markers based GWAM vs. Candidate gene based GWAM



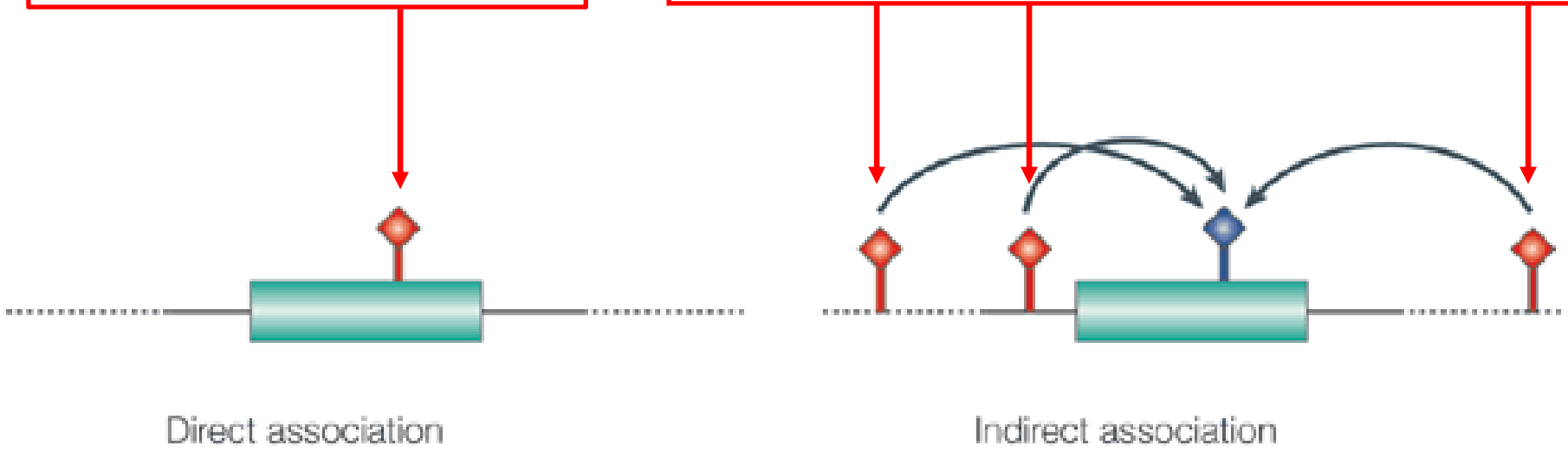
Extent of linkage disequilibrium

Linkage Disequilibrium (LD) is a nonrandom association of alleles at linked loci

SNP Association Testing: Direct or Indirect

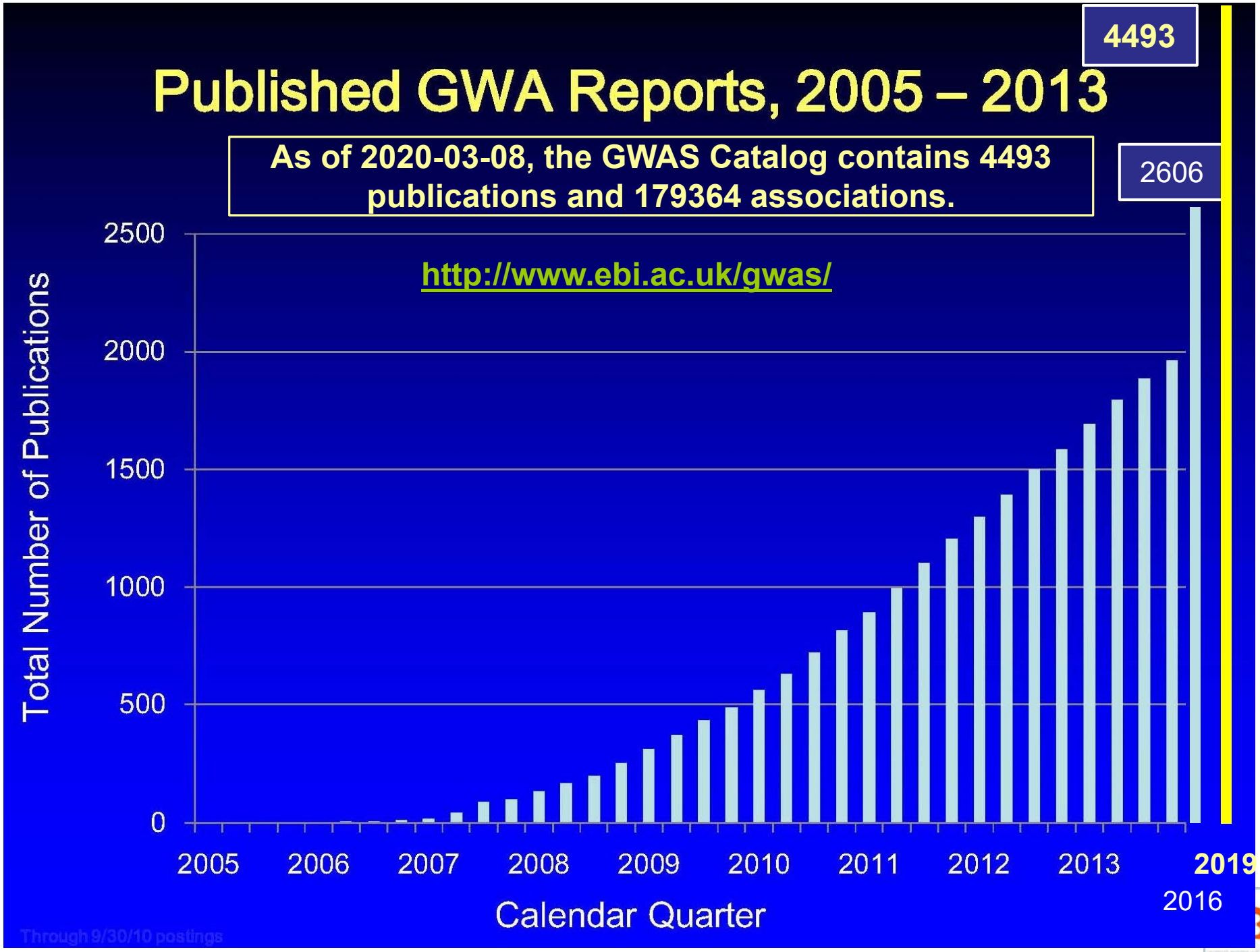
A candidate, functional SNP is directly tested for association with disease

SNPs to be genotyped are chosen randomly or on the basis of linkage disequilibrium patterns to provide as much information about other SNPs as possible



Nature Reviews | Genetics

2005; 6:95-108


Published GWA Reports, 2005 – 2013
4493
As of 2020-03-08, the GWAS Catalog contains 4493 publications and 179364 associations.
2606
http://www.ebi.ac.uk/gwas/
Total Number of Publications
2500
2000
1500
1000
500
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2019
2016
Calendar Quarter
Through 9/30/10 postings

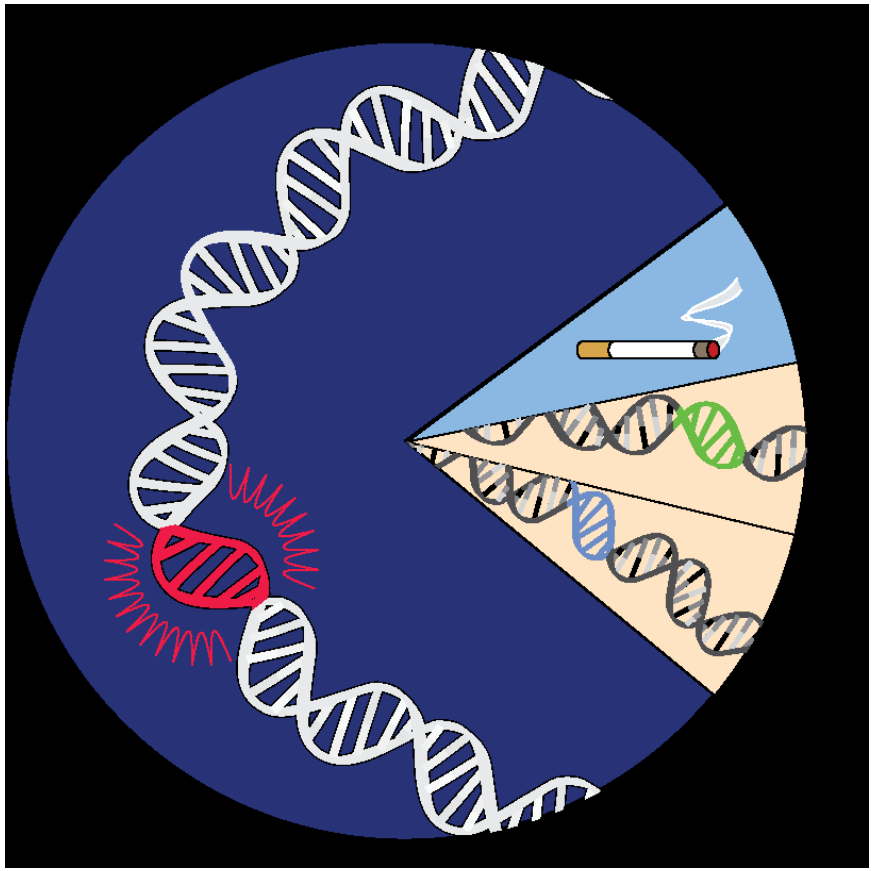
Association Mapping Components

- **Phenotypes**
  - trait values
- **Numerous Molecular Markers**
  - SNPs
    - ✓ SNP genotyping assays based on preselected SNPs (now)
    - ✓ SNP genotyping by sequencing (future)
- **Statistical Models**
  - Linear model: phenotype as response and genotype as predictor

To search for association due to linkage strong linkage disequilibrium (LD) is required between causal factor and SNP(s)

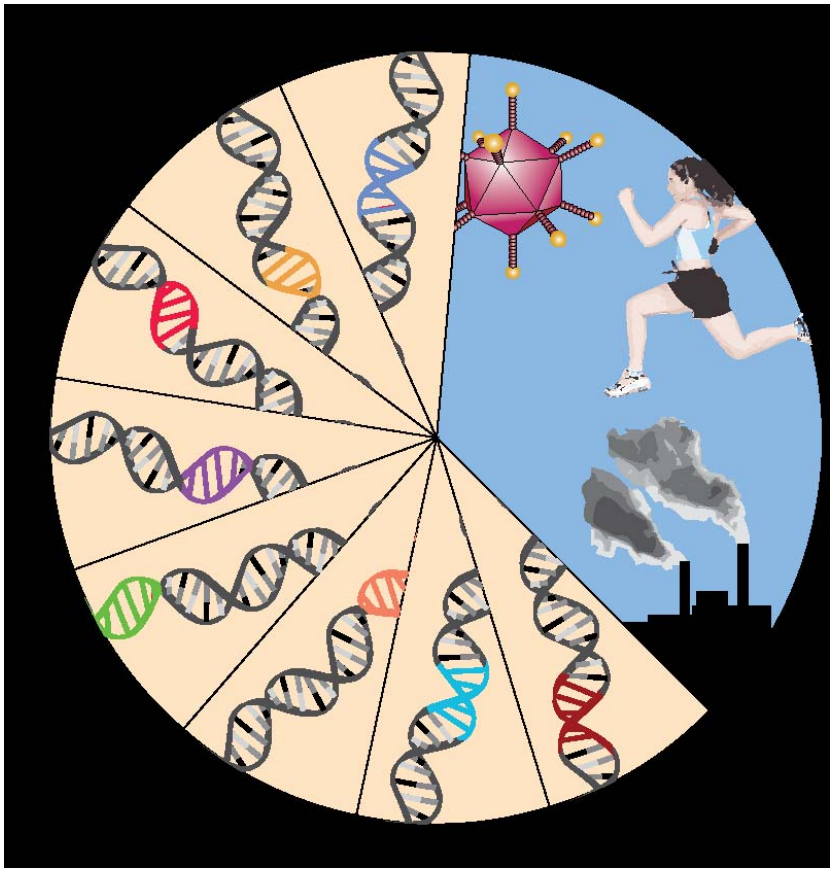
- Detection power heavily rely upon either existence of strong LD between SNP marker(s) and casual factor, or causal factor itself
- Increasing marker density has contributed to increased chance of including a marker, which is in strong LD with the disease allele
- But number of individuals sequenced often is still insufficient
- Negative results do not exclude possibility of a significant gene effect in that region. The SNP marker(s) may be in modest LD or no LD with the causal factor, and the casual variant(s) will not be detected.

Genomic Architecture of Genetic Diseases



- **monogenic, Mendelian…**
- **simple**
- **rare**
- **mostly *protein coding* mutations**

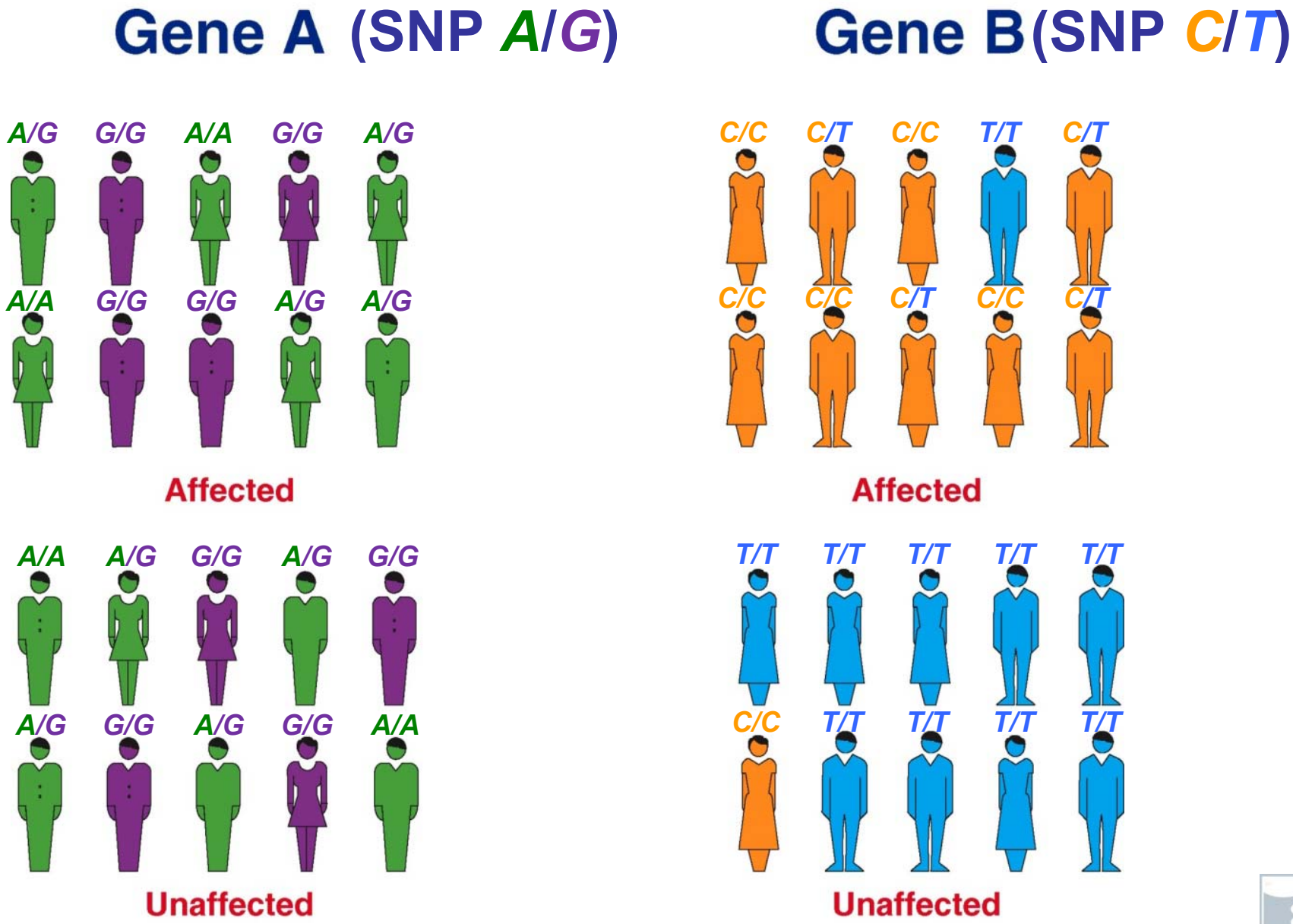
e.g., Duchenne muscular dystrophy, phenylketonuria, sickle cell disease



- **multigenic, non-Mendelian…**
- **complex**
- **common**
- **mostly *regulatory* mutations**

e.g., heart disease, diabetes, schizophrenia, some cancers

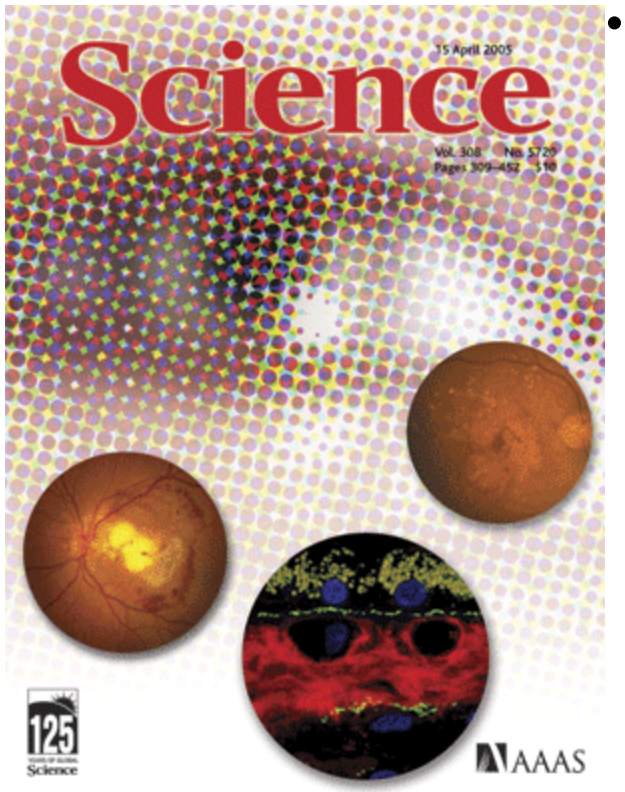
Genome-Wide Association Mapping (GWAS)



The First GWAM Success Story:
Возрастная макулодистрофия

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7†] ***Science* (2005)**



- Because of high costs, initial high-density screens are often conducted on a few hundred cases and controls
  - Age-Related Macular Degeneration: 96 cases & 50 controls, 105,980 SNPs analyzed (Klein et al. 2005 Science 308: 385-389)
  - Breast Cancer: 1st stage - 390 invasive breast cancer cases & 364 controls, 227,876 SNPs; 2nd stage - 3,990 cases & 3,916 controls, 12,711 SNPs; 3rd stage - 21,860 cases & 22,578 controls, 30 SNPs (Easton et al. 2007 Nature 447: 1087-1095)
  - Coronary Heart Disease: 1st stage - 322 cases & 312 controls, 100,000 SNPs; 2nd stage - an independent sample of 311 cases & 326 controls, 2586 SNPs; 3rd stage – 1,347 cases & 9,054 controls, 50 SNPs; plus additional validation (McPherson et al. 2007 Science 316: 1488-1491)

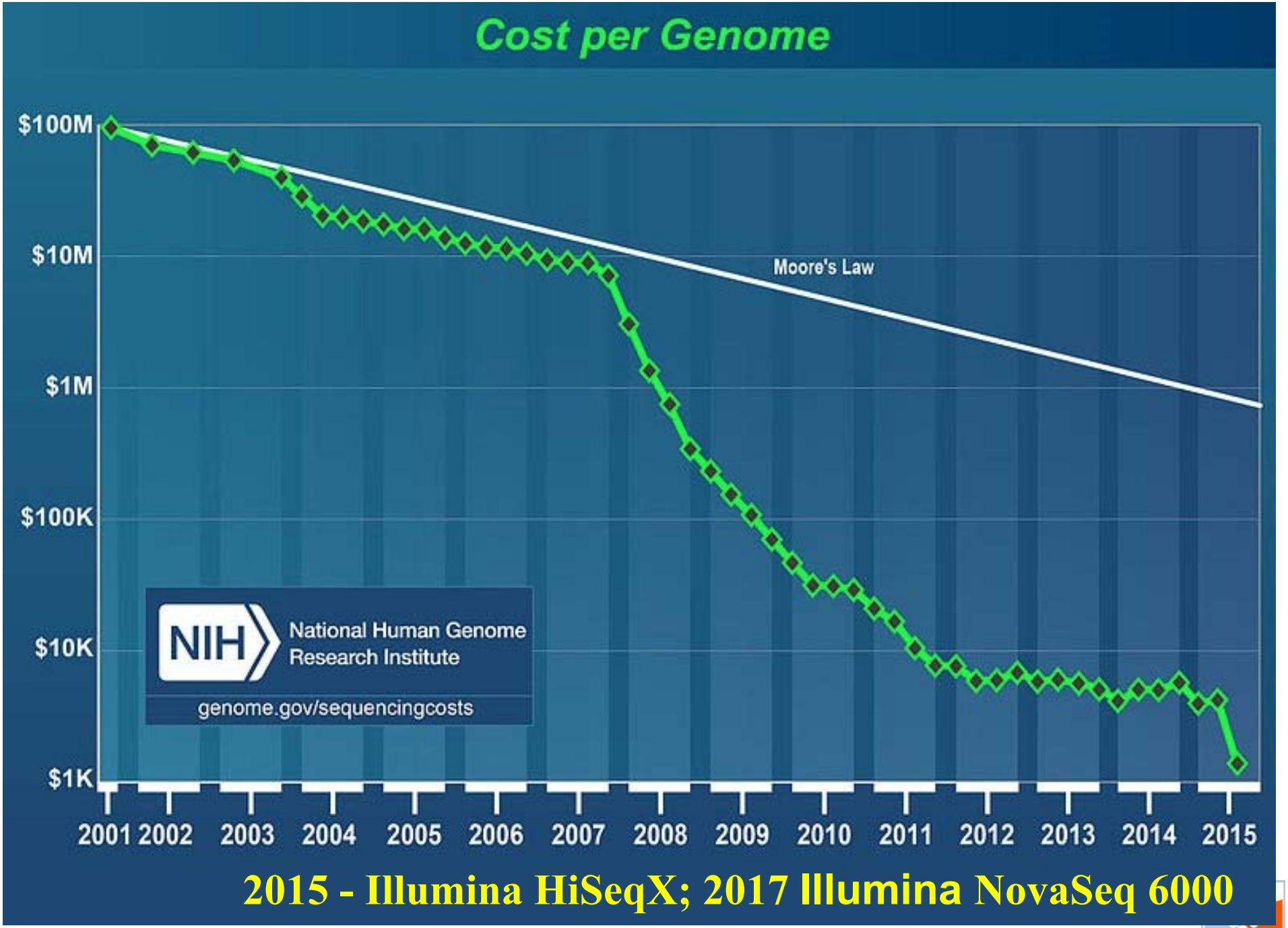
Lack of power in the initial GWA screening step

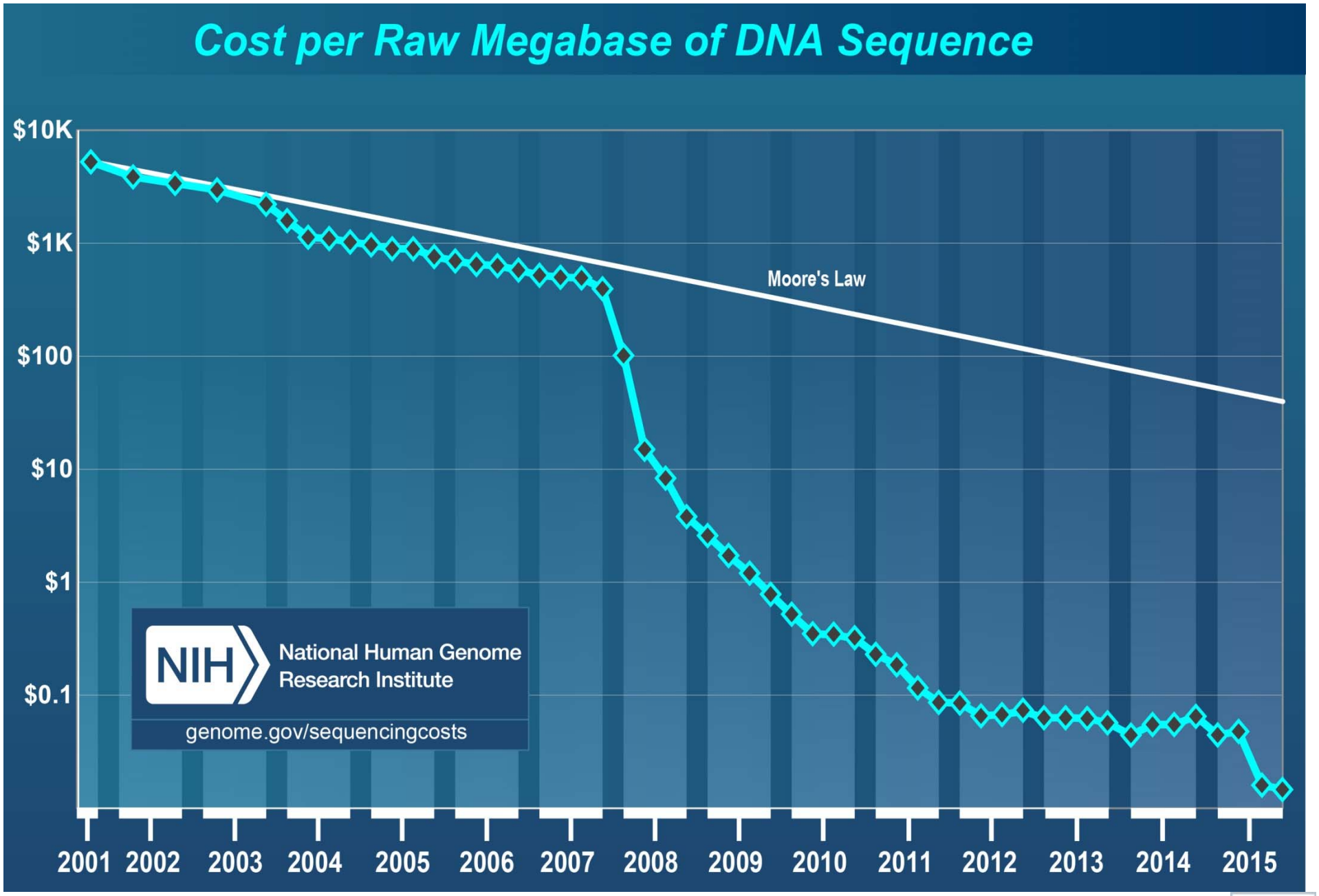
- Most studies in initial screen have insufficient power in high-density panels (Shriner et al. Problems with Genome-Wide Association Studies. Science 2007, 316: 1840-1842)

- Replication step is required that can involve 1000's of individuals
  - For example, 38,759 individuals were involved in 13 obesity studies (Frayling et al. A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science 2007, 316: 889-894)

- But power of GWAM to detect small-to-modest effect variants depends on initial screening step not on replication step

- Insufficient power in initial screen results often in testing false positive markers and not testing false negative markers in replication step


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
2015 - Illumina HiSeqX; 2017 Illumina NovaSeq 6000


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

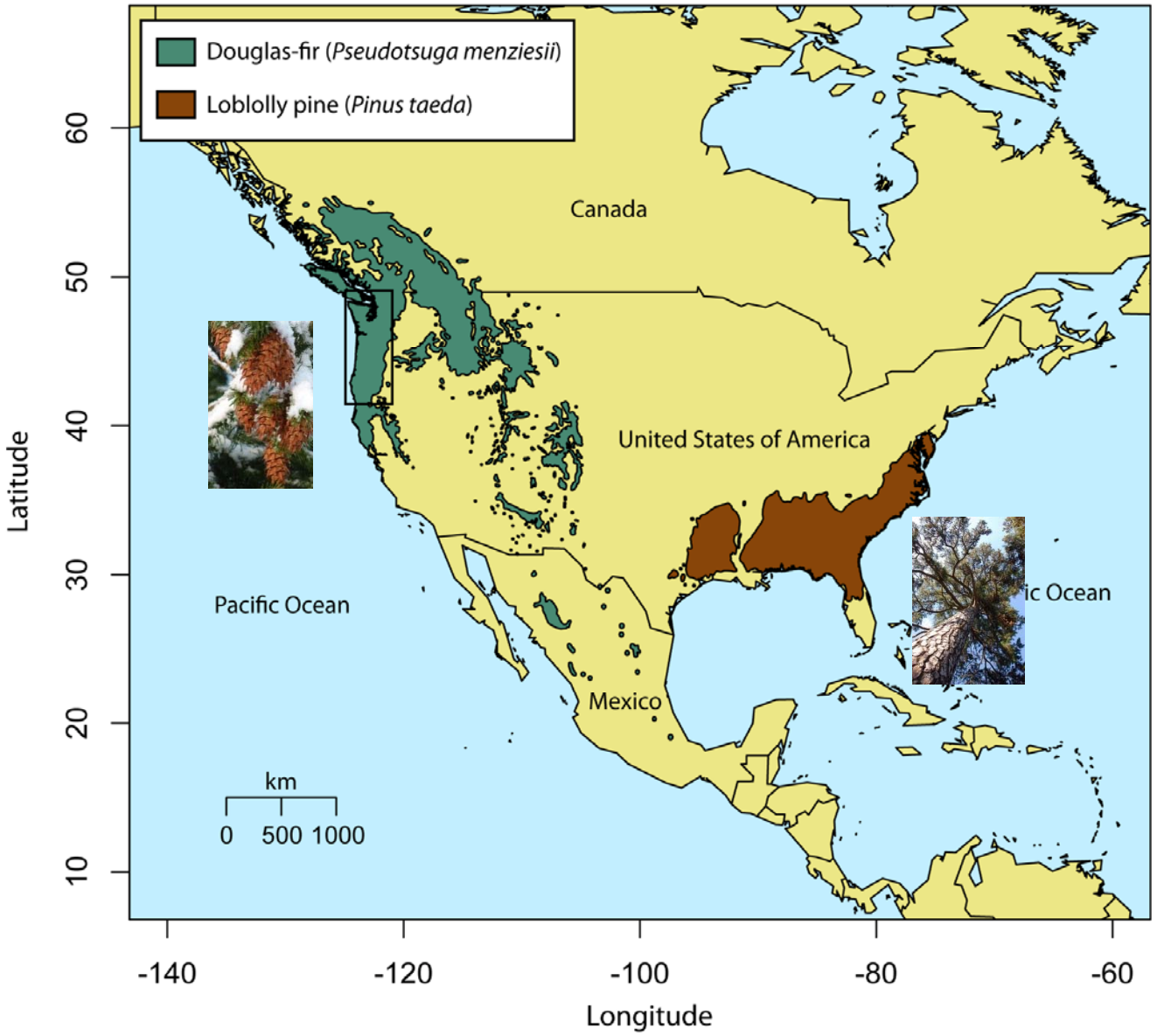
Douglas-fir and Loblolly pine case studies

- United States Department of Agriculture (USDA) National Research Initiative Competitive Grant Program (NRICGP), Plant Genome, Bioinformatics, and Genetic Resources, # 2004-35300-14670, PI: D. Neale, CoPIs: K. Krutovsky, G. Howe, and J.B. St. Clair, 3 years, 2004-2007, $ 490,000, "Association mapping of adaptive traits in Douglas-fir (*Pseudotsuga menziesii* Mirb. Franco)" (https://nealelab.ucdavis.edu/dfpg).

- USDA NRICGP Plant Genome Program / National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) Competitive Grants Program, Applied Plant Genetics Coordinated Agricultural Project (CAP), #CA-D-PLS-2038-CG, Project Directors: D. Neale, T. Byram, D. Harry, G. Howe, D. Huber, F. Isik, S. McKeand, C. Nelson, J.B. St. Clair, J. Wegrzyn, N. Wheeler, R. Whetten and others, 2004-2011, $6,000,000; “Conifer Translational Genomics Network” (http://dendrome.ucdavis.edu/ctgn).

- USDA NIFA AFRI Competitive Grants Program, CAP, Climate Change Program 1: Regional Approaches to Climate Change, Program Area Code – A3101, #2011-68002-30185, PI: Timothy Martin, CoPIs: R. Abt, D. Adams, G. Boyd, R. Boyles, H. Burkhart, T. Byram, D. Carter, W. Cropper, F. Cubbage, J. Davis, J.-C. Domec, T. Fox, J. Gan, D. Grebner, S. Grunwald, T. Hennessey, J. Holliday, W. Hubbard, D. Huber, J. Idassi, F. Isik, K. Johnsen, E. Jokela, J. Jones, M. Kane, J. King, K. Krutovsky, C. Loopstra, D. Markewitz, S. McKeand, S. McNulty, M. Megalos, M. Monroe, C. D. Nelson, A. Noormets, G. Peter, G. Powell, R. Rubilar, L. Samuelson, J. Seiler, S. Sriharan, J. Stape, B. Strahm, G. Sun, E. Taylor, R. Teskey, J. Vogel, R. Whetten, R. Will, D. Wilson, R. Wynne, 5 years, 3/1/2011-2/28/2016, $19,976,825; “Integrating research, education and extension for enhancing southern pine climate change mitigation and adaptation” (http://pinemap.org).

Douglas-fir and Loblolly pine species



Identification of Genes in Douglas-fir and Loblolly pine

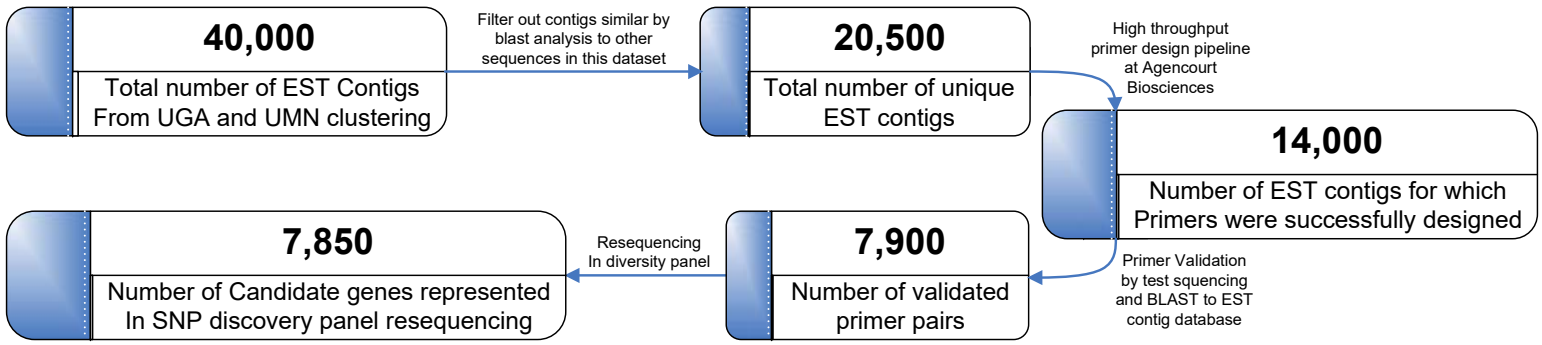
EST based unigene approach to design PCR primers for amplicon based Sanger sequencing of multiple genes:

Candidate gene approach in Douglas-fir:

- Candidate genes for *Arabidopsis*: **939**
- Douglas-fir ESTs selected from BLASTx against candidate genes in *Arabidopsis* with scores < E-10: **553**
- PCR primers designed (automated and manual): **378**
- Finally selected genes: **121**
- SNPs found: **933**
- SNPs genotyped in a range-wide association population: **384**

Eckert, Wegrzyn, Pande, Jermstad, Lee, Liechty, Tearse, Krutovsky & Neale 2009 *Genetics* 183: 289-298

Genome-wide approach in Loblolly pine:



- **7216** SNPs were finally selected out of total ~23000 SNPs to design Illumina Infinium SNP genotyping assay and were genotyped in multiple association mapping and breeding population (>4500 trees)

SNP genotyping in populations

	Douglas-fir 	Loblolly pine 
Number of trees	**700**	**4500**
Sequencing technique for SNP discovery	Sanger, amplicon based	Sanger, amplicon based
SNPs selected	384 (out of 933)	7216 (out of ~23000)
SNA genotyping technique	Illumina GoldenGate	Illumina Infinium
SNPs genotyped	280	5,379


Bud flush
Propensity to lammas flush
Length of lammas flush
ADAPTIVE TRAITS IN DOUGLAS-FIR
Cold damage to buds, needles and stems
Bud set
Drought

Douglas-fir Genome Project (DFGP)

DFGP
Douglas-fir Genome Project

Home | Overview | Members | Publications | Neale Lab

Research
- Overview
- Principle Investigator

Research Projects
- ACE-SAP
- ADEPT
- ADEPT 2
- CCGP
- CRSP
- Dendrome
- DFGP
  - Agenda 2020
  - Adapt
- EGFHN
- LPGP
- PBGP
- SSGP
- WHISP
- WPGP

Research Publications
- Publications
- Meeting Abstracts

Research Staff
- Current Staff
- Alumni

Research Facilities
- UCDavis
- UCD Department of Plant Sciences
- Institute of Forest Genetics
- Visiting UCDavis

Contact
- Contact

Research Education
- Genetics Graduate Group
- UC Davis Plant Breeding
- Plant Biology Graduate Group
- Population Biology Graduate Group
- Horticulture and Agronomy Graduate Group

collaborators

David Neale
University of California, Davis
Department of Plant Sciences
One Shields Avenue
Davis, CA 95616
Phone: (530) 754-8431
Fax: (530) 754-9366
e-mail: dbneale@ucdavis.edu

Kathleen D. Jermstad
Institute of Forest Genetics
USDA Forest Service - Pacific Southwest Research Station
2480 Carson Road
Placerville, CA 95667
phone: (530) 622-1225
Fax: (530) 622-2633
e-mail: kdjermstad@ucdavis.edu

Konstantin (Kostya) V. Krutovsky
Department of Forest Science - Texas A&M University
2135 Tamu
College Station, TX 77843-2135
Phone: (979) 458-1417
Fax: (979) 458-0159
e-mail: k-krutovsky@tamu.edu

Glenn Howe
Department of Forest Science
321 Richardson Hall
Oregon State University
Corvallis, OR 97331-5752
Tel: (541) 737-9001
Fax: (541) 737-1393
e-mail: Glenn.Howe@oregonstate.edu

Marilyn Cherry
Department of Forest Science
321 Richardson Hall
Oregon State University
Corvallis, OR 97331-5752
Tel: (541) 737-6579
Fax: (541) 737-1393
e-mail: Marilyn.Cherry@oregonstate.edu

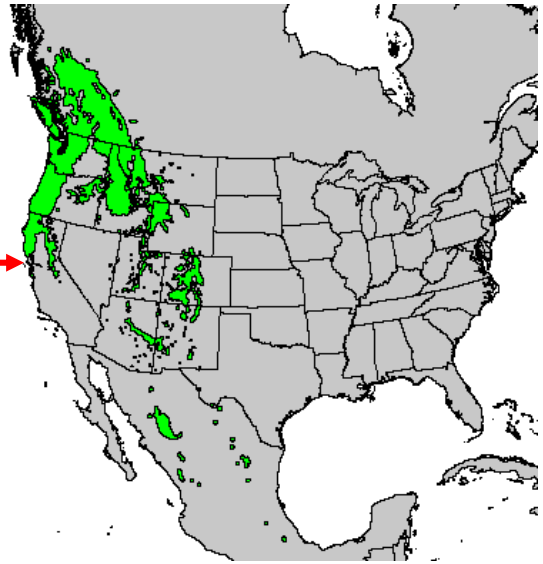
Brad St Clair
USDA Forest Service - Pacific Northwest Research Station
3200 SW Jefferson Way
Corvallis, OR 97331-4401
Phone: (541) 750-7294
Fax: (541) 750-7329
e-mail: BSTCLAIR@fs.fed.us

Sally Aitken
Department of Forest Sciences - Faculty of Forestry
The University of British Columbia
3rd Floor, Forest Sciences Centre #3041 - 2424 Main Mall
Vancouver, British Columbia
Canada V6T 1Z4
Tel: (604) 822-6020
Fax: (604) 822-9102
e-mail: aitken@interchg.ubc.ca

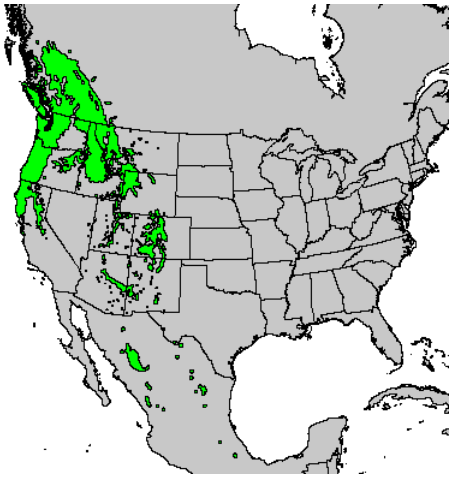
Nicholas C. Wheeler
Molecular Tree Breeding Services, LLC
21040 Flumerfelt Rd SE
Centralia, WA 98531
Phone: 360-278-3535
Fax: 360-278-3092
e-mail: nickwheeler@scattercreek.com

David Harry
Genetic Foundations
PO Box 1019
Philomath, OR 97370
Phone: (541) 740-7444
e-mail: deharry@peak.org
www.geneticfoundations.com

University of California, Davis



Douglas-fir Genome Project (DFGP)



Neale Lab

Home | Education | Staff | TGG | Research | Publication | Research Centers | Location

Summary of DFGP

Over the past decade we have used quantitative trait locus (QTL) mapping to dissect adaptive and economic traits in Douglas-fir, loblolly pine and poplar. Though highly informative, these studies do not reveal which genes are responsible for controlling traits of interest.

In this study we propose to use a population genomic approach called association mapping to identify the specific genes (i.e. loci and alleles) that are responsible for phenotypic differences in adaptive traits of Douglas-fir.

More

Resources

- Dendrome
- IUFRO Calendar of Events
- UCD Department of Plant Sciences
- Institute of Forest Genetics
- Directory of Forest Geneticists
- Adapt
- Agenda2020
- Douglas-fir QTL mapping

Research

- Genetic Mapping
- cDNA sequencing
- QTL mapping
- Pedigree

Members

- Members area

Partcipating Organizations

UCDAVIS PLANT SCIENCES | OSU Oregon State | USFS

Upcoming Events

June 28-30, 2010
High Five Conference: The Future of High-Elevation Five-Needle White Pines in Western North America
Missoula, MT

More information

July 1, 2010
2010 WPBR Workshop
Missoula, MT

More information

July 4-9, 2010
FESPB 2010: XVII Congress of the Federation of European Societies of Plant Biology (FESPB)
Conference Center (Palacio de Congresos), Valencia, Spain

More information

Latest Publications

Eckert A.J., Liechty J.D., Tearse B.R., Pande B., Neale D.B. (2010) DnaSAM: Software to perform neutrality testing for large datasets with complex null models. Molecular Ecology Resources. 10:542-545. Full Text. Supplemental.

Krutovsky K.V., Saich R., Hipkins V.D., Neale D.B. (2009) Estimation of population structure in coastal Douglas-fir [Pseudotsuga menziesii (Mirb.) Franco var. menziesii] using allozyme and microsatellite markers. Tree Genetics and Genomes. 5:641-658. Full Text.

Eckert A.J., Bower A.D., Wegrzyn J.L., Pande B., Jermstad K.D., Krutovsky K.V., St Claire., Neale D.B. (2009) Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits. Genetics. 182:1289-1302. Full Text.

https://nealelab.ucdavis.edu/dfpg

DFGP Members : Consortium Members

http://dendrome.ucdavis.edu/NealeLab/dfgp/members.php

- Collaboriters
- Member Area

David Neale

University of California, Davis
Department of Plant Sciences
One Shields Avenue
Davis, CA 95616
Phone: (530) 754-8431
Fax: (530) 754-9366
e-mail: dbneale@ucdavis.edu

Kathleen D. Jermstad

Institute of Forest Genetics
USDA Forest Service - Pacific Southwest Research Station
2480 Carson Road
Placerville, CA 95667
phone: (530) 622-1225
Fax: (530) 622-2633
e-mail: kdjermstad@ucdavis.edu

Andrew Eckert

University of California, Davis
Department of Evolution and Ecology
One Shields Avenue
Davis, CA 95616
Phone: (530) 754-5743
Fax: (530) 754-9366
e-mail: ajeckert@ucdavis.edu

Konstantin (Kostya) V. Krutovsky

Department of Forest Science - Texas A&M University
2135 Tamu
College Station, TX 77843-2135
Phone: (979) 458-1417
Fax: (979) 458-0159
e-mail: k-krutovsky@tamu.edu

Glenn Howe

Department of Forest Science
321 Richardson Hall
Oregon State University
Corvallis, OR 97331-5752
Tel: (541) 737-9001
Fax: (541) 737-1393
e-mail: Glenn.Howe@oregonstate.edu

Marilyn Cherry

Department of Forest Science
321 Richardson Hall
Oregon State University
Corvallis, OR 97331-5752
Tel: (541) 737-6579
Fax: (541) 737-1393
e-mail: Marilyn.Cherry@oregonstate.edu

Acknowledgements

Brad St Clair

USDA Forest Service - Pacific Northwest Research Station
3200 SW Jefferson Way
Corvallis, OR 97331-4401
Phone: (541) 750-7294
Fax: (541) 750-7329
e-mail: BSTCLAIR@fs.fed.us

Sally Aitken

Department of Forest Sciences - Faculty of Forestry
The University of British Columbia
3rd Floor, Forest Sciences Centre #3041 - 2424 Main Mall
Vancouver, British Columbia
Canada V6T 1Z4
Tel: (604) 822-6020
Fax: (604) 822-9102
e-mail: aitken@interchg.ubc.ca

Nicholas C. Wheeler

Molecular Tree Breeding Services, LLC
21040 Flumerfelt Rd SE
Centralia, WA 98531
Phone: 360-278-3535
Fax: 360-278-3092
e-mail: nickwheeler@scattercreek.com

David Harry

Genetic Foundations
PO Box 1019
Philomath, OR 97370
Phone: (541) 740-7444
e-mail: deharry@peak.org
www.geneticfoundations.com

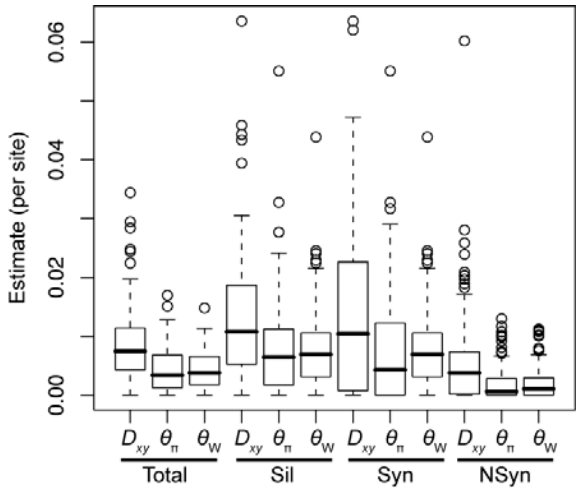
Valerie Hipkins
National Forest Genetics Laboratory
Institute of Forest Genetics, USDA Forest Service
Placerville, CA 95667
530-622-1609 (phone)
vhipkins@fs.fed.us

Population Genomics Study of Adaptation in Douglas-fir

- **Gene discovery**: Selection of candidate genes for adaptive traits using expressed sequence tags (ESTs) and unigenes (**transcriptomics**)
- **Allele and SNP discovery**: Direct sequencing and nucleotide variation analysis in populations
- **Discovery alleles under selection**:
  - QTL mapping using candidate genes
  - Neutrality tests
  - Association mapping ("**Linking Genotypes to Phenotypes**")
  - F_{ST}-outliers
  - Alleles with clinal variation and associated with environmental variables
- **Verification of associations** in clonally replicated populations

Nucleotide Diversity in Douglas-fir

- 121 candidate genes were sequenced in
- 24 trees (sampled in Oregon & Washington)
- 933 SNPs were used for inferences of:
  (1) selection by neutrality tests
  (2) genotype-phenotype correlations



SNP type	#SNPs	θ	π	D_{xy}
All	933	0.005	0.004	0.009
Silent	732	0.008	0.008	0.013
Syn	254	0.008	0.008	0.014
Nonsyn	201	0.002	0.002	0.006

Eckert, Wegrzyn, Pande, Jermstad, Lee, Liechty, Tearse, Krutovsky & Neale 2009 *Genetics* 183: 289-298

Neutrality tests

- Application of several different tests for neutrality, including those that incorporated demographic models (instantaneous growth and two bottleneck models), revealed signatures of selection consistent with selective sweeps at three to eight loci, depending upon the severity of a bottleneck event and the method used to detect selection.
- *GRAM-containing/ABA-responsive protein, cold-regulated plasma membrane protein, dehydrin-like protein, lumenal binding protein* (compound test that included Tajima's *D* based on different demographic models, Fay and Wu's normalized *H*, and the Ewens–Watterson test of neutrality to detect positive selection);
- *cyclosporin A-binding protein, GRAM-containing/ABA-responsive protein, transcription regulation protein* (polymorphism-to-divergence Hudson–Kreitman–Aguade (HKA) test);
- *thaumatin-like protein, LRR receptor-like protein kinase* (synonymous-to-nonsynonymous divergence, K_a/K_s).
- Putative homologs in Arabidopsis act primarily to stabilize the plasma membrane and protect against denaturation of proteins at freezing temperatures.

Eckert, Wegrzyn, Pande, Jermstad, Lee, Liechty, Tearse, Krutovsky & Neale 2009 *Genetics* 183: 289-298

Neutral Theory of molecular evolution

- provides quantitative predictions for levels of variation within and between species
- the null hypothesis for examining the amount and pattern of molecular genetic variation
- sequence data can be used to determine whether the patterns of molecular genetic variation is consistent with the neutral theory (**Neutrality Tests!**)

Neutral Theory: Neutrality tests

- Ewens-Watterson Test
- Tajima Test
- Hudson-Kreitman-Aguade (HKA) Test
- McDonald-Kreitman (MK) Test
- Synonymous/Nonsynonymous Ratio Test

Ewens-Watterson homozygosity test of neutrality

- **The Ewens-Watterson homozygosity test of neutrality** (Ewens 1972; Watterson 1978) compares the homozygosity expected from the experimental data assuming HW vs. the equilibrium homozygosity expected under random neutral mutations and genetic drift (neutrality) for a given sample size and observed number of alleles
- The homozygosity expected under Hardy-Weinberg proportions (f_e) is computed from experimental data as the sum of the squares of allele frequencies $f_e = \sum \hat{p}_i^2$
- The homozygosity expected under neutrality (f_{eq}), for the same sample size ($2N$) and number of unique alleles (k), is obtained by simulation $f_{eq} = \frac{1}{4N_e u + 1}$
- The *normalized deviate of the homozygosity* is the difference between f_e and f_{eq} divided by the square root of the variance of the expected homozygosity under neutrality (also obtained by simulations; Salamon et al. 1999)
- This test can tell us whether there is:
  - **no selection (HW homozygosity ≈ homozygosity expected under neutrality),**
  - **balancing selection (HW homozygosity < homozygosity expected under neutrality)** or
  - **directional selection (HW homozygosity > homozygosity expected under neutrality)** operating on a particular locus across populations.
- The standard current implementation of the test uses a Monte-Carlo implementation of the exact test written by Slatkin (Slatkin 1994; Slatkin 1996). A Markov-Chain Monte Carlo method is used to obtain the probability of the homozygosity under neutrality. The *p*-value is the probability of the observed homozygosity under the null hypothesis of neutrality, and can be interpreted as a two-tailed test.

Ewens-Watterson homozygosity test of neutrality

Markow *et al.* (1993) examined the distribution of two HLA loci in the Havasupai, a small tribe (< 600) that inhabits an isolated side canyon of the Grand Canyon in Arizona. In the sample of 122 individuals ($2N = 244$), for *HLA-A* and *HLA-B* three (see Parham *et al.*, 1997) and eight different alleles, respectively, were observed (Table 8.6). For both loci, the distribution of alleles was more even than expected from neutrality,

TABLE 8.6 Allele frequencies, the expected Hardy–Weinberg and equilibrium homozygosity, and the statistical significance level for two *HLA* loci in the Havasupai (Markow *et al.*, 1993; Parham *et al.*, 1997).

HLA-A (human leukocyte antigen loci)		*HLA-B*	
Allele	*Frequency*	*Allele*	*Frequency*
A2	0.545	B5v	0.119
A24	0.184	B27	0.037
A31	0.270	B35	0.164
		B39	0.061
		B48	0.422
		B51	0.086
		B60	0.102
		B61	0.008
$f_e = 0.404$		$f_e = 0.242$	
$f_{eq} = 0.714$		$f_{eq} = 0.611$	
$P = 0.04$		$P = 0.11$	

$f_e < f_{eq}$

Ewens-Watterson homozygosity test of neutrality

- Be aware that a bottleneck generally reduces the number of alleles, particularly rare ones, faster than the amount of heterozygosity
- As a result, testing of a recently bottlenecked population can give a pattern similar to a population with balancing selection, that is, given the number of alleles and the sample size, there could be a more even distribution of allele frequencies than would be expected by the neutral theory
- On the other hand, a recent selective sweep can give a very uneven allele frequency distribution because of the replacement of a large proportion of the population variation by the positively selected mutant, similar to that expected for purifying selection
- It is always a good idea to use neutral or genome-wide markers such as microsatellites or synonymous SNPs (genotyped in the same individuals as other markers) as a control in the tests

Tajima's test of neutrality

From theoretical neutral molecular evolution model nucleotide diversity $\theta = 4N_e u$, where N_e – effective population size, and u - the mutation rate per nucleotide (bp) per generation

There are two estimators for the parameter θ, or two fundamental types of nucleotide diversity statistics:

(1) one is based on the **observed** numbers of differences between sequences. This statistic is called π (or "pi", based on '*p*' for *p*airwise?)

$$\pi = \frac{1}{n(n-1)} \sum_{ij} p_{ij}$$

where p_{ij} (or π_{ij} is the proportion of different nucleotides **observed** between the *i*th and *j*th variant and n is the total number of sequence comparisons (Nei 1987)

2) another is θ_S (theta) based on the number of segregating sites (S), which is the number of polymorphic sites. Assuming an infinite allele model of evolution it is **expected** and calculated by the formula:

$$\theta = \frac{S}{\alpha_1} \text{ , where } \quad a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

Nucleotide sequence variation

- **Example:** Assume that the following nucleotide sequences were found in four individuals:

  ACTTGTCCTA
  AGTTCTCCTA
  TCTTGACGTA
  AGTTGACCTA

- **Expected nucleotide polymorphism or diversity:** $\theta = \frac{S}{\alpha_1}$

number of segregating substitution sites divided by sequence length ***S***=5/10=0.5

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} = 1.833 \qquad \theta = \frac{S}{\alpha_1} = \frac{0.5}{1.833} = 0.273$$

- **Observed nucleotide polymorphism or diversity:**

seq#	1	2	3	4
1		2/10	3/10	2/10
2	2/10		5/10	2/10
3	3/10	5/10		3/10
4	2/10	2/10	3/10	

$$\pi = \frac{1}{n(n-1)} \sum_{ij} \pi_{ij}$$

$\boldsymbol{\pi}$ =(1/12)$\sum\pi_{ij}$ = 0.083(0.2+0.3+0.2+0.2+0.5+0.2+0.3+0.5+0.3+0.2+0.2+0.3) = 0.083×3.4 = **0.283** $D = (\pi - \theta_S)/\sqrt{V_d} = 0.372$ $P > 0.10$

Tajima's *D* test of neutrality

- Developed by a Japanese researcher Fumio Tajima (1989)
- It compares the observed nucleotide diversity (π) with the measure of nucleotide diversity based on the number of segregating sites (S) per site (θ_S). (A site is considered segregating if there are two or more nucleotides at that site in a comparison of m sequences.)
- The observed nucleotide diversity π is defined as the average number of nucleotide differences per site between two sequences, but under the IAM and neutrality the expected nucleotide diversity can be estimated as $\theta_S = S/\alpha_1$
- π and θ_S reflect different types of information, but they are expected to be equal under the neutral theory
- π is affected mainly by the polymorphic alleles, while θ_S counts all variable sites equally and can be strongly influenced by rare alleles
- Fumio Tajima (1989) suggested a test that compares the difference $d = \pi - \theta_S$ between these estimates: $D = (\pi - \theta_S)/\sqrt{V_d} = d/\sqrt{V_d}$
- If the population is at neutral equilibrium, then Tajima's $D = 0$
- Rare alleles would not affect π, but would inflate θ_S (negative D)
- Overdominance would increase π, but reduce θ_S (positive D)

Tajima's test of neutrality

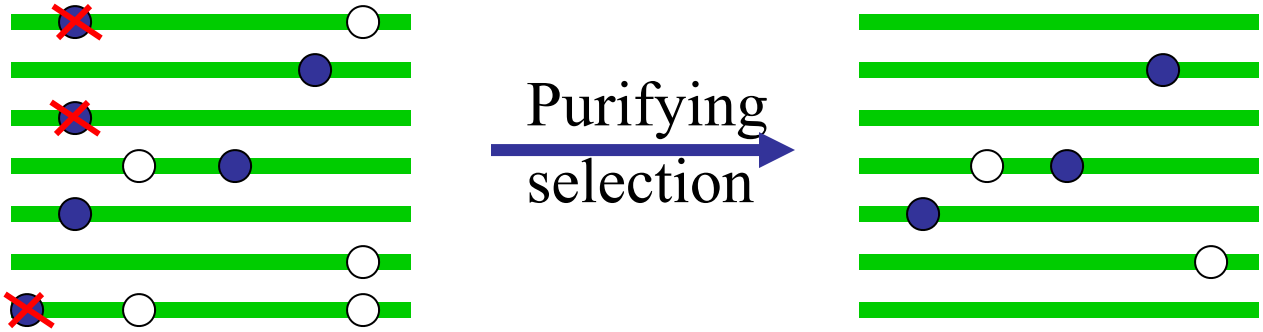
- **negative *D*** is a sign of an excess of the rare alleles and could mean **purifying ("negative") selection** or relatively recent **"positive" selection**
- **positive *D*** is a sign of high proportion of polymorphic alleles and could mean **heterozygote advantage (overdominance)**
- However, demographic events can affect the test
- If the population is growing from an equilibrium situation, then ***S*** and therefore θ_S would increase faster than π (negative ***D***) and would imitate **purifying ("negative") selection**
- If the population is fast expanding after a bottleneck, then π would exceed θ_S and would imitate **heterozygote advantage (overdominance)**
- However, distinguishing between the contributions of selection and demographic events may be very difficult
- The test is likely limited to detecting particularly strong **heterozygote advantage (overdominance)** or relatively recent **"positive" selection**
- If the assumption of population equilibrium is strongly violated, then
  - low ***D*** values may result from selective sweeps, population growth, weak purifying selection or limited sampling
  - high ***D*** values may result from balancing selection or bottlenecks

Deviations from the neutral model expectation $\pi = \theta$ ($D \approx 0$)

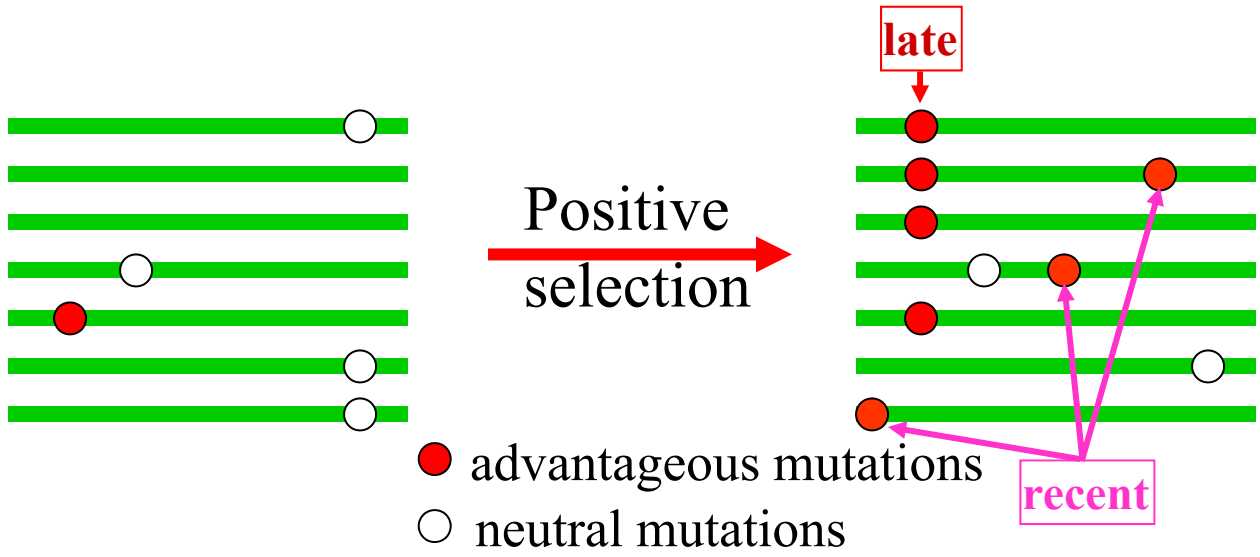
- **Selection**
  - Directional positive selection
  - Purifying selection
  - Balancing selection

- **Demographic effects**
  - Bottleneck
  - Expansion

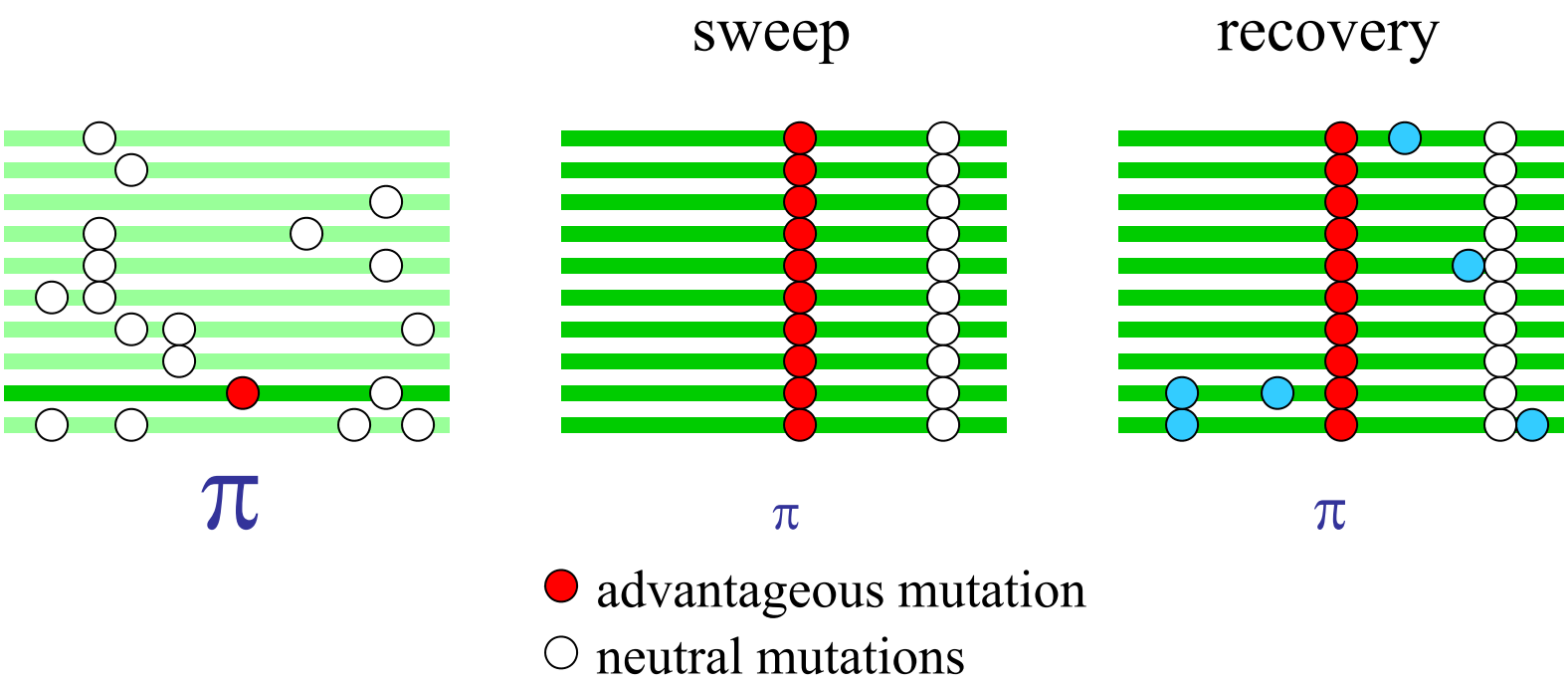
Under negative or purifying selection mutations will be mostly at low frequency, causing uneven allele frequencies and $\pi < \theta$ ($D < 0$)



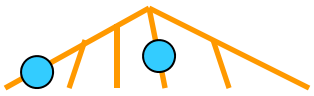
However, recent or late positive selection will also keep alleles at uneven allele frequencies, causing $\pi < \theta$ ($D < 0$)



Strong positive selection ("selective sweep") will decrease both measures of nucleotide diversity π and θ ($D \sim 0$ or > 0)

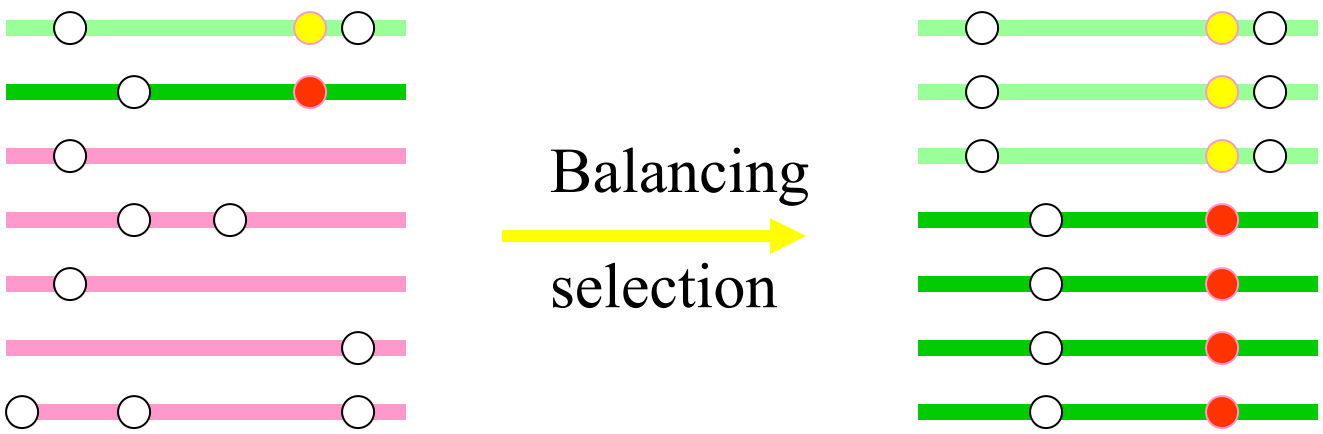


The strong evidence for positive selection is when an advantageous mutation arises repeatedly and independently in different lineages



HyPhy (Hypothesis Testing Using Phylogenies) software package
http://www.hyphy.org

Balancing selection will keep alleles at more evenly distributed polymorphic frequencies and increase π faster than θ ($D > 0$)



π

π

‘balanced’ mutation

neutral mutation

Tajima's test of neutrality

- The nucleotide data from human populations mostly show significant or near significant negative ***D*** values
- How to explain it?
- Relatively recent positive selection?
- Recent expansion of the human population in the last 10,000 years?
- Genome-wide analysis helps to distinguish selection and demographic events because demographic events affect all genes in the genome and, therefore, have genome-wide effects
- Other statistical tests, such as Fu & Li's ***D**** and ***F**** (1993), Fu's ***F*** (1997) and Fay and Wu's ***H*** statistics (2003) take into account phylogenetic information and helps to distinguish balancing selection (increased number of internal branch mutants) from purifying selection (increased number of external branch mutants or singletons), or recent population expansion (excess of any rare pre-existing alleles) from directional selection (excess of recently derived alleles)
- The Fu & Li's ***D**** test statistic is based on the differences between the total number of mutations (η) and the number of singletons (η_s, mutations appearing only once among the sequences).

$$D^* = \frac{\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}{\sqrt{u_{D^*}\eta + v_{D^*}\eta^2}}$$

- The Fu & Li's ***F**** test statistic is based on the differences between the average number of nucleotide differences between pairs of sequences (Π) and the number of singletons (η_s).

$$F^* = \frac{\Pi_n - \frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta + v_{F^*}\eta^2}},$$

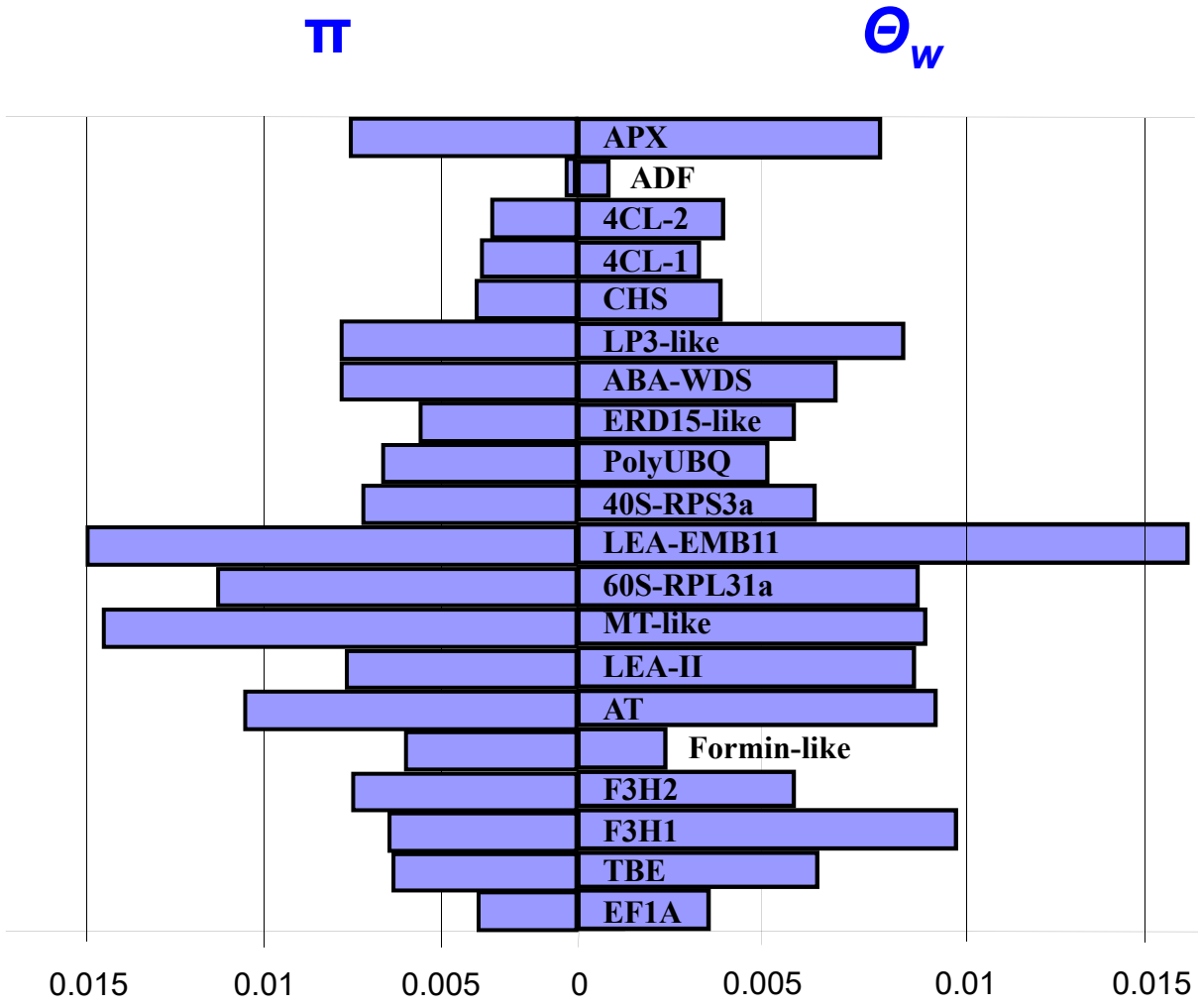
Nucleotide diversity in 20 Douglas-fir candidate genes

Gene product	Gene	Total sites, bp	SNPs	bp per SNP	Pars. SNPs	*H*	π	Θ	Tajima's *D*
Translation elongation factor-1, alpha subunit	EF1A	1072	14	77	9	0.940	0.00274	0.00339	-0.656
Thiazole biosynthetic enzyme	TBE	2954	58	51	36	0.963	0.00516	0.00626	-0.723
Flavanone-3-hydroxylase	F3H1	365	14	26	4	0.690	0.00528	0.00988	-1.576*
Flavanone-3-hydroxylase	F3H2	647	14	46	12	0.828	0.00629	0.00562	0.150
Formin-like protein AHF1	Formin-like	337	3	112	3	0.585	0.00480	0.00229	1.498
Alpha tubulin	AT	2578	93	28	66	0.966	0.00936	0.00935	-0.037
Late embryogenesis abundant type II dehydrin-like protein	LEA-II	504	18	28	13	0.884	0.00647	0.00878	-0.862
Metallothionein-like protein	MT-like	579	20	29	20	0.907	0.01334	0.00911	1.639*
60S ribosomal protein L31a	60S-RPL31a	609	21	29	18	0.701	0.01011	0.00891	0.479
Late embryogenesis abundant EMB11-like protein	LEA-EMB11	545	33	17	26	0.950	0.01378	0.01594	-0.593
40S ribosomal protein S3a	40S-RPS3a	500	12	42	10	0.810	0.00601	0.00617	-0.336
Polyubiquitin	PolyUBQ	898	17	53	15	0.840	0.00544	0.00494	0.357
Early response to dehydration protein	ERD15-like	646	14	46	12	0.598	0.00438	0.00563	-0.757
Abscisic acid, water deficit stress and ripening inducible protein	ABA-WDS	344	9	38	5	0.825	0.00662	0.00672	-0.048
Water deficit inducible protein	LP3-like	481	16	30	13	0.866	0.00662	0.00848	-0.713
Chalcone synthase	CHS	762	11	69	5	0.569	0.00281	0.00371	-1.011
4-coumarate:CoA ligase 1	4CL-1	628	8	79	3	0.841	0.00268	0.00316	-0.460
4-coumarate:CoA ligase 2	4CL-2	629	10	63	7	0.814	0.00237	0.00378	-1.128
Actin depolymerizing factor	ADF	634	2	317	0	0.140	0.00023	0.00081	-1.511
Ascorbate peroxidase	APX	867	26	33	17	0.884	0.00636	0.00789	-0.700
Mean		**829.0**	**20.7**	**40**	**14.7**	**0.780**	**0.00604**	**0.00654**	**-0.349**
Total		**16579**	**413**		**294**				*$P < 0.05$

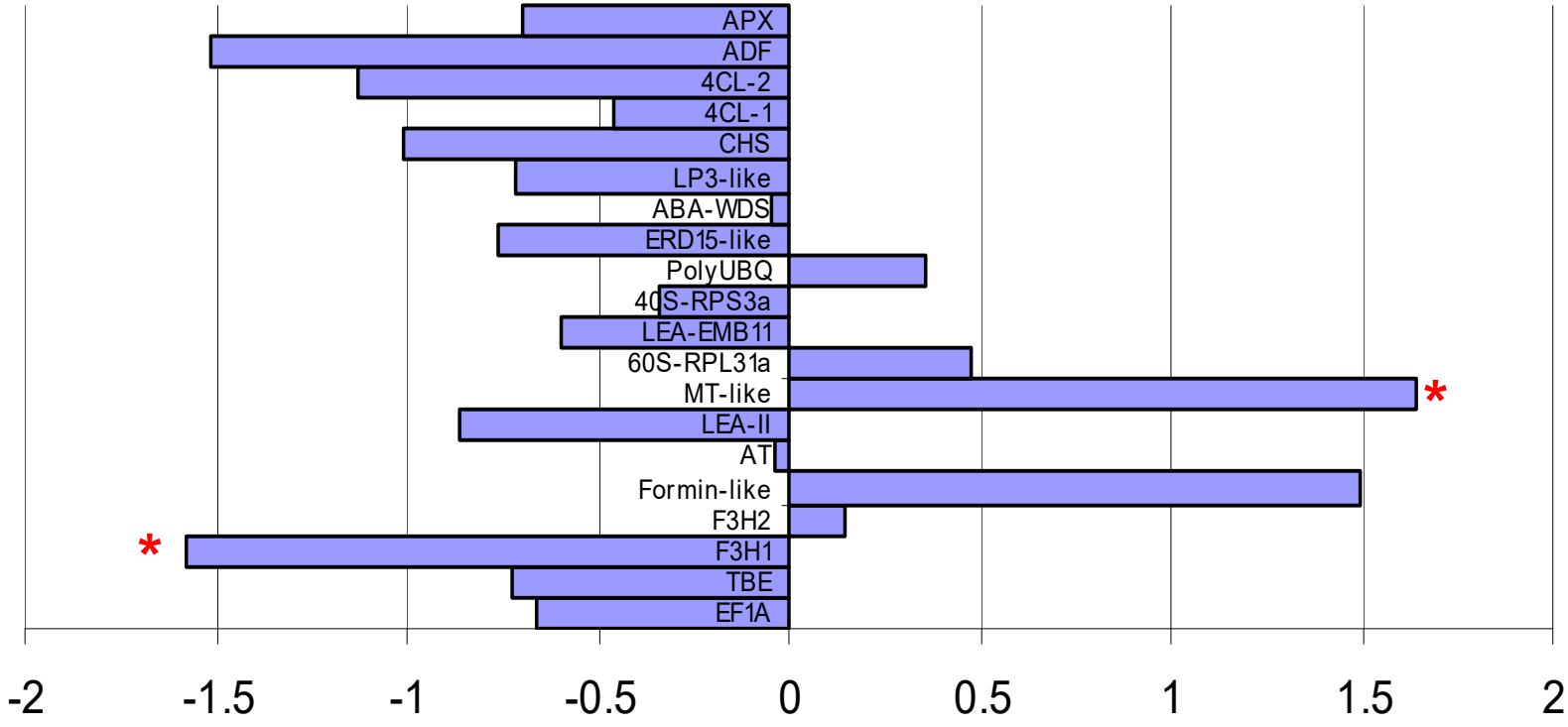
Krutovsky & Neale 2005 *Genetics* 171:2029-2041

Coalescence simulations without recombination were used to test deviations of the observed π and Θ estimates from average values (Hudson 1991)

Nucleotide diversity



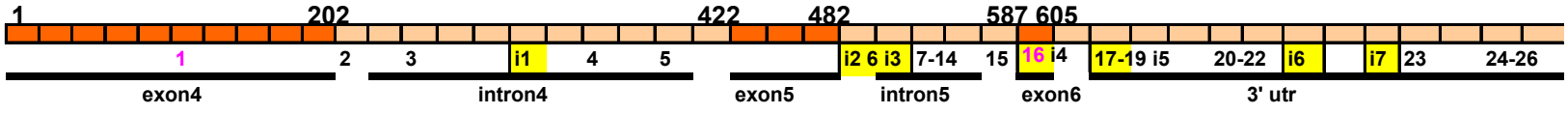
Testing for selection using Tajima's *D*



*P < 0.05

Ascorbate peroxidase (APX) gene (partial)

26 single nucleotide polymorphic (SNP) sites and 7 indels in 12 haplotypes representing 24 individual *Pseudotsuga menziesii* trees from 6 regions and two parents



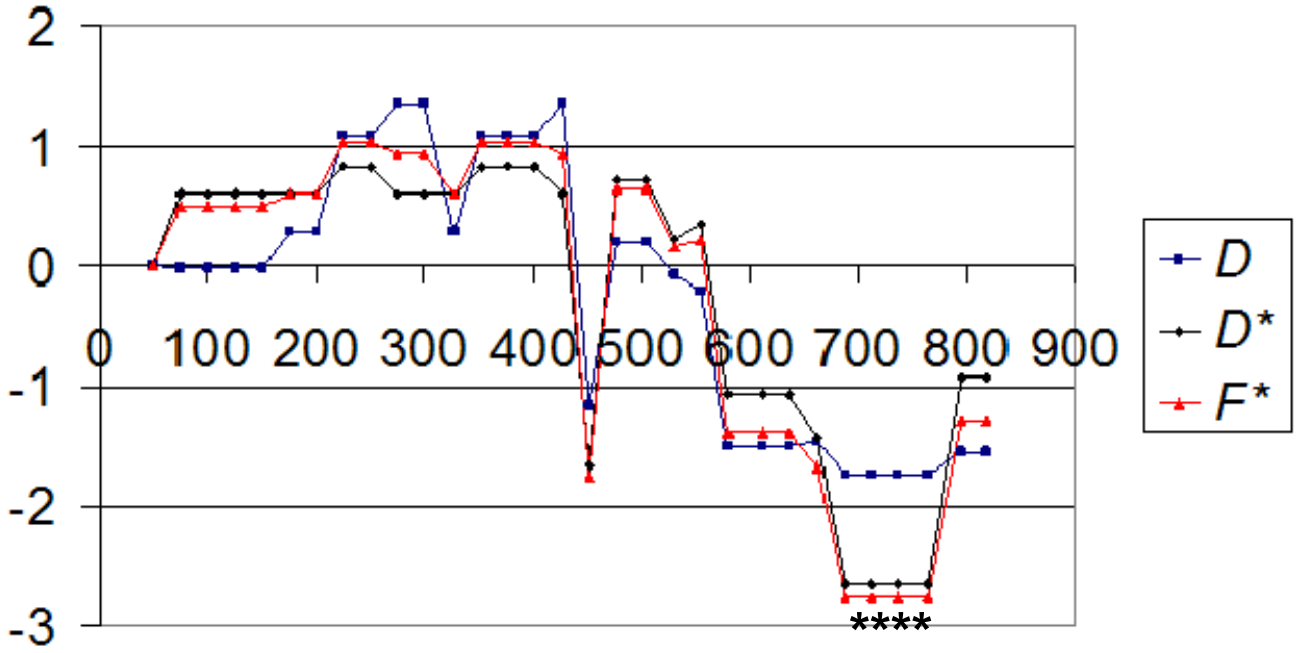
#	1	2	3	ind1	4	5	i2	6	i3	7	8	9	10	11	12	13	14	15	16	indel4	17	18	19	i5	20	21	22	i6	indel7	23	24	25	26	
1_1	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
2_3	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
3_4	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
4_4	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
5_1	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
5_4	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
6_4	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
m1=m4	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	A
2_1	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	T	A	-	G	C	A	T	- - - - - - -	G	T	G	A	B
3_1	A	A	A	T	A	C	A	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	T	A	-	G	C	A	T	- - - - - - -	G	T	G	A	B
2_2	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	C
2_4	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	C
5_2	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	C
6_3	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	C
m2=m3	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	C
1_2	A	A	G	-	-	C	G	A	T	T	-	T	C	T	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	D
3_2	A	A	G	-	-	C	G	A	T	T	-	T	C	T	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	D
6_1	A	A	G	-	-	C	G	A	T	T	-	T	C	T	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	D
1_3	G	T	G	-	-	T	G	A	T	T	-	C	A	A	A	A	A	A	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	E
4_3	G	T	G	-	-	T	G	A	T	T	-	C	A	A	A	A	A	A	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	E
3_3	G	T	G	-	-	T	G	A	T	T	-	C	C	A	A	A	A	A	A	C	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	T C A C T T	G	G	T	G	A	F
d2=d3	G	T	G	-	-	T	G	A	T	T	-	C	C	A	A	A	A	A	A	C	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	T C A C T T	G	G	T	G	A	F
d1	A	T	G	-	-	T	G	A	T	T	-	C	C	A	T	A	T	T	A	T	A	T	A	T	C	A	G	C	T	A	-	G	C	A	T	- - - - - - -	G	T	G	A	G
1_4	A	A	G	-	-	C	G	A	T	C	C	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	T	T	H
4_1	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	T	T	I
4_2	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	T	T	T	A	T	A	T	A	T	C	A	G	G	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	J
5_3	A	A	G	-	-	C	G	-	-	T	-	T	C	A	T	T	T	T	A	T	G	T	A	T	C	A	G	C	A	T	-	G	C	A	T	- - - - - - -	G	T	G	A	K
6_2	A	A	G	-	-	C	G	A	T	T	-	T	C	A	T	A	T	T	T	T	A	-	-	-	-	-	-	C	A	T	T	T	G	T	- - - - - - -	T	A	G	A	L

Protein (100 a/a):

DKDIVALSGAHTLGRCHKERSGFEGAWTSNPLIFDNSYFKELLSGEKEGLLQLPSDKALLEDPVFRSYVEKYAADEDAFFADYAEAHLKLSELGFAEEYQ

Testing for selection using *D*, *D** and *F**)

- Sliding window analysis of neutrality test statistics (Tajima's ***D***, Fu & Li's ***D**** and Fu's ***F****) distribution along the ascorbate peroxidase (**APX**) gene (**P*<0.05)



Coalescence simulations without recombination were used to test the significance of the ***D***, ***D**** and ***F**** statistics (Hudson 1991)

DnaSP (DNA Sequence Polymorphism) software package
http://www.ub.es/dnasp

Interspecific neutrality tests

- **Hudson, Kreitman and Aguadé's (1987) test (HKA test):** based on the neutral theory of molecular evolution prediction that **regions of the genome that highly divergent between species will also present high levels of polymorphism within species**

- **McDonald and Kreitman (1991) test (MK test):** based on the neutral theory of molecular evolution prediction that **the ratio of nonsynonymous to synonymous substitutions fixed between species should be the same as the ratio of nonsynonymous to synonymous substitutions polymorphic within species**

Hudson, Kreitman and Aguadé's (HKA) test

- tests whether intraspecific nucleotide variation (within species) is the same as interspecific variation (between species)
- this is a goodness-of-fit test that requires data from one interspecific comparison of at least two regions of the genome, and also data of the intraspecific polymorphism in the same regions of at least one species
- one of the regions in the comparison should be a putative neutral region
- both the amount of polymorphism within species and the amount of divergence between species should correlate

Hudson, Kreitman and Aguadé's (HKA) test

- Data for the 3rd position codon sites at the *Adh* gene and the noncoding 5' flanking region:

TABLE Estimates of the amount of variation within species and the amount of divergence between species for *D. melanogaster* and *D. sechellia* at the *Adh* gene and the 5′ flanking region. Data from R. R. Hudson, *et al.*, 1987.

	Adh		**5'***Flanking region*	*Ratio (Adh/flanking)*
Variation within species	0.101	>	0.022	4.59
Divergence between species	0.056	≈	0.052	1.08
Ratio variation/divergence	1.80	>	0.42	
expected:	**1.00**		**1.00**	

neutrality is rejected

- The 5' flanking region was chosen under assumption that non-coding regions are more likely to be neutral (we now know that 5' utr can be under strong selection)

Sliding windows of the polymorphism/divergence ratio

- **McDonald (1996, 1998) developed statistical tests to evaluate sliding windows of the polymorphism/divergence ratio when a gene or a region cannot be divided *a priori* into discrete regions**

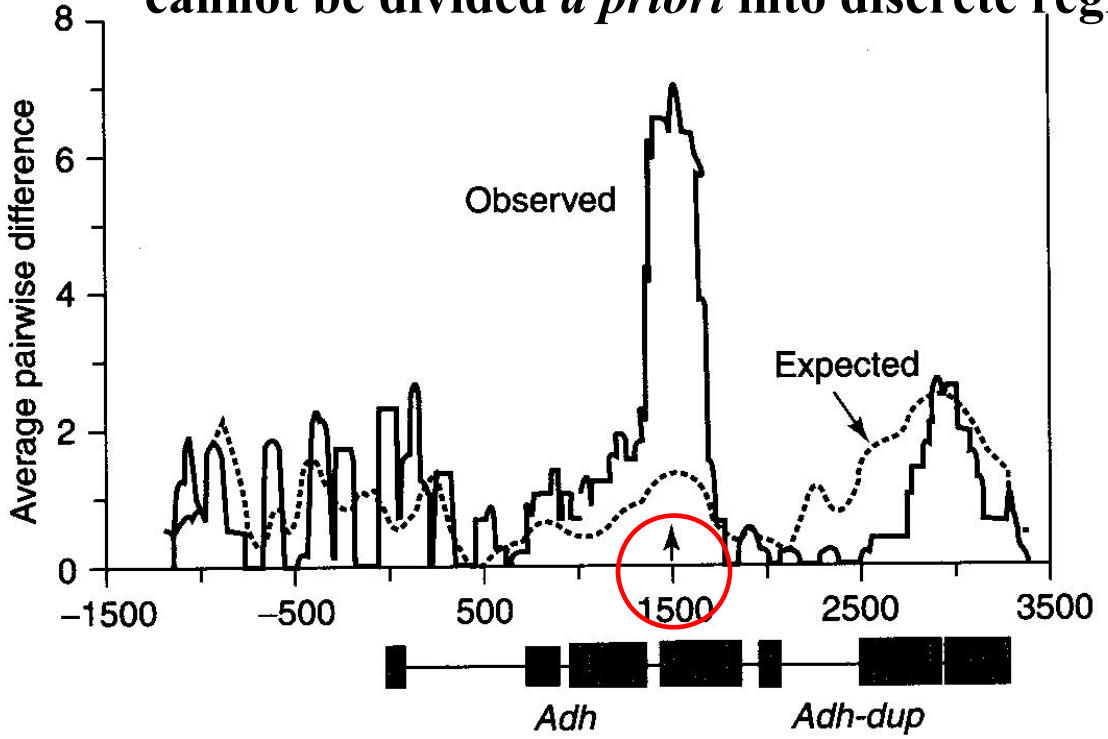


FIGURE A sliding window of the observed and expected genetic variation over the 5′ flanking region, the *Adh* gene, and the *Adh-dup* gene in *Drosophila*. The expected values are based on a no-selection model calculated from between-species divergence. **The arrow indicates the position of the amino acid polymorphism.** Adapted from M. Kreitman and R. R. Hudson, 1991.

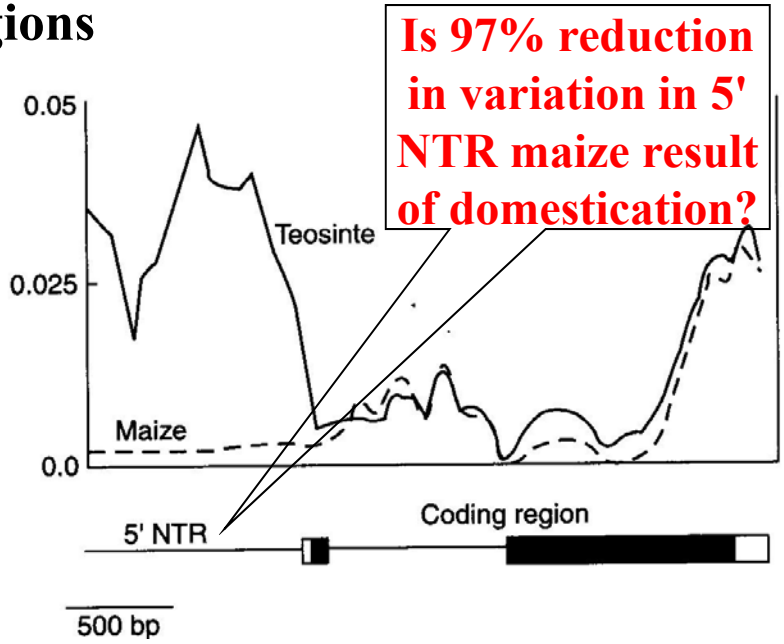


FIGURE A sliding window of the variation of the *teosinte branched 1* gene for maize and teosinte. Adapted from R.-L. Wang *et al*., 1999.

McDonald and Kreitman (MK) test

- based on a comparison of nonsynonymous (replacement) and synonymous variation within and between species
- under neutrality, **the ratio of nonsynonymous to synonymous substitutions fixed between species should be the same as the ratio of nonsynonymous to synonymous substitutions polymorphic within species**
- patterns of variation at the same gene are compared
- MK is statistically more powerful than HKA

McDonald and Kreitman (MK) test

TABLE 1 Variable nucleotides from the coding region of the *Adh* locus in *D. melanogaster*, *D. simulans* and *D. yakuba*

	Con.	*D. melanogaster* a b c d e f g h i j k l	*D. simulans* a b c d e f	*D. yakuba* a b c d e f g h i j k l		
781	G	T T T T T T T T T T T T	- - - - - -	- - - - - - - - - - - -	Repl.	Fixed
789	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
808	A	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Repl.	Fixed
816	G	T T T T - - - - - - - T	T T T T T T	- - - - - - - - - - - -	Syn.	Poly.
834	T	- - - - - - - - - - - -	C C - - - C	- - - - - - - - - - - -	Syn.	Poly.
859	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Repl.	Fixed
867	C	- - - - - - - - - - - -	- - - - - -	G G G G G A G G G G G G	Syn.	2 Poly.
870	C	T T T T T T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Fixed
950	G	- - - - - - - - - - - -	- A - - - -	- - - - - - - - - - - -	Syn.	Poly.
974	G	- - - - - - - - - - - -	T - T T T T	- - - - - - - - - - - -	Syn.	Poly.
983	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1019	C	- - - - - - - - - - - -	- - - - - -	- - - - A - - - - - - -	Syn.	Poly.
1031	C	- - - - - - - - - - - -	- - - - - -	- - - - - - - - A - - -	Syn.	Poly.
1034	T	- - - - - - - - - - - -	- - - - - -	- C C C C C - - C - C C	Syn.	Poly.
1043	C	- - - - - - - - - - - -	- - - - - -	- - - - A - - - - - - -	Syn.	Poly.
1068	C	T T - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1089	C	- - - - - - - - - - - -	A A A A A A	- - - - - - - - - - - -	Repl.	Fixed
1101	G	- - - - - - - - - - - -	- - - - - -	A A A A A A A A A A A A	Repl.	Fixed
1127	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1131	C	- - - - - - - - - - - -	- - - - - -	- - - - T - - - - - - -	Syn.	Poly.
1160	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1175	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1178	C	- - - - - - - - - - - -	- - - - - -	- - - - - - - - - - A -	Syn.	Poly.
1184	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Syn.	Fixed
1190	C	- - - - - - - - - - - -	- - - - - -	- - A - - - - - - - - -	Syn.	Poly.
1196	G	- - - - - - - - - - - -	- - - - - -	T T T T - T T T - T - -	Syn.	Poly.
1199	C	- T - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1202	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1203	C	- - - - - - - - - - - -	- - T - - -	- - - - - - - - - - - -	Syn.	Poly.
1229	T	- - C C C C C C C C C C	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1232	T	- - - - - - - - - - - -	- - - - - -	A A A A A A A A A A A A	Syn.	Fixed
1235	C	- - - - - - - - - A - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1244	C	- - - - - - - - - - - -	- - - - - -	- - A - - - - - - - - -	Syn.	Poly.
1265	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Syn.	Fixed
1271	A	- - - - - - - - - - - -	- T - T - -	- - - - - - - - - - - -	Syn.	Poly.
1277	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1283	C	A A - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1298	C	- - - - - - - - - - - -	- - - - - -	T T T T T T T T T T T T	Syn.	Fixed
1304	C	- - - - - - - - - - - -	- - - - T -	- - - - - - - - - - - -	Syn.	Poly.
1316	C	- - - - - - - - - - - -	- - T T - -	T T T T T T T T T T T T	Syn.	Poly.
1425	C	A A - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1431	T	C C - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Poly.
1443	C	- - - - - - G G G G G G	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1452	C	- - - - - T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
* 1490	A	- - - - - - C C C C C C	- - - - - -	- - - - - - - - - - - -	Repl.	Poly.
1504	C	T T T T T T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Fixed
1518	C	- - - - - T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1524	T	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Syn.	Fixed
1527	C	T T T T T T - - - - - -	- - - - - -	- - - - T - - - - - - -	Syn.	Poly.
1530	G	- - - - - - - - - - - -	- - - - - -	- - - A - - - - - - - -	Syn.	Poly.
1545	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1548	C	- - - - - - - - - - - -	- - - - - -	- - - - A - - - - - - -	Syn.	Poly.
1551	C	- - - - - - - - - - - -	- - - T - -	- - - - - - - - - - - -	Syn.	Poly.
1555	C	- - - - - - - - - - - T	- - - - - -	- - - - - - - - - - - -	Repl.	Poly.
1557	C	A A A A A - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1560	G	- - - - - - - - - - - -	- - - A - -	- - - - - - - - - - - -	Syn.	Poly.
1573	G	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Repl.	Fixed
1581	C	- - - - - - - - - - - -	- - - - - -	T T T T T T T T T T T T	Syn.	Fixed
1584	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G -	Syn.	Poly.
1590	C	T T T T T T T T T T T T	T T T - - -	- - - - - - - - - - - -	Syn.	Poly.
1596	G	- - A A - A - - - - - A	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1611	A	- - - - - - - - - - - -	- - - - - -	T T T T T T T T T T T T	Syn.	Fixed
1614	C	- - - - - - - - - - - -	- G - - - -	- - - - - - - - T - - -	Syn.	2 Poly.
1635	C	- - - - - - - - - - - -	- - - - - -	- - T - T - - - - - - -	Syn.	Poly.
1657	A	- - - - - - - - - - - -	- - - - - -	T T T T T T T T T T T T	Repl.	Fixed
1665	C	- - - - - - - - - - - -	- - - - - -	- - - - - - - - - - T -	Syn.	Poly.

The first column contains the nucleotide positions, numbered according to ref. 6, so that the first base of the coding region of exon 2 is 778, of exon 3 is 942 and exon 4 is 1,417. The second column contains the consensus (con.) nucleotide for each site. Nucleotides identical to the consensus are shown as a dash. Several *D. yakuba* individuals were heterozygous; sites which are underlined contained both the nucleotide shown and the consensus nucleotide. The *D. melanogaster* sequences are six Adh^S (a-f) and five Adh^F (g-k) alleles[6] and one Adh^{FCnD} allele[10]. *D. simulans* sequences a and b were also from the literature[11,12]. *D. simulans* sequences c-f were of cloned alleles from flies from Australia, Canada, France and the Congo. They were sequenced by the dideoxy chain termination method directly from supercoiled plasmids[13]. The *D. yakuba* alleles of single flies from Brazzaville, Congo were sequenced directly from products of amplification by the polymerase chain reaction[14,15]. Each of these sequences is from one fly from a separate isofemale line. Every allele was completely sequenced in both directions. Each substitution relative to the consensus nucleotide at a site is classified as either fixed or polymorphic. A substitution that is fixed in one species and polymorphic in another is classified as a polymorphism. Because we do not know whether a substitution that is polymorphic in more than one species represents one or two mutations, it is classified as a single polymorphism. Syn, synonymous; Repl., replacement; Poly., polymorphic.

John H. McDonald & Martin Kreitman (1991)
Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature 351: 652-654 (see also Table 6.10 in the texbook)

TABLE Nucleotide sequence data for the nine nonsynonymous (replacement) differences for the *Adh* gene in three species of *Drosophila*. In the right column, the status, fixed for differences between the species or polymorphic in one of the species, is indicated. A polymorphism at position 1490, indicated by an asterisk, results in the difference in the *F* and *S* allozyme alleles within *D. melanogaster*. Data from J. McDonald and M. Kreitman, 1991.

Substitution	Fixed	Poly.
Nonsynonymous (Repl.)	7	2
Synonymous (Syn.)	17	42
Ratio	**0.41** >	**0.05**

Fisher's Exact Test (*P*=0.007)
http://www.langsrud.com/fisher.htm

McDonald and Kreitman (MK) test

TABLE 6.11 The number of nonsynonymous (replacement) and synonymous substitutions for fixed differences between species and polymorphism within species (a) in general, (b) for *Adh* in three *Drosophila* species (McDonald and Kreitman, 1991), and (c) for *G6pd* in *D. melanogaster and D. simulans* (Eanes *et al.*, 1993). Below, data from humans and chimpanzees are given for (d) mtDNA gene *ND3* (Nachman *et al.*, 1996), (e) *G6pd* (Verrelli *et al.*, 2002), and (f) *HLA-B* (Garrigan and Hedrick, 2003).

positive selection?

	(a) *General*		(b) *Adh*		(c) *G6pd*	
	Fixed	*Polymorphic*	*Fixed*	*Polymorphic*	*Fixed*	*Polymorphic*
Nonsynonymous	N_F	N_P	7	2	21	2
Synonymous	S_F	S_P	17	42	26	36
Ratio	$N_F/S_F = N_P/S_P$		0.41 >	0.05	0.81 >	0.06

purifying selection?

balancing selection?

	(d) *ND3*		(e) *G6pd*		(f) *HLA-B*	
Nonsynonymous	4	8	0	5	0	76
Synonymous	31	10	44	23	0	49
Ratio	0.13 <	0.8	0.0 <	0.28	— <	1.61

Synonymous vs. Nonsynonymous Nucleotide Substitutions

- If all substitutions are neutral, than the ratio between nonsynonymous (K_A or K_N) and synonymous (K_S) nucleotide substitutions should be equal $K_A/K_S = 1$
- Synonymous substitutions (usually in the second and mostly the third position in codon) and substitutions in noncoding regions represent "silent" substitutions that are likely selectively neutral
- Unlike "silent" substitutions nonsynonymous nucleotide substitutions are likely under selection (not neutral)
- If nonsynonymous nucleotide substitutions are under purifying or "negative" selection, then the ratio $K_A/K_S < 1$
- If they are under adaptive or "positive" (Darwinian) selection, then the $K_A/K_S > 1$

Synonymous vs. Nonsynonymous Nucleotide Substitutions

- There are two main approaches to calculate the ratio between nonsynonymous and synonymous nucleotide substitutions

- Li et al. 1985: K_A/K_S where $K_A = N_d/N$ and is calculated as the number of nonsynonymous substitutions (N_d) per total number of nonsynonymous sites (N), and $K_S = S_d/S$ and is calculated as the number of synonymous substitutions (S_d) per total number of synonymous sites (S)

- Nei & Gojobori 1986: K_A/K_S is calculated as d_N/d_S where d_N is the average diversity, divergence or pairwise distance between sequences calculated only for nonsynonymous sites ($p_N=N_d/N$), and d_S is the average diversity, divergence or pairwise distance between sequences calculated only for synonymous sites ($p_S=S_d/S$):

$$d_S = -\frac{3}{4}\ln(1-\frac{4p_S}{3}) \quad \text{and} \quad d_N = -\frac{3}{4}\ln(1-\frac{4p_N}{3})$$

Synonymous vs. Nonsynonymous Nucleotide Substitutions

TABLE 8.12 The number of potentially synonymous sites (s_j) in parentheses for the DNA genetic code. The amino acid for each codon is given by the one-letter symbol as in Table 6.14

	Second position							
	T		*C*		*A*		*G*	
First position	TTT (0.333)	F	TCT (1.0)	S	TAT (1.0)	Y	TGT (0.5)	C
	TTC (0.333)	F	TCC (1.0)	S	TAC (1.0)	Y	TGC (0.5)	C
	TTA (0.667)	L	TCA (1.0)	S	TAA (Stop)		TGA (Stop)	
	TTG (0.667)	L	TCG (1.0)	S	TAG (Stop)		TGG (0.0)	W
	CTT (1.0)	L	CCT (1.0)	P	CAT (0.333)	H	CGT (1.0)	R
	CTC (1.0)	L	CCC (1.0)	P	CAC (0.333)	H	CGC (1.0)	R
	CTA (1.333)	L	CCA (1.0)	P	CAA (0.333)	Q	CGA (1.5)	R
	CTG (1.333)	L	CCG (1.0)	P	CAG (0.333)	Q	CGG (1.333)	R
	ATT (0.667)	I	ACT (1.0)	T	AAT (0.333)	N	AGT (0.333)	S
	ATC (0.667)	I	ACC (1.0)	T	AAC (0.333)	N	AGC (0.333)	S
	ATA (0.667)	I	ACA (1.0)	T	AAA (0.333)	K	AGA (0.833)	R
	ATG (0.0)	M	ACG (1.0)	T	AAG (0.333)	K	AGG (0.667)	R
	GTT (1.0)	V	GCT (1.0)	A	GAT (0.333)	D	GGT (1.0)	G
	GTC (1.0)	V	GCC (1.0)	A	GAC (0.333)	D	GGC (1.0)	G
	GTA (1.0)	V	GCA (1.0)	A	GAA (0.333)	E	GGA (1.0)	G
	GTG (1.0)	V	GCG (1.0)	A	GAG (0.333)	E	GGG (1.0)	G

For example, for the leucine codon TTA, $s_j = 0.333 + 0 + 0.333 = 0.667$ because all nucleotide changes at the second position result in nonsynonyomous changes, and only one of the three changes at both the first and the third positions results in a synonymous change.

Synonymous vs. Nonsynonymous Nucleotide Substitutions

- There are also now numerous maximum likelihood methods to analyze coding sequence data based on explicit codon substitution models
- It takes into account codon usage

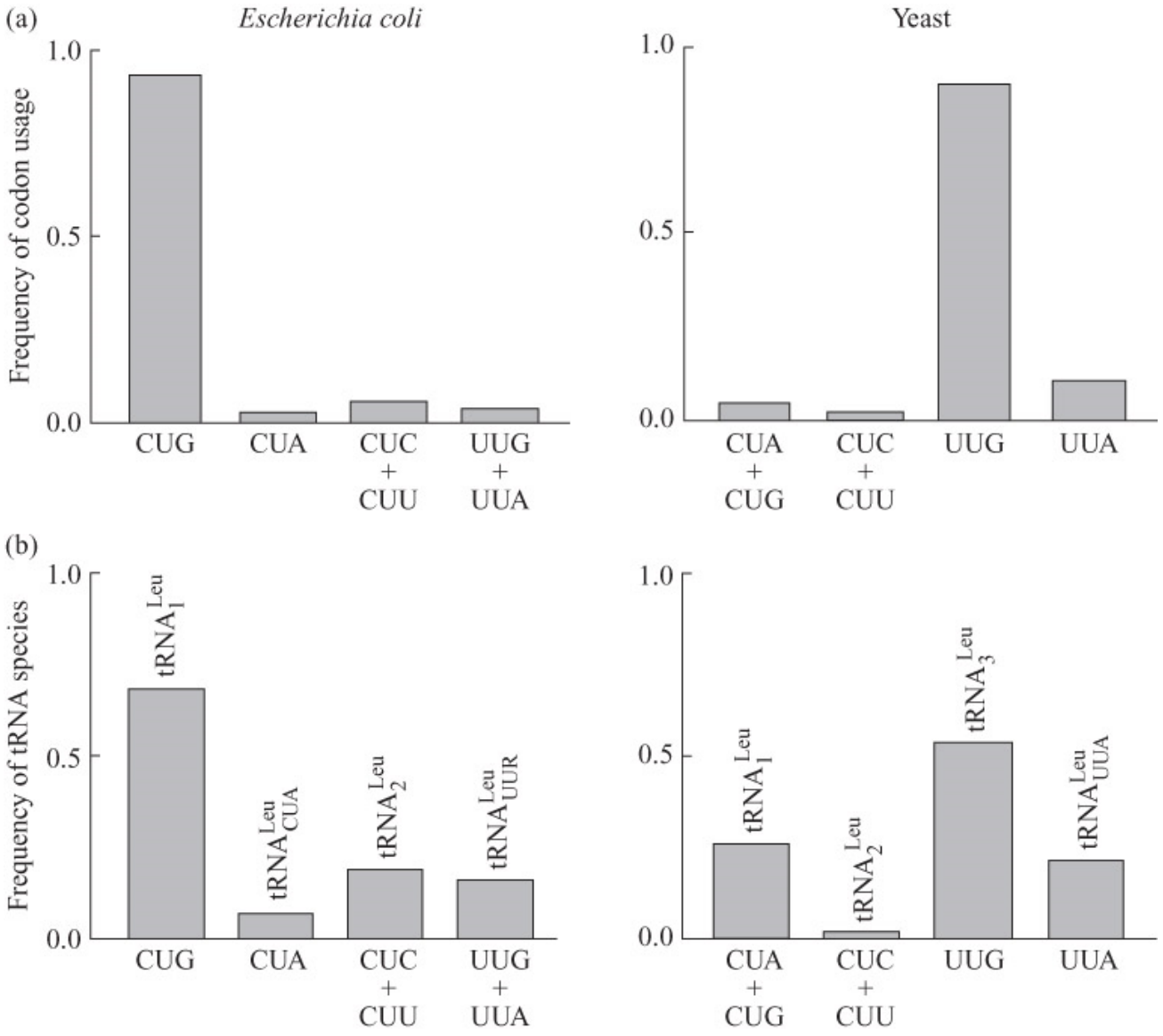


FIGURE
(a) The frequencies of codon usage for leucine codons in highly expressed genes and

(b) the frequencies of the different leucine tRNA species in *E. coli* (left) and yeast (right). Here the RNA code is used where pyrimidine T (thymine) is replaced by the pyrimidine U (uracil). Adapted from W.-H. Li and D. Graur, 1991.

Synonymous vs. Nonsynonymous Nucleotide Substitutions

- Codon usage bias

(a) 1133 genes in *D. melanogaster*

(b) 1586 genes in *E. coli*

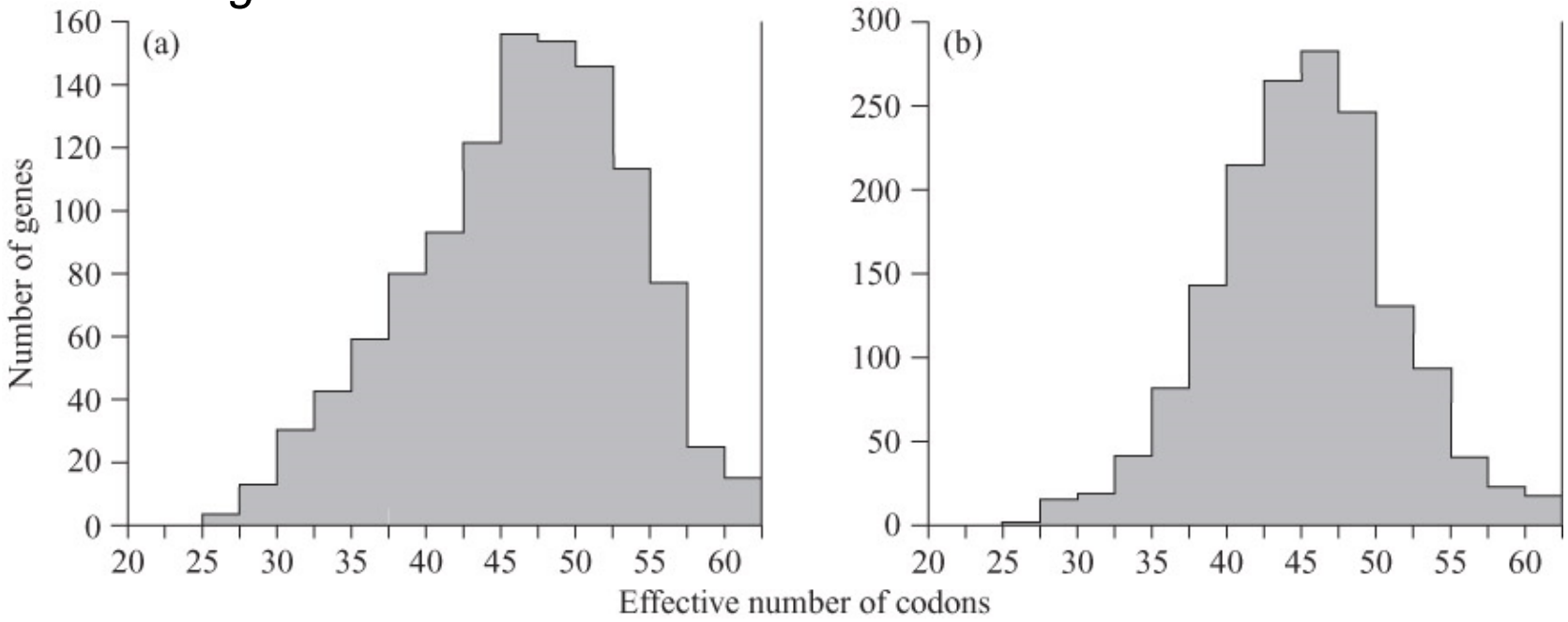


FIGURE The distribution of codon usage bias as measured by the effective number of codons over (a) 1133 genes in *D. melanogaster* and (b) 1586 genes in *E. coli*. Adapted from J. R. Powell and E. N. Moriyama, 1997.

Nonsynonymous to synonymous nucleotide substitution ratio (d_N/d_S), its confidence intervals (*CI*) and *P*-values for $d_N/d_S = 1$ (neutrality test) for 18 Douglas-fir candidate genes estimated under different codon and nucleotide substitution models

Gene	Nonsynonymous		Synonymous		*P* (Fisher's exact test)	mNG86×JC69			MG94×F81			MG94×HKY85		
	SNPs	Sites	SNPs	Sites		d_N/d_S	Z-test	*P*	d_N/d_S	*CI*	*P*	d_N/d_S	*CI*	*P*
EF1A	**0**	**564**	**7**	**178**	**0.000** [b]	**0.000**	**2.300**	**0.023**	**0.000** [c]	**0.000,0.082**	**0.000**	**0.000**	**0.000,0.106**	**0.000**
TBE	**7**	**775**	**13**	**254**	**0.000**	**0.118**	**2.772**	**0.007**	**0.149**	**0.069,0.283**	**0.000**	**0.153**	**0.070,0.285**	**0.000**
F3H1	6	209	2	58	1.000	0.493	0.575	0.566	0.818	0.328,1.686	0.810	0.931	0.370,1.902	0.950
F3H2	3	345	3	96	0.354	0.522	0.668	0.506	0.277	0.071,0.738	0.129	0.301	0.075,0.784	0.162
Formin	2	262	1	74	0.529	0.683	0.304	0.762	0.284	0.048,0.896	0.222	0.316	0.053,0.977	0.274
AT	**0**	**1026**	**30**	**327**	**0.000**	**0.000**	**4.687**	**0.000**	**0.000**	**0.000,0.019**	**0.000**	**0.000**	**0.000,0.017**	**0.000**
LEA-II	3	192	4	57	0.107	0.309	1.058	0.292	**0.217**	**0.055,0.572**	**0.048**	0.260	0.065,0.678	0.087
MT-like	2	162	2	42	0.197	0.259	1.006	0.317	0.208	0.036,0.670	0.134	0.221	0.055,0.573	0.083
60S-RPL31a	**0**	**263**	**6**	**76**	**0.000**	**0.000**	**2.315**	**0.022**	**0.000**	**0.000,0.086**	**0.001**	**0.000**	**0.000,0.103**	**0.000**
LEA-EMB11	**6**	**191**	**7**	**64**	**0.046**	0.332	1.660	0.100	**0.273**	**0.119,0.539**	**0.014**	0.375	0.161,0.730	0.063
40S-RPS3a	0	131	1	40	0.238	0.000	0.951	0.343	**0.000**	**0.000,0.283**	**0.014**	**0.000**	**0.000,0.211**	**0.010**
PolyUBQ	**0**	**528**	**9**	**159**	**0.000**	**0.000**	**2.853**	**0.005**	**0.000**	**0.000,0.064**	**0.000**	**0.000**	**0.000,0.068**	**0.000**
ERD15-like	4	309	3	93	0.362	0.623	0.536	0.593	0.399	0.126,0.944	0.247	0.738	0.228,1.725	0.695
ABA-WDS	**4**	**258**	**5**	**84**	**0.049**	0.252	1.379	0.170	**0.223**	**0.081,0.485**	**0.011**	**0.299**	**0.093,0.694**	**0.014**
LP3-like	**4**	**257**	**8**	**79**	**0.002**	**0.135**	**2.288**	**0.024**	**0.136**	**0.043,0.320**	**0.000**	**0.174**	**0.054,0.405**	**0.003**
4CL-1	4	420	1	144	1.000	7.250	1.386	0.168	1.620	0.581,3.488	0.643	2.306	0.815,4.892	0.419
4CL-2	4	423	3	141	0.376	0.423	0.863	0.390	0.415	0.130,0.971	0.266	0.387	0.120,0.900	0.239
APX	2	213	1	65	1.000	0.714	0.241	0.810	0.596	0.101,1.877	0.639	1.153	0.191,3.575	0.908
Mean	**2.8**	**363**	**5.9**	**113**	**0.009**	0.209			0.312			0.423		

mNG86×JC69 - modified Nei and Gojobori 1986 (mNG86) method with Jukes Cantor 1969 (JC69) model (Nei and Kumar 2000); MG94×F81 and MG94×HKY85 - Muse and Gaut 1994 codon model (MG94) with Felsenstein 1981 (F81) and Hasegawa, Kishino and Yano 1985 (HKY85) nucleotide substitution models, respectively. (Krutovsky & Neale 2005 Genetics 171:2029-2041)

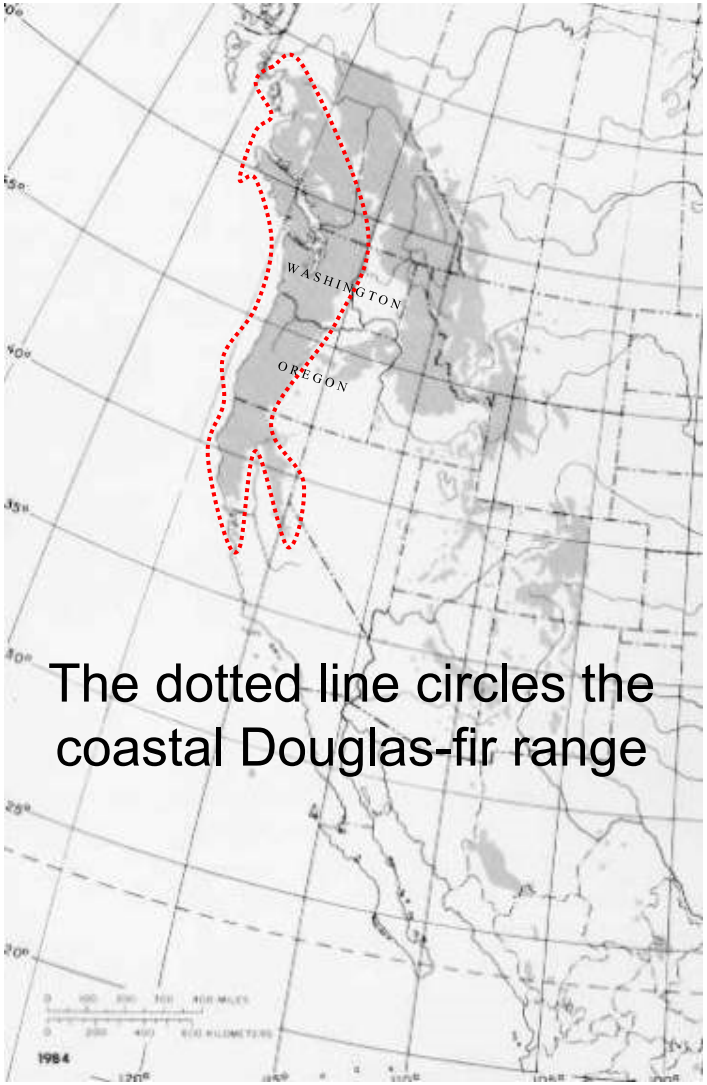
Positively and negatively selected sites identified using likelihood-based methods for estimating the rates of nonsynonymous and synonymous substitutions at each site along the phylogeny in 18 Douglas-fir candidate genes

Gene	Positively selected sites						Negatively selected sites					
	nonsynonymous			synonymous			nonsynonymous			synonymous		
	SLAC	ALRS	FL	SLAC	ALRS	FL	SLAC	ALRS	FL	SLAC	ALRS	FL
EF1A	-	-	-	-	-	-	-	-	-	-	3	-
TBE	-	-	-	-	-	-	-	-	-	1	7	13
F3H1	-	-	-	-	-	-	-	-	-	-	2	2
F3H2	-	-	-	-	-	-	-	-	-	-	1	-
Formin	-	-	-	-	-	-	-	-	-	-	1	1
AT	-	-	-	-	-	-	-	-	-	2	5	-
LEA-II	-	-	-	-	-	-	-	-	-	-	1	1
MT-like	-	-	-	-	-	-	-	-	-	-	-	-
60S-RPL31a	-	-	-	-	-	-	-	-	-	-	2	-
LEA-EMB11	-	-	6	-	-	-	-	-	-	-	-	-
40S-RPS3a	-	-	-	-	-	-	-	-	-	-	1	-
PolyUBQ	-	-	-	-	-	-	-	-	-	-	1	-
ERD15-like	-	-	-	-	-	-	-	-	-	-	1	-
ABA-WDS	-	-	2	-	-	-	-	-	-	1	1	1
LP3-like	-	-	-	-	-	-	-	-	1	-	4	7
4CL-1	-	-	4	-	-	-	-	-	-	-	-	-
4CL-2	-	-	-	-	-	-	-	-	-	-	1	-
APX	-	-	-	-	-	-	-	-	-	-	1	1

SLAC - Single Likelihood Ancestor Counting, ALRS - Approximate Likelihood Ratio, FL - Full Likelihood (Kosakovsky-Pond & Frost 2004)

HyPhy (Hypothesis Testing Using Phylogenies) software package http://www.hyphy.org

Candidate gene based association mapping in Douglas-fir

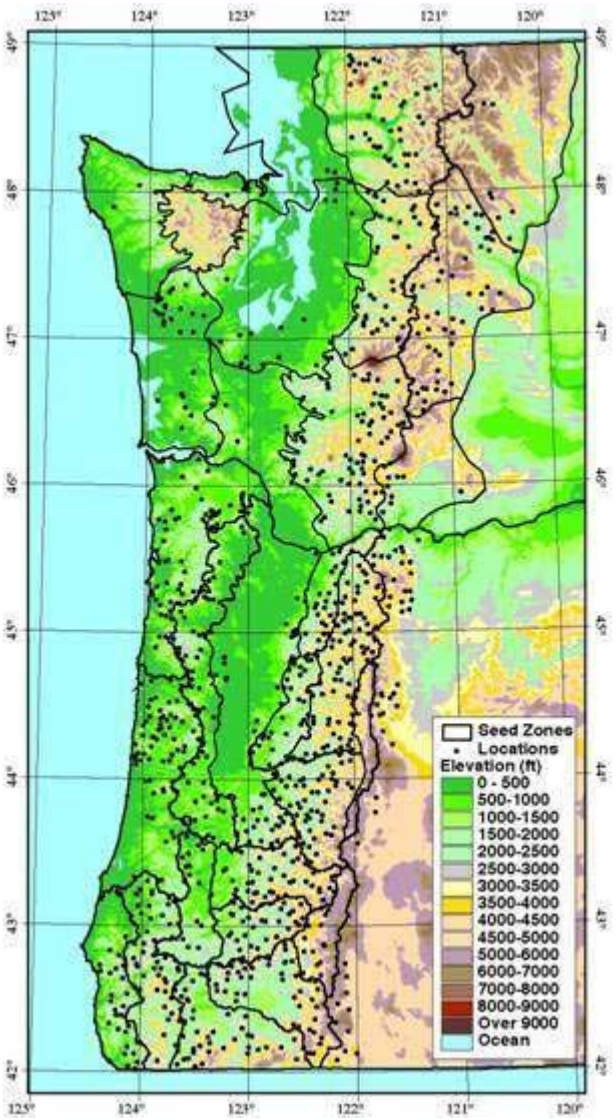


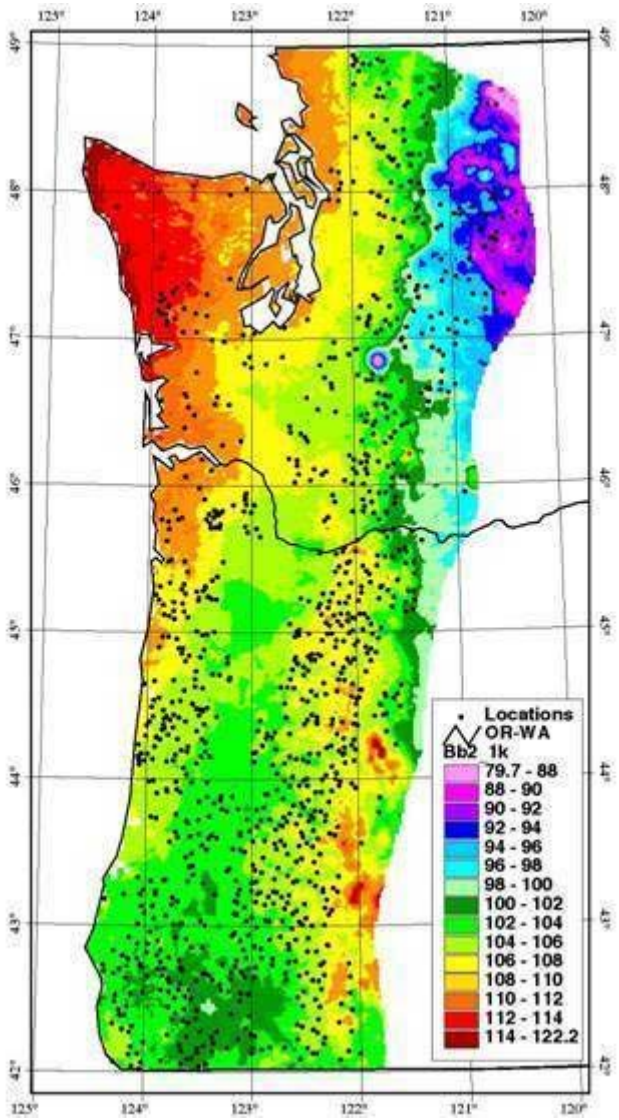
The dotted line circles the coastal Douglas-fir range

- ~700 trees in the association mapping population representing the entire range
- phenotyped for 21 growth rhythm (bud flush, bud set) and cold-hardiness related traits
- genotyped for 384 SNPs in 117 candidate genes using Illumina GoldenGate genotyping assay

Eckert, A.J., A.D. Bower, J.L. Wegrzyn, B. Pande, K.D. Jermstad, K.V. Krutovsky, J.B. St. Clair and D.B. Neale, 2009 Association genetics of coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* 182: 1289-1302

GIS derived maps of parent trees in the common-garden population





Elevation and seed zones

Julian day for family mean bud flush

(Brad St. Clair *et al.* 2005, 2006)

Candidate gene-based association mapping in Douglas-fir

- ~**700** individuals phenotyped for **21** cold-hardiness related traits were genotyped for **384 SNPs** in **117** candidate genes using Illumina GoldenGate genotyping assay.

- **30** highly significant genetic associations for **12** candidate genes (*4CL*, *LEA*, *F3H2*, *MADS-box* and *MYB-like TFs*, *a-expansin*, etc.) and **10** were discovered.

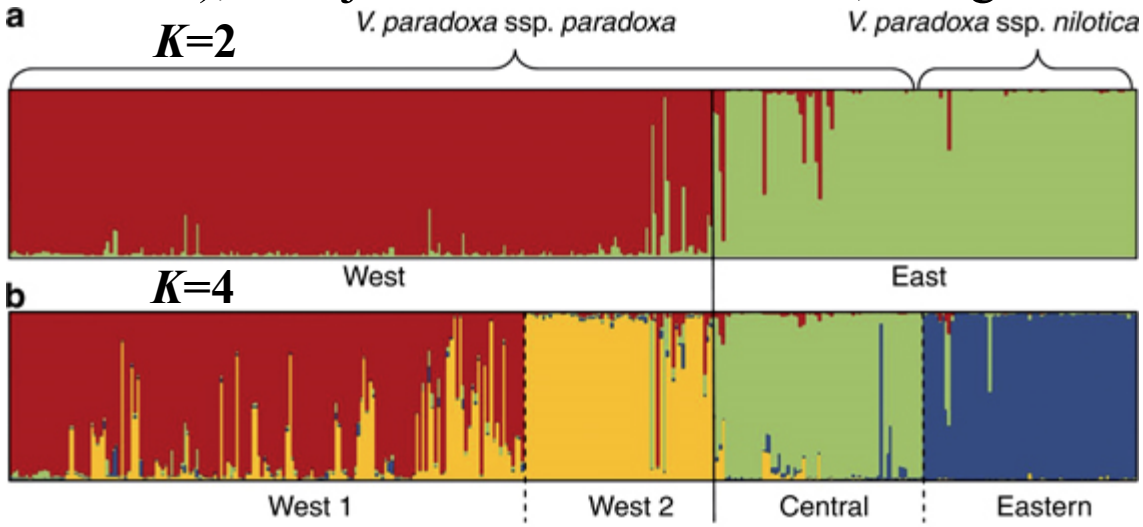
- **7** markers had elevated levels of differentiation between sampling sites situated across the Cascade crest in northeastern Washington.

- Marker effects were small (**1% < r^2 < 3.6%**) and within the range of those published previously for forest trees, but **6 SNPs** explained **17**% of the phenotypic variance in cold damage to stems.

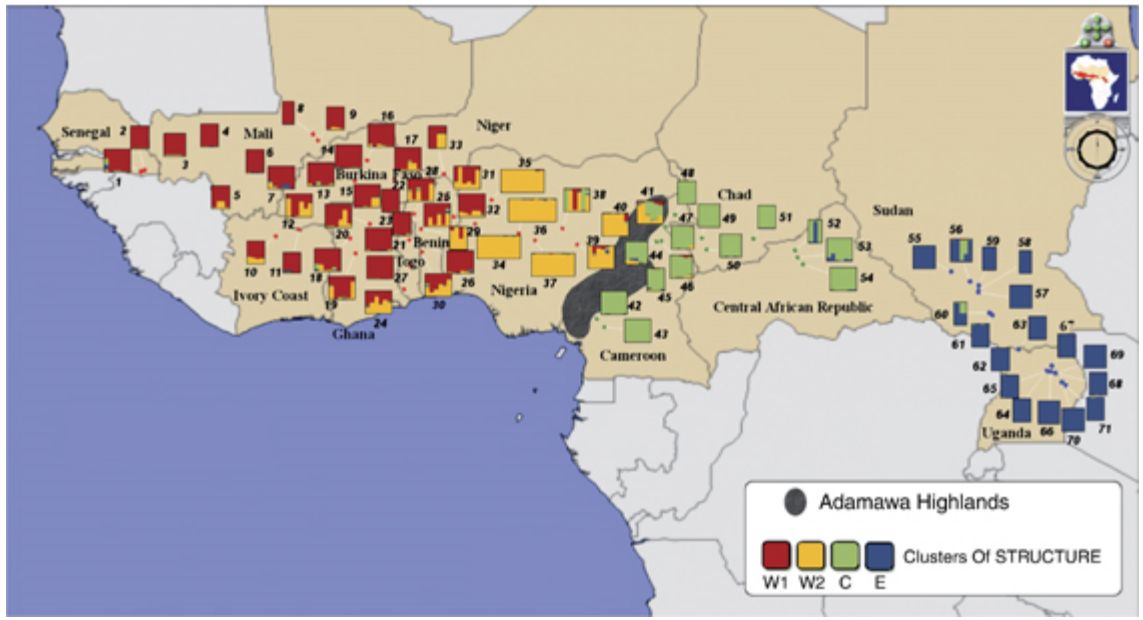
Eckert, A.J., A.D. Bower, J.L. Wegrzyn, B. Pande, K.D. Jermstad, K.V. Krutovsky, J.B. St. Clair and D.B. Neale, 2009 Association genetics of coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* **182**: 1289-1302.

Landscape / Spatial Genomics

- Individual *Q*-values and population clusters found in *Vitellaria paradoxa* (shea tree), a major African savanna tree, using STRUCTURE program



c **Distribution of individual *Q*-values highlighted by different colors**



Genetic structure of nuclear microsatellites across 374 individuals from 71 populations of *Vitellaria paradoxa*:

(a) Bar plot showing clustering of individuals by STRUCTURE with ***K*=2** (Pritchard *et al.*, 2000).

(b) Bar plot showing clustering of individuals by STRUCTURE with ***K*=4**.

(c) Geographical plot of clustering of individuals by STRUCTURE with ***K*=4**: W1: 'West 1'; W2: 'West 2'; C: 'Central'; E: 'Eastern'.

Allal et al. (2011) Past climate changes explain the phylogeography of *Vitellaria paradoxa* over Africa. *Heredity* **107:** 174–186.

Landscape / Spatial Genomics

- Individual *Q*-values and 8 population clusters found in *Ctenomys minutus* (subterranean rodent) from the coastal plain of Rio Grande do Sul, southern Brazil, using STRUCTURE program

Distribution of individual *Q*-values

***K*=8**

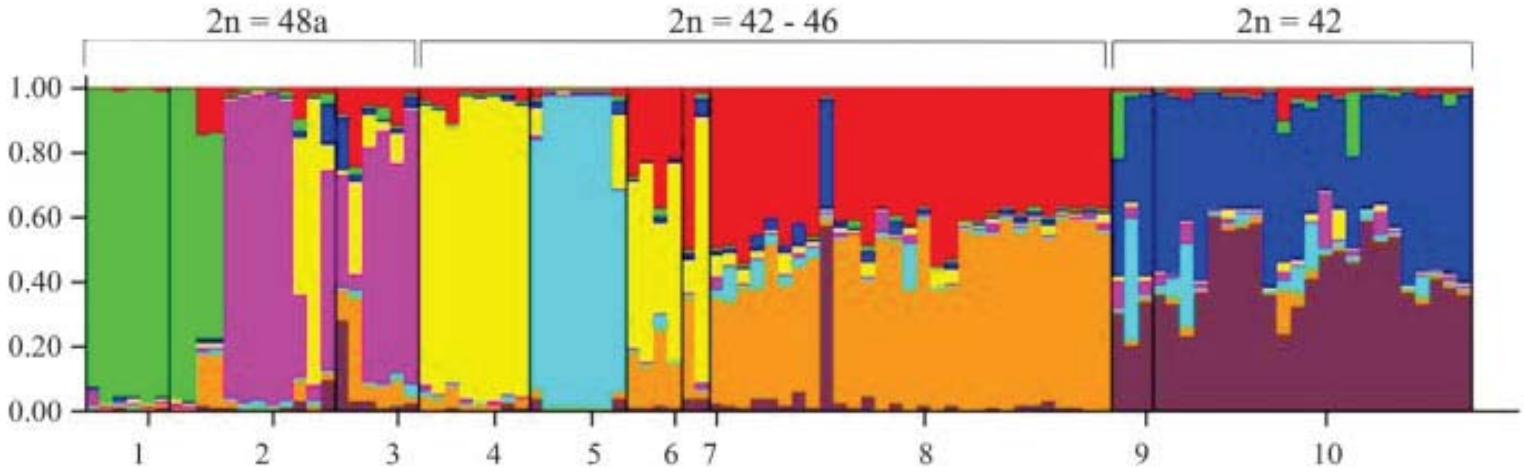


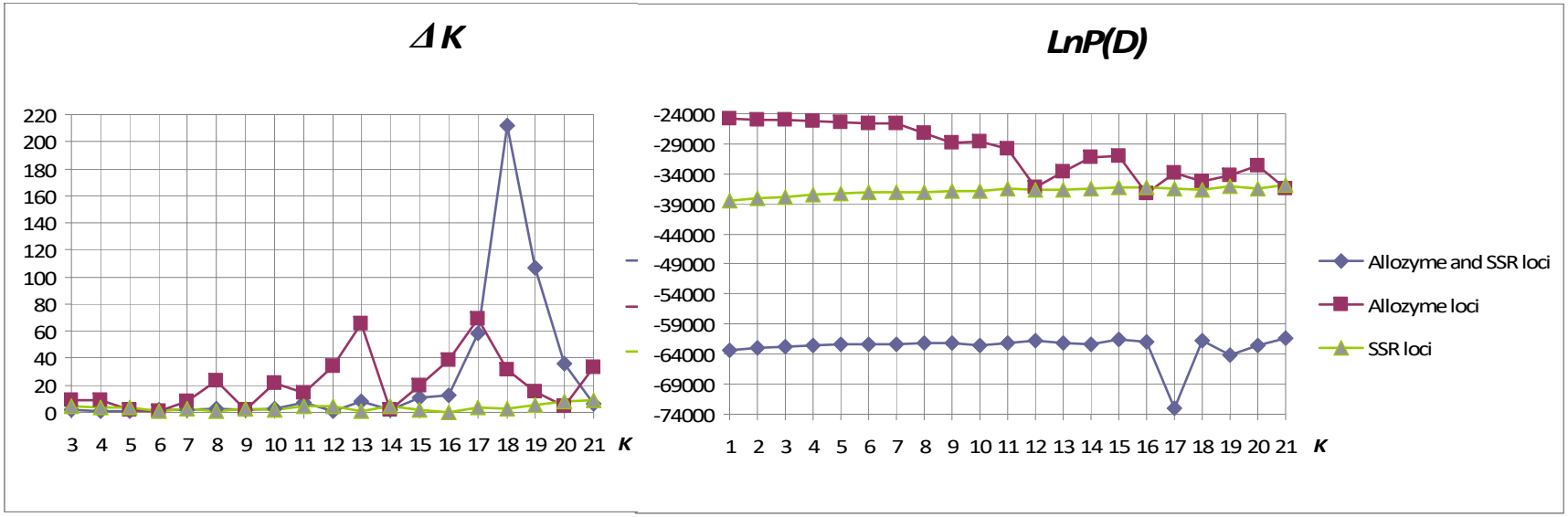
Figure 2 - Structure bars plot showing the eight genetic populations identified by the analysis. The numbers 1-10 correspond to the sample spots in Figure 1 and Table 1.

These results suggest chromosomal rearrangements are only of minor significance in the establishment of reproductive barriers for this species

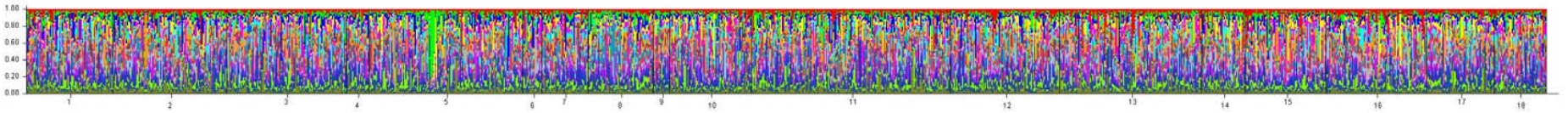
Castilho et al. A hybrid zone of the genus *Ctenomys*: a case study in southern Brazil. Genet. Mol. Biol. 2012 35: 990-997.

Landscape / Spatial Genomics of coastal Douglas-fir

- 18 population clusters were found using STRUCTURE program

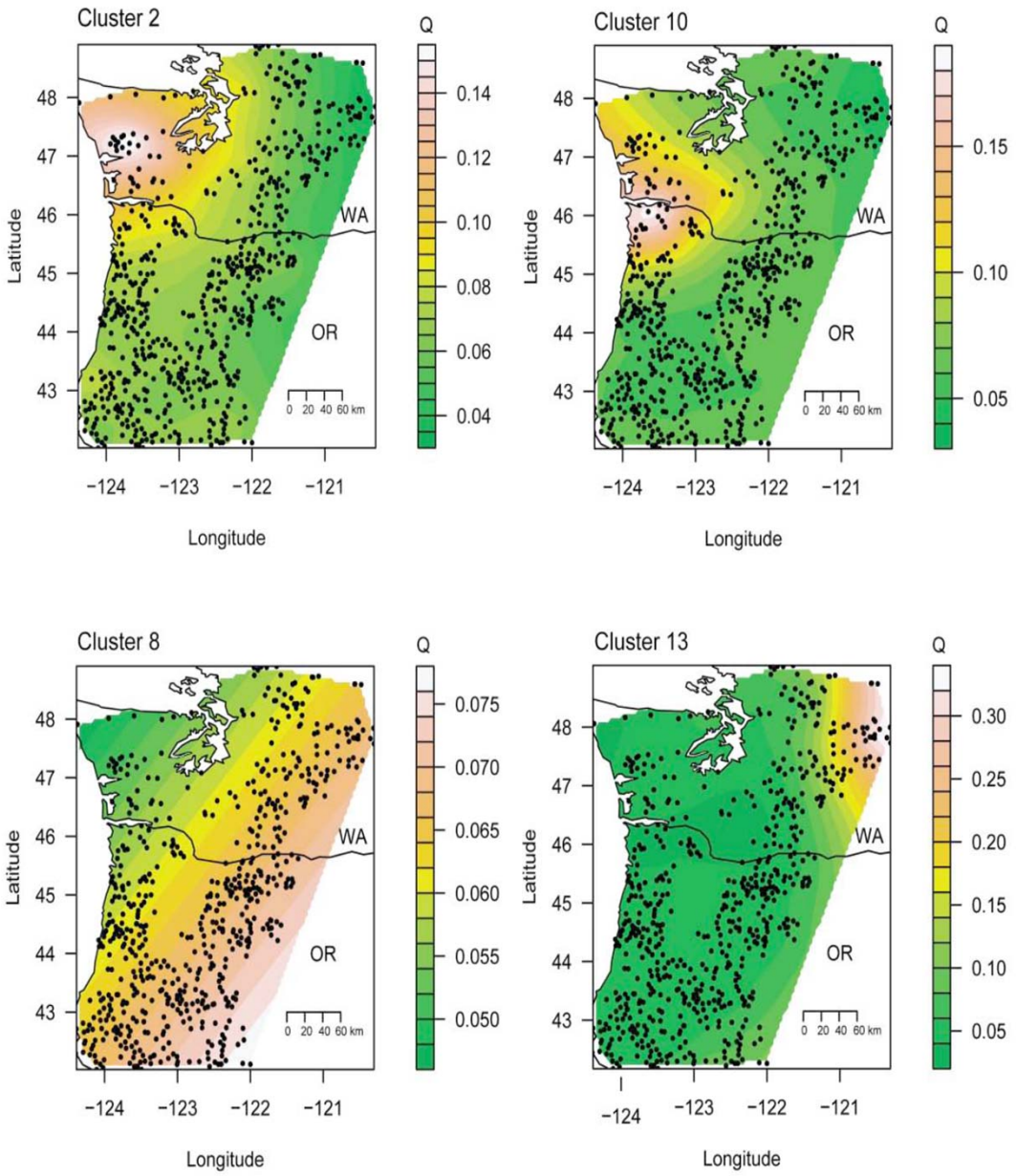


***K*=18**



Krutovsky, K. V., J. B. St.Clair, R. Saich, V. D. Hipkins and D. B. Neale, 2009 Estimation of population structure in coastal Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco var. *menziesii*] using allozyme and microsatellite markers. *Tree Genetics and Genomes* **5**(4): 641–658.

Landscape / Spatial Genomics of coastal Douglas-fir



Dots mark sampled trees (N = 700)

- 18 population clusters were found using STRUCTURE program
- Geographic trends in population structure were observed.
- For each of the 18 clusters, *Q*-values were smoothed with universal Kriging interpolation.
- Two patterns are apparent – a southwest to northeast trend (Clusters 8 and 13) and a cluster centered on the coast of Washington (Clusters 2 and 10).

Eckert, A.J., A.D. Bower, J.L. Wegrzyn, B. Pande, K.D. Jermstad, K.V. Krutovsky, J.B. St. Clair and D.B. Neale, 2009 Association genetics of coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* 182: 1289-1302

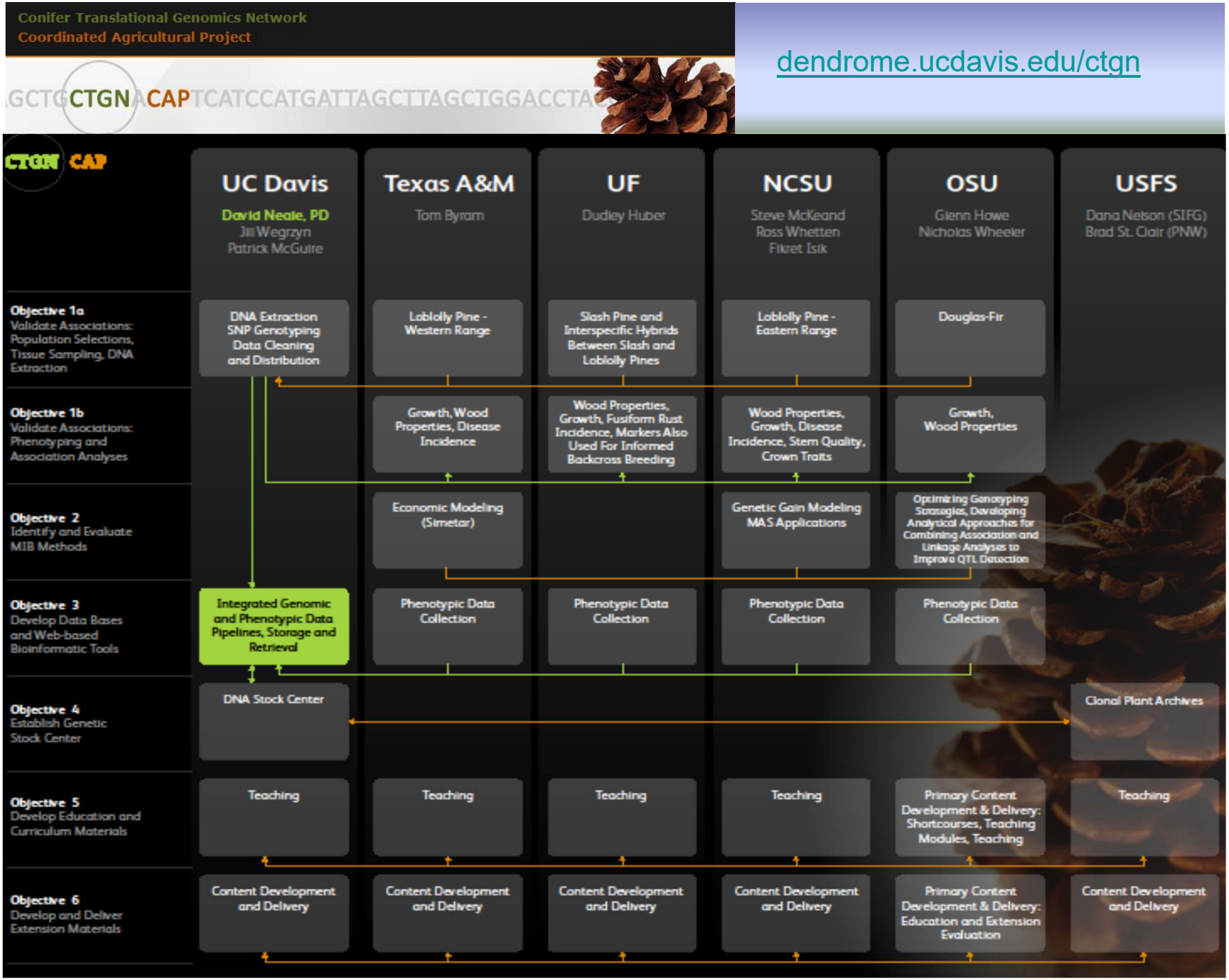
Conifer Translational Genomics Network (CTGN)

- CTGN is a multi-state, multi-institution Coordinated Agricultural Project (CAP) funded by USDA and USDA Forest Service; $5.9M (2007-2011)
- Project goal: Bringing Genomic Assisted Breeding to Application in Tree Improvement by linking experimental research with tree breeding

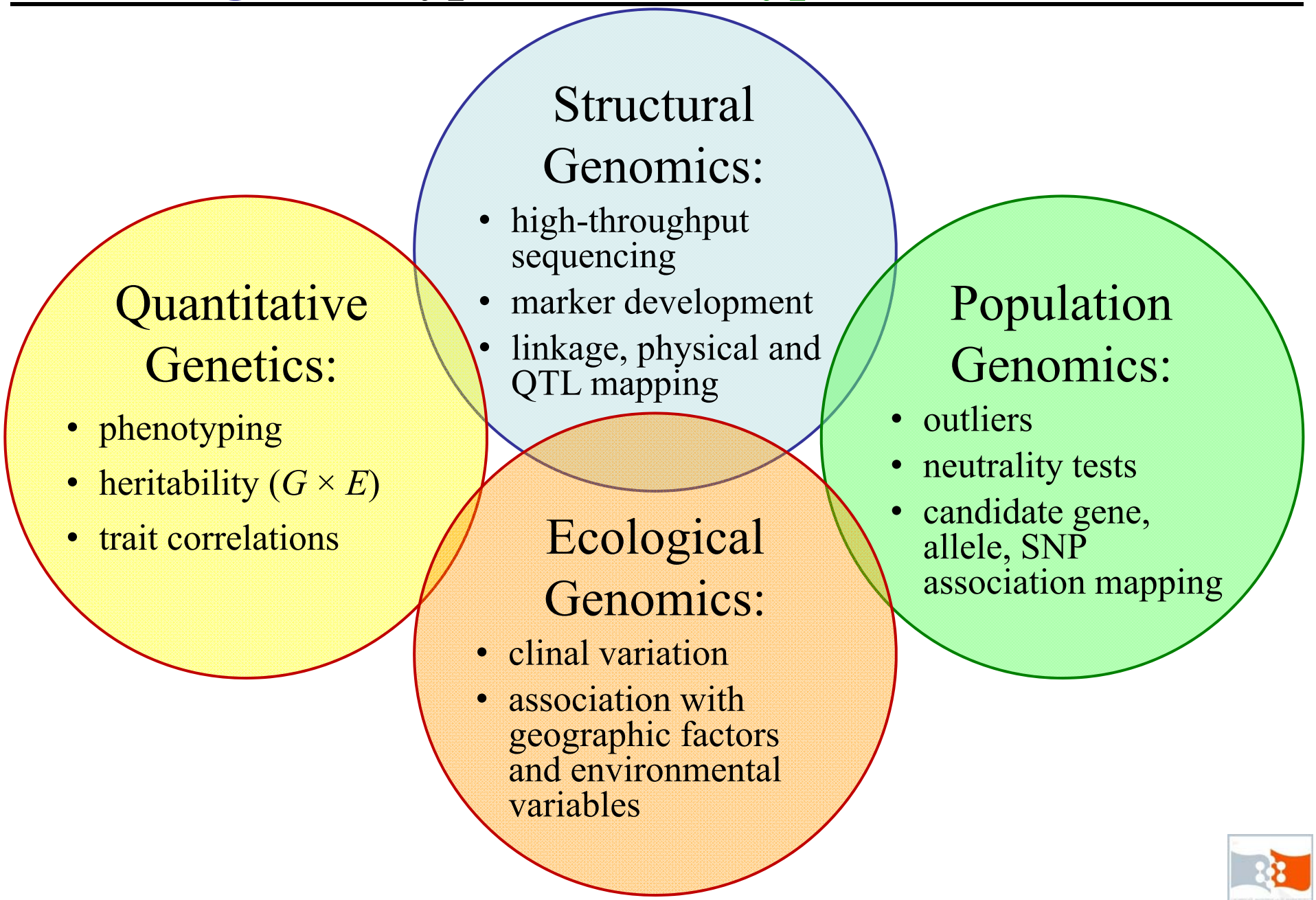
- Project Approach:
  - large scale genotyping in elite loblolly pine and Douglas-fir populations belonging to tree improvement cooperatives
  - validating genetic marker / phenotypic trait associations
  - modeling, outlining and implementing optimal approaches for incorporating markers in breeding programs

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

www.pinegenome.org/ctgn

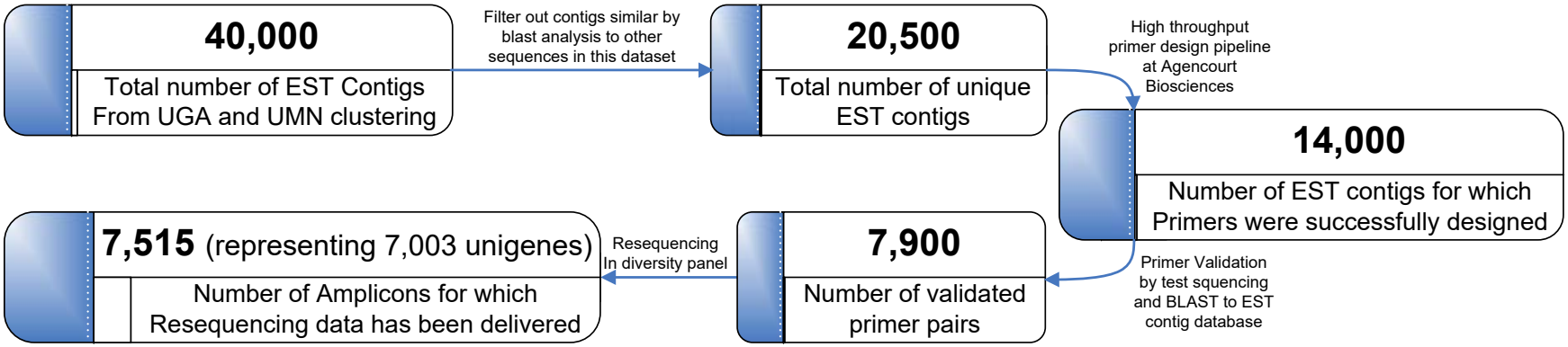

Conifer Translational Genomics Network
Coordinated Agricultural Project
CTGN CAP
dendrome.ucdavis.edu/ctgn
UC Davis
David Neale, PD
Jill Wegrzyn
Patrick McGuire
Texas A&M
Tom Byram
UF
Dudley Huber
NCSU
Steve McKeand
Ross Whetten
Fikret Isik
OSU
Glenn Howe
Nicholas Wheeler
USFS
Dana Nelson (SIFG)
Brad St. Clair (PNW)
Objective 1a
Validate Associations: Population Selections, Tissue Sampling, DNA Extraction
DNA Extraction SNP Genotyping Data Cleaning and Distribution
Loblolly Pine - Western Range
Slash Pine and Interspecific Hybrids Between Slash and Loblolly Pines
Loblolly Pine - Eastern Range
Douglas-Fir
Objective 1b
Validate Associations: Phenotyping and Association Analyses
Growth, Wood Properties, Disease Incidence
Wood Properties, Growth, Fusiform Rust Incidence, Markers Also Used For Informed Backcross Breeding
Wood Properties, Growth, Disease Incidence, Stem Quality, Crown Traits
Growth, Wood Properties
Objective 2
Identify and Evaluate MIB Methods
Economic Modeling (Simetar)
Genetic Gain Modeling MAS Applications
Optimizing Genotyping Strategies, Developing Analytical Approaches for Combining Association and Linkage Analyses to Improve QTL Detection
Objective 3
Develop Data Bases and Web-based Bioinformatic Tools
Integrated Genomic and Phenotypic Data Pipelines, Storage and Retrieval
Phenotypic Data Collection
Phenotypic Data Collection
Phenotypic Data Collection
Phenotypic Data Collection
Objective 4
Establish Genetic Stock Center
DNA Stock Center
Clonal Plant Archives
Objective 5
Develop Education and Curriculum Materials
Teaching
Teaching
Teaching
Teaching
Primary Content Development & Delivery: Shortcourses, Teaching Modules, Teaching
Teaching
Objective 6
Develop and Deliver Extension Materials
Content Development and Delivery
Content Development and Delivery
Content Development and Delivery
Content Development and Delivery
Primary Content Development & Delivery: Education and Extension Evaluation
Content Development and Delivery

Linking Genotype to Phenotype & Environment



Sequencing & SNP Genotyping

NSF Allele Discovery of Economic Pine Traits 2 (ADEPT2) resequencing & SNP discovery project



~23,000 SNPs discovered in ~7,000 partly amplified unique genes sequenced in 18 loblolly pine haploid megagametophytes

USDA Conifer Translational Genomics Network (CTGN) project

5,379 SNPs were genotyped in multiple association and breeding populations using Illumina Infinium platform;

4,264 SNPs were polymorphic in East Texas populations

Nucleotide Diversity in Loblolly pine

- 5,772 genes
- SNP discovery based on 18 trees sampled (TX to VA)
- SNPs were used for inferences of:
  (1) selection
  (2) genotype-phenotype correlations

SNP type	#SNPs	θ	π	D_{xy}
All	22,620	0.004	0.003	0.007
Silent	7,676	0.006	0.005	0.010
Syn	3,233	0.007	0.006	0.012
Nonsyn	2,914	0.002	0.001	0.003

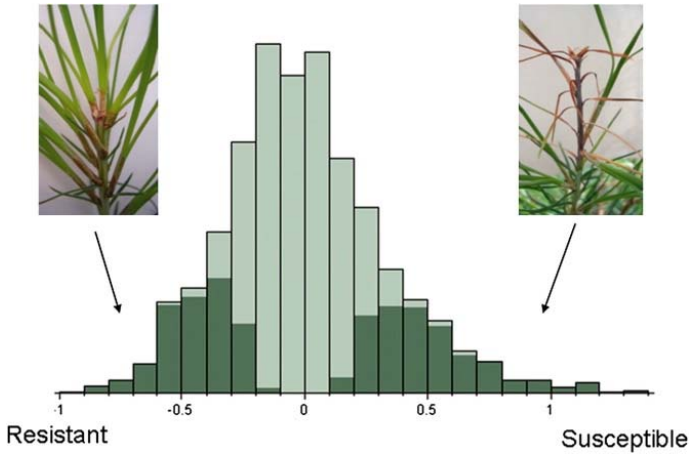
Eckert et al. *in prep.*

Ecological & Population genomic data for loblolly pine (*Pinus taeda* L.)

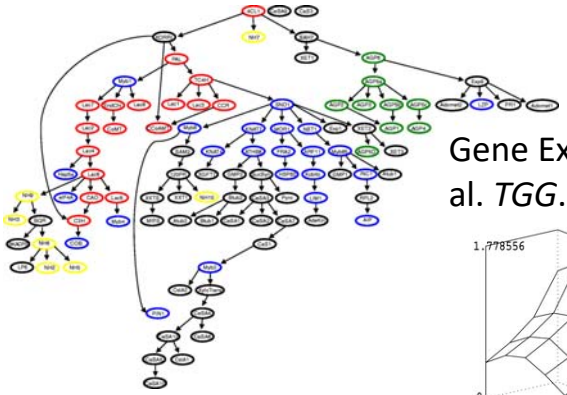
- SNPs in 5,772 genes were genotyped in >4500 trees sampled from numerous natural and breeding populations covering the full-range of the species
- Allelic correlations with geography, temperature, growing degree-days, precipitation and aridity.
- Significant associations represented a diverse sets of genes including abiotic stress response genes ranging from trans-membrane proteins to proteins involved in sugar metabolism and transcription factors
- Numerous SNP outliers
- Multiple allele candidates for local adaptation.

Eckert *et al*. 2010 *Genetics* 185: 969–982;
Eckert *et al*. 2010 *Molecular Ecology* 19: 3789–3805
Chhatre *et al*. 2013 *Tree Genetics and Genomes* 9: 1161–1178

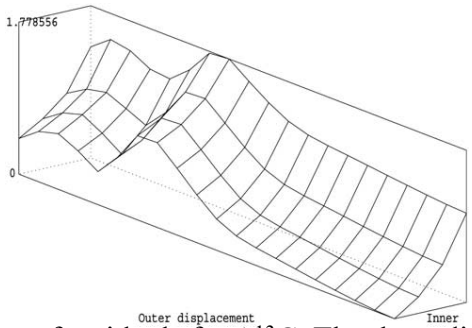
Loblolly pine - Associations



Disease resistance (Quesada et al. *Genetics*. 2010)



Gene Expression (Palle et al. *TGG*. 2010)



Drought-tolerance (Cumbie et al. *Heredity.* 2011)

Variogram plot of residuals for $\Delta^{13}C$. The three dimensional plot reflects the similarity of trees because of environmental trends in a row and column spatial configuration.

- Phenotypes:
  - gene expression
  - metabolome
  - wood properties
  - drought-tolerance
  - disease resistance
- Genotypes:
  - 7,216 SNPs (3,938)
- Associations:
  - 1,020 total
  - 842 unique SNPs
  - 119 traits

Association mapping of SNPs and phenotypic variation in adaptive and breeding traits (such as growth rate, wood density, disease resistance, drougth tolerance, etc.)

Significant associations were found, for example, for:

- Arabinofuranosidase
- Xylosidase
- Protein kinases
- Chloroplast proteins
- Metallothionein-like protein
- Chlorophyll binding protein
- Glucuronase 4-epimerate
- Clavata-like receptor
- RNA polymerases
- Decarboxylases
- Sodium simporter family protein
- Acyl CoA synthetase
- Tubulin beta-chains
- NBS disease resist. protein - P. taeda Transmembrane protein
- Universal stress protein
- Cyclin-D like protein
- cdc2 protein kinases
- Synaptotagmins etc.

Ecological Genomics Questions

- What are some of the genes that putatively underlie ecologically relevant traits?
- Are these genes also correlated to environmental variables?
- Are genes that harboring an association under selection?

Environmental data

- **Climate data** were gathered from the WORLDCLIM 2.5 minute geographical information system (GIS) layer using Diva-GIS ver. 5.4 (Hijimans *et al.* 2005; http://www.diva-gis.org/).

- Monthly minimum and maximum **temperatures**, monthly **precipitation** and 19 **bioclimatic variables** were obtained from this layer.

- The temperature and precipitation data were used to estimate potential **evapotranspiration** (PET) with the method of Thornthwaite (1948).

- An **aridity** index (AI) was defined as the ratio of precipitation to PET (File S1), with this ratio being defined quarterly. Annual quarters were defined starting with January 1st through March 31st as quarter one and are labeled as AI1 through AI4. We focus on aridity because it encapsulates water availability as a function of temperature and precipitation.

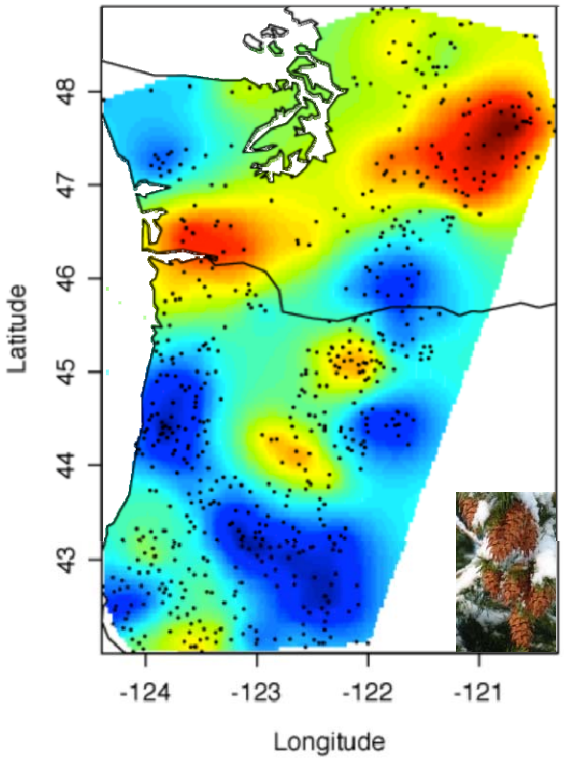
(Eckert *et al*. 2010 *Genetics*, 185: 969–982)

- **Multivariate measures** of climate corresponding to geography, temperature, growing degree-days, precipitation and aridity.

(Eckert *et al*. 2010 *Molecular Ecology* 19: 3789–3805)

Environmental Correlations

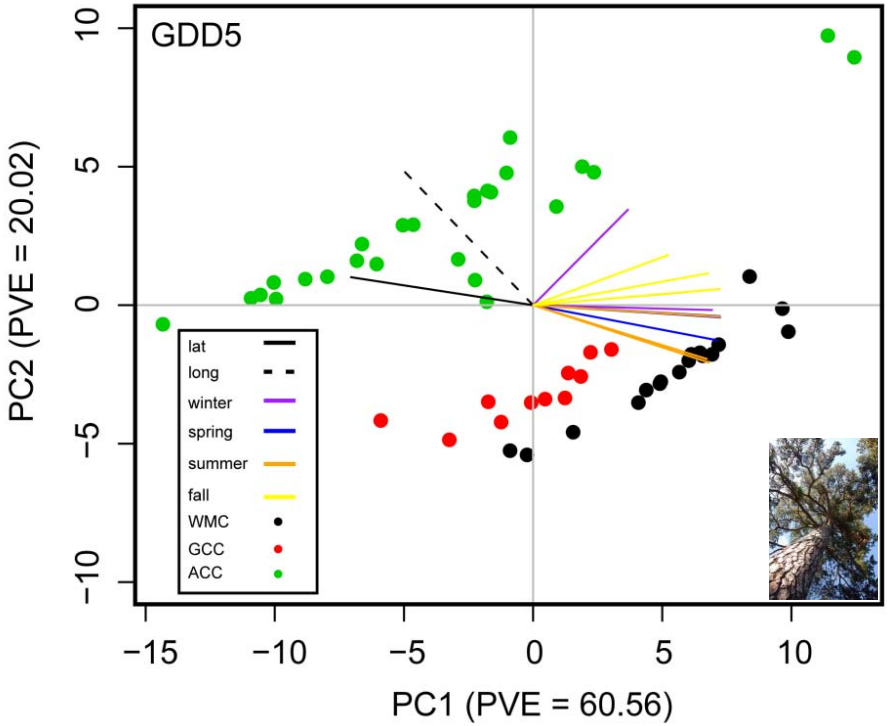
Douglas-fir



Neale and Kremer. *Nat. Rev. Genet.* (2011).

Population structure is not well-correlated with environment or geography

Loblolly pine



Eckert et al. *Mol. Ecol.* (2010b).
Eckert et al. *Genetics.* (2010a).

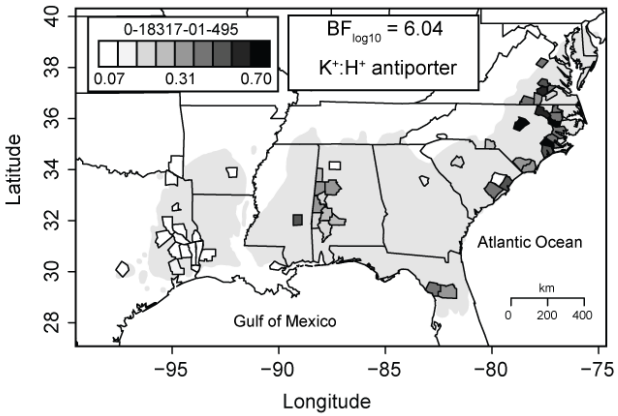
Population structure is correlated with environment or geography

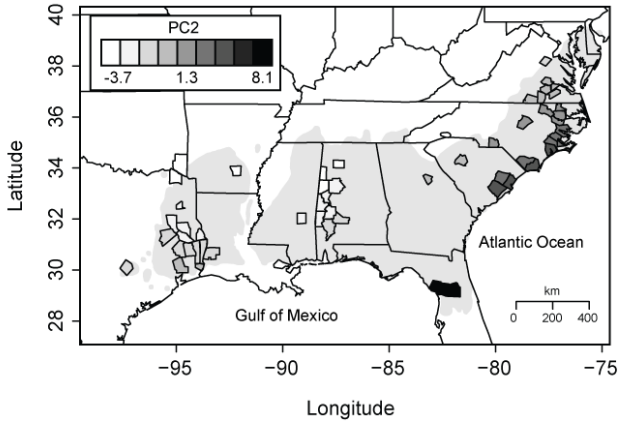
Environmental Correlations

Douglas-fir

- 18 highly correlated SNPs
- 14 are associated

Loblolly pine



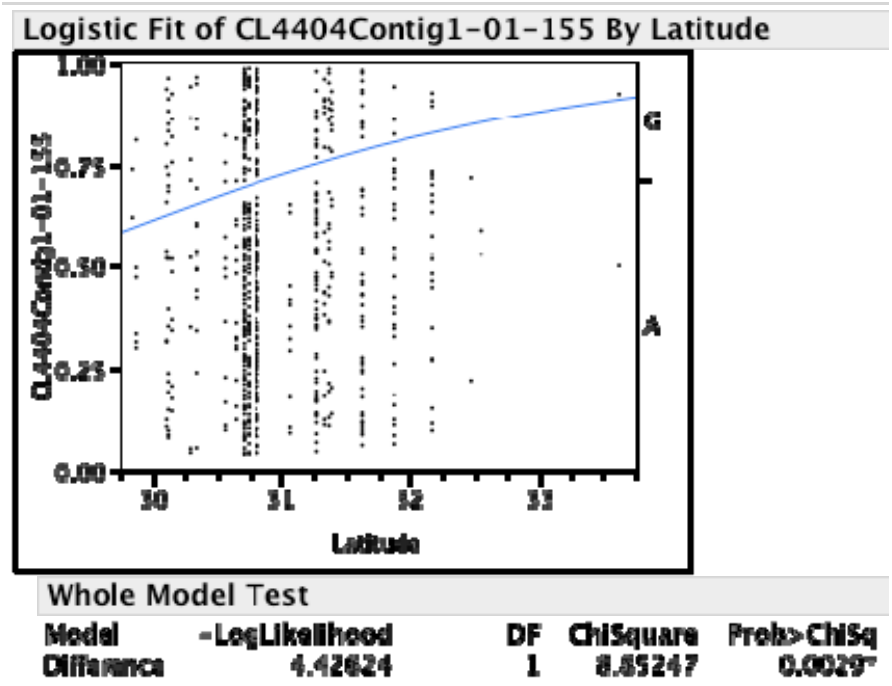


- 48 highly correlated SNPs
- 12 are associated

Eckert et al. *Mol. Ecol.* (2010b).

Clinal variation: Logistic regression (LR) data

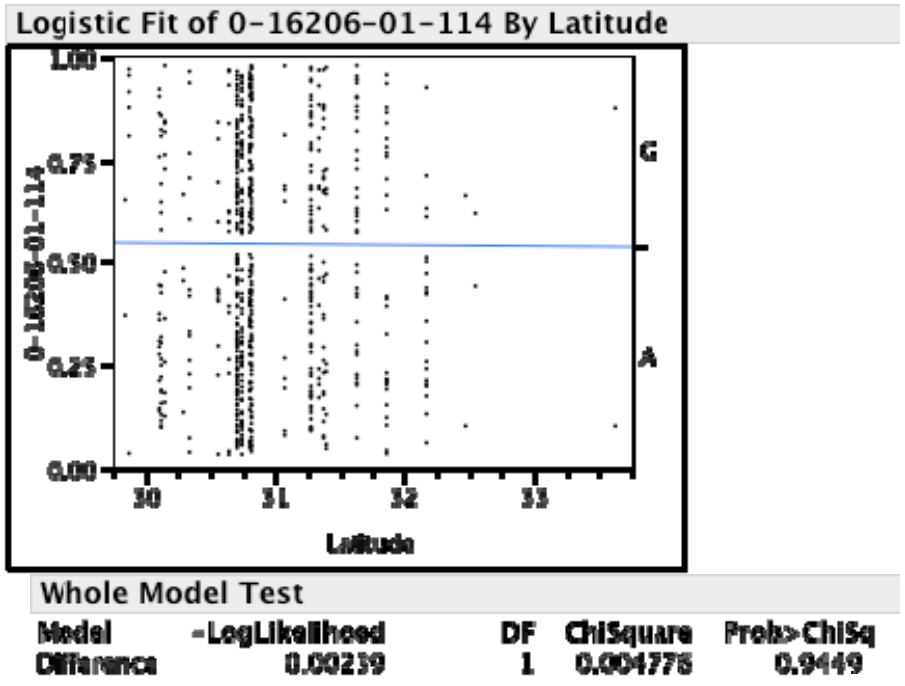
Significant association of the A/G alleles of the CL4404Contig1-01-155 SNP with Latitude:



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	4.42624	1	8.85247	0.0029*

Insignificant association of the A/G alleles of the 0-16206-01-114 SNP with Latitude:

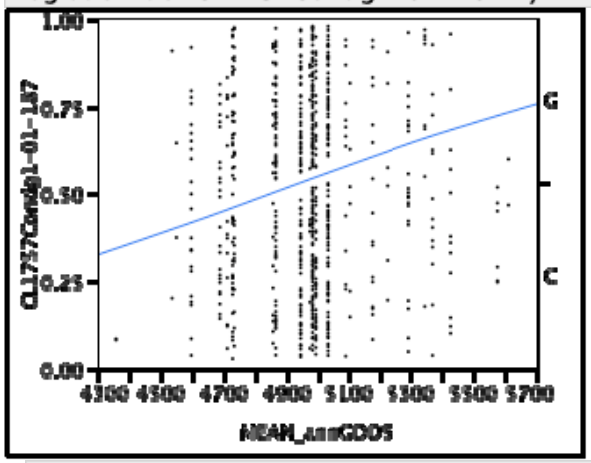


Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.00239	1	0.004776	0.9449

Association with environmental variables: LR data

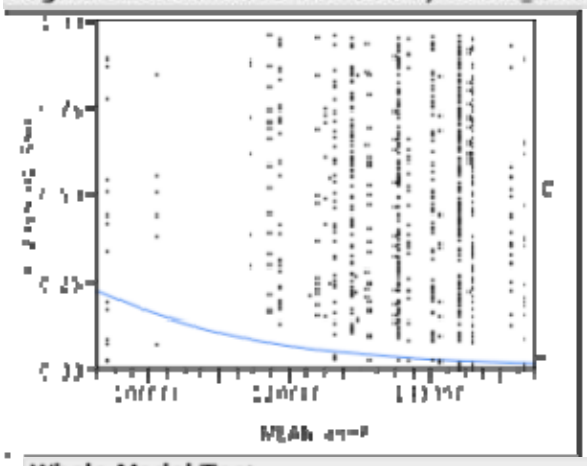
Logistic Fit of CL1757Contig1-01-187 By MEAN_annGDD5



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	5.44054	1	10.88108	0.0010*

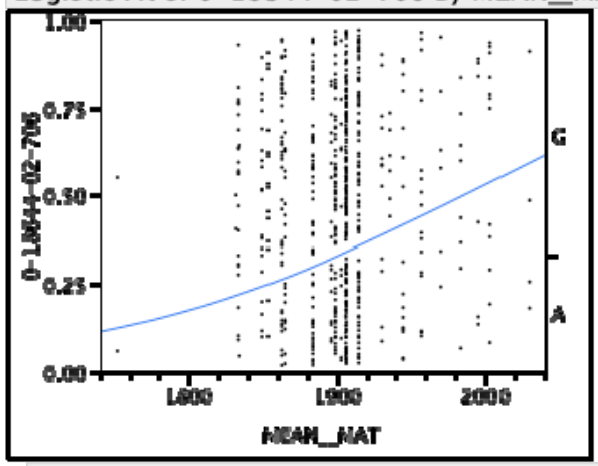
Logistic Fit of 0-2039-01-580 By MEAN_annP



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	[illegible]	1	[illegible]	[illegible]

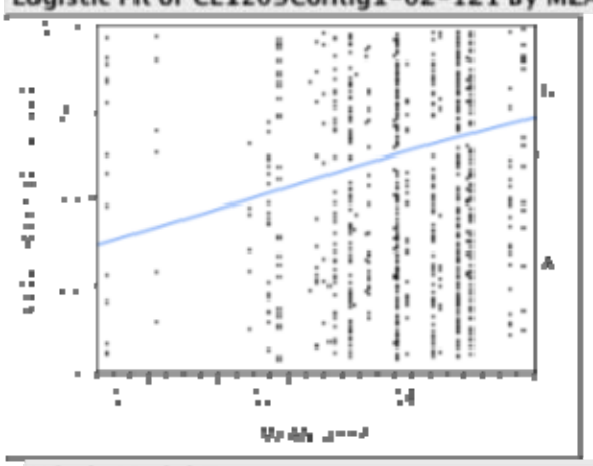
Logistic Fit of 0-18644-02-706 By MEAN_MAT



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.40695	1	12.81391	0.0003*

Logistic Fit of CL1209Contig1-02-121 By MEAN_annP

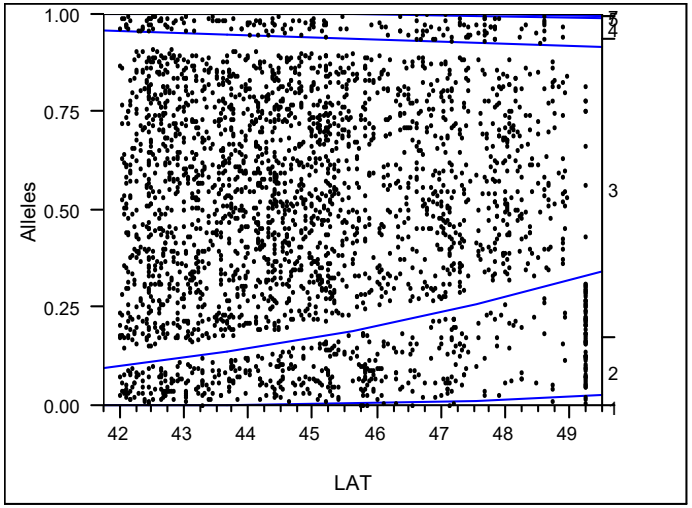


Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	[illegible]	[illegible]	[illegible]	[illegible]

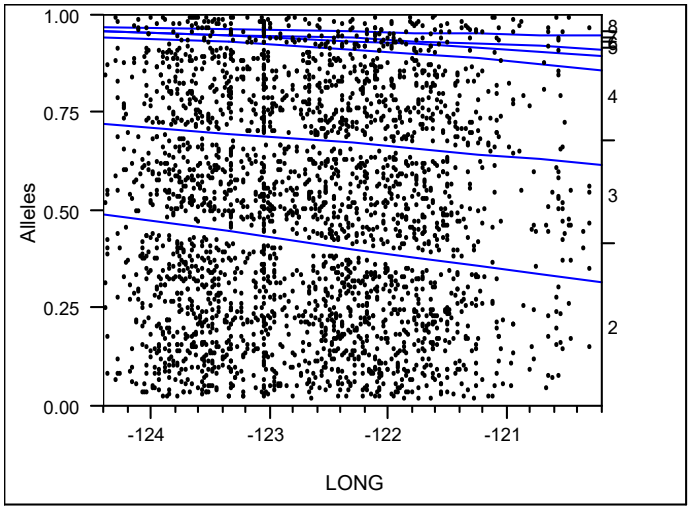
Clinal allelic variation (logistic regression analysis)

Logistic fit of PGM-1 alleles by LATITUDE



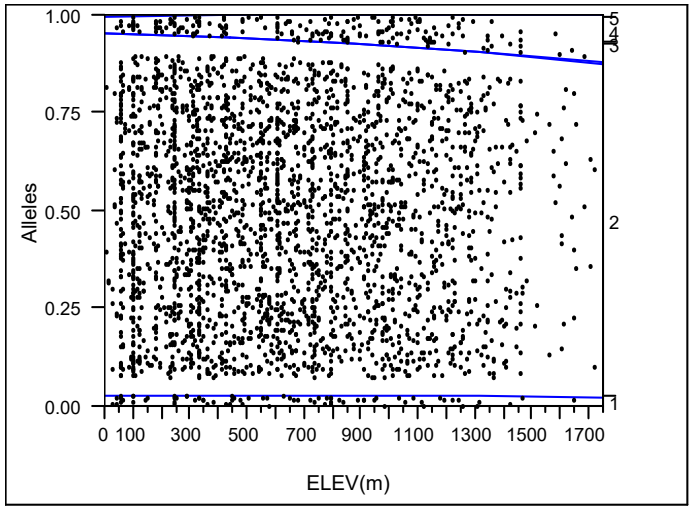
Whole Model Test: LL=43.5; df=5; χ^2 = 87.07; *P* <0.001

Logistic fit of LAP-1 alleles by LONGITUDE



LL=12.3; df=6; χ^2 = 24.51; *P* <0.001

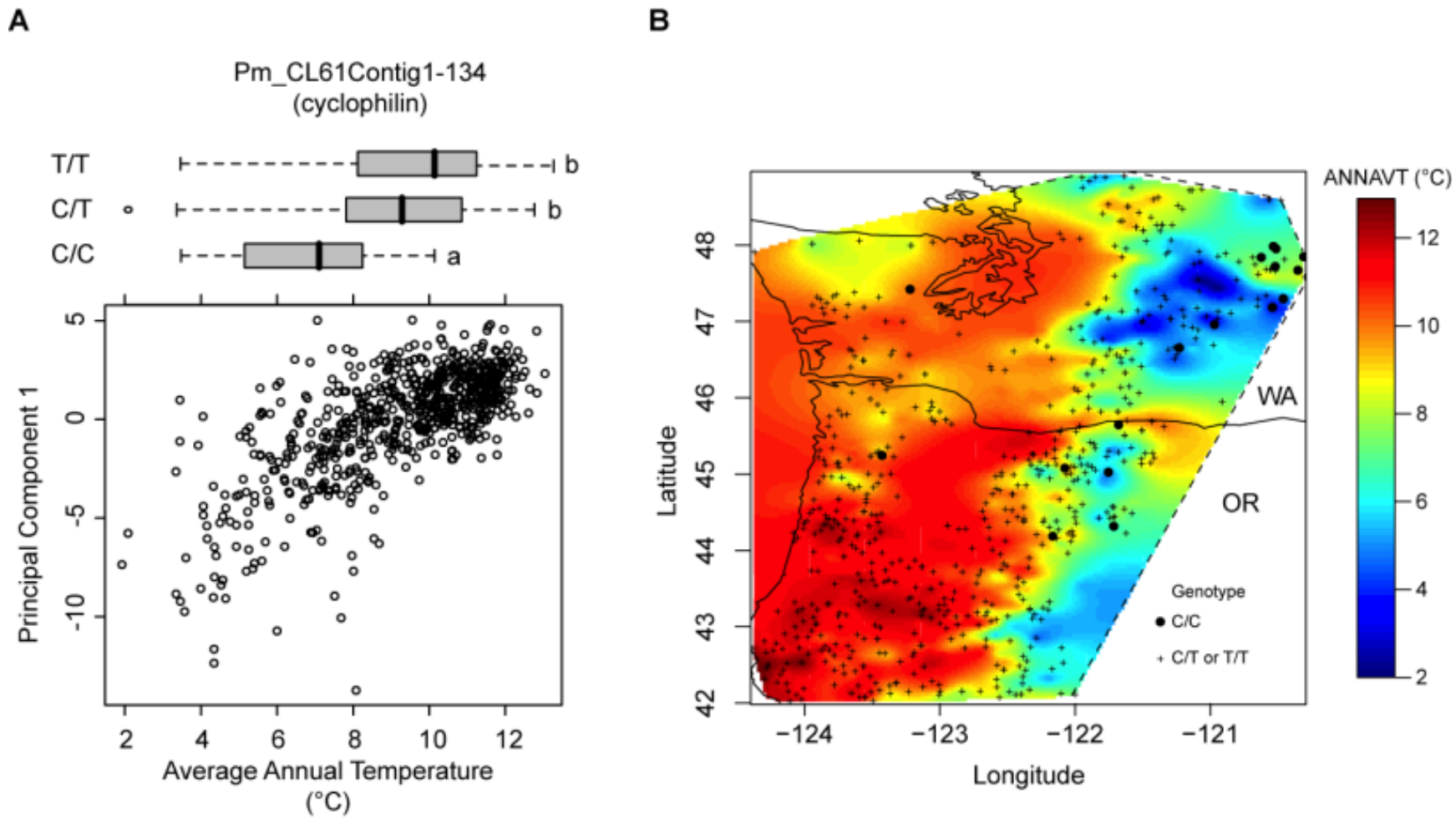
Logistic fit of 6-PGD alleles by ELEVATION (*m*)



LL=7.4; df=4; χ^2 = 14.81; *P* <0.005

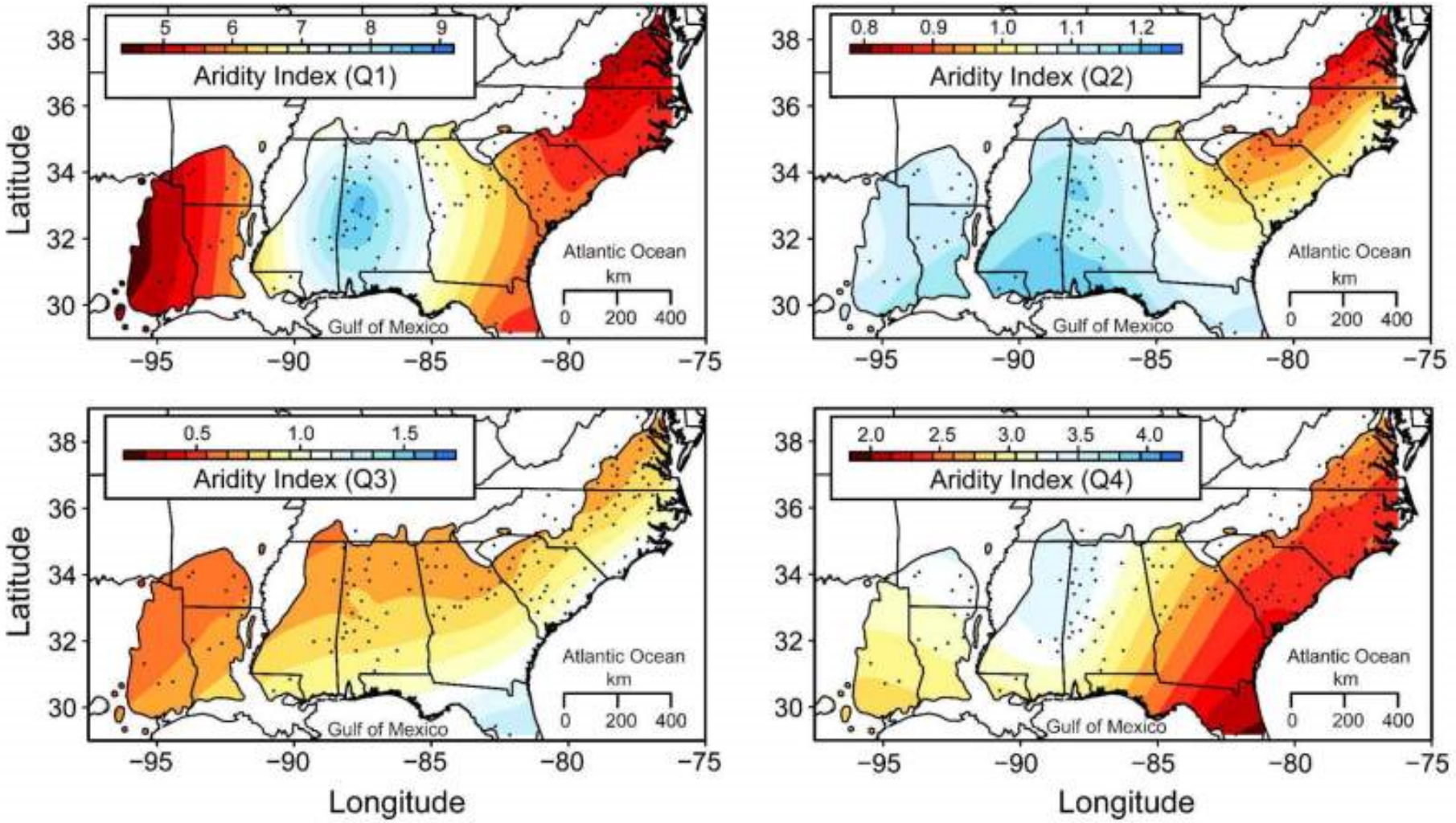
Krutovsky *et al.* (2009) *Tree Genetics and Genomes* 5: 641–658

Marker Effects along Environmental Gradients



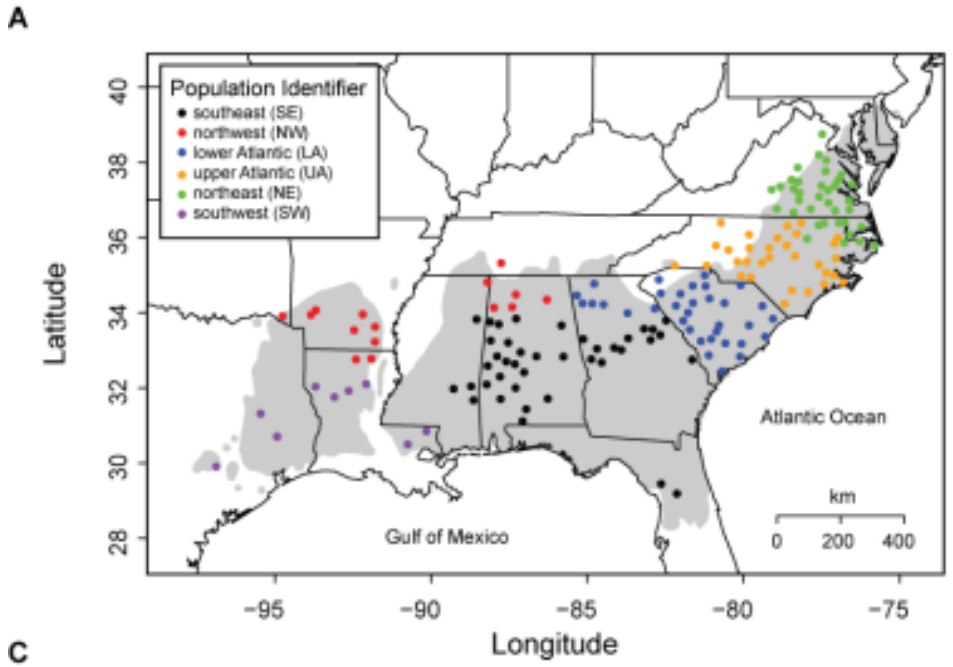
Eckert, A.J., A.D. Bower, J.L. Wegrzyn, B. Pande, K.D. Jermstad, K.V. Krutovsky, J.B. St. Clair and D.B. Neale, 2009 Association genetics of coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* 182: 1289-1302

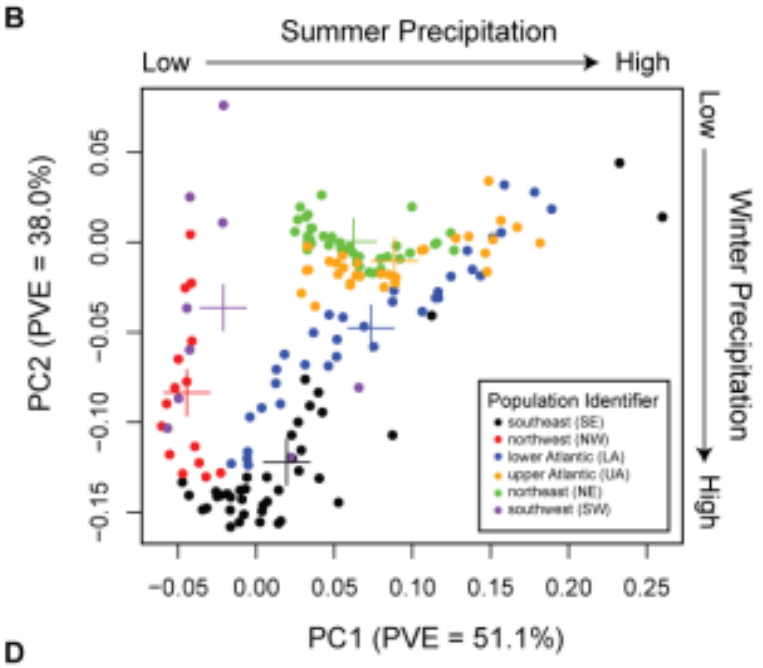
The distribution of loblolly pine, sampling localities (black points) and aridity gradients (color gradients) used to assess patterns of population structure and environmental associations

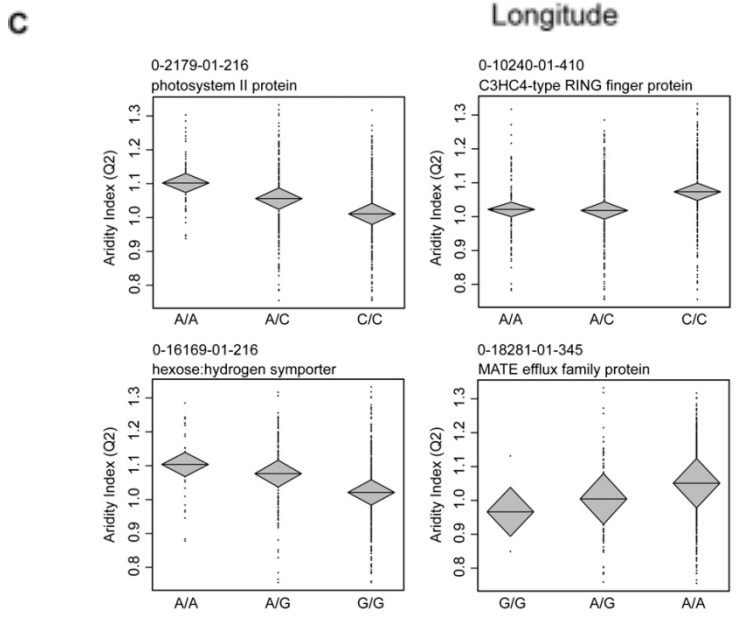


- Aridity gradients are shown by annual quarter (Q1: Jan-Mar, Q2: Apr-Jun, Q3: Jul-Sep, Q4: Oct-Dec). (Eckert *et al*. 2010 *Genetics*, 185: 969–982)

Adaptive population and gene differentiation







Example of significant genotypic associations for four loci with the aridity index during the second quarter (Q2 season)

(Eckert *et al*. 2010 *Genetics*, 185: 969–982)

Clinal variation and association with environmental variables : Logistic regression (LR) data

Number of SNPs genotyped in 463 trees from 27 populations in East Texas:

- total = **5379**
- polymorphic = **4264**
- used for LR = **3667**

Clinal variation - significant correlation with:

- Latitude = **210**
- Longitude = **293**
- both latitude and longitude = **34**

Environmental variables- significant correlation with:

- Monthly mean total annual temperature above 5˚C or Growing Degree Days (MEAN_annGDD5)* = **245**
- Mean Annual Precipitation (MEAN_annP)** = **268**
- Mean Annual Temperature (MEAN_MAT)*** = **259**
- Aridity index = *in progress*

* Monthly temperature estimates were averaged for 1971-2000 years, threshold temperature of 5˚C was subtracted, multiplied by number of days in the month, and then added up for all 12 months to get annual estimate following Rehfeldt (2006) and Eckert *et al*. (2010).

** Monthly precipitation estimates in *mm* were averaged for 1971-2000 years, rounded to two decimal points, multiplied by 100, and then added up for all 12 months to get annual estimate following Eckert *et al*. (2010).

*** Monthly minimal and maximum temperature estimates were averaged for 1971-2000 years, rounded to two decimal points, multiplied by 100, and then added up for all 12 months to get annual estimate following Eckert *et al*. (2010).

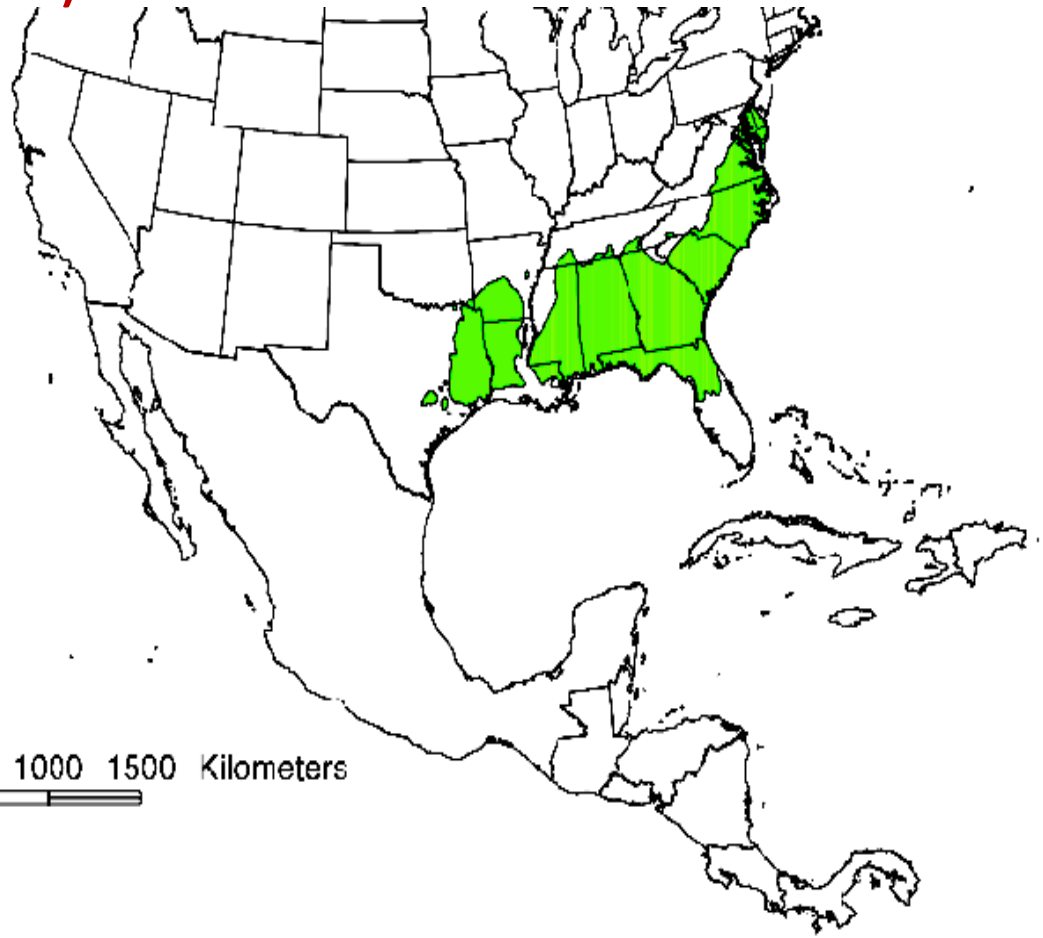
PINEMAP project:

Pine Integrated Network: Education, Mitigation and Adaptation Project

Subproject:

Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.)

Loblolly pine range



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

USGS

Objectives

Exome genotyping by sequencing (GBS) & SNP discovery

- Exome-wide genotyping by sequencing of 375 clonally propagated and phenotyped trees representing the entire loblolly pine and identify SNPs

Lu, M., **K. V. Krutovsky**, C.D. Nelson, T. E. Koralewski, T. D. Byram, and C. A. Loopstra, 2016 Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). *BMC Genomics* 17:730 (http://www.biomedcentral.com/content/pdf/s12864-016-3081-8.pdf)

Exome-wide Association mapping to discover the genes with adaptive variation

- Associate the genotyped SNPs with growth and adaptive traits

Lu, M., **K. V. Krutovsky**, C.D. Nelson, J. B. West, N. A. Reilly, and C. A. Loopstra, 2017 Association genetics of adaptive traits in a clonally tested loblolly pine (*Pinus taeda* L.) population. *Tree Genetics and Genomes* 13(3): 57 (http://link.springer.com/article/10.1007/s11295-017-1140-1)

- Associate the genotyped SNPs with gene expression (transcripts) and metabolite level

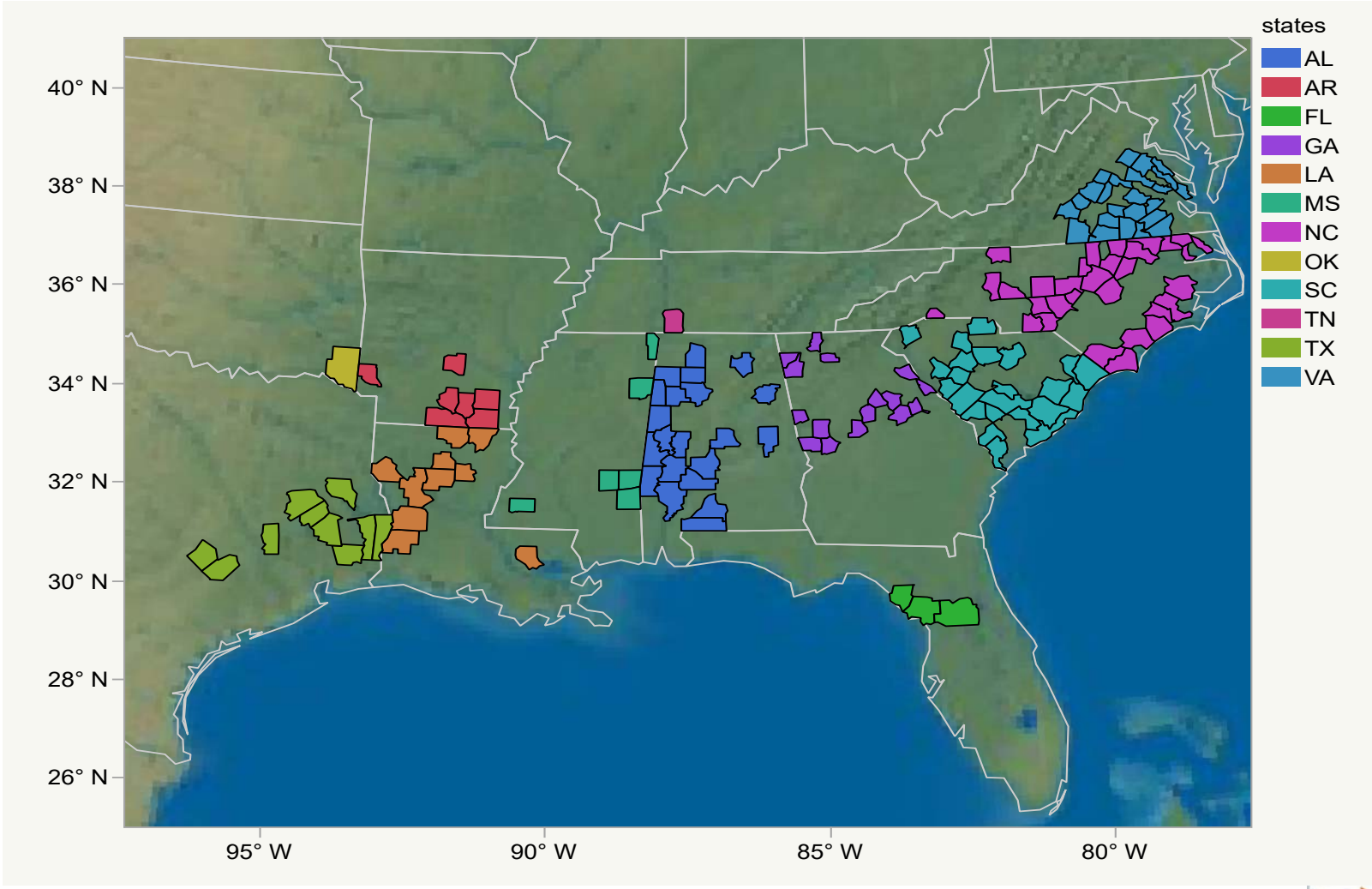
Lu, M., C. M. Seeve, **K. V. Krutovsky**, and C. A. Loopstra, 2018 Exploring the genetic basis of gene transcript abundance and metabolite level in loblolly pine (*Pinus taeda* L.) using association mapping and network construction. *BMC Genetics* 19:100 (https://doi.org/10.1186/s12863-018-0687-7)

- Associate the genotyped SNPs with environmental variables

Lu, M., **K. V. Krutovsky**, and C. A. Loopstra, 2019 Detecting the genetic basis of local adaptation in loblolly pine (*Pinus taeda* L.) using whole exome-wide genotyping and an integrative landscape genomics analysis approach. *Ecology and Evolution* 9(12): 6798-6809 (https://doi.org/10.1002/ece3.5225)

Plant material

The counties of origin of the maternal trees colored by states



Plant material

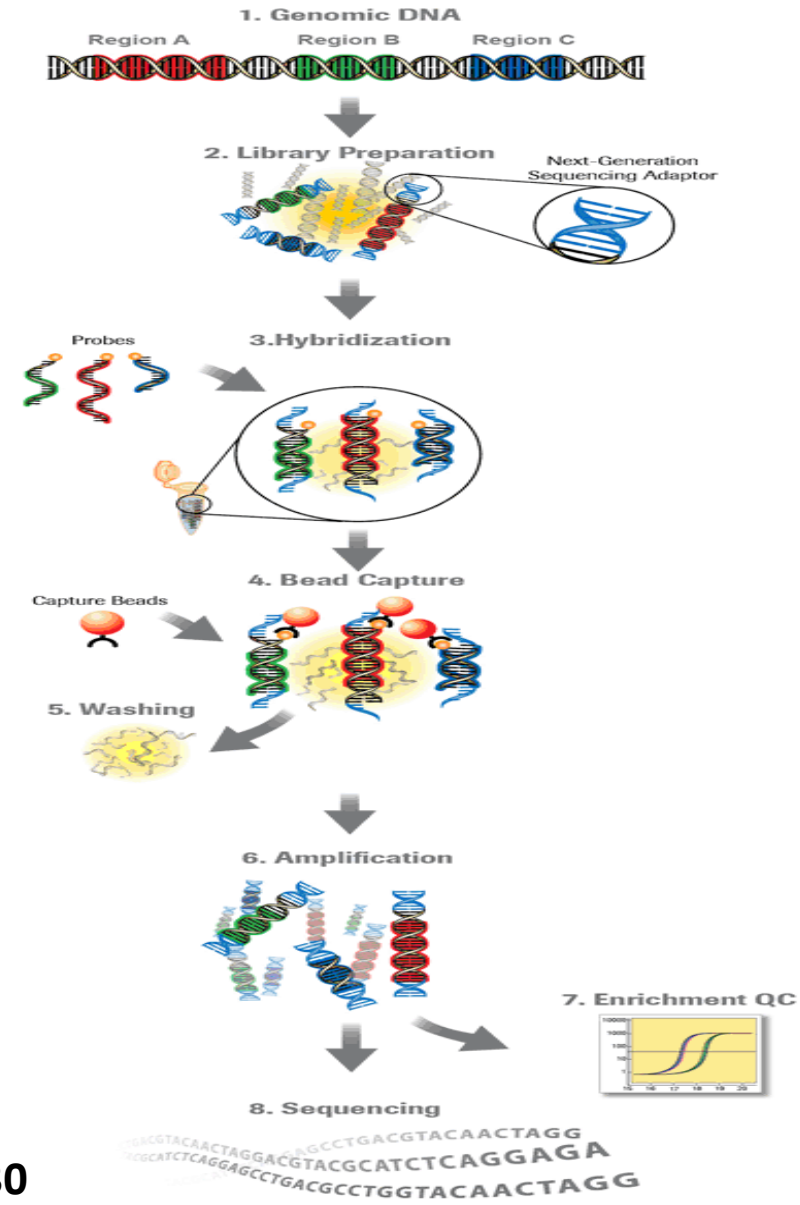
384 × 3 unrelated trees were planted

Lu *et al*. (2016) *BMC Genomics* 17:730

Exome enrichment and sequencing

Workflow of the NimbleGen SeqCap EZ system

http://www.nimblegen.com/products/seqcap/ez/choice/index.html



Lu *et al*. (2016) *BMC Genomics* 17:730

Exome enrichment and sequencing

Capture probes

Design	Nimblegen Probes
Exons partitioned from reference assembly v.1.01	199,723 exon regions (≈ 49 Mbp) in 48,391 high quality tentative genes
Total probes number	≈ 2.1 million
Target Region Size mapped on reference assembly v1.01	≈ 46 Mbp
Oligonucleotide probes	55~105 bp DNA single strands

Lu *et al*. (2016) *BMC Genomics* 17:730

SNP discovery

Bi-allelic sites
10X sequencing depth in 90 % of the individuals
MAF ≥ 0.05

972,720 SNPs

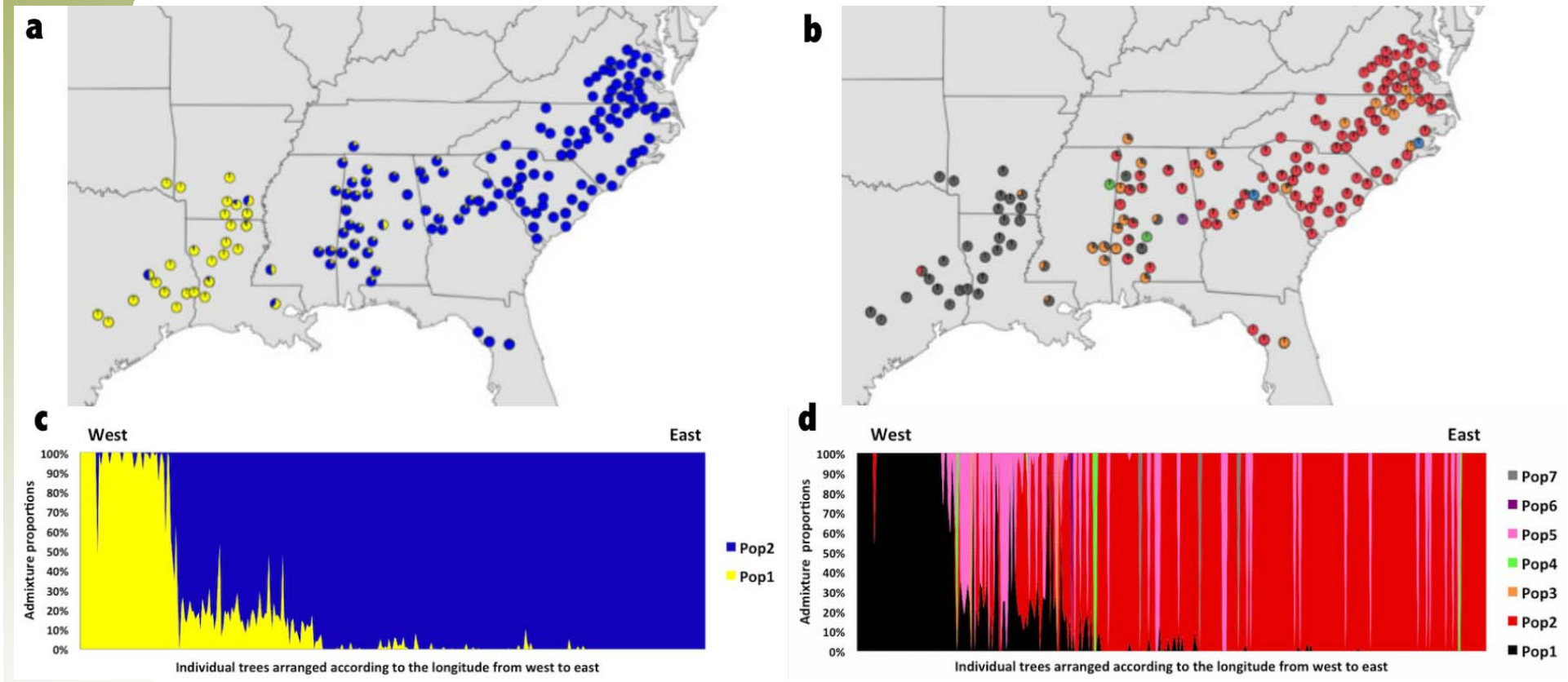
Bi-allelic sites
5X sequencing depth for all individuals without missing data
MAF ≥ 0.01

2,822,609 SNPs

Lu *et al*. (2016) *BMC Genomics* 17:730

Genetic structure analysis

K = 2 and *K* = 7 were chosen to explain the population structure



Lu *et al*. (2016) *BMC Genomics* 17:730

Association of SNPs with adaptive and growth phenotypic traits

Growth traits:

- Total height (2014, 2015Before, 2015After)
- Diameter (DIA)

Crown structure traits:

- Specific leaf area (SLA)
- Crown width (CW)
- Branch angle (BA)

Physiology and resistance traits:

- Carbon isotope discrimination ($\Delta^{13}C$)
- Nitrogen concentration (N)
- Pitch canker disease resistance (Quesada et al., 2011)

Lu *et al*. (2016) *Tree Genetics and Genomes* (ms)

SNP-Trait associations

Traits	SNPs
Specific leaf area	5
Branch angle	2
Crown width	3
Stem diameter	4
Total height	9
Carbon isotope discrimination	4
Nitrogen concentration	2
Pitch canker resistance	7

Lu *et al*. (2016) *Tree Genetics and Genomes* (ms)

Epistasis

Traits	Number of SNP-SNP interaction
Branch angle	1
Crown width	2
Total height in 2014	2
Carbon isotope discrimination	2
Nitrogen concentration	1
Pitch canker resistance	3

Lu *et al*. (2016) *Tree Genetics and Genomes* (ms)

Statistical Models

- Estimation of confounding issues
  - population structure (*Q*-matrix)
  - family structure (*K*-matrix)

http://web.stanford.edu/group/pritchardlab/structure.html

- Model choice: general linear model, mixed linear model, multi-SNP linear models
- Multiple test corrections
- TASSEL software for GWAM

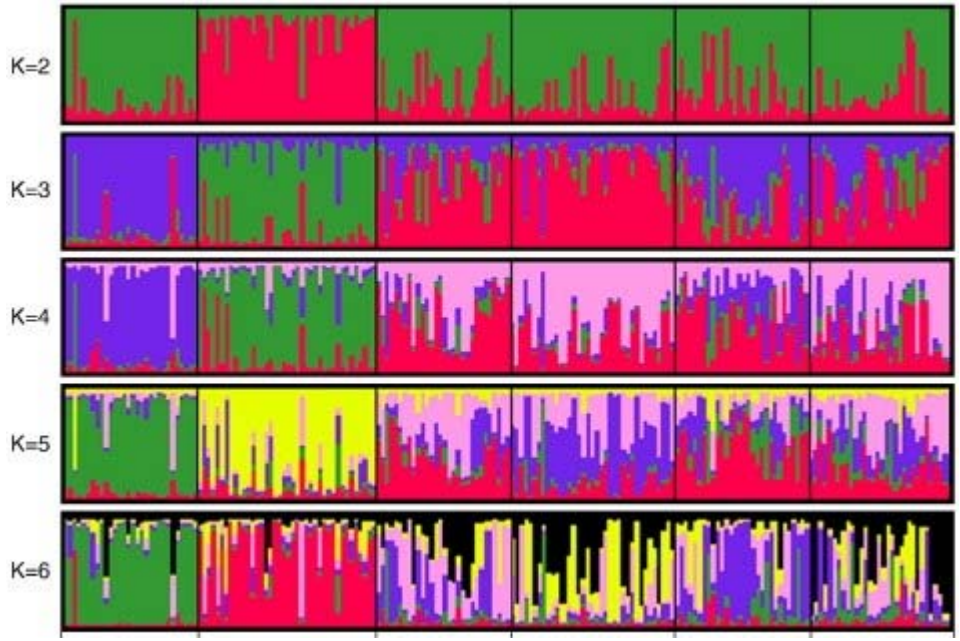
https://www.maizegenetics.net/tassel

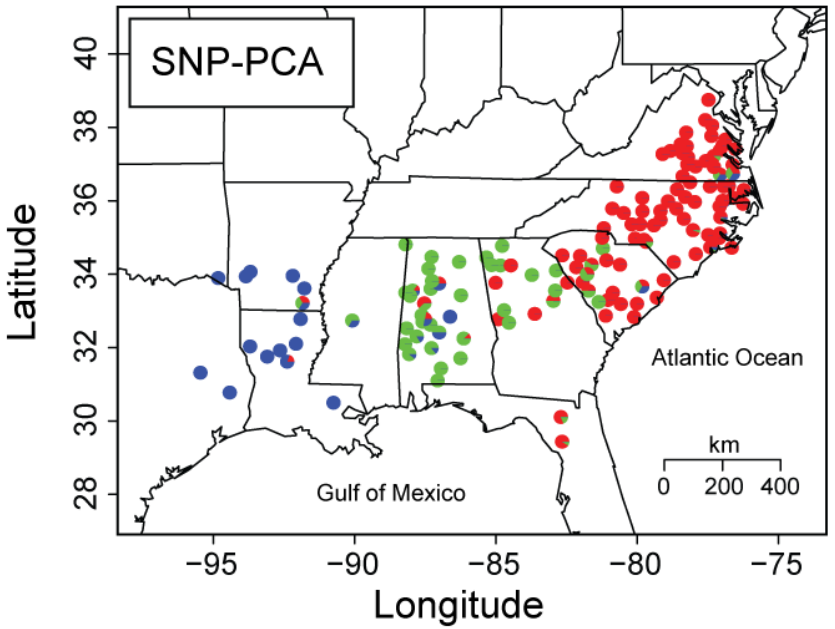
Population structure

- Allele frequencies of genetic markers vary among populations or lines due to genetic drift, historical processes or breeding design (selectively neutral factors that can create selectively neutral population structure).
- Most quantitative traits also vary among populations due to 1) selectively neutral factors and 2) natural selection.
- In some cases association of variation in allele frequencies of genetic markers with variation of quantitative traits can produce false positive correlations.
- Selectively neutral population structure should be estimated in the studied material and used as confounding factor in association analysis.

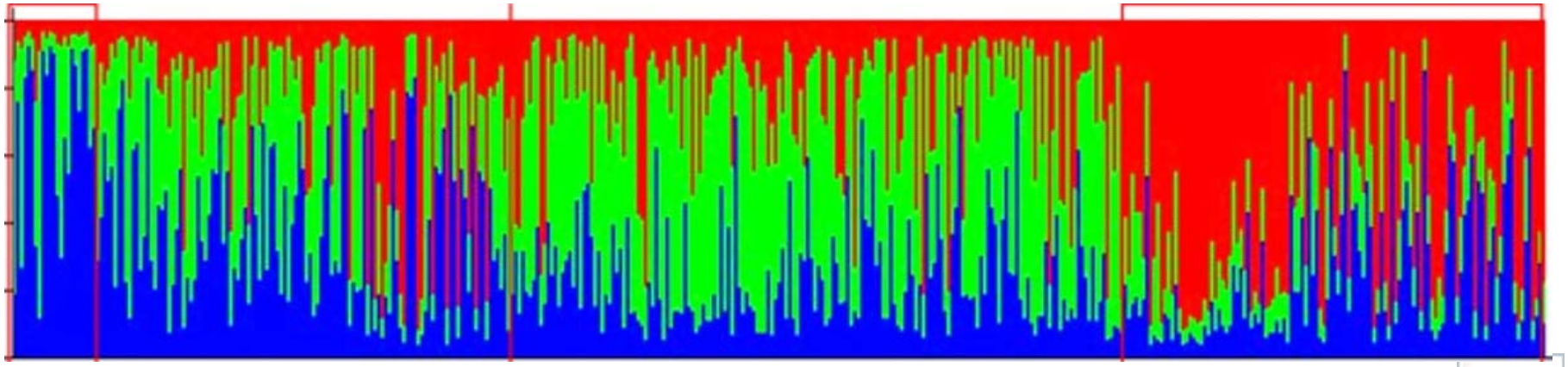
STRUCTURE software for inferring Population structure – Example with Loblolly pine populations

http://pritchardlab.stanford.edu/structure.html





***K*=3**



Trees (West → East)

Eckert et al. *Genetics.* (2010a)

Multiple Testing Problem

- Number of tests = number of markers × number of phenotypes = BIG NUMBER!
- How do we deal with false positives?
  - "Correct" our significance threshold (Bonferroni)
  - Base significance on probability of false discovery (FDR)
  - Permutation analysis – minimum *P*-value distributions
  - Use Bayesian inference
- How do we deal with false negatives?
  - Maximize power with study design
  - Adhere to limits of multiple test corrections (e.g. tests must be independent)

Ассоциативное картирование SNPs с изменчивостью средовых факторов: Ландшафтная геномика

Geographic variables (clinal variation):

- latitude
- longitude
- both latitude and longitude

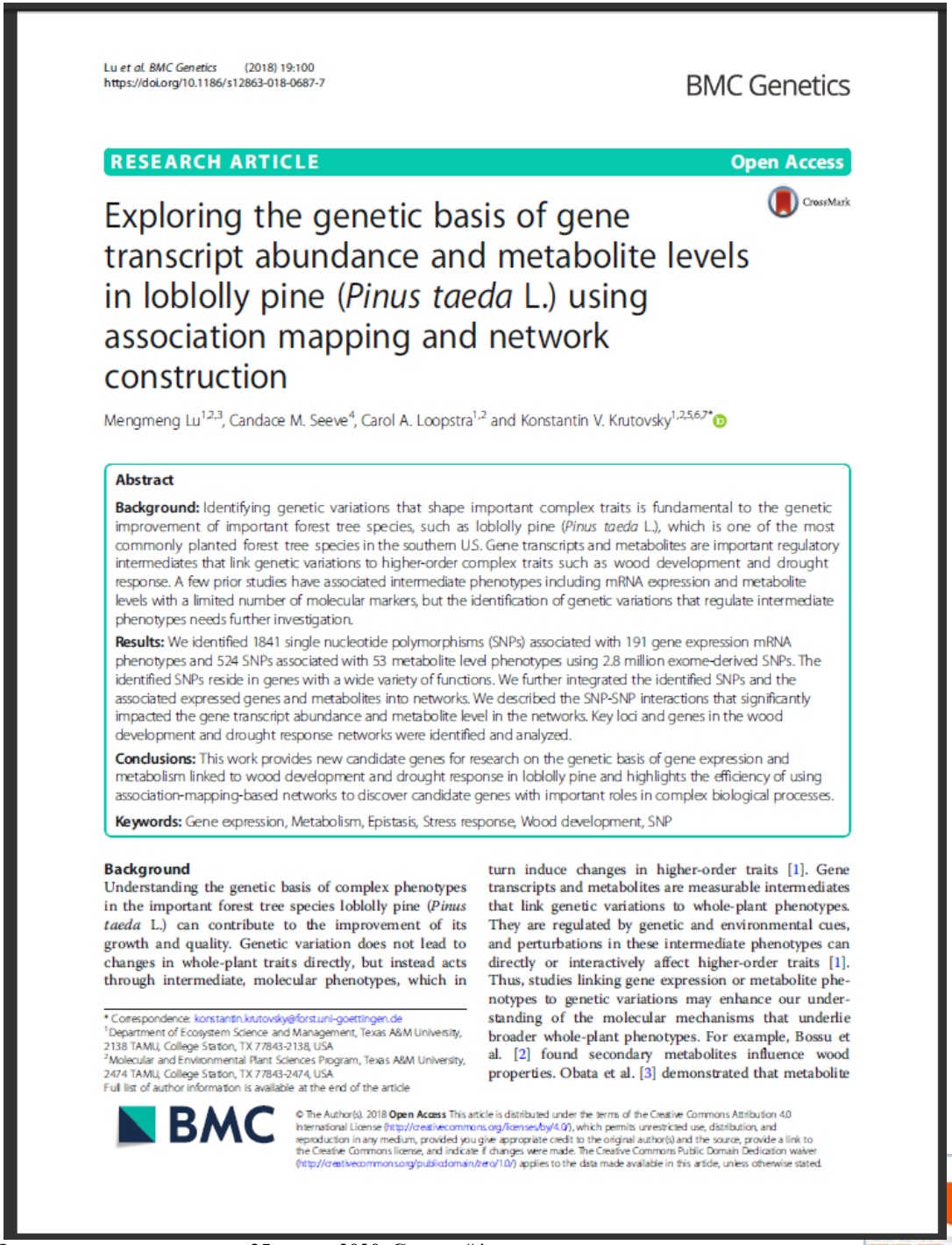
Environmental variables (and 19 bioclimatic variables; http://www.worldclim.org/bioclim):

- Monthly Mean total Annual temperature above 5˚C or Growing Degree Days (MEAN_annGDD5)
- Mean Annual Precipitation (MEAN_annP)
- Mean Annual Temperature (MEAN_MAT)
- Aridity index
- *etc.*

Complex geographic × environmental variables (= environmental resistance)

Lu M, Loopstra CA, **Krutovsky KV (**2019) Detecting the genetic basis of local adaptation in loblolly pine (*Pinus taeda* L.) using whole exome-wide genotyping and an integrative landscape genomics analysis approach. *Ecology and Evolution*

Ассоциативное картирование SNPs с уровнем экспрессии генов (мРНК) и вторичных метаболитов



Lu et al. BMC Genetics (2018) 19:100
https://doi.org/10.1186/s12863-018-0687-7

BMC Genetics

RESEARCH ARTICLE Open Access

Exploring the genetic basis of gene transcript abundance and metabolite levels in loblolly pine (*Pinus taeda* L.) using association mapping and network construction

Mengmeng Lu[1,2,3], Candace M. Seeve[4], Carol A. Loopstra[1,2] and Konstantin V. Krutovsky[1,2,5,6,7*]

Abstract

Background: Identifying genetic variations that shape important complex traits is fundamental to the genetic improvement of important forest tree species, such as loblolly pine (*Pinus taeda* L.), which is one of the most commonly planted forest tree species in the southern U.S. Gene transcripts and metabolites are important regulatory intermediates that link genetic variations to higher-order complex traits such as wood development and drought response. A few prior studies have associated intermediate phenotypes including mRNA expression and metabolite levels with a limited number of molecular markers, but the identification of genetic variations that regulate intermediate phenotypes needs further investigation.

Results: We identified 1841 single nucleotide polymorphisms (SNPs) associated with 191 gene expression mRNA phenotypes and 524 SNPs associated with 53 metabolite level phenotypes using 2.8 million exome-derived SNPs. The identified SNPs reside in genes with a wide variety of functions. We further integrated the identified SNPs and the associated expressed genes and metabolites into networks. We described the SNP-SNP interactions that significantly impacted the gene transcript abundance and metabolite level in the networks. Key loci and genes in the wood development and drought response networks were identified and analyzed.

Conclusions: This work provides new candidate genes for research on the genetic basis of gene expression and metabolism linked to wood development and drought response in loblolly pine and highlights the efficiency of using association-mapping-based networks to discover candidate genes with important roles in complex biological processes.

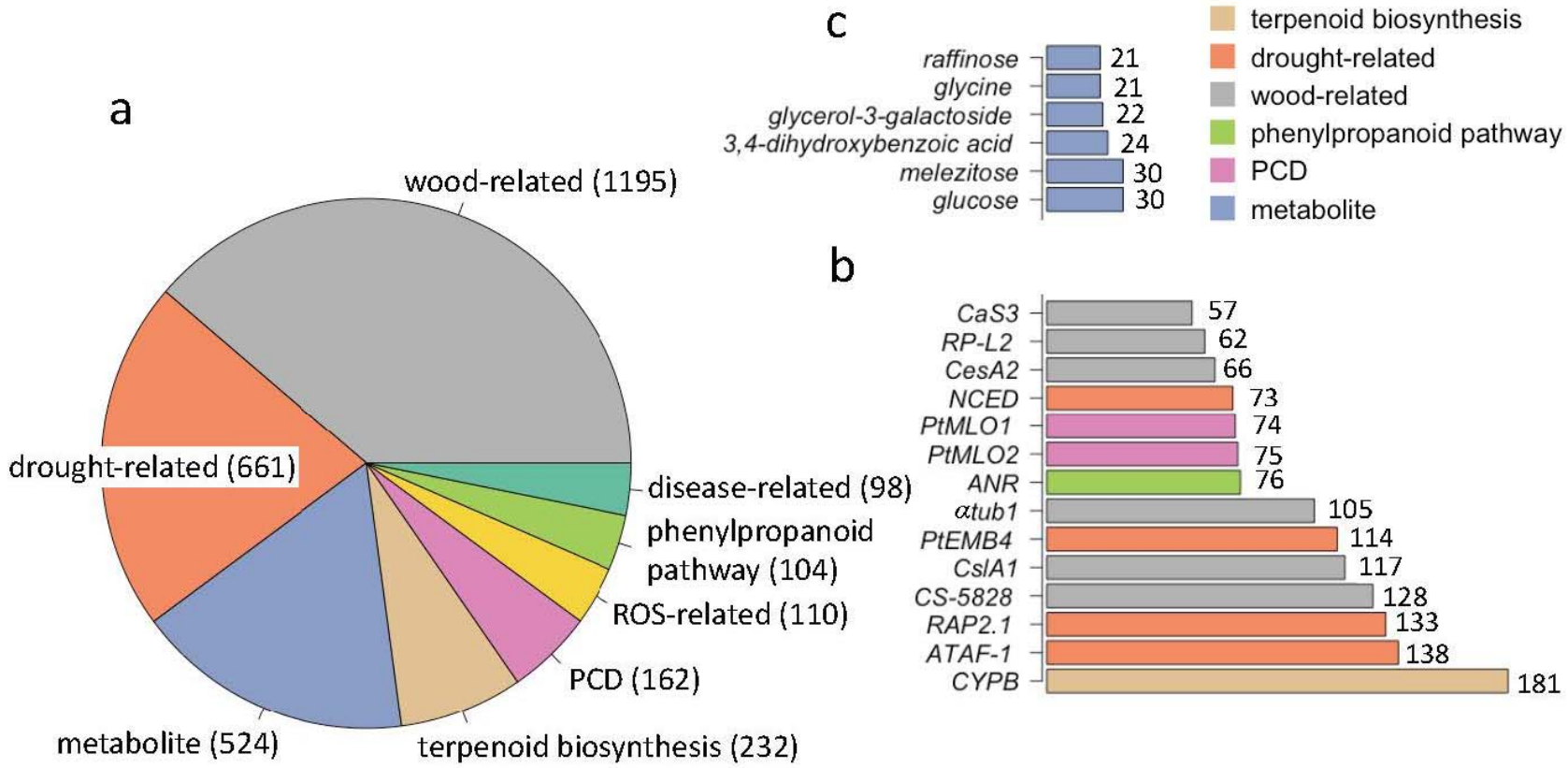
Keywords: Gene expression, Metabolism, Epistasis, Stress response, Wood development, SNP

Background

Understanding the genetic basis of complex phenotypes in the important forest tree species loblolly pine (*Pinus taeda* L.) can contribute to the improvement of its growth and quality. Genetic variation does not lead to changes in whole-plant traits directly, but instead acts through intermediate, molecular phenotypes, which in turn induce changes in higher-order traits [1]. Gene transcripts and metabolites are measurable intermediates that link genetic variations to whole-plant phenotypes. They are regulated by genetic and environmental cues, and perturbations in these intermediate phenotypes can directly or interactively affect higher-order traits [1]. Thus, studies linking gene expression or metabolite phenotypes to genetic variations may enhance our understanding of the molecular mechanisms that underlie broader whole-plant phenotypes. For example, Bossu et al. [2] found secondary metabolites influence wood properties. Obata et al. [3] demonstrated that metabolite

* Correspondence: konstantin.krutovsky@forst.uni-goettingen.de
[1]Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843-2138, USA
[2]Molecular and Environmental Plant Sciences Program, Texas A&M University, 2474 TAMU, College Station, TX 77843-2474, USA
Full list of author information is available at the end of the article

Ассоциативное картирование SNPs с уровнем экспрессии генов (a, b) и вторичных метаболитов (c)



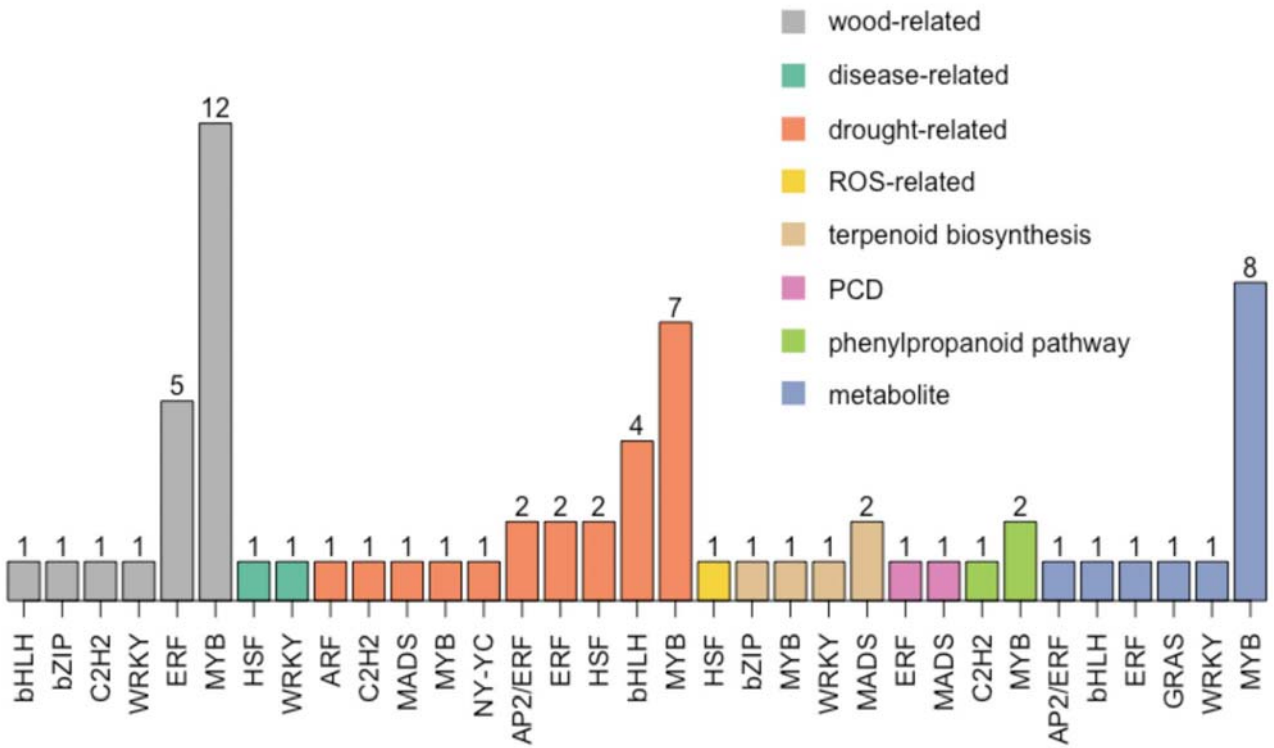
Number of significant SNP-phenotype associations:

(a) 8 different functional groups of genes, for which expression level (мРНК) was used as a phenotypic trait in the SNP association study

(b) 14 genes with the most SNP associations

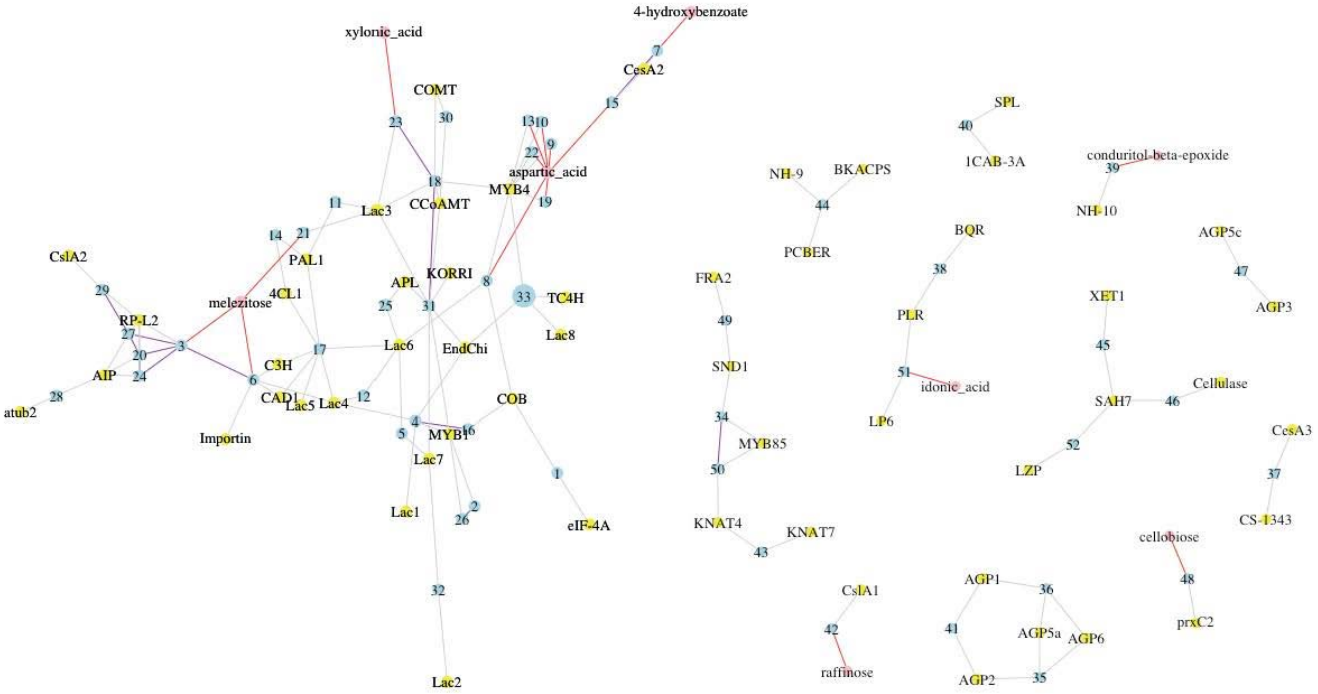
(c) 6 metabolites with the most SNP associations

Ассоциативное картирование SNPs с уровнем экспрессии генов (мРНК) и вторичных метаболитов



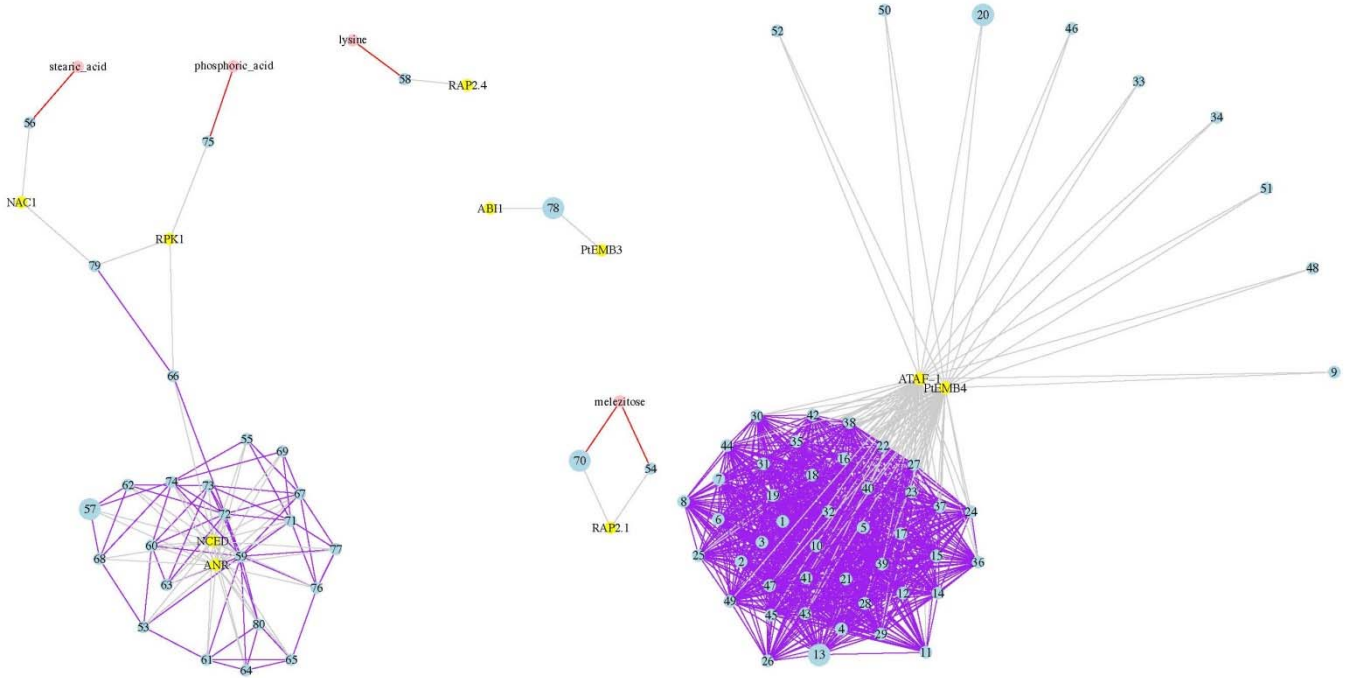
- **Number of significant SNP associations with gene expression and metabolite concentration level phenotypes for 59 transcription factor (TF) genes representing 12 TF families** AP2/ERF, ARF, bHLH, bZIP, C2H2, ERF, GRAS, HSF, MADS, MYB, NY-YC, and WRKY with 69 SNPs associations, in total.
- The gene with expression phenotypes were classified into **seven different functional groups**: wood-related, disease-related, drought-related, reactive oxygen species (ROS)-related, terpenoid biosynthesis, programmed cell death (PCD), and phenylpropanoid pathway. The numbers above each bar represent the numbers of identified SNPs associated with gene expression or metabolite level phenotypes

Ассоциативное картирование SNPs с уровнем экспрессии генов (мРНК) и вторичных метаболитов



Gene networks comprised of SNPs and their associated wood-related gene expression and metabolite level phenotypes. The blue dot nodes represent SNPs. Details of the SNPs and the genes containing them are presented in Table 2. The large blue dot node represents a SNP that resides in a GAMYB transcription factor (TF) gene. The yellow dot nodes represent genes, for which expression level was used as a phenotype trait in the SNP association analysis. The pink dot nodes represent metabolites, for which concentration level was used as a phenotype trait in the SNP association analysis. The grey and red edges represent SNP-gene-expression and SNP-metabolite-level associations, respectively. The purple edges represent SNP-SNP interactions that significantly impact the phenotypes. Expressed genes in the network include arabinogalactan-protein and cell wall protein genes (*AGP1-6*), cell expansion genes (*COB* and *KORRI*), cell wall related (resistance related) genes (*CslA1*), cellulose and callose synthase genes (*CesA3*, *CslA2*, and *CS-1343*), lignin biosynthesis enzyme genes (*4CL1*, *C3H*, *CAD1*, *CCoAMT*, *COMT*, *Lac1-8*, *PAL1*, and *TC4H*), α-tubulin gene (*αtub2*), wood development enzyme genes (*BKACPS*, *BQR*, *Cellulase*, *EndChi*, *Importin*, *LP6*, *PCBER*, *PLR*, *prxC2*, *SAH7*, *SPL*, and *XET1*), wood development protein genes (*1CAB*-3A, *NH-10*, *NH-9*, and *RP-L2*), wood development TF genes (*SND1*, *AIP*, *APL*, *eIF-4A*, *FRA2*, *KNAT4*, *KNAT7*, *LZP*, *MYB1*, *MYB4*, and *MYB85*)

Ассоциативное картирование SNPs с уровнем экспрессии генов (мРНК) и вторичных метаболитов

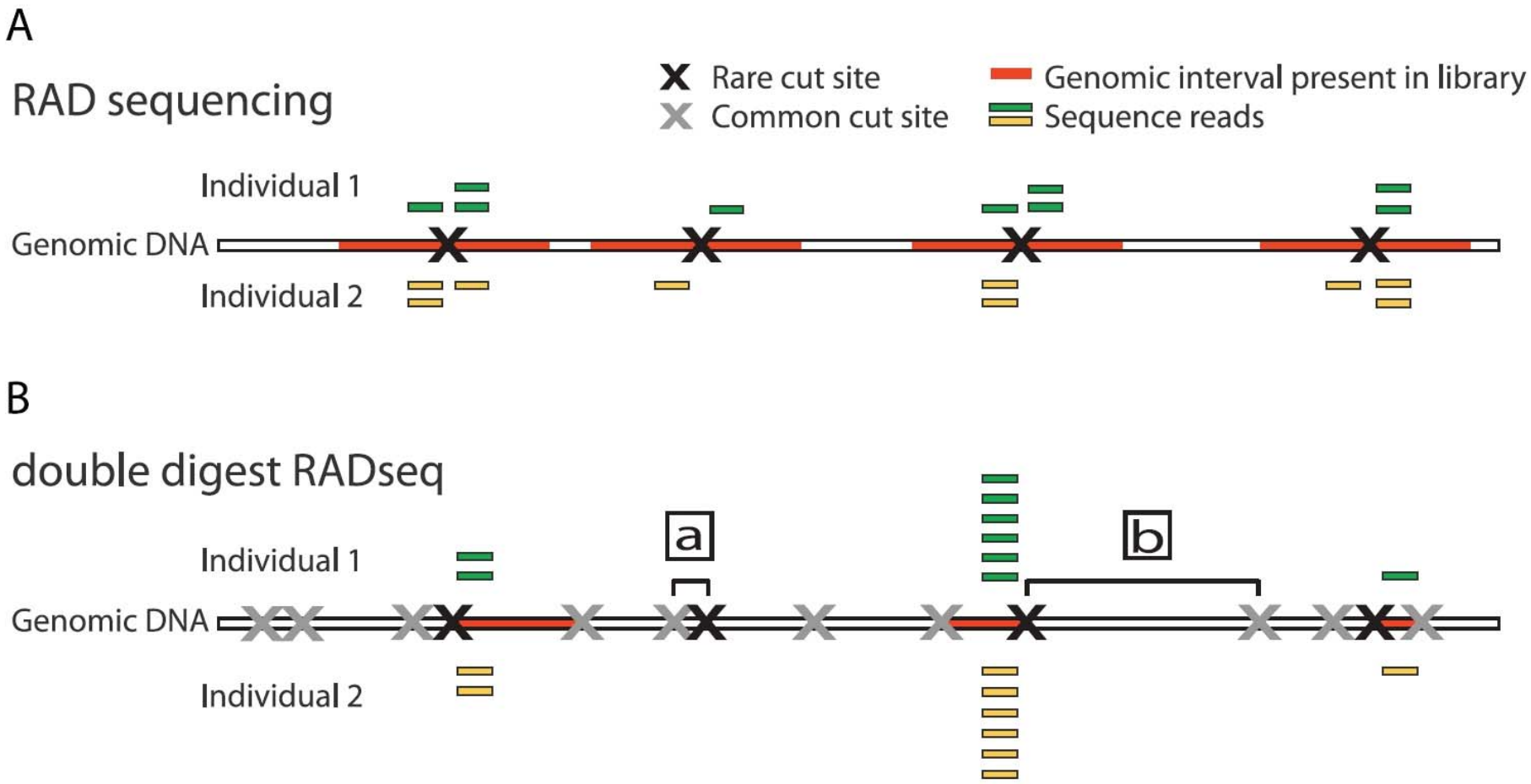


Gene networks comprised of SNPs and their associated drought-related gene expression and metabolite level phenotypes. Blue dot nodes represent SNPs. Details of the SNPs and the genes containing them are presented in Table 3. The blue dot nodes 13, 20, 57, 70, and 78 with a larger size represent the SNPs that reside in transcription factor (TF) genes. The yellow dot nodes represent genes, for which expression level was used as a phenotype trait in the SNP association analysis. The pink dot nodes represent metabolites, for which concentration level was used as a phenotype trait in the SNP association analysis. The grey and red edges represent SNP-gene-expression and SNP-metabolite-level associations, respectively. The purple edges represent SNP-SNP interactions that significantly impact the phenotypes. Expressed genes in the network include drought signaling genes (*ABI1*, *NCED*, and *RPK1*), drought-responsive TF genes (*NAC1*, *RAP2.1*, *RAP2.4*, and *ATAF-1*), late embryogenesis abundant protein genes (*PtEMB3*-4), phenylpropanoid pathway gene (*ANR*)

PINEMAP проект: Результаты

- **2,822,609 SNPs генотипированы путём секвенирования в почти 40,000 генах ладанной сосны** в Texas A&M University в этом проекте путём прямого секвенирования экзомной части геномной ДНК, обогащённой экзонами с помощью гибридизации тотальной ДНК с 600 млн олигонуклетидных проб, представляющих почти полный транскриптом (~40 тыс. экспрессируемых генов) ладанной сосны
- **375 деревьев со всего ареала, профенотипированных по большому числу адаптивных и селекционно-ценных признаков**, а также изученных по большому числу средовых факторов **прогенотипированы по всем обнаруженным SNPs** и обнаружено множество аллелей и гаплотипов связанных с изменчивостью адаптивных и селекционно-ценных признаков, с экспрессией генов, концентрацией вторичных метаболитов, а также с устойчивостью к средовым факторам. Обнаружены эпистатические взаимодействия между генами и генные сети
- фактически, это означает **переход от отдельных маркёров к полному генотированию через секвенирование!**
- **эра маркёров заканчивается – наступает эра полногеномного секвенирования!**
- **популяционная геномика вместе с молекулярной экологией (экогеномикой) позволят**:
  - обнаружить гены и аллели ответственные за адаптацию
  - связать генотипы с адаптивными фенотипами и средой

Restriction site Associated DNA Sequencing RAD-seq



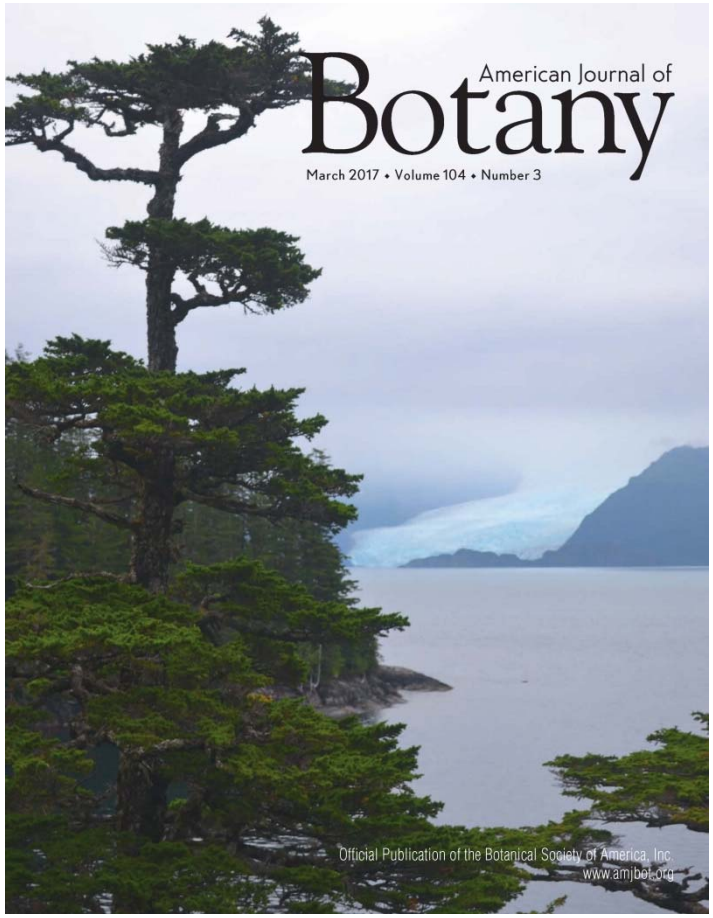
Peterson *et al.* 2012

ddRAD-seq

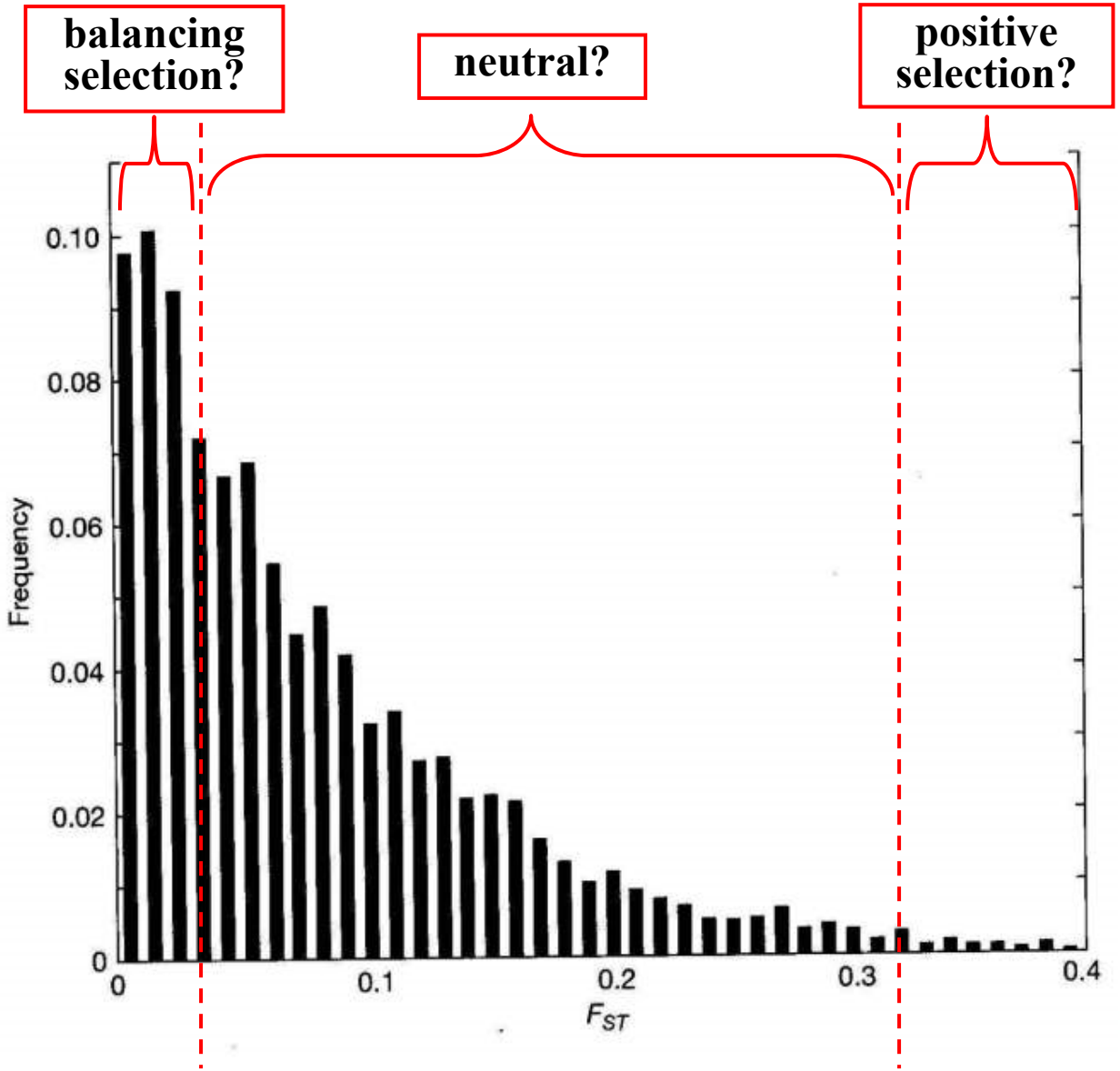
Johnson JS, P Chhetri, **KV Krutovsky**, DM Cairns (2017) Growth and its relationship to individual genetic diversity of mountain hemlock (*Tsuga mertensiana*) at alpine treeline in Alaska: combining dendrochronology and genomics. *Forests* 8(11): 418; http://www.mdpi.com/1999-4907/8/11/418

Johnson JS, KD Gaddis, DM Cairns, K Konganti, **KV Krutovsky** (2017) Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. *American Journal of Botany* 104(3): 439-450; http://www.amjbot.org/content/104/3/439.short

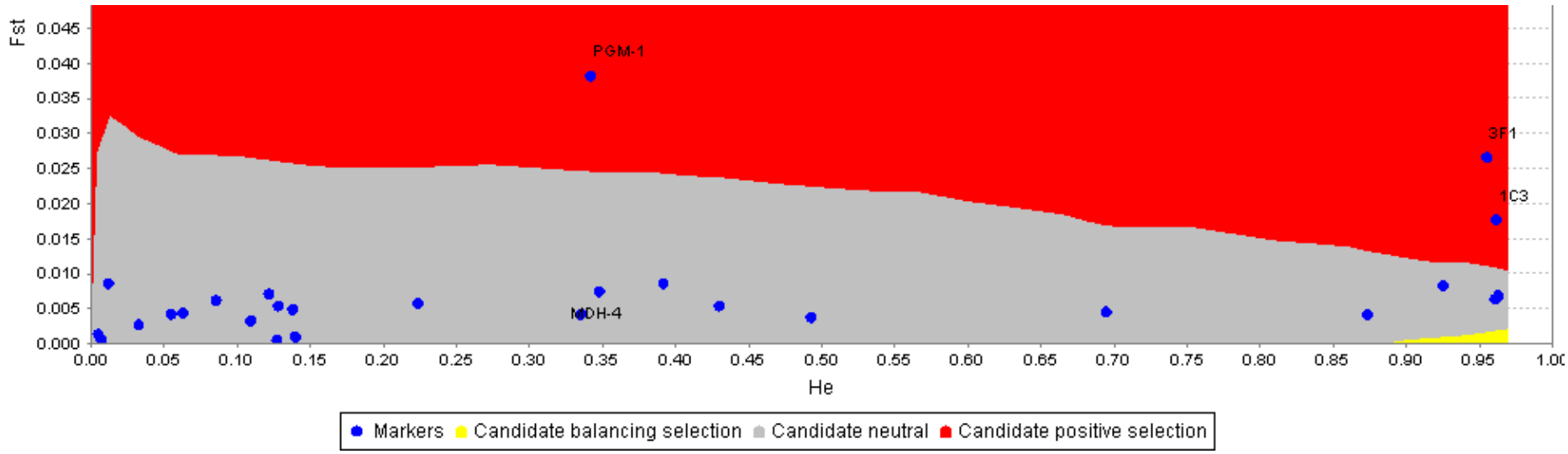
Johnson JS, KD Gaddis, DM Cairns, **KV Krutovsky** (2017) Seed dispersal at alpine treeline: an assessment of seed movement within the alpine treeline ecotone. *Ecosphere* 8(1):e01649; http://onlinelibrary.wiley.com/doi/10.1002/ecs2.1649/full



Individual gene F_{ST} coefficients to measure population differentiation and search for outliers

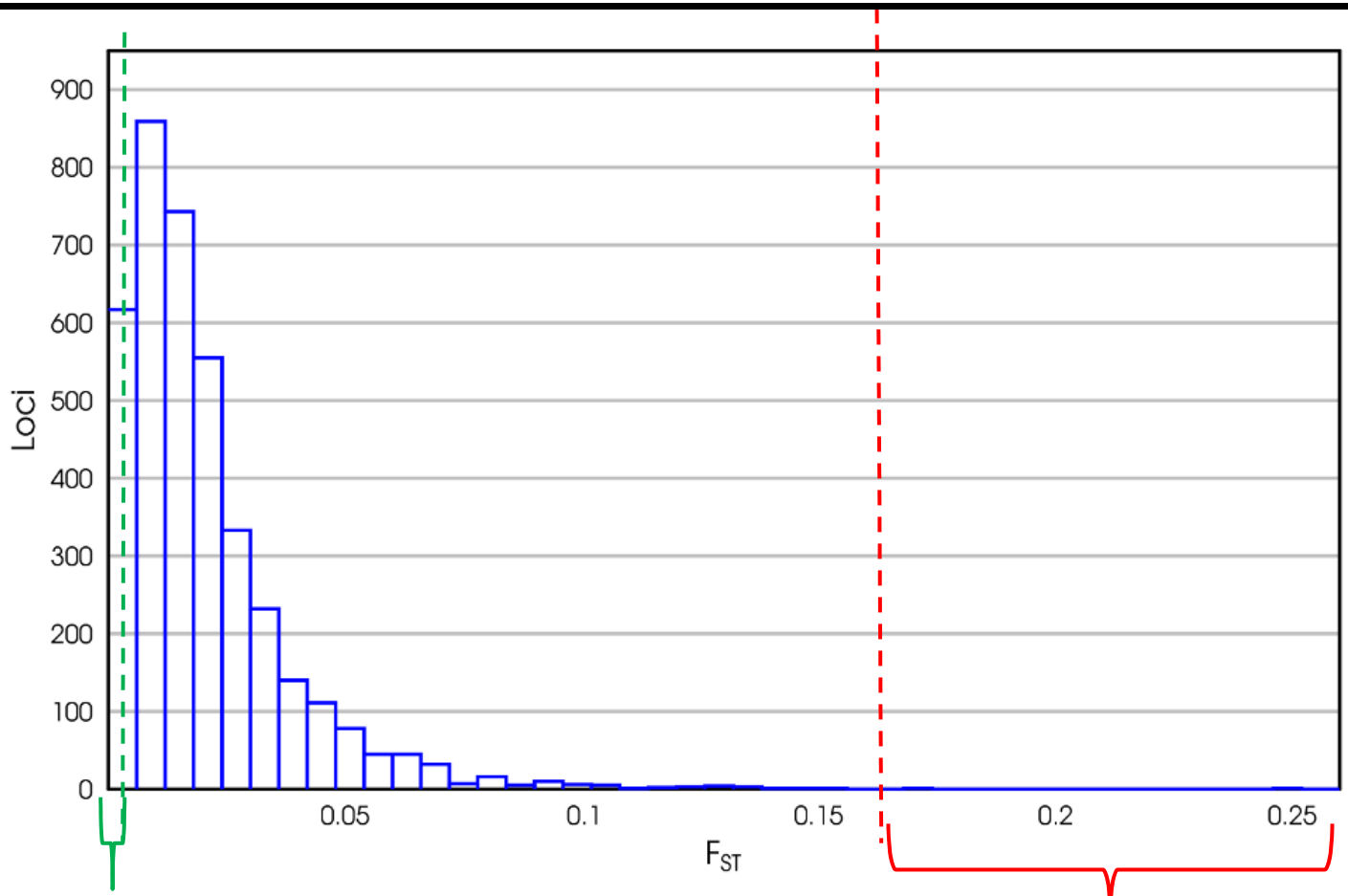


Selection detection based on a F_{ST}-outlier detection method



- A graphical output generated by the LOSITAN software (Antao *et al*. 2008) with the simulated confidence area for neutral loci (middle gray color zone) calculated using 31 supposedly neutral SSR loci from 18 Douglas-fir populations (represented as dots) using method described in Beaumont (2005) and in Beaumont & Nichols (1996) and implemented in the fdist program.
- This method evaluates the relationship between F_{ST} and H_e (expected heterozygosity) in the Wright's island model of migration with neutral markers.
- This distribution is used to identify outlier loci that have excessively high (red zone) or low (yellow zone) F_{ST} compared to neutral expectations.
- The outliers are tagged with labels and are candidates for being subject to selection.

Genetic differentiation (F_{ST}) for 4,264 SNPs in the loblolly pine populations in East Texas

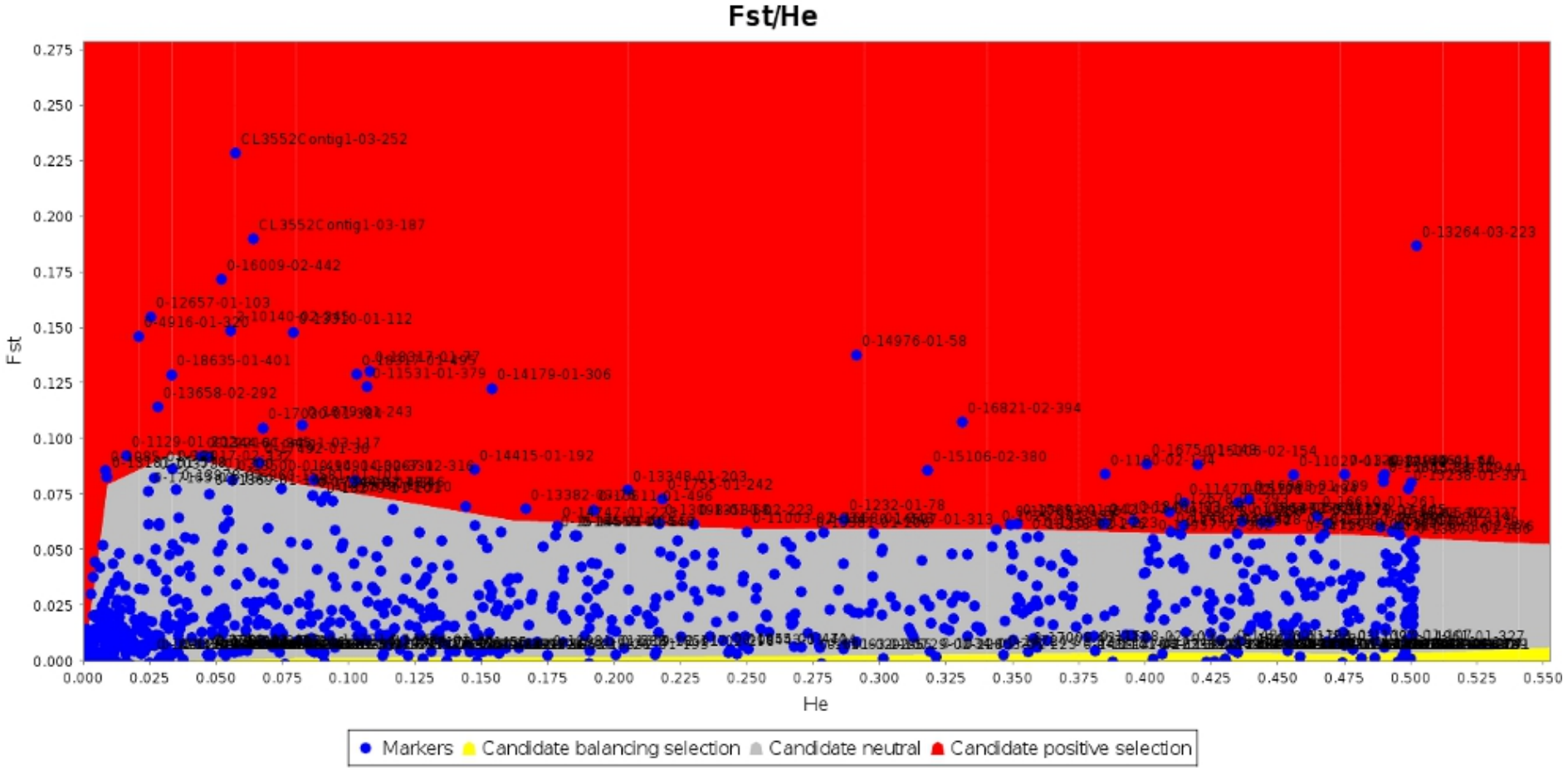


Very low differentiation; under balancing or stabilizing selection?

Very differentiated; under positive or divergent selection?

Search for outlier SNPs under selection and supposedly neutral SNPs using coalescence simulation to find thresholds for selectively neutral markers

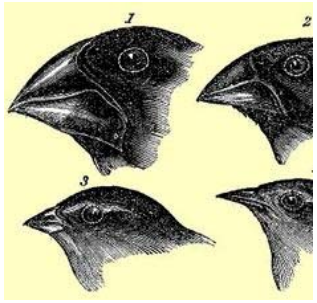
4,264 SNPs, distribution of F_{ST} & expected heterozygosity (H_e) in the loblolly pine populations (1st generation selections), 95K simulations to find thresholds



Beaumont *et al.* (1996) Royal Soc. B; Antao *et al.* (2008) Bioinformatics

F_{ST} «outliers»

Examples of Candidate genes after correction for false positive:



Balancing selection (244):

Arabinofuranodisase, Glycoprotease protein, Dehydrin, Lypoxygenase, Cytokinin oxidase, Transmembrane transporters, Endoglucanase, MYB transcription factor, Pinus taeda Heme oxygenase I, etc.

Dirigent protein, Homeobox-leucine zipper, Cytochrome p450, Gras transcription factor, Gigantia protein, Ethylene-forming enzymes Histone H4, Reductases, etc.

Directional selection (74):

Geranyldiphosphate, Disease resistance proteins, Arabinogalactan-like proteins, Pinus Expansin, Pinus alpha-Xyloxidase, etc.

Potassium/proton antiporter, Laccase 90DProtein kinases, Histone H3.2, etc.

Chhatre, V., T. Byram, D.B. Neale, J.L. Wegrzyn, and **K.V. Krutovsky**, 2013 Genetic structure and association mapping of adaptive and selective traits in the East Texas loblolly pine (*Pinus taeda* L.) breeding populations. *Tree Genetics and Genomes* **9**(5): 1161-1178.

F_{ST} outliers and neutrality tests

Gene	Map: LG,cM	SNP	Het	Fst	*P*(Simul Fst<sample Fst = 0.010)	Neutrality test
4cl (*4-coumarate:CoA ligase*)	-	0-7767-01-191	0.0073	0.0000	0.7356	
	-	UMN-CL379Contig1-12-117	0.0013	0.0000	0.0000	Tajima's *D* +
ccoaomt(*caffeoyl CoA O-methyltransferase 1*)	V, 32.7	CL544Contig1-03-112	0.4992	0.0097	0.5488	Tajima's *D* +
ccr1(*cinnamoyl CoA reductase*)	-	CL594Contig1-06-236	0.4643	0.0051	0.4217	
comt2(*caffeate O-methyltransferase 2*)	-	0-10914-02-331	0.0397	0.0174	0.7236	
	-	0-10914-02-55	0.0097	0.0022	0.8055	
cpk-3 (*calcium-dependent protein kinase*)	-	CL2332Contig1-01-175	0.1571	0.0000	0.0086	
	-	CL2332Contig1-01-314	0.4127	0.0229	0.8561	
lp3-3(*water-stress inducible protein 3*)	-	CL1740Contig1-03-78	0.3703	0.0351	0.9621	
pal1(*phenylalanine ammonia-lyase 1*)	-	CL863Contig1-03-164	0.0093	0.0005	0.7761	
ppap12 (*putative wall-associated protein kinase*)	-	CL3898Contig1-04-256	0.0329	0.0165	0.7109	Tajima's *D* +
ptlim1 (*LIM domain protein 1 (LIM transcription factor)*)	-	CL1905Contig1-06-353	0.0142	0.0000	0.3565	
	-	CL1905Contig1-03-377	0.0142	-0.0007	0.3565	
ptlim2 (*LIM domain protein 2 (LIM transcription factor)*)	II, 3.5	CL711Contig1-04-212	0.1766	0.0010	0.2830	

Koralewski, T.E., Brooks, J.E. and K.V. Krutovsky, 2014 Molecular evolution of drought tolerance and wood strength related candidate genes in loblolly pine (*Pinus taeda* L.). *Silvae Genetica* **63**(1-2): 59-66.

Conclusions

- Many genetic associations between genes and phenotypic traits and diseases have been discovered
- These associations often involve loci also correlated to the environment
- Significant associations were found between alleles in multiple SNPs and environmental variables (Growing Degree Days, temperature, and precipitation) and geographical factors (latitude and longitude).
- Significant associations represented a diverse set of genes including abiotic stress response genes ranging from transmembrane proteins to heat shock proteins and transcription factors.
- These loci are also often appear as F_{ST}-outliers and non-neutral – both for positive and negative selection

Key terms and definitions

- **Quantitative Trait Locus (QTL) mapping**: A QTL is a chromosomal region suspected to contain a gene (or cluster of genes) that contributes to the variation observed at a quantitative trait. QTLs are detected through linkage mapping experiments using progeny usually obtained in experimental crosses or pedigrees that segregate for both quantitative traits and genetic markers. QTL and genetic markers that are close together on a chromosome will tend to co-segregate.
- **Association mapping**: As in QTL mapping, the goal of association mapping is to find a statistical association between genetic markers and a quantitative trait. However, unlike QTL mapping, which is performed in the context of a pedigree, association mapping is performed at the population level: the genotypes of the candidate gene markers and the phenotypes of the corresponding trait are determined in a set of unrelated or distantly-related individuals sampled from a population. Association mapping relies on linkage disequilibrium (LD) between the markers and the actual causative genes (i.e., the actual polymorphism that causes the differences in the phenotypic trait). Hence association mapping is also referred to as 'LD mapping'. For association to be detected the genetic markers usually must be closely linked to genes (lie within or directly upstream or downstream of them) that contribute to the variation in that trait, and the goal is to identify the actual genes affecting that trait, rather than just (relatively large) chromosomal segments. Since population genetic structure (genetic differences that accumulate between populations) can cause LD even at unlinked loci, association analyses must account for population genetic structure whenever it is present in the population from which your sample has been drawn (Pritchard et al. 2000; Thornsberry et al. 2001).

Key terms and definitions

- **Amino-acid or nucleotide multiple sequence alignment**: is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.

- **Contig**: (from *contiguous*) is a set of overlapping DNA sequences in a **multiple sequence alignment** that together represent a **consensus** region of DNA. In sequencing projects, a contig refers to overlapping sequence reads or to the overlapping clones that form a physical map of the genome that is used to guide sequencing and assembly. Contigs can thus refer both to overlapping DNA sequence and to overlapping physical segments (fragments) contained in clones depending on the context.

- **Consensus sequence**: is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a **multiple sequence alignment**. It represents the results of a **multiple sequence alignment,** in which related sequences are compared to each other, and most frequent residues are calculated.

- **Basic Local Alignment Search Tool (BLAST)**: is an algorithm for comparing amino-acid or nucleotide sequences using their alignment. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

- **Expressed Sequence Tag (EST):** is a short sub-sequence of a mRNA/cDNA sequence. They are used to identify gene transcripts, and are instrumental in gene discovery, gene sequence determination and differential gene expression analysis (transriptome profiling).

- **Unigene:** is a supposedly unique transcript that represents the same transcription locus (expressed gene or pseudogene), often inferred as a **consensus sequence** from **EST** based **multiple sequence alignment.**

- **SNP**: stands for *S*ingle *N*ucleotide *P*olymorphism. This refers to a particular nucleotide (or "base") in a DNA sequence that is variable within a species (or between related species). For example, at a certain position in a DNA sequence there may be a **C** (cytosine) present in some individuals but a **T** (thymine) present in others (**C**/**T** polymorphism). SNPs represent the most basic form of genetic polymorphism. There are tens of millions of SNPs present in the genome of a typical organism and can be used as genetic markers (SNP markers).