UNIVERSITY OF HOHENHEIM

Signatures and tests of selection

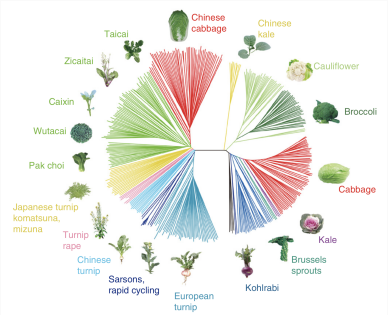
3502-470 Plant Genetic Resources

Prof. Karl Schmid

SS 2025

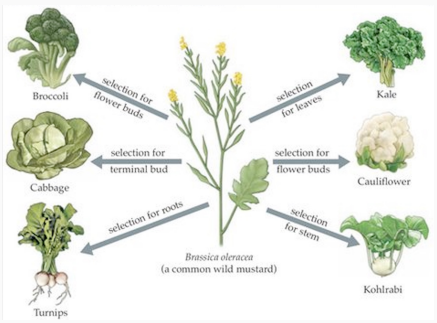
Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

Phylogeny of Brassica vegetable crops



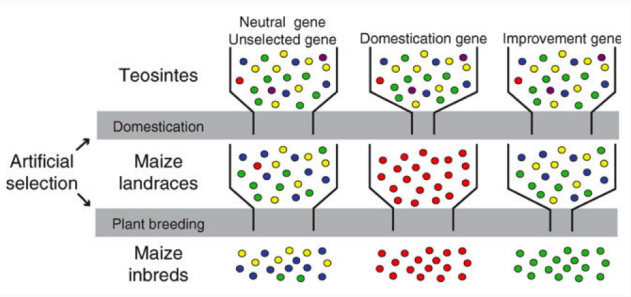
Cheng et al. Nature Genetics (2016)

Traits selected during Brassica domestication



Source: Unknown

Loss of diversity during crop evolution

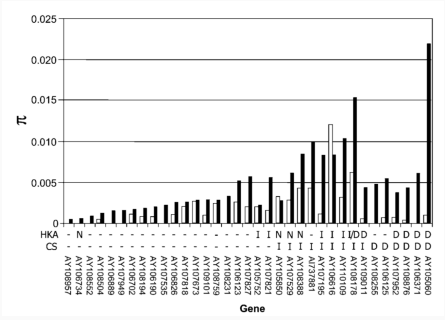


Yamasaki et al. Plant Cell (2005)

Identification of selected genes

- Genetic mapping: QTL mapping of crosses or genome-wide association studies of diverse panels
- Selection scans: Tests of selection throughout genome

Loss of diversity during maize evolution



Yamasaki et al. Plant Cell (2005)

Applications of coalescent theory in plant genetic resources

- Core collections: Establish a subset of accessions with high genetic diversity
- Geographic analysis: Identify accessions adapted to different climates
- Suitable populations for introgression: Identify exotic germplasm
- Heterotic groups: Identify genetically differentiated individuals for hybrid breeding
- Allele mining: Identify new alleles for useful genes (i.e., resistance genes)

The Neutral Theory of Molecular Evolution

Motoo Kimura
(1924-1994)

Basic assumptions:

- Most segregating variation is neutral or nearly neutral
- Why?
  - Strongly advantageous mutations are fixed rapidly
  - Strongly deleterious mutations are removed rapidly
- Consequence: Simple predictions about the level and patterns of genetic variation

Use neutral model as null hypothesis for neutrality tests.

Application of the theory

Simple predictions in the neutral model:

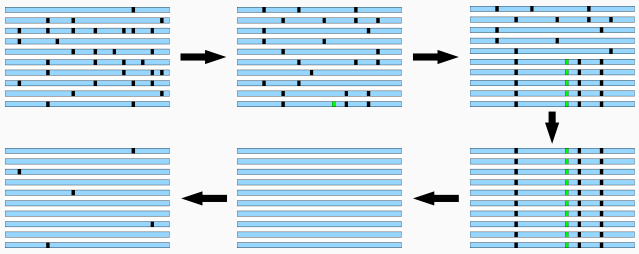
- We look at the genetic variation in a sample of n sequences
- Recall $\theta = 4N_e\mu$, the scaled mutation rate. Let S be the number of segregating sites and Π the number of mismatches per pair of sequences (see genetic diversity lecture)
- Expected values can be computed from coalescent theory,
  - $E(S) = \underbrace{\sum_{i=1}^{n-1} i^{-1}}_{:=a_n} \theta$ (see coalescent lecture)
  - $E(\Pi) = \theta$
- 2 estimates for θ: Observe S and Π in a sample, compute estimates $\theta_W = S/a_n$ (Watterson's estimator), $\theta_\Pi = \Pi$ (Tajima's estimator)
- Other properties can also be derived, e.g. expected values of the site frequency spectrum (SFS) of polymorphisms.

Types of natural selection

- Purifying or negative selection: Deleterious/disadvantageous alleles are lost.
- Positive (directional) selection ("Positive Darwinian selection"): Advantageous/beneficial alleles increase in frequency and eventually become fixed.
- Balancing selection: Different alleles are maintained.
- Epistatic selection: Interaction between different genes

All types of selection leave (distinguishable) footprints in the genome and that can be detected.

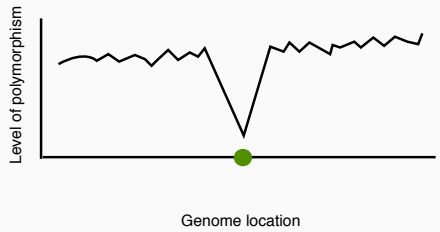
Outline of a selective sweep



The different stages of a selective sweep

- Early stage: Intermediate, "normal" linkage disequilibrium (LD)
- Intermediate stage: Increased LD, high proportion of derived polymorphisms, strong differentiation between haplotypes of ancestral and derived (selected) polymorphism
- Final stage: Lack of polymorphism, very strong LD
- After sweep: Excess of rare polymorphisms

The "genomic neighbourhood" of selective sweeps



- "Sliding window" analysis of genetic variation
- Linked neutral polymorphism decreases around a selected polymorphism
- The length of the window with reduced variation depends on:
  - the rate of recombination, c
  - the strength of selection, s

Outline of the neutrality test statistic Tajima's D (Tajima, 1989)

1. Sequence genes or genomes from different individuals of the same population.
2. Compare two estimators of nucleotide diversity, θ_W and θ_Π.
3. Should be close under neutral evolution (expected values are equal).
4. Calculated as standardized difference:
$$D = \frac{\theta_\Pi - \theta_W}{\sqrt{\widehat{Var}(\theta_\Pi - \theta_W)}}$$
5. $D < 0$: Positive selection (or exponential population growth)
6. $D > 0$: Balancing selection (or population admixture)
7. Simulate a neutral model to generate a null distribution
8. Compare simulated Tajima's D values with observed data

Tajima's D

Hypothesis testing:

- H_0: Neutral evolution
- H_1: Non-neutral evolution

- We test for non-neutrality, not specifically for selection
- To distinguish between several explanations for non-neutrality (selection vs. demography), we need more information (data from different loci, different populations, LD information, results from other analyses, ...)
- Tajima's D doesn't use all information present in the sequences (e.g., no linkage information)
- Neutral model has strong assumptions: No recombination, no population size change, ...
  Coalescent simulations can be adjusted to these factors

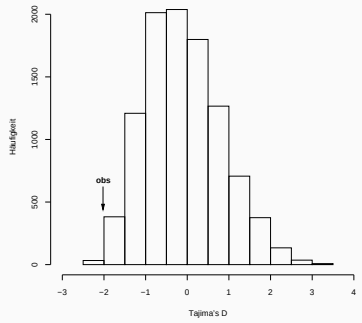
Example: Polymorphisms in *Drosophila melanogaster*

Type of polymorphism	S	θ_S	θ_Π	D
RFLP	53	11.21	11.92	0.213
Small insertions/deletions	40	8.46	10.02	0.607
Large insertions/deletions	15	3.17	0.94	-2.071

Test of significance by coalescence simulations (Hudson's ms)

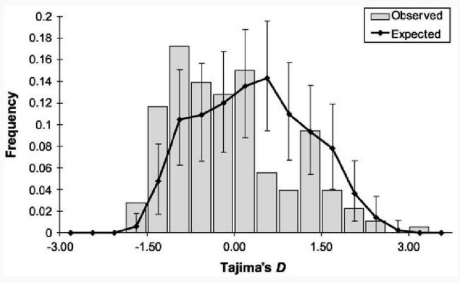
Polymorphism	Observed	2.5% Percentile	97.5% Percentile
RFLP	0.213	-1.618	2.172
Small indels	0.607	-1.672	1.758
Large indels	-2.071	-1.400	2.539

Comparison of simulated with observed data



Tajima's *D* distribution in *Arabidopsis thaliana*

Analysis of 150 short genomic loci in 12 individuals



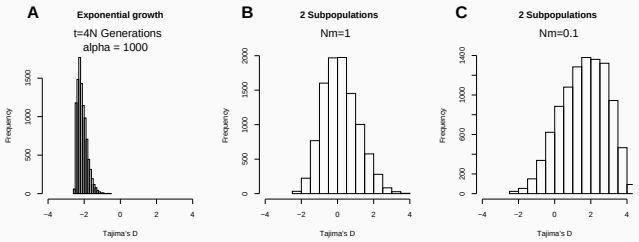
Schmid et al., Genetics (2005)

Studies in other plant species

Species	Accessions	Loci	Length	π_P/L	Average Tajima's D	P Value
Arabidopsis thaliana	96	846	583	0.007	-0.8	***
Arabidopsis thaliana	12	185	414	0.010	-0.4	***
Arabidopsis lyrata	140	77	530	0.0135	0.32	*
Boechera drummondii	46	86	591	0.0041	-0.46	
Inbred Maize	14	774	?	?	0.04	n.s.
Teosinte	16	774	?	?	-0.50	***
Sorghum bicolor	16	204	671	0.0038	-0.08	

Table 1: Distribution of Tajima's D values in various plant species

Effect of demographic history on Tajima's D



- 10,000 simulation runs
- Left: Strong exponential growth
- Middle: Two subpoplations with migration rate $NM = 1$
- Right: Two subpopulation with migration rate $NM = 0.1$

Summary

- There are three major types of selection on the molecular level: Purifing, positive and balancing selection.
- Several tests of neutral evolution were developed that are based on the comparison of nucleotide variation within and between species, and on the comparison of divergence between species.
- Tajima's D is an important test for the analysis of selection acting on individual genes. This summary statistic measures the frequency distribution of polymorphisms.
- Coalescent simulations are a powerful tool to compare the likelihood of different evolutionary models given the observed data.

Further reading

- Hartl and Clark, Principles of population genetics, Chapter 4
- Jensen et al. (2007) Approaches for identifying targets of positive selection. Trends in Genetics 23:568-577 (2007)
- Walsh, Using molecular markers for detecting domestication, improvement and adaptation genes. Euphytica 161:1-17 (2008)
- Nielsen, Molecular signatures of natural selection. Annu Rev Genet 39:197–218 (2005)

References i

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. Trends in Genetics 23(11):568–577, DOI 10.1016/j.tig.2007.08.009, URL http://www.sciencedirect.com/science/article/pii/S0168952507002971