# Conifer Translational Genomics Network
# Coordinated Agricultural Project

## Genomics in Tree Breeding and Forest Ecosystem Management

---

## Module 10 – Linkage Disequilibrium

***Nicholas Wheeler & David Harry – Oregon State University***

# Moving from family-based to population-based QTL discovery

- Linkage and QTL mapping using pedigreed families
  - *QTL, when located, are on large chromosomal blocks*
  - *With only a few generations, the amount of recombination is limited*
- Association genetics: Identifying QTL using populations comprising unrelated individuals or mixed relationships
  - *QTL are located on small chromosomal blocks. These locations are mapped with great precision relative to closely linked markers*
  - *Linkage blocks are shaped by historical recombination*
  - *Population histories reflect 10's – 1000's of generations*

# Chromosome blocks in families and populations

- Family-based linkage mapping (a) involves tracking a QTL, here denoted as “m”, over a few generations in larger chromosomal blocks
- Population-based association mapping (b) tracks “m” on smaller chromosomal segments, taking advantage of historical recombination

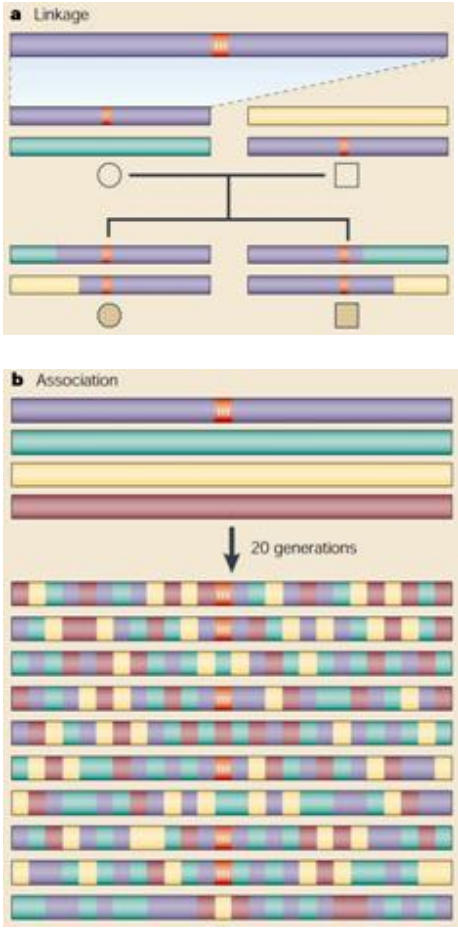



Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Cardon and Bell, 2001.

# It is a question of resolution

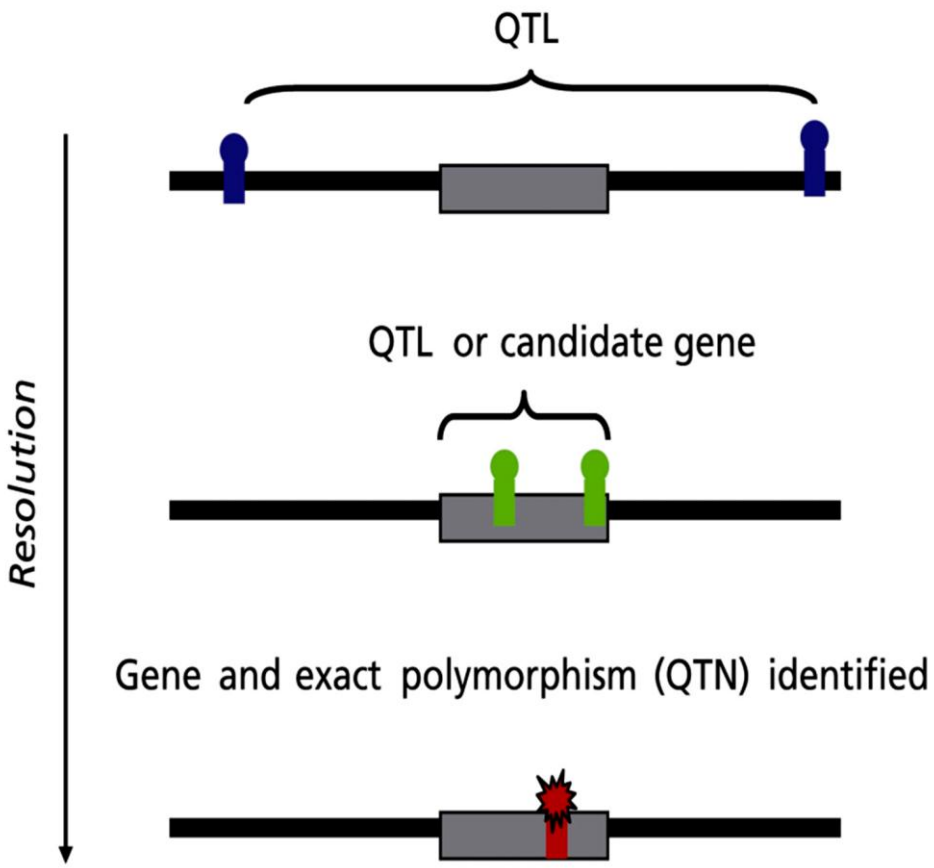



Figure Credit: Modified from Grattapaglia. 2007

# From families to populations: Linkage to linkage disequilibrium

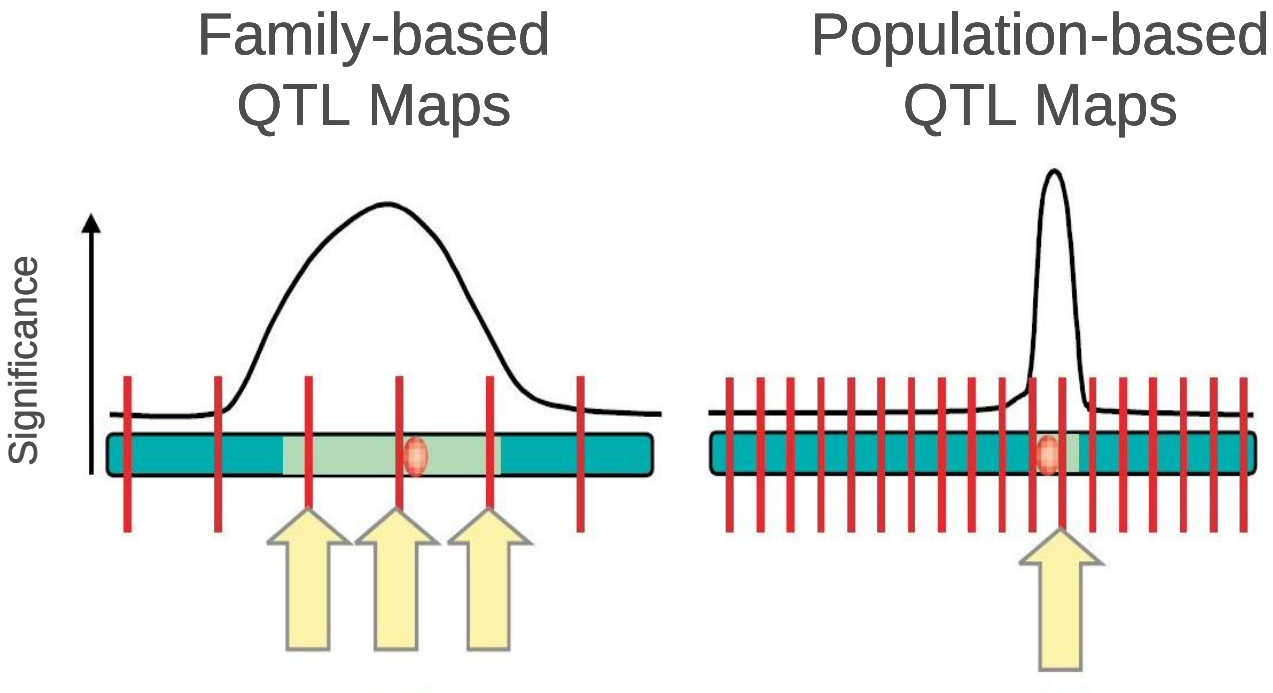



Figure Credit: Modified from Rafalski, 2002

# Comparing approaches

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

# Linkage disequilibrium (LD): The foundation of association genetics

- LD measures non-random associations among alleles at different loci (or non-random associations among SNPs)
- LD is the basis for associating markers with traits. It is the “glue” that binds them
- LD also provides insights into population history, which helps in selecting experimental populations for marker-trait associations
- Estimating LD and understanding how it is organized in populations is crucial for deciding how to sample marker genotypes
- Knowing how population history can affect LD is essential for avoiding pitfalls and spurious false-positives

# A conceptual view of LD

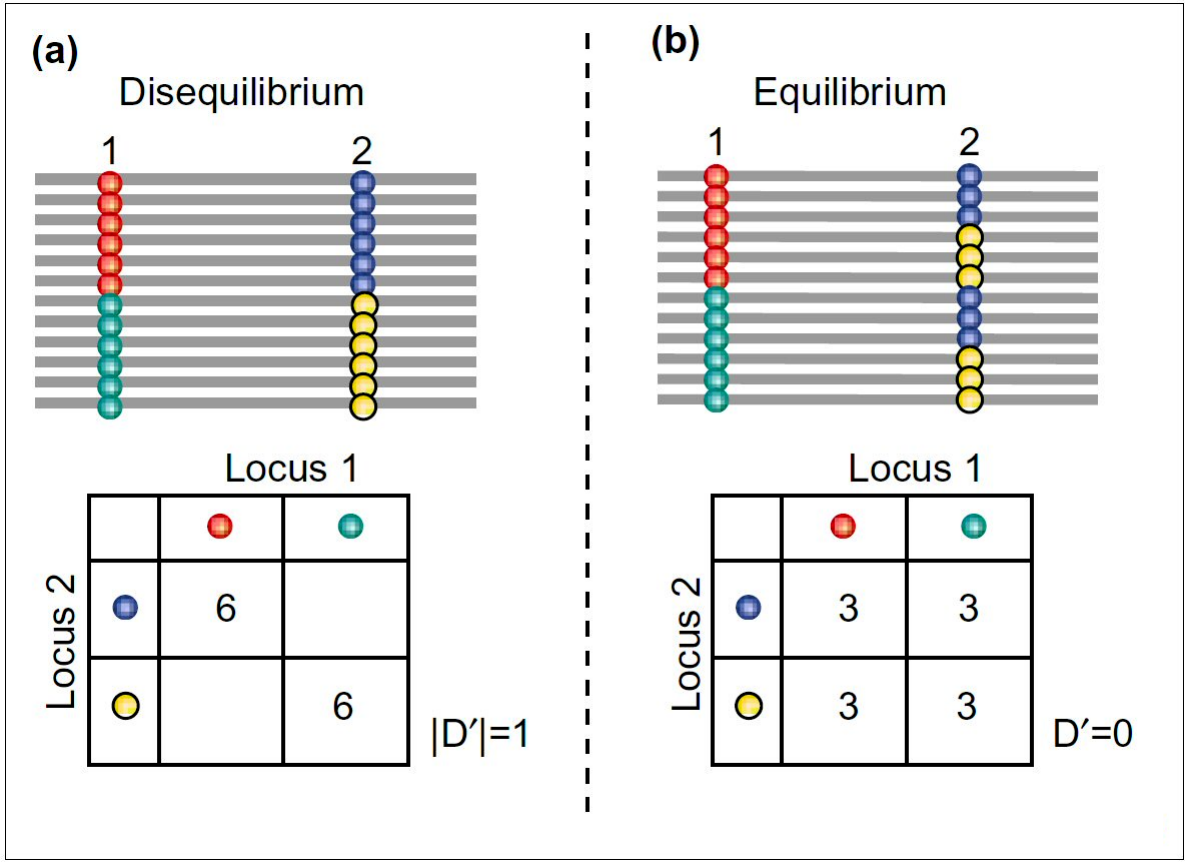



Figure Credit: Reprinted from Current Opinion in Plant Biology, 5, Rafalski, Applications of single nucleotide polymorphisms in crop genetics, 94-100, copyright 2002, with permission from Elsevier.

# Calculating LD (for biallelic loci)

- Pair-wise single-locus allele frequencies predict frequencies for each of four gamete types (left)
- D = 0 (center) implies that predicted = observed gamete frequencies
- D measures the degree to which observed and predicted gamete frequencies differ (right)

Allele & Gamete Frequencies

| | B | b | |
|---|---|---|---|
| A | $P_{AB}$ | $P_{Ab}$ | $p_A$ |
| a | $P_{aB}$ | $P_{ab}$ | $p_a$ |
| | $p_B$ | $p_b$ | |

Observed Freq, D = 0
$P_{AB} = p_A p_B$

| | B | b | |
|---|---|---|---|
| A | .25 | .25 | .50 |
| a | .25 | .25 | .50 |
| | .50 | .50 | |

Observed Freq, D ≠ 0
$P_{AB} \neq p_A p_B$

| | B | b | |
|---|---|---|---|
| A | .40 | .10 | .50 |
| a | .10 | .40 | .50 |
| | .50 | .50 | |

$D = P_{AB} - p_A\, p_B$

$D = P_{AB}\, P_{ab} - P_{Ab}\, P_{aB}$

$D = 0.40 - 0.5*0.5 = 0.15$

$D = 0.4*0.4 - 0.1*0.1 = 0.15$

Figure Credit: David Harry, Oregon State University

# LD can be positive (+) or negative (-)

$D = P_{AB} - p_A\, p_B$

$D = P_{AB}\, P_{ab} - P_{Ab}\, P_{aB}$

$D = 0.40 - 0.5*0.5 = 0.15$

$D = 0.4*0.4 - 0.1*0.1 = 0.15$

| | B | b | |
|---|---|---|---|
| A | .40 | .10 | .50 |
| a | .10 | .40 | .50 |
| | .50 | .50 | |

$D = P_{AB} - p_A\, p_B$

$D = P_{AB}\, P_{ab} - P_{Ab}\, P_{aB}$

$D = 0.10 - 0.5*0.5 = -0.15$

$D = 0.1*0.1 - 0.4*0.4 = -0.15$

| | B | b | |
|---|---|---|---|
| A | .10 | .40 | .50 |
| a | .40 | .10 | .50 |
| | .50 | .50 | |

Figure Credit: David Harry, Oregon State University

# Standardized measures for LD

- Our definition of LD means that its magnitude depends on allele frequencies
- D values of 0.01 in one population may be small, and yet in another, may be large — depending on allele frequencies
- From our previous example
  - $D = P_{AB} - p_A\, p_B$
  - $D = 0.40 - 0.5*0.5 = 0.15$
- How large is D = 0.15?
- Consequently, two standardized measures of LD were created
  - $D'$ and $r^2$

# Standardized measures for LD: D’

$$|D'| = \frac{D_{AB}}{\min(p_A p_b, p_a p_B)} \qquad \text{When } D_{AB} > 0$$

$$|D'| = \frac{D_{AB}}{\max(-p_A p_B, -p_a p_b)} \qquad \text{When } D_{AB} < 0$$

- Read “D prime”, D' ranges from 0 to 1
- D' is maximized (D' = 1) whenever a gamete type is missing, as would happen for a recent mutation
- However, D' is unstable when alleles are rare, as often happens for recent mutations
- D' can be made more reliable by establishing a minimum threshold frequency for minor alleles, e.g. MAF ≥ 0.05; or MAF ≥ 0.10

# Standardized measures for LD: $r^2$

$$r^2 = \frac{D_{AB}^2}{p_A p_a p_B p_b}$$

- D is the covariance between alleles at different loci
- Can consider $r^2$ to be the square of the correlation coefficient
- Note that $r^2$ can only attain a value of 1 when allele frequencies at the two loci are the same
- Like a correlation coefficient, $r^2$ can be used to assess to what extent variation in one marker explains variation in a second
- Both measures are often used, as D´ and $r^2$ are sensitive to different factors (e.g., recombination, haplotype history, allele frequencies)

*Devlin and Risch, 1995*

# LD in populations: Determining phase

- LD metrics such as $r^2$ or D' are based on counts or frequencies of gametes or haplotypes (e.g., $P_{AB}$ vs. $P_{Ab}$)
- Diploid genotypes create challenges: When individuals are heterozygous for two loci, how do we know which alleles are associated?
- In the following example, phase is unknown

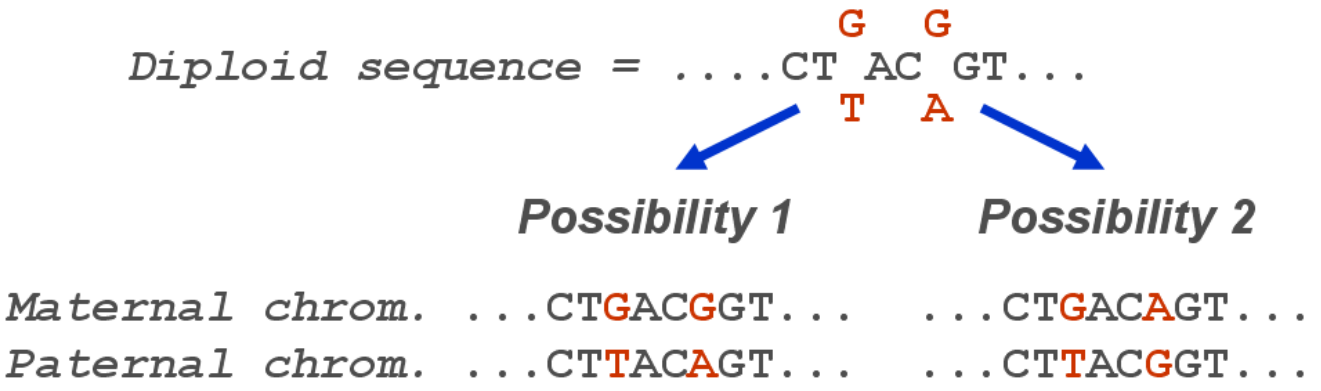





Figure Credit: Glenn Howe, Oregon State University

# Approaches for determining phase

- Phase can be observed directly in haploids (best approach)
  - *Single sperm*
  - *Conifer megagametophytes*
- Determine sequence (hence phase) using cloned DNA
  - *Cloned fragments are copies of individual chromosomes*
  - *Larger clones yield more extensive information on phase*
- Statistically infer phase from population data
  - *Determine haplotype frequencies from unambiguous genotypes, e.g., AB/AB; AB/Ab; Ab/Ab; aB/aB; etc*
  - *Use these estimates to infer haplotypes for ambiguous genotypes (AB/ab and Ab/aB)*
- Computer programs exist to make these calculations

# Statistical tests for LD

- As with many such measures, statistical significance depends on sample sizes, allele frequencies, and strength of association. How can we assess the significance of LD?
- LD between two loci with two alleles/locus
  - *D* } *Fisher's exact test or*
  - *D'* } *Likelihood ratio test*
  - $r^2$ } $\chi 2$
- LD can also be calculated for loci with more than two alleles, for unknown linkage phase of double heterozygotes, and for samples of rare alleles, but that goes well beyond what we need to know here

# Biology of linkage disequilibrium

- What does LD mean biologically?
- What promotes LD
  - *Linkage*
  - *Population admixture*
  - *Selection / epistasis*
- What affects LD
  - *How is LD maintained?*
  - *How does LD change?*

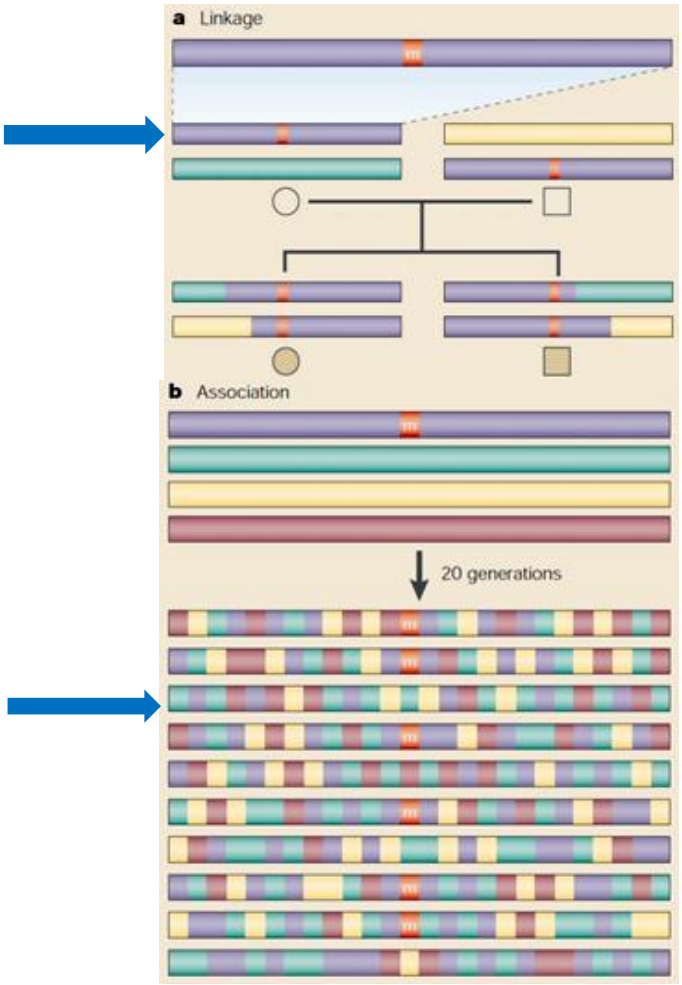



Figure Credit: Modified from Cardon and Bell, 2001

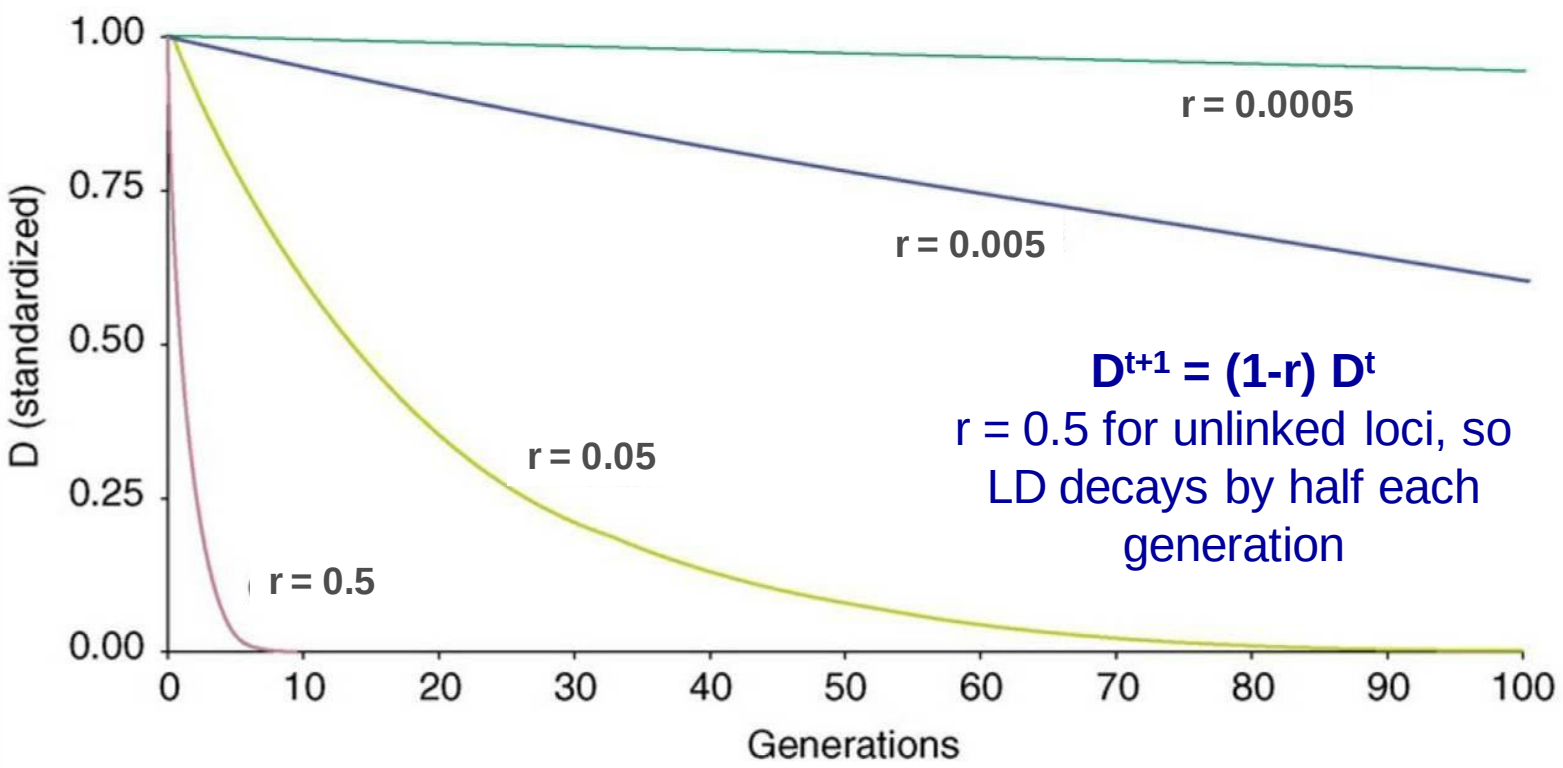
# LD and random mating

- HWE and LD (or LE) both pertain to random (or non-random) associations of alleles and genotypes
  - *HWE describes associations of alleles at the same locus*
  - *LD (or LE) measures associations of alleles at different loci*
- HW proportions are restored by one generation of random mating
- However, once established, LD persists for some time, even in random mating populations
- How quickly LD dissipates depends on several factors

# Factors affecting the decay of LD

- Recombination rate — describes how often linked loci tend to recombine
  - *Closely linked loci rarely recombine*
- Selfing — decreases the frequency of double heterozygotes, which decreases the opportunity for creation of new recombinants
- Small populations or population bottlenecks — mechanism is analogous to the reduction of heterozygosity in small populations, so double heterozygotes are also less common
- Selection — can increase the frequency of certain haplotypes, counteracting LD decay from recombination
  - *Selection favoring one or a few haplotypes (positive selection)*
  - *Selection favoring heterozygotes (or genotypic combinations in different environments, balancing selection)*

# Rate of LD decay driven by recombination (r)




$D^{t+1} = (1-r) D^t$

r = 0.5 for unlinked loci, so LD decays by half each generation

*D is expressed in standardized units as D' or $r^2$*

Figure Credit: Modified from Mackay and Powell, 2007.

# Effect of mating system on LD decay

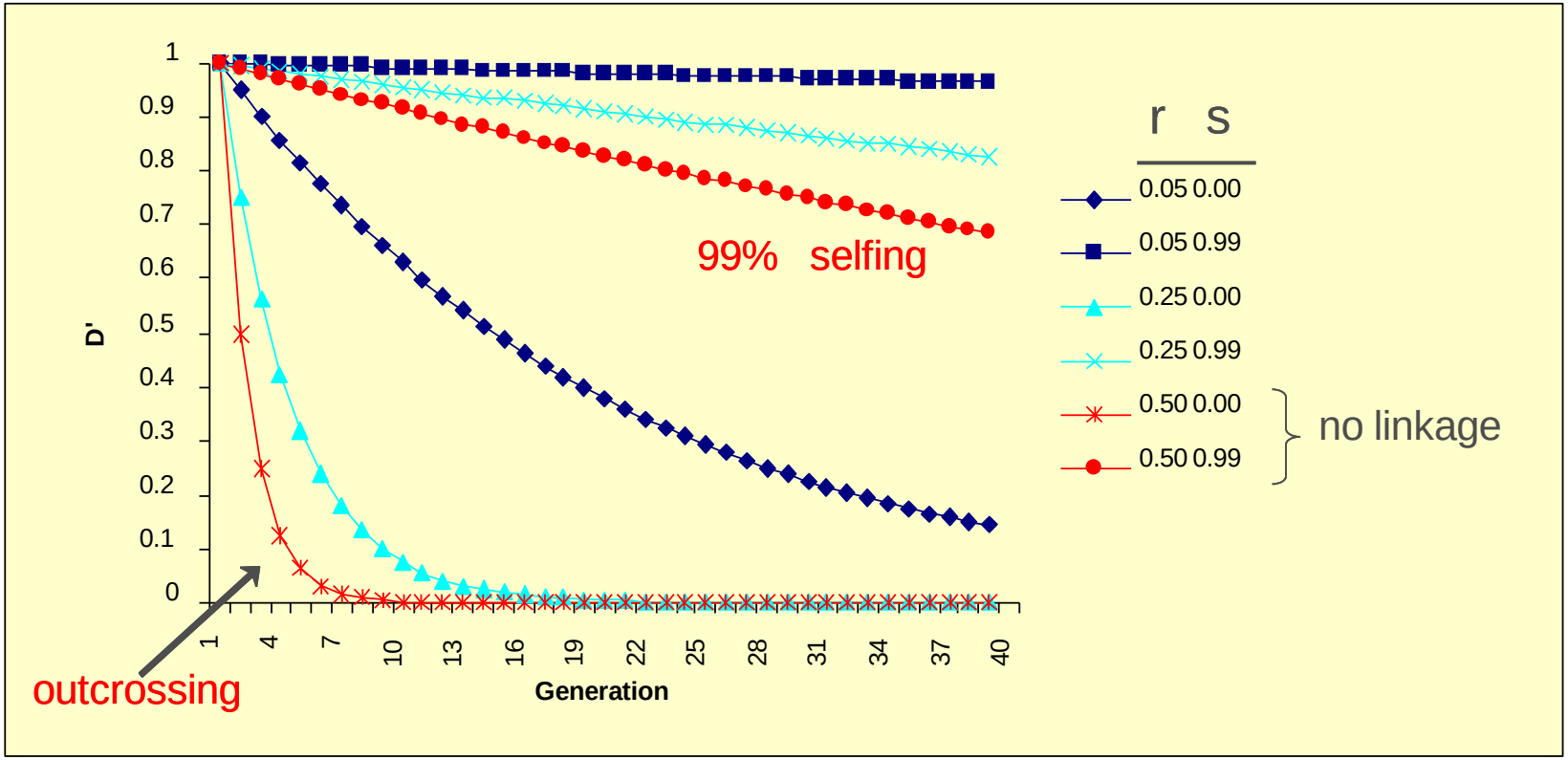



Figure Credit: Jennifer Kling, Oregon State University

# Average decay for LD in *Pinus taeda*

- Conifers are primarily outcrossing and have large Ne
- Therefore, LD decays rapidly
- Figure shows average decay of LD over 19 candidate genes in loblolly pine (*Pinus taeda*)
- LD decays to ~$r^2$ = 0.2 within ~1500 bp

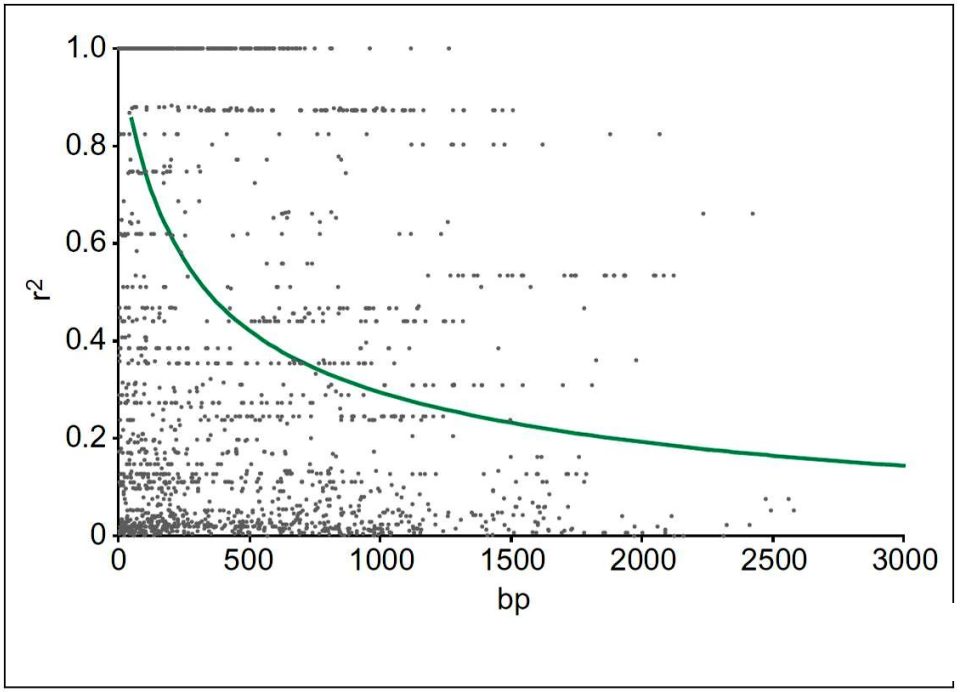



Figure Credit: Reprinted from Trends in Plant Science Vol. 9, Neale, D. B., and O. Savolainen, Association genetics of complex traits in conifers, Pages: 325-330, 2004, with permission from Elsevier.

# Decay of LD in *Eucalyptus*

- Rapid decay of intragenic linkage disequilibrium in the cinnamyl-alcoholdehydrogenase (cad) gene in two *Eucalyptus* species

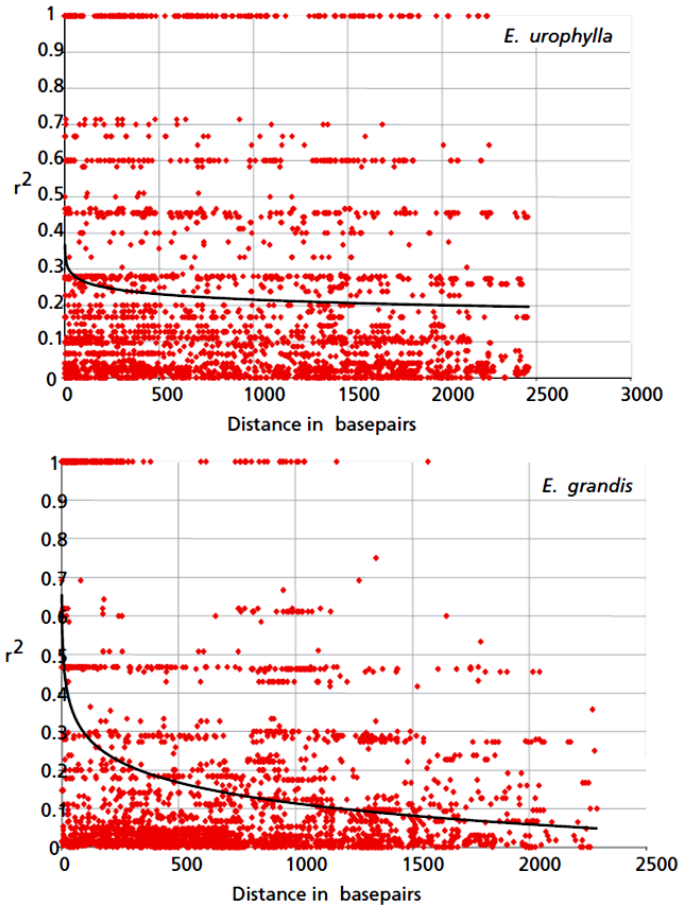



Figure Credit: Grattapaglia and Kirst, 2008. Used with permission of Wiley and Sons

# Extent of LD in various plants

| Species | Mating system | LD range | Reference |
|---|---|---|---|
| Maize | Outcrossing | 0.5–7.0 kb | Remington *et al.* (2001), Ching *et al.* (2002), Palaisa *et al.* (2003) |
| | Outcrossing | 0.4–1.0 kb | Tenaillon *et al.* (2001) |
| Barley | Selfing | 10–20 cM | Stracke *et al.* (2003), Kraakman *et al.* (2004) |
| Tetraploid wheat | Selfing | 10–20 cM | Maccaferri *et al.* (2004) |
| Rice | Selfing | 100 kb | Garris *et al.* (2003) |
| Sorghum | Selfing | < 4 cM | Deu and Glaszmann (2004) |
| | Selfing | ≤10 kb | Hamblin *et al.* (2004) |
| Sugarcane | Outcrossing/Vegetative propagation | 10 cM | Jannoo *et al.* (1999) |
| *Arabidopsis* | Selfing | 250 kb | Nordborg *et al.* (2002) |
| Soybean | Selfing | > 50 kb | Zhu *et al.* (2003) |
| Sugar beet | Outcrossing | < 3 cM | Kraft *et al.* (2000) |
| Potato | Selfing | 0.3–1.0 cM | Gebhardt *et al.* (2004), Simko (2004) |
| Lettuce | Selfing | ~200 kb | van der Voort *et al.* (2004) |
| Grape | Vegetative propagation | > 500 bp | Rafalski and Morgante (2004) |
| Norway spruce | Outcrossing | ~100–200 bp | Rafalski and Morgante (2004) |
| Loblolly pine | Outcrossing | 100–150 bp | González-Martínez (2004) |
| Loblolly pine | Outcrossing | ~1500 bp | Neale and Savolainen (2004) |

Figure Credit: With kind permission from Springer Science+Business Media: Plant Molecular Biology, Linkage disequilibrium and association studies in higher plants: Present status and future prospects, 57, 2005, page 475, Gupta, P. K., R. Rustgi, and P. L. Kulwal, Table 2.

# Tools for visualizing LD: Haploview

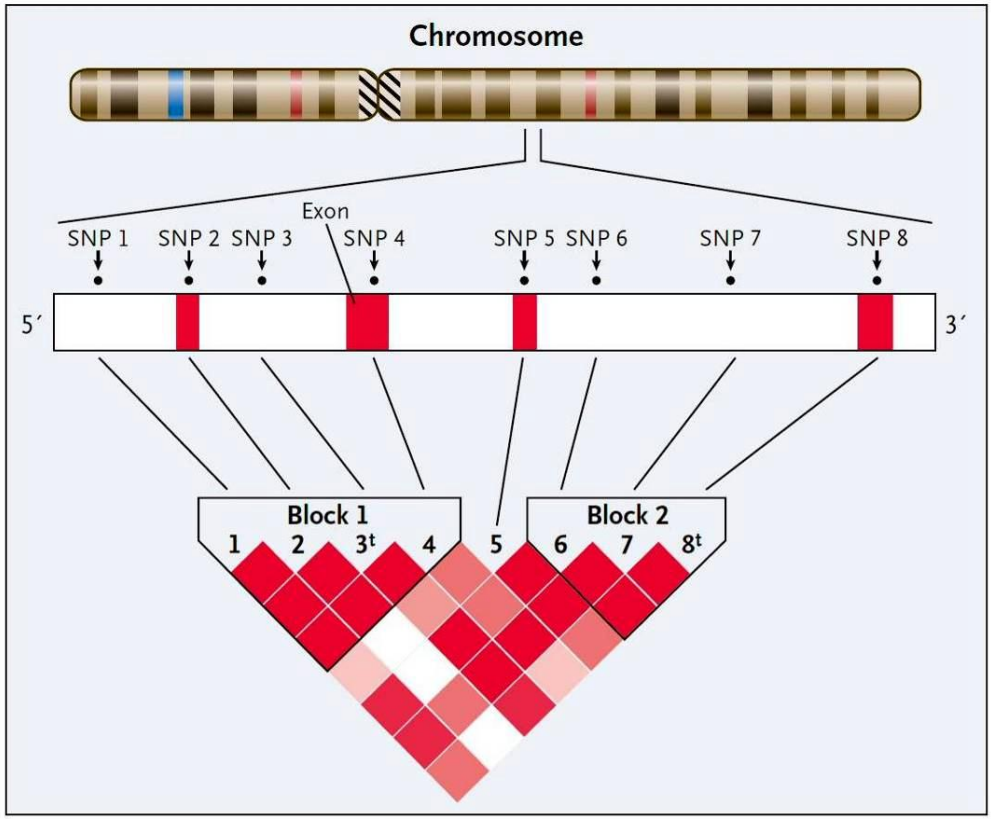



Figure Credit: Christensen and Murray, 2007. Reprinted with permission of the Massachusetts Medical Society.

# Recombination and demography shape haploblock structure

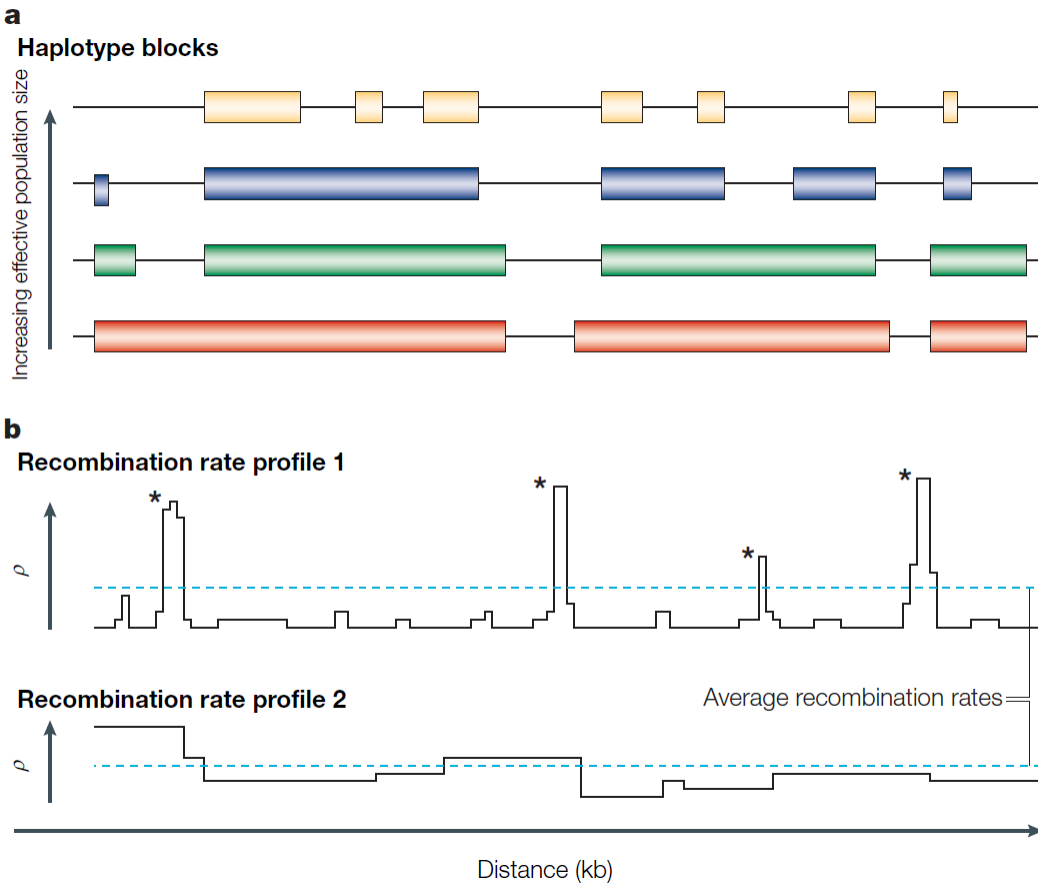



Figure Credit: Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Stumpf and McVean, 2003.

# Recombination “hotspots” delineate haplotype boundaries in human populations

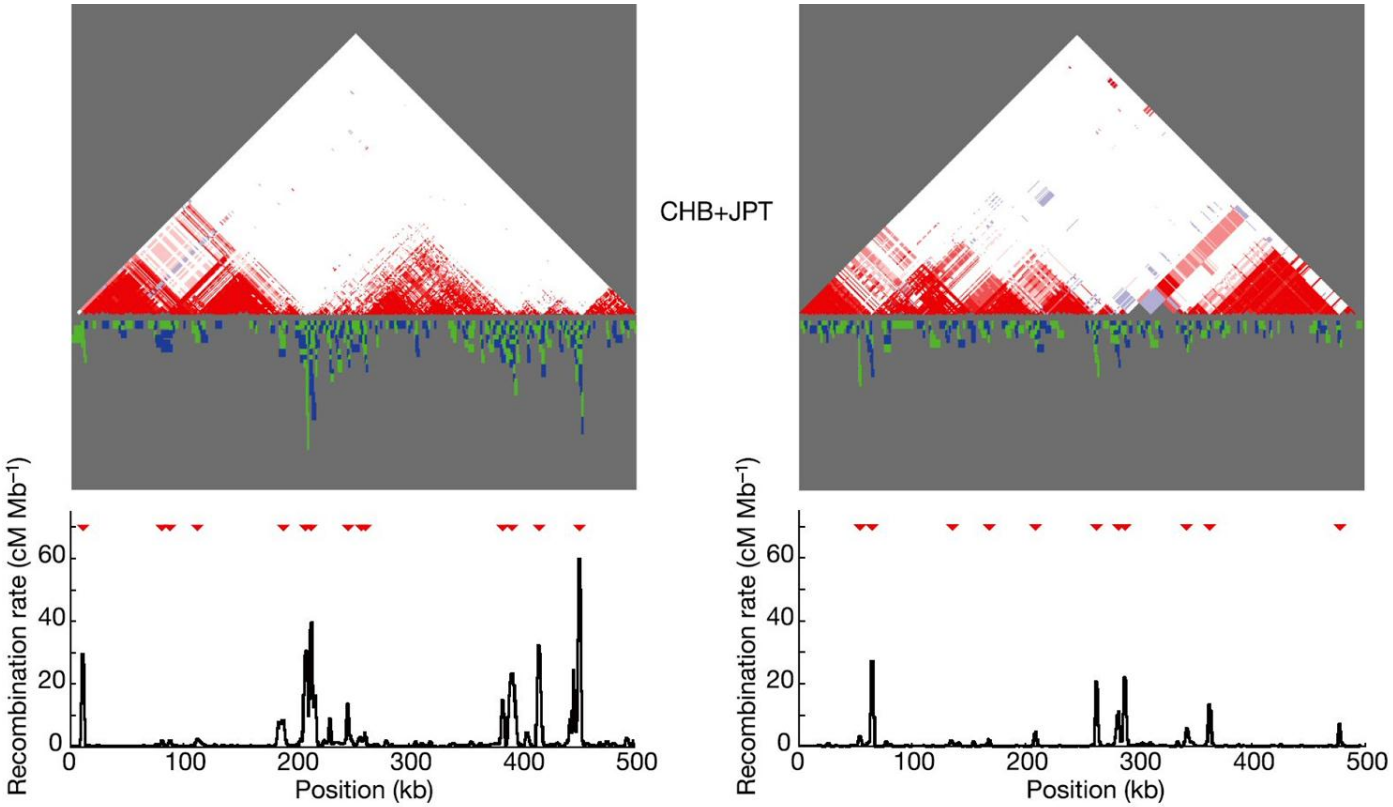



Figure Credit: Modified from HapMap Consortium, 2005.

# LD within and among nearby genes in *P. taeda*

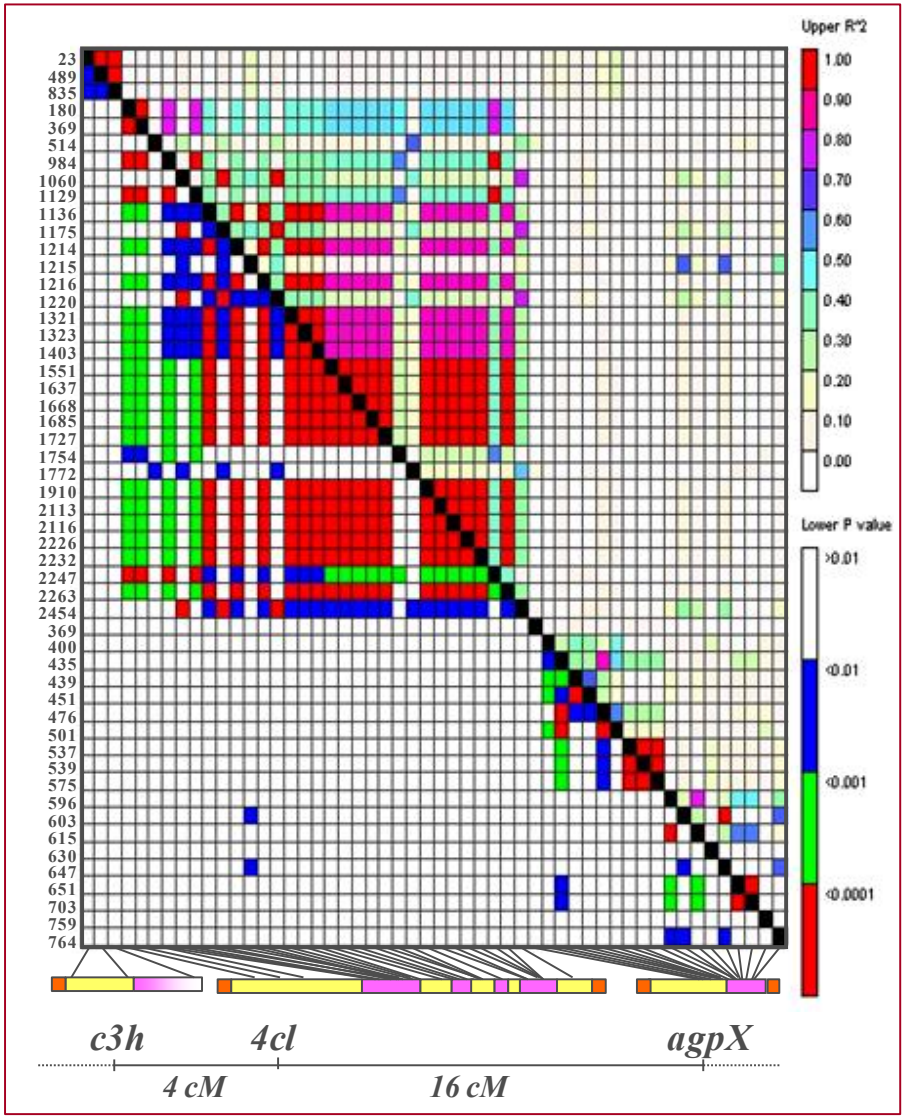
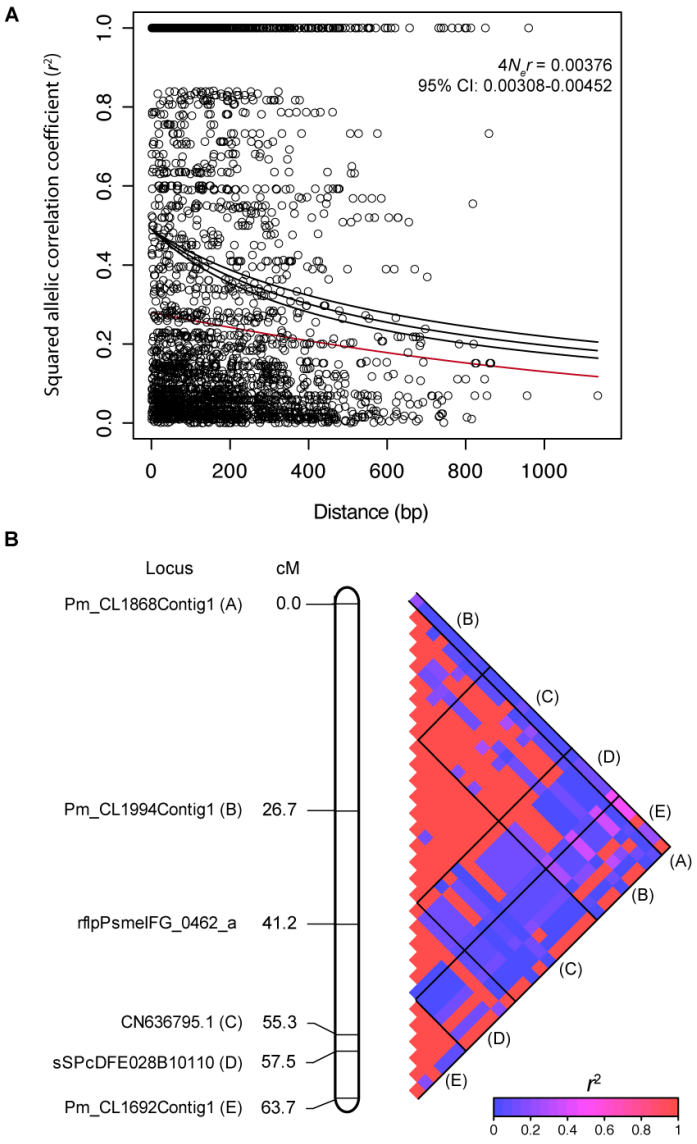



Figure Credit: David Neale, University of California, Davis.

# Patterns of intra and interlocus LD for coastal Douglas-fir



# Haplotype genealogy and LD

- Colored circles are polymorphic sites (e.g. SNPs) located along haplotypes with evolutionary histories shown on the left
- LD reflects mutational events bound by history
- Areas of LD (circled) don't tell us about the presence or nature of selection
- LD is reduced by recombination
- Amount of reduction depends when recombination occurs relative to haplotype history

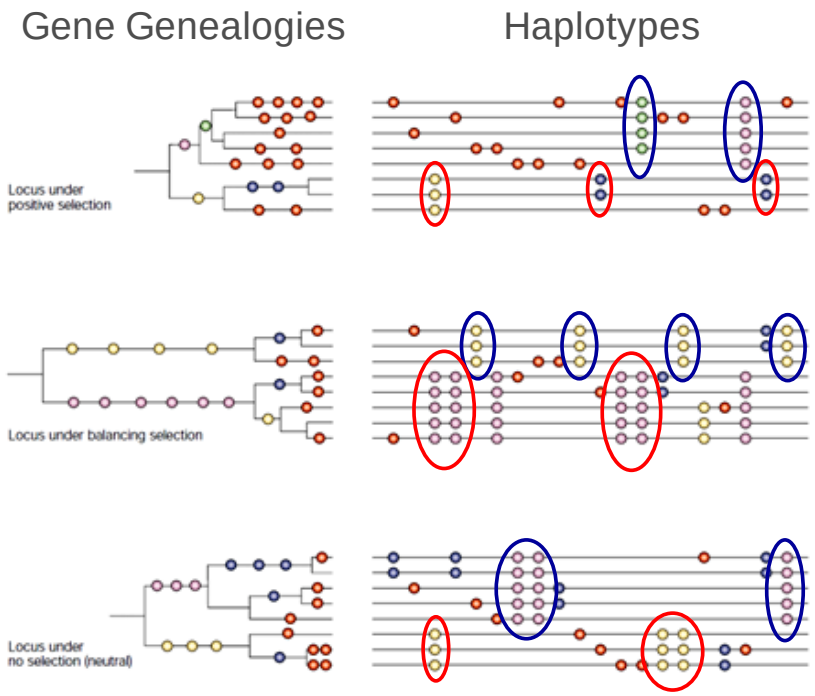



Figure Credit: Modified from Bamshad and Wooding, 2003

# References cited


- Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. J. Daly, P. Donnelly, R. A. Gibbs, et al. 2005. Nature 437: 1299-1320. (Available online at: http://dx.doi.org/10.1038/nature04226) (verified 1 June 2011).
- Bamshad, M., and S. Wooding. 2003. Signatures of natural selection in the human genome. Nature Reviews Genetics 4: 99-111. (Available online at: http://dx.doi.org/10.1038/nrg999) (verified 1 June 2011).
- Cardon, L., and J. Bell. 2001. Association study designs for complex diseases. Nature Reviews Genetics 2: 91-99. (Available online at: http://dx.doi.org/10.1038/35052543) (verified 1 June 2011).
- Christensen, K., and J. C. Murray. 2007. What genome-wide association studies can do for medicine. New England Journal of Medicine 356: 1094-1097. (Available online at: http://dx.doi.org/10.1056/NEJMp068126) (verified 1 June 2011).
- Devlin, B., and N. Risch. 1995. A comparison of linkage disequilibrium measures for fine scale mapping. Genomics 29: 311-322. (Available online at: http://dx.doi.org/10.1006/geno.1995.9003) (verified 1 June 2011).
- Eckert, A. J., J. L. Wegrzyn, B. Pande, K. D. Jermstad, J. M. Lee, J. D. Liechty, B. R. Tearse, K. V. Krutovsky, and D. B. Neale. 2009. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal douglas fir (*Pseudotsuga menziesii var. menziesii*) Genetics 183: 289-298. (Available online at: http://dx.doi.org/10.1534/genetics.109.103895) (verified 1 June 2011).

# References cited


- Gibson, G., and S. Muse. 2009. A primer of genome science. Sinauer Associates, Sunderland, MA.
- Grattapaglia, D. 2007. Marker–assisted selection in Eucalyptus. p. 251-281. *In*. E. P. Guimaraes, J. Ruane, B. D. Scherf, A. Sonnino, and J. D. Dargie (ed.) Marker assisted selection: Current status and future perspectives in crops, livestock, forestry and fish. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.
- Grattapaglia, D. and M. Kirst. 2008. *Eucalyptus* applied genomics: from gene sequences to breeding tools. New Phytologist 179: 911-929. (Available online at: http://dx.doi.org/10.1111/j.1469-8137.2008.02503.x) (verified 1 June 2011).
- Gupta, P. K., R. Rustgi, and P. L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Molecular Biology 57: 461-485. (Available online at: http://dx.doi.org/10.1007/s11103-005-0257-z) (verified 1 June 2011).

# References cited


- Mackay, I., and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. Trends In Plant Science 12: 57-63. (Available online at: http://dx.doi.org/10.1016/j.tplants.2006.12.001) (verified 1 June 2011).
- Neale, D. B., and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends in Plant Science 9: 325-330. (Available online at: http://dx.doi.org/10.1016/j.tplants.2004.05.006) (verified 1 June 2011).
- Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5: 94-100. (Available online at: http://dx.doi.org/10.1016/S1369-5266(02)00240-6) (verified 1 June 2011).
- Stumpf, M.P.H., and G.A.T. McVean. 2003. Estimating recombination rates from population-genetic data. Nature Reviews Genetics 4: 959-968. (Available online at: http://dx.doi.org/10.1038/nrg1227) (verified 1 June 2011).

# Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UCDAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture