Conifer Translational Genomics Network Coordinated Agricultural Project

Genomics in Tree Breeding and Forest Ecosystem Management

Module 7 – Measuring, Organizing, and Interpreting Marker Variation

Nicholas Wheeler & David Harry – Oregon State University

Molecular population genetics

- "…the focus of population genetics has changed absolutely – from inquiring what deductions can be made about the evolutionary process from the abstract principles of Mendelian inheritance and Darwinian selection, to inquiring what inferences can be made about the evolutionary process from the analysis of sequences of real genes sampled from actual evolving populations"
  - *Hartl. 2000. A primer of population genetics*
- Deductive reasoning: Given a set of general principles, determine what would happen under a specific set of conditions
- Inductive reasoning: Given specific information, infer or induce some general principles that apply to all cases

Neutral theory of molecular evolution

- The **neutral theory of molecular evolution** states that the vast majority of evolutionary change at the molecular level is caused by random drift of selectively neutral mutants
- The fundamental population genetic parameter affecting diversity under the neutral theory is estimated by θ
- The effective number of new mutants per generation, where

$$\theta = 4N_e\,\mu$$

N_e = **number of reproductive individuals in the population (effective pop size) and μ is the mutation rate**

Coalescent theory and gene genealogies

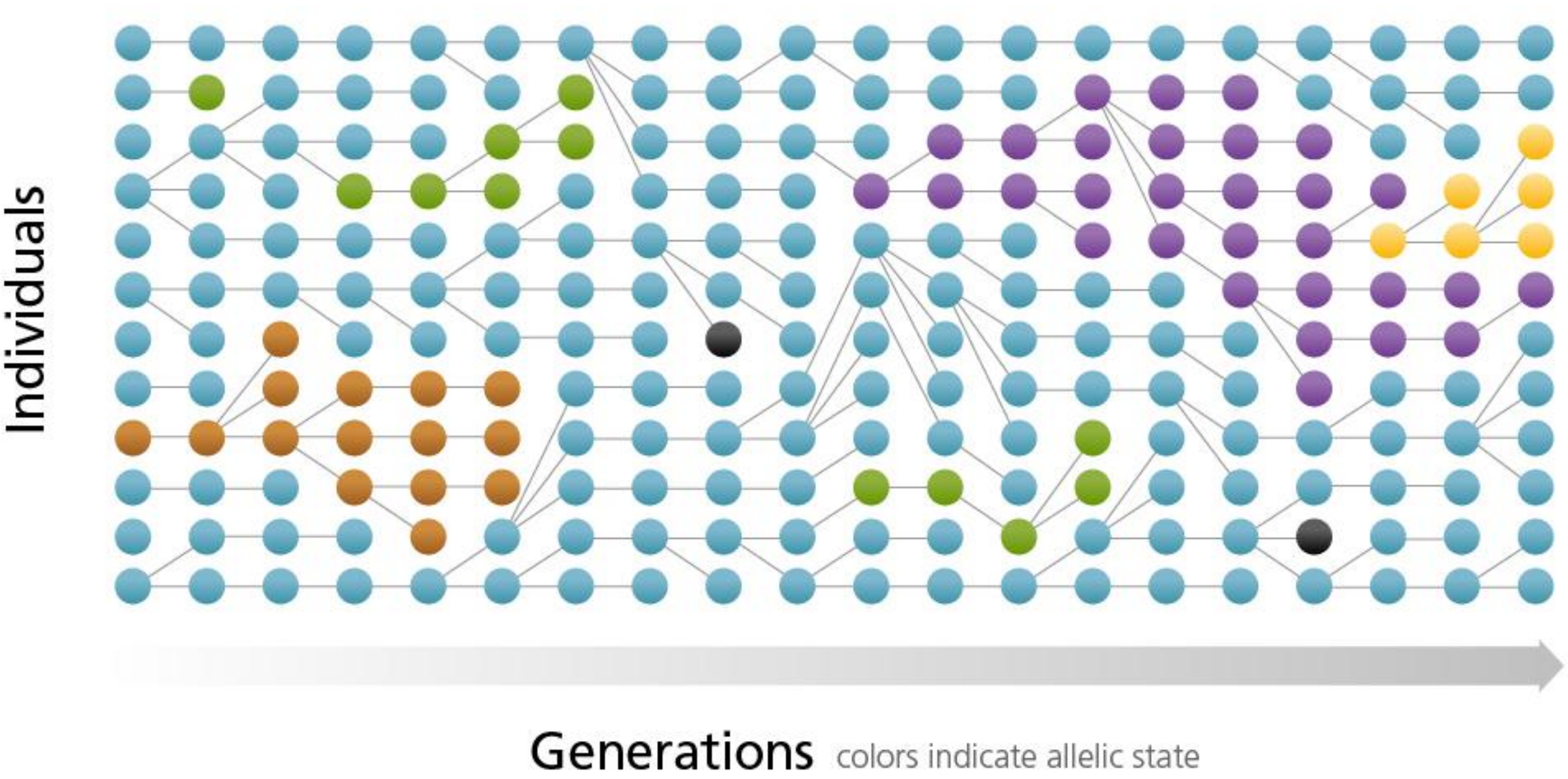
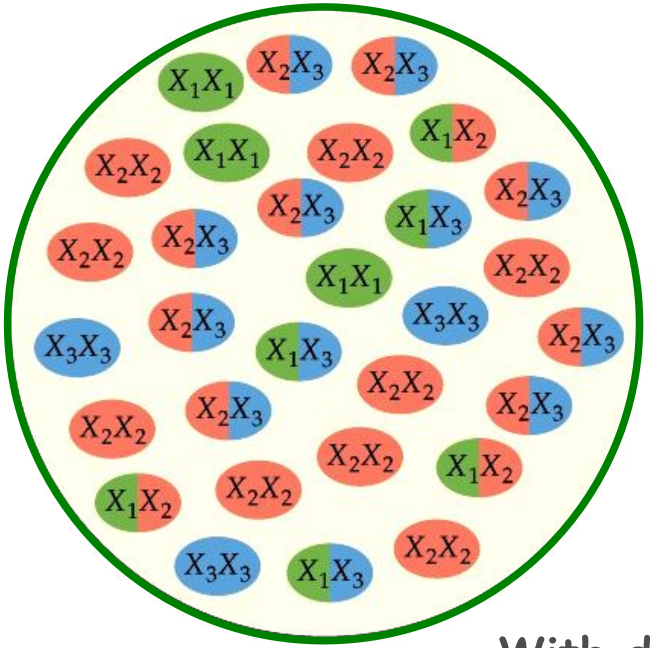


Figure Credit: Nicholas Wheeler, Oregon State University

How do geneticists measure diversity?

Typical descriptive statistics

Locus ‘X’ in pop #1



Allele	Frequency
A_1	0.2
A_2	0.5
A_3	0.3
	Total = 1.0

	Genotype	Frequency
	$A_1\ A_1$	0.1
➡	$A_1\ A_2$	0.1
➡	$A_1\ A_3$	0.1
	$A_2\ A_2$	0.3
➡	$A_2\ A_3$	0.3
	$A_3\ A_3$	0.1
		Sum = 1.0

A (# alleles) = 3

H_o (observed heterozygosity) = 0.5

With data from more loci, you an also calculate,
P (% polymorphic loci) = % of loci with >1 allele

Figure Credit: Glenn Howe, Oregon State University

Segregating sites (S)

```
              1         2         3         4
     1234567890123456789012345678901234567890
A    ACGATCGAGGCATCGACAACGAGTAGCGAGGGATCGACAG
B    ACGATCGAGGCATCGACAACGAGTAGCGCGGGATCGACAG
C    ACGAGCGAGGCATCGACAACGAGTAGCGAGGGATCGACAG
D    ACGATCGAGCCATCGACATCGAGTAGCGTGGGATCGACAG
E    ACGATCGAGCCATCGACATCGAGTAGCGAGGGATCGACAG
SNPs     *    *        *         *
```

(Number of polymorphic sites)

Proportion of segregating sites:

= (# SNPs)/(Total # NT bases)

= 4/40

S = 0.1

Figure credit: David Harry, Oregon State University

Nucleotide polymorphism (θw or θs)

- We begin with the **number of polymorphic sites**, as in the slide before

$$\theta_w = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} bp^{-1}$$

Nucleotide polymorphism (θ_w):

= (0.1)/[1+1/2+1/3+1/4]

θ_w = 0.048

Nucleotide diversity (π, or θ_π)

```
               1         2         3         4
      1234567890123456789012345678901234567890
A     ACGATCGAGGCATCGACAACGAGTAGCGAGGGATCGACAG
B     ACGATCGAGGCATCGACAACGAGTAGCGCGGGATCGACAG
C     ACGAGCGAGGCATCGACAACGAGTAGCGAGGGATCGACAG
D     ACGATCGAGCCATCGACATCGAGTAGCGTGGGATCGACAG
E     ACGATCGAGCCATCGACATCGAGTAGCGAGGGATCGACAG
#diff     4    6        6         7
```

(Number of pairwise differences)

Proportion of Pairwise Differences:

= (# Pairwise Diff's)/(Total # Pairwise Comparisons)

= (4 + 6 + 6 + 7)/(10 x 40)

= 23/400

π = 0.0575

Figure Credit: David Harry, Oregon State University

Sample-based estimators of θ

Estimator	Sensitivity	Source
$\theta_W = \frac{1}{\sum_{i=1}^{n-1}\frac{1}{i}}\sum_{i=1}^{n-1}\xi_i$	low	Watterson (1975)
$\theta_\pi = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)\xi_i$	intermediate	Tajima (1989)
$\theta_{\xi_e} = \xi_e = \xi_1$	singleton	Fu and Li (1993)
$\theta_H = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i^2\xi_i$	high	Fay and Wu (2000)
$\theta_L = \frac{1}{n-1}\sum_{i=1}^{n-1} i\xi_i$	high	Zeng et al. (2006)

Sensitivity = the frequency of observed polymorphisms that makes estimates using a given estimator large relative to the others

Table Credit: Andrew Eckert, University of California, Davis

Are all SNPs equal?

- Not likely! Simply consider where they occur
  - *Within genes vs. between genes vs. upstream regulatory elements*
  - *Within exons vs. within introns*
  - *At synonymous sites in a codon vs. at non-synonymous sites*

- Consequently, tests have been created to determine if the characteristics of a SNP (location, frequency, etc), as measured by some of the parameters discussed here, suggests the polymorphism violates the neutral hypothesis. That is, can we detect signatures of selection? One such test is Tajima's D

Nucleotide diversity (θ_π) by species

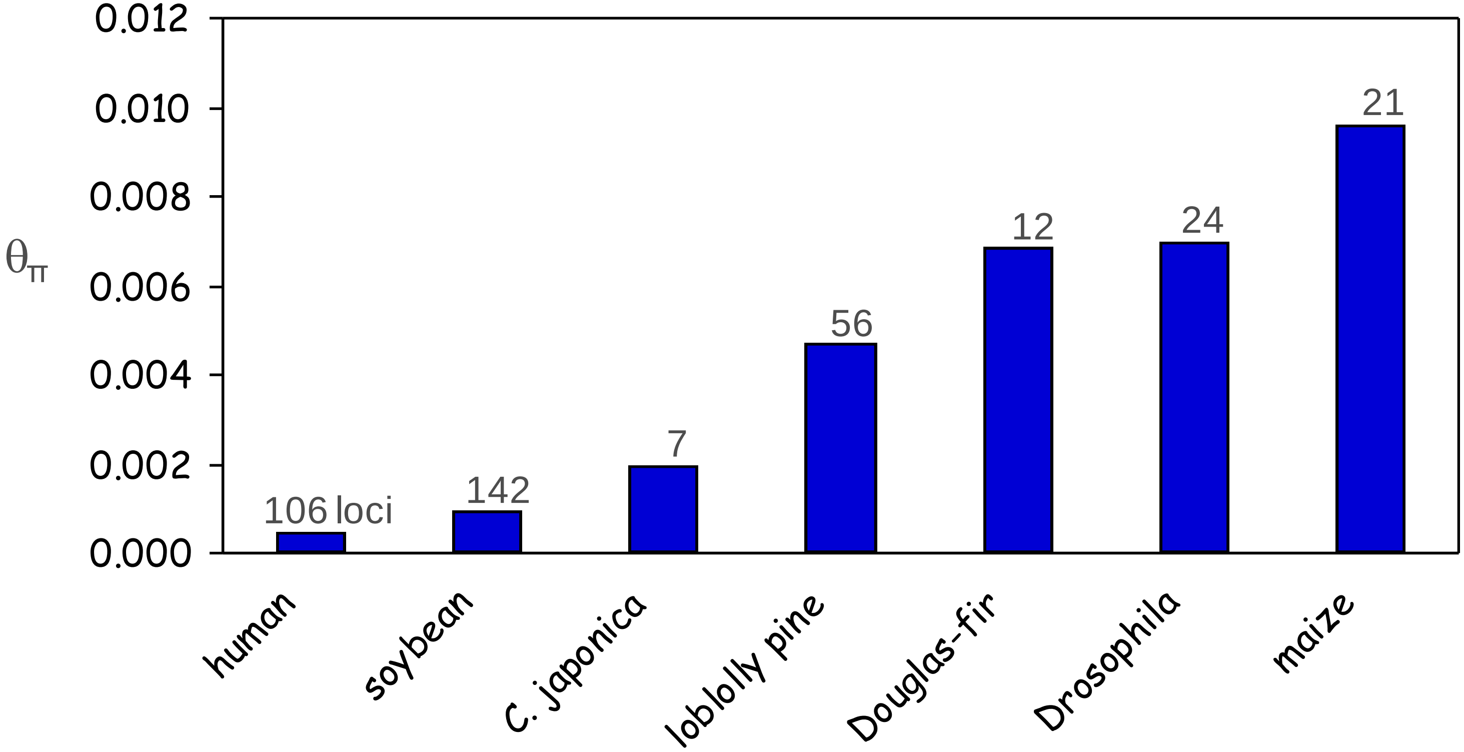
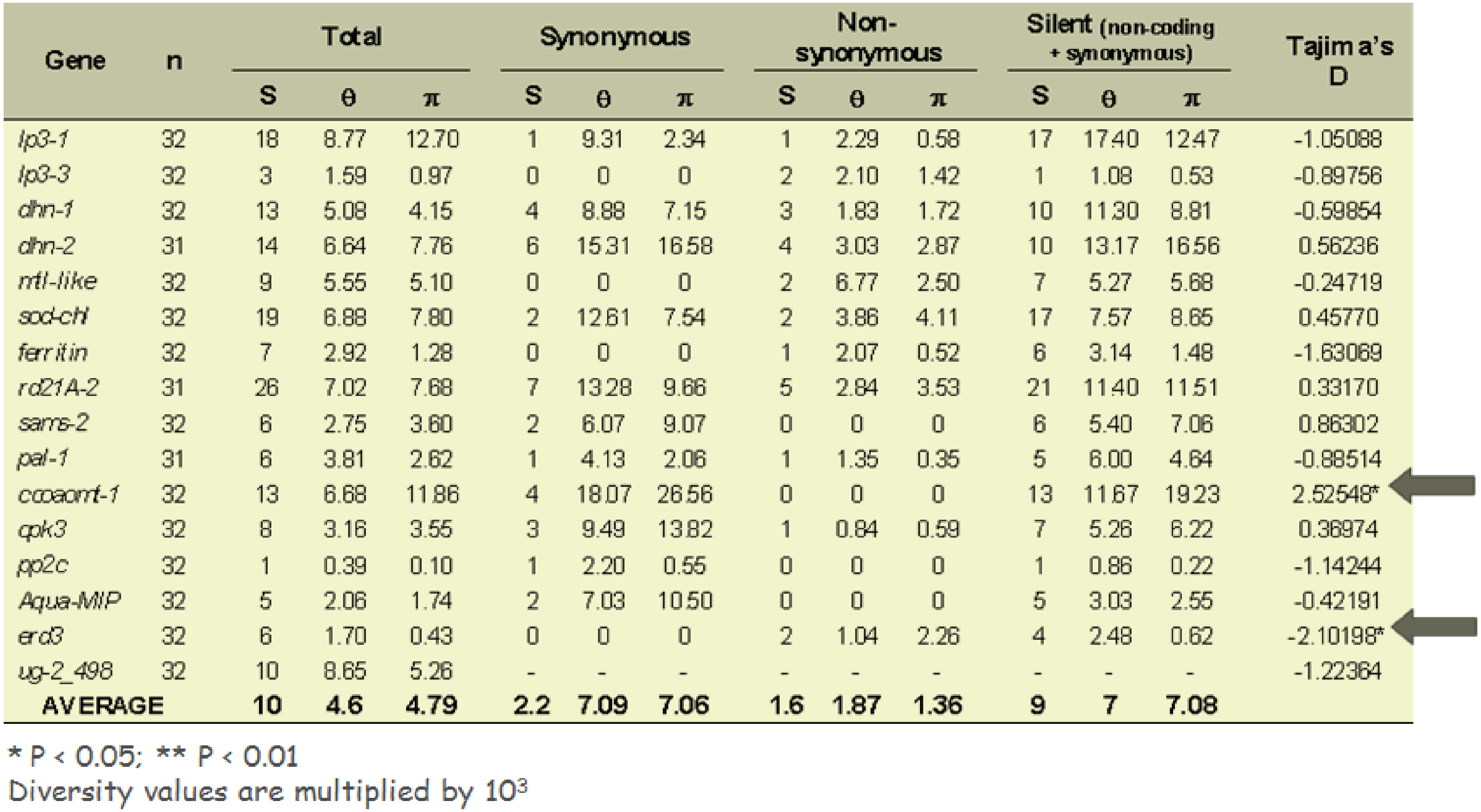


Figure Credit: Andrew Eckert, University of California, Davis

Nucleotide diversity in loblolly pine candidate genes for drought



Gene	n	Total			Synonymous			Non-synonymous			Silent (non-coding + synonymous)			Tajima's D
		S	θ	π	S	θ	π	S	θ	π	S	θ	π	
lp3-1	32	18	8.77	12.70	1	9.31	2.34	1	2.29	0.58	17	17.40	12.47	-1.05088
lp3-3	32	3	1.59	0.97	0	0	0	2	2.10	1.42	1	1.08	0.53	-0.89756
dhn-1	32	13	5.08	4.15	4	8.88	7.15	3	1.83	1.72	10	11.30	8.81	-0.59854
dhn-2	31	14	6.64	7.76	6	15.31	16.58	4	3.03	2.87	10	13.17	16.56	0.56236
mtl-like	32	9	5.55	5.10	0	0	0	2	6.77	2.50	7	5.27	5.68	-0.24719
sod-chl	32	19	6.88	7.80	2	12.61	7.54	2	3.86	4.11	17	7.57	8.65	0.45770
ferritin	32	7	2.92	1.28	0	0	0	1	2.07	0.52	6	3.14	1.48	-1.63069
rd21A-2	31	26	7.02	7.68	7	13.28	9.66	5	2.84	3.53	21	11.40	11.51	0.33170
sams-2	32	6	2.75	3.60	2	6.07	9.07	0	0	0	6	5.40	7.06	0.86302
pal-1	31	6	3.81	2.62	1	4.13	2.06	1	1.35	0.35	5	6.00	4.64	-0.88514
ccoaomt-1	32	13	6.68	11.86	4	18.07	26.56	0	0	0	13	11.67	19.23	2.52548*
cpk3	32	8	3.16	3.55	3	9.49	13.82	1	0.84	0.59	7	5.26	6.22	0.36974
pp2c	32	1	0.39	0.10	1	2.20	0.55	0	0	0	1	0.86	0.22	-1.14244
Aqua-MIP	32	5	2.06	1.74	2	7.03	10.50	0	0	0	5	3.03	2.55	-0.42191
erd3	32	6	1.70	0.43	0	0	0	2	1.04	2.26	4	2.48	0.62	-2.10198*
ug-2_498	32	10	8.65	5.26	-	-	-	-	-	-	-	-	-	-1.22364
AVERAGE		**10**	**4.6**	**4.79**	**2.2**	**7.09**	**7.06**	**1.6**	**1.87**	**1.36**	**9**	**7**	**7.08**	

* P < 0.05; ** P < 0.01

Diversity values are multiplied by 10^3

Table used with permission of the Genetics Society of America from "DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in Pinus taeda L", Gonzalez-Martinez et al. Genetics 172. 2006; permission conveyed through Copyright Clearance Center, Inc.

Organizing genetic diversity

- Wright's F statistics
- AMOVA
- STRUCTURE

Defining Wright's F statistics

We begin by discussing heterozygosity at different levels

- **H_I:** Observed heterozygosity within subpopulations
- **H_S:** Expected heterozygosity within subpopulations
- **H_T:** Expected heterozygosity if the combined population (metapopulation) were random mating. This would be $H_T = 2p_{avg}q_{avg}$ (average allele frequencies over metapopulation)

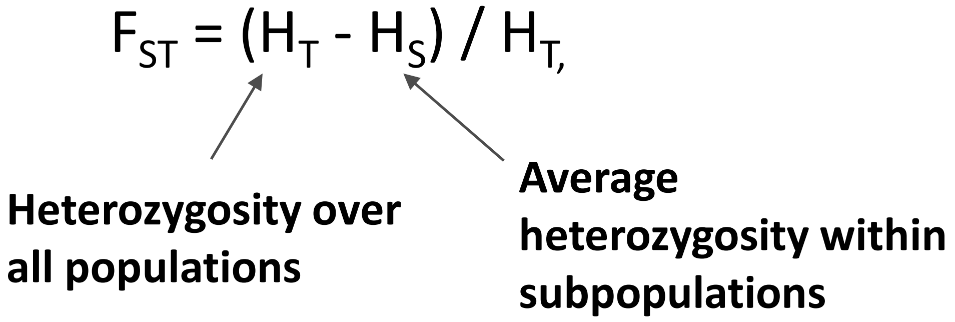
F statistics are defined in terms of H

- $F_{IS} = (H_S – H_I)/H_S$ (measuring departures from HW within subpopulations or local inbreeding)
- $F_{ST} = (H_T - H_S)/H_T$ (measuring departures from HW due to population differences, which is the same as measuring admixture or Wahlund’s effect)
- $F_{IT} = (H_T – H_I)/H_T$ (includes both local inbreeding and population structure)
- Together, they are related as: $(1 – F_{IS})\ (1 – F_{ST}) = (1 – F_{IT})$
- Of these measures, F_{IS} and F_{ST} are the most meaningful since they partition local inbreeding vs. population subdivision and describe how variation is proportioned

Wright's F_{ST}

A measure of the proportion of variation among populations

- Reduction of heterozygosity compared to random mating
- Measure of the probability that two gene copies chosen at random from different subpopulations are identical-by-descent (1 - F_{ST})
- Scale: 0 (heterozygosity identical across populations) to 1 (populations maximally different)



$$F_{ST} = (H_T - H_S) / H_{T,}$$

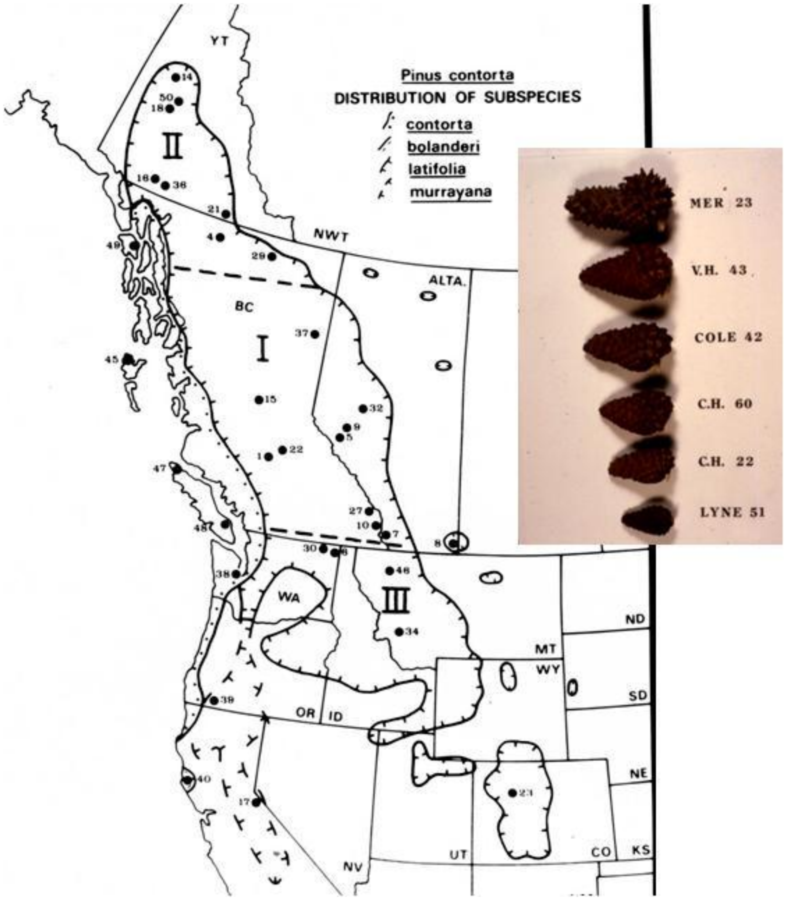
Heterozygosity over all populations

Average heterozygosity within subpopulations

- Significance detected by permutation

Population structure: Lodgepole pine

- How was genetic diversity distributed? (F stats)
  - *Within populations: 90.7%*
  - *Among populations within subspecies: 6.1%*
  - *Among subspecies: 3.2%*
- How was morphometric trait variation distributed?
  - *Within populations: 43.9%*
  - *Among populations within subspecies: 18.6%*
  - *Among subspecies: 37.6%*



Figure/Image Credit: Nicholas Wheeler, Oregon State University

Analysis of molecular variance (AMOVA)

- Nucleotide diversity can be partitioned in a manner analogous to F statistics for simple allelic variation. Consider this familiar looking equation

$$\Phi_{st} = (\pi_t - \pi_s) / \pi_t$$

- Where π_t is the nucleotide sequence diversity across the entire set of populations and π_s is the average nucleotide sequence diversity within populations
- AMOVA uses all information available in molecular markers (frequencies, measures of difference, etc)

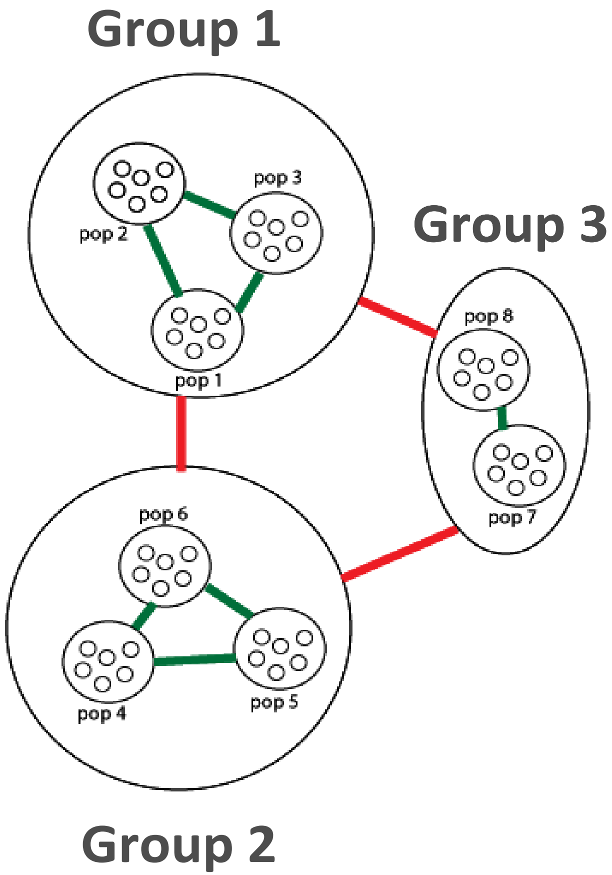
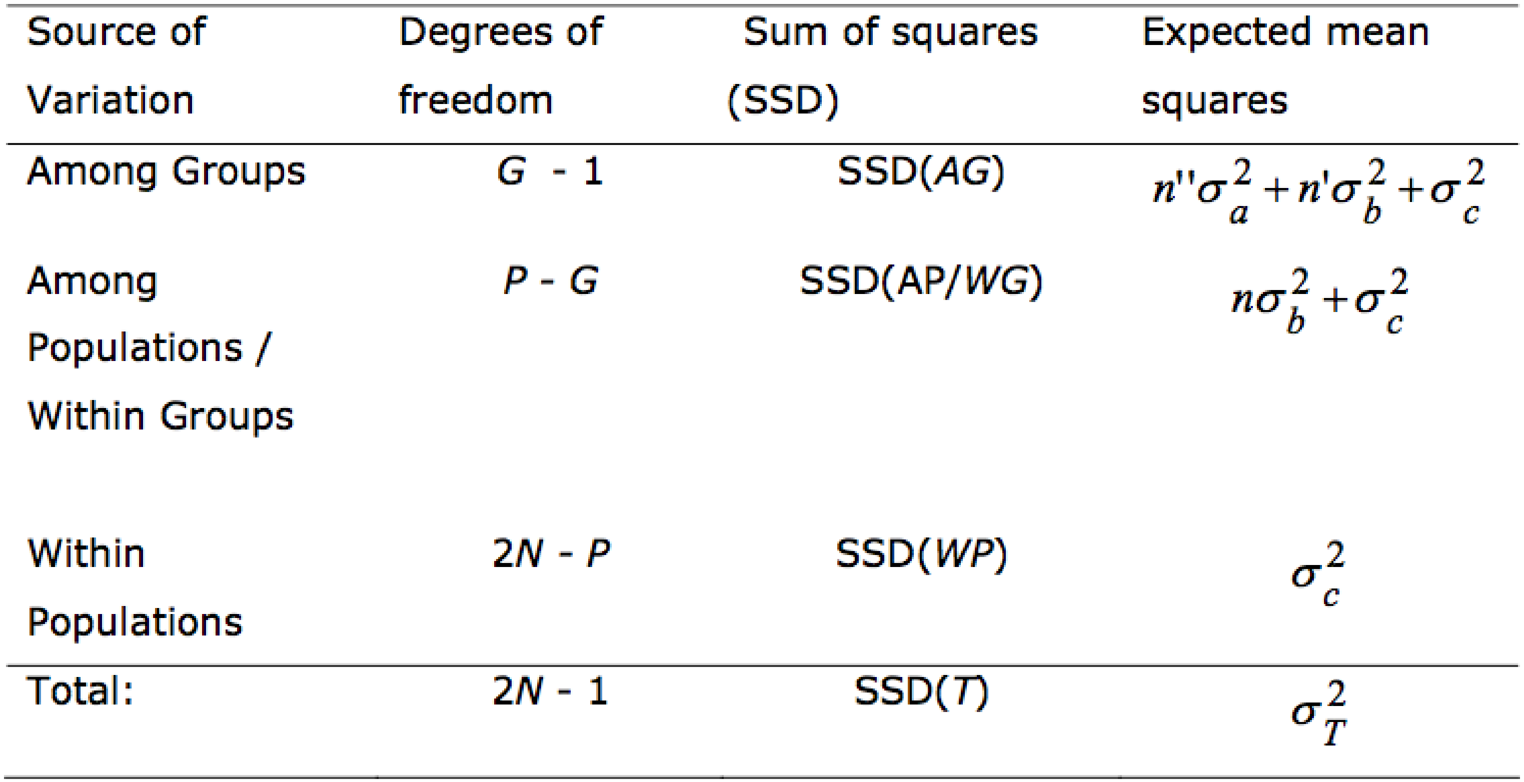


Figure Credit: Glenn Howe, Oregon State University

AMOVA: The layout



Source of Variation	Degrees of freedom	Sum of squares (SSD)	Expected mean squares
Among Groups	$G - 1$	SSD(AG)	$n''\sigma_a^2 + n'\sigma_b^2 + \sigma_c^2$
Among Populations / Within Groups	$P - G$	SSD(AP/WG)	$n\sigma_b^2 + \sigma_c^2$
Within Populations	$2N - P$	SSD(WP)	σ_c^2
Total:	$2N - 1$	SSD(T)	σ_T^2

Table Credit: Arlequin User Manual, http://cmpg.unibe.ch/software/arlequin35/

Bayesian clustering: Structure (and others)

- Inference on population structure using multi-locus genotype data

STRUCTURE
V2.1

- Goals of Bayesian clustering
  – *Assign individuals to populations on the basis of their genotypes, while simultaneously estimating population allele frequencies*
  – *Infer number of populations "K" in the process*

Structure is a model-based method

- Model with or without admixture
  - ***Without:*** *Each individual is assumed to originate in one (only one) of K populations*
  - ***With:*** *Each individual is assumed to have inherited some proportion of its ancestry from each of K populations*

- Linkage model
  - *"Blocks" of chromosomes are derived as intact units from one or another K population*
  - *All allele copies on the same linkage block derive from the same population*

- F model
  - *Populations all diverged from a common ancestral population at the same time, but allows that the populations may have experienced different amounts of drift since the divergence event*

Understanding variation in DNA sequences

- Population geneticists measure variation using
  - *Individual markers (e.g. allozymes, RAPDs, SSRs, SNPs)*
  - *Sequenced DNA fragments (ordered collection of bases)*

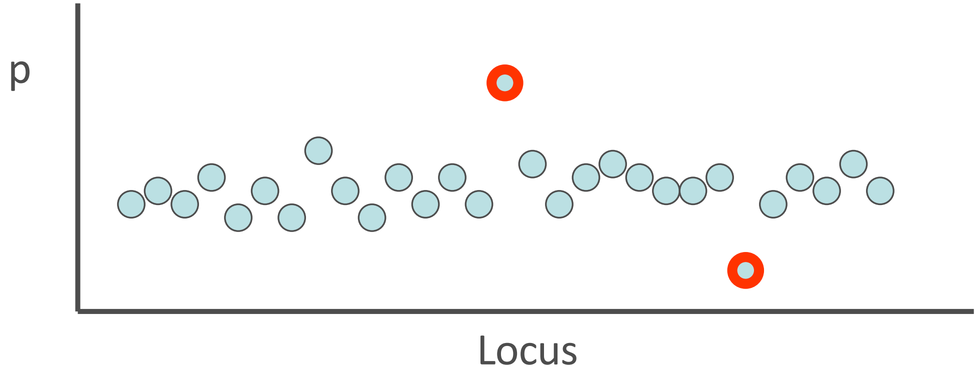
- Measures of marker variation include
  - *Number of alleles*
  - *Proportion of polymorphic loci*
  - *Heterozygosity*

- Once aligned, nucleotide sequences allow other comparisons, e.g.
  - *For individual nt bases: SNPs*
  - *Along the length of a sequenced fragment*
    - *How many nt bases match or mismatch?*
    - *What proportion of nt bases match or mismatch?*

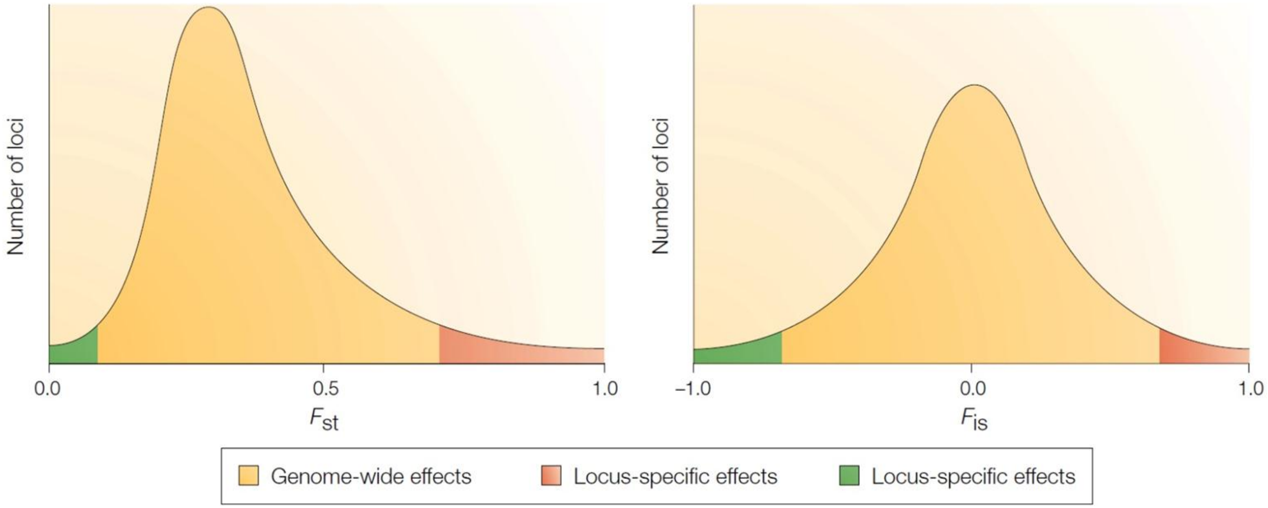
- Conceptually familiar metrics (i.e. resembling allelic variation)
  - *Segregating sites (nucleotide polymorphisms), S*
  - *Nucleotide diversity, θ_{π}*
  - *Nucleotide polymorphism, θ_{w}*

Lewontin-Krakauer test

- Proposed in 1973 to detect selection at isozyme loci
- Hypothesis: Selection creates excessive values of F_{st} at a selected locus
  - *Procedure: Estimate F_{st} for each locus*
  - *Distribution of (n-1)(locus specific F_{st})/(Mean F_{st}) is chi-square with n-1 df (n = # of subpopulations)*
  - *"Outliers" are those with significant chi-square*
  - *See Excoffier 2009*



Diversity outliers and locus-specific effects



- By screening many markers, baseline levels of within (F_{is}) and among (F_{st}) population diversity can be established. Unusually high or low diversity indices may signal a locus-specific response, which can be used to identify candidate genes for follow-up studies. For example, high F_{st} may suggest genes undergoing diversifying selection, possibly in response to environmental gradients

Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Luikart et al., 2003.

Commonly used neutrality tests

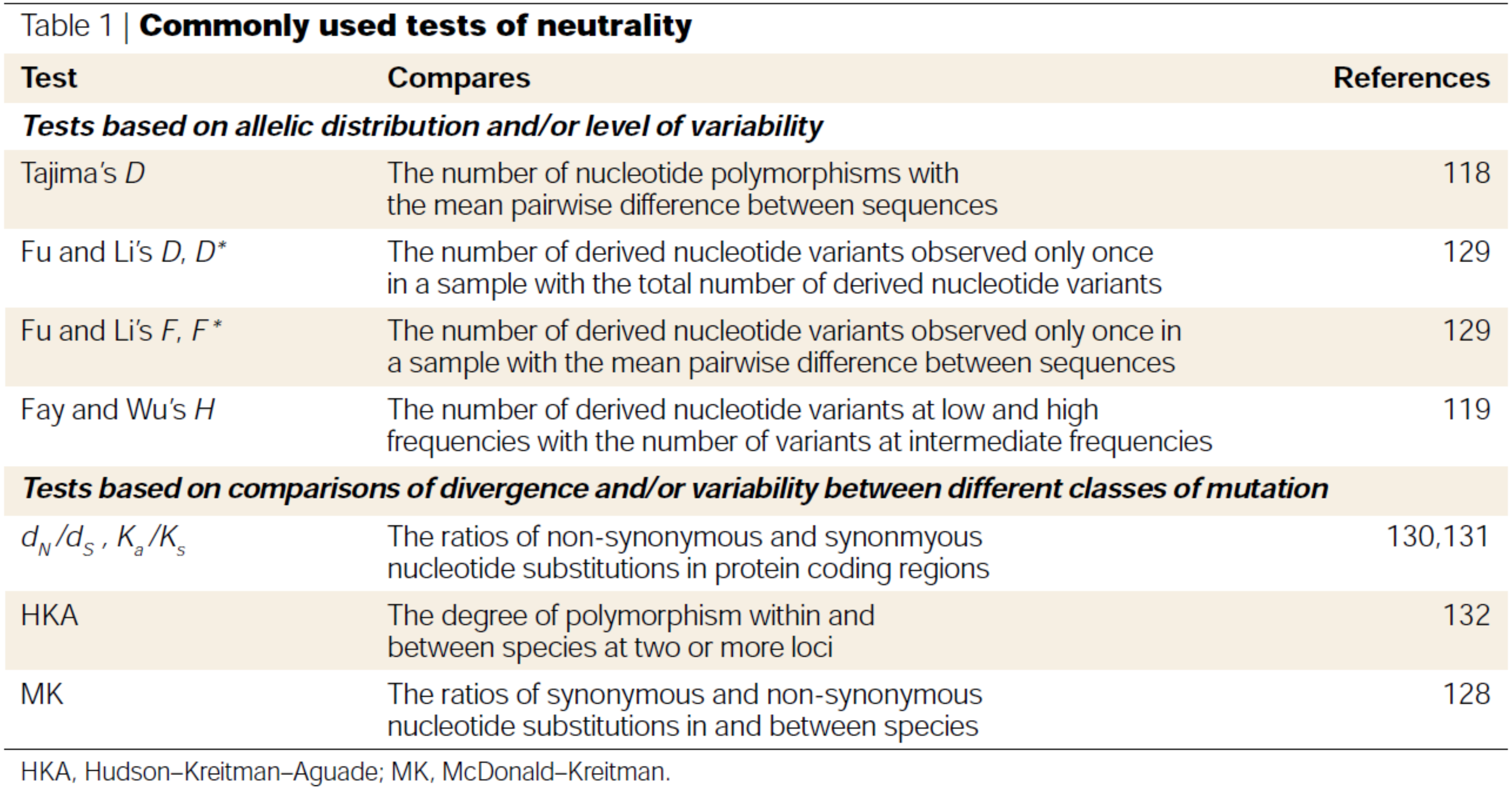


Table 1 | **Commonly used tests of neutrality**

Test	Compares	References
Tests based on allelic distribution and/or level of variability		
Tajima's *D*	The number of nucleotide polymorphisms with the mean pairwise difference between sequences	118
Fu and Li's *D, D**	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	129
Fu and Li's *F, F**	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	129
Fay and Wu's *H*	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	119
Tests based on comparisons of divergence and/or variability between different classes of mutation		
d_N/d_S, K_a/K_s	The ratios of non-synonymous and synonmyous nucleotide substitutions in protein coding regions	130,131
HKA	The degree of polymorphism within and between species at two or more loci	132
MK	The ratios of synonymous and non-synonymous nucleotide substitutions in and between species	128

HKA, Hudson–Kreitman–Aguade; MK, McDonald–Kreitman.

Table Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Bamshad and Wooding, 2003

Calculating Tajima’s D

- Two components for estimating Tajima’s D

$$\hat{\theta}_\pi = \pi \qquad \hat{\theta}_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

- Deviation of these two indicates deviation from neutral expectations

$$d = \hat{\theta}_\pi - \hat{\theta}_S$$

- **Tajima’s D:** $$D = \frac{d}{\sqrt{V(d)}}$$ where V(d) is variance of d

Tests based on diversity and divergence

- The MK test removes effects of the genealogy by dividing the types of polymorphisms at a given locus into four types
  - *Synonymous/polymorphic*
  - *Synonymous/fixed*
  - *Nonsynonymous/polymorphic*
  - *Nonsynonymous/fixed*
- The HKA test uses the differences in measures of polymorphism within and between species at two or more loci to draw inferences on neutrality

Signatures of selection: Summary

- A number of measures or statistics have been developed to identify signatures of selection
- Terminology and nuanced differences among estimators is confusing. Note citations for publications that help sort this out
- Measures noted here are of academic interest and lend scientific credibility to association genetic tests
- Positive results of such tests are not required to use significant associations, however

DnaSAM: Bioinformatics tools help summarize nucleotide diversity

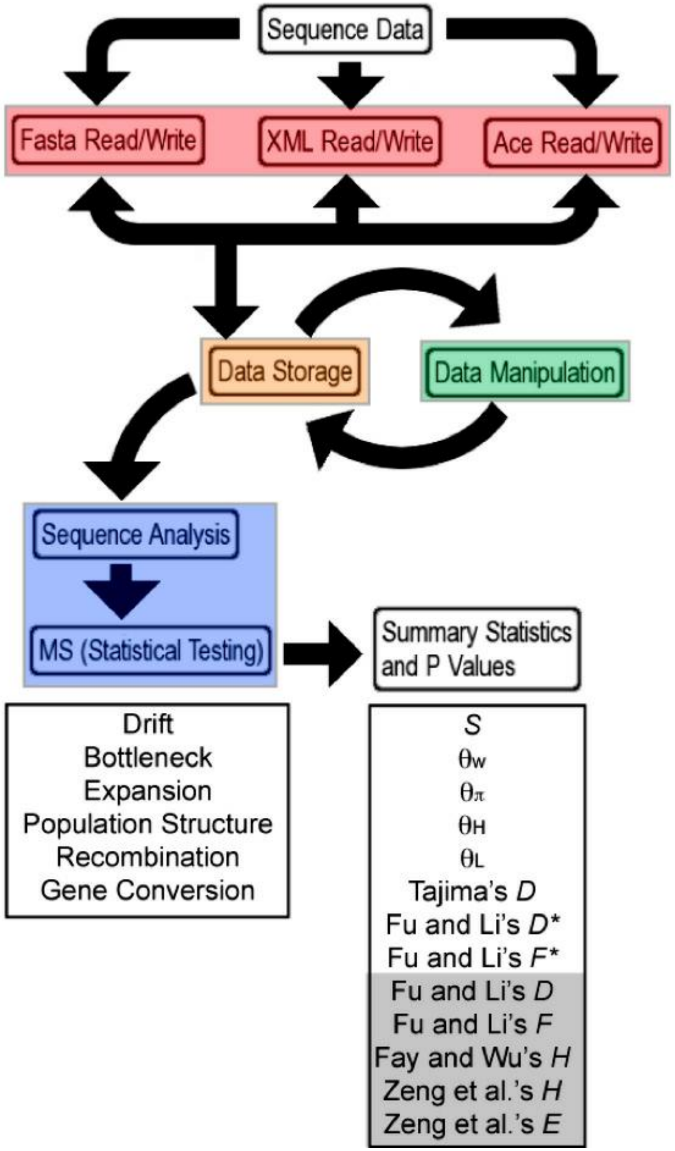


Figure Credit: Andrew Eckert, University of California, Davis

References cited

- Bamshad, M., and S. P. Wooding. 2003. Signatures of natural selection in the human genome. Nature Reviews Genetics 4: 92-111. (Available online at: http://dx.doi.org/10.1038/nrg999) (verified 8 March 2011).
- Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131: 479-491.
- Excoffier, L. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103: 285-298. (Available online at: http://dx.doi.org/10.1038/hdy.2009.74) (verified 8 March 2011).
- Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and corrected allele frequencies. Genetics 164: 1567-1587.
- Hartl, D. L. 2000. A primer of population genetics. Sinauer Associates, Sunderland, MA.
- Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet. 2003. The power and promise of population genomics. Nature Reviews Genetics 4: 981-994. (Available online at: http://dx.doi.org/10.1038/nrg1226) (verified 8 March 2011).

References cited (cont'd)

- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
- González-Martínez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler, and D. B. Neale. 2006. DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in Pinus taeda L. Genetics 172: 1915-1926. (Available online at: http://dx.doi.org/10.1534/genetics.105.047126) (verified 8 March 2011).
- Wheeler, N. C. 1981. Genetic variation in Pinus contorta Doubl. And related species of the subsection Contortae Ph. D. Thesis, University of Wisconsin, Madison, WI.

External links

- Arlequin ver 3.5.1.2 [Online]. Population Genetics, CMPG Lab, Institute of Ecology and Evolution, University of Bern. Available at: http://cmpg.unibe.ch/software/arlequin35/ (verified 8 March 2011).
- DnaSAM: DNA sequence analysis and manipulation [Online]. David Neale Lab, University of California Davis. Available at: http://dendrome.ucdavis.edu/NealeLab/adept2/dnasam/index.php (verified 8 March 2011).

Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UC DAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture