Molecular markers in plant systematics and population biology

2. Overview of applications and questions

Tomáš Fér

tomas.fer@natur.cuni.cz

Molecular markers overview

1. proteins – isozymes
2. DNA markers
   - RFLP (**R**estriction **F**ragment **L**ength **P**olymorphism)
   - PCR based – analysis of DNA fragments
     - order of nucleotides – DNA sequences
     - „whole genome“ analysis – fragment length polymorphism
       - RAPD (**R**andom **A**mplified **P**olymorphic **D**NA)
       - AFLP (**A**mplified **F**ragment **L**ength **P**olymorphism)
       - ISSRs (**I**nter **S**imple **S**equence **R**epeats)
     - information from specific genome regions
       - PCR-RFLP (**P**olymerase **C**hain **R**eaction – RFLP)
       - microsatellites (**S**imple **S**equence **R**epeats – SSRs)
       - SSCP (**S**ingle **S**train **C**onformation **P**olymorphism)…
   - whole genome markers – SNP, whole genome sequencing
     - RADseq
     - Hyb-Seq (target enrichment)
     - de novo sequencing, re-sequencing
     - RNA-seq (transcriptome)

Utility of markers in different types of studies

		Fragment-based			Sequencing				NGS		
	Allo-zymes	RAPD	AFLP	SSR	nDNA	cpDNA	mtDNA (plant)	mtDNA (animal)	Hyb-Seq	RADseq	resequencing
Genetic diversity	++	++	++	++	+++	++	+	++	+++	+++	+++
Population differentiation	+++	++	++	++	+++	++	++	+++	++?	+++	+++
Gene flow	++	(+)	(+)	+++	+++	++	(+)	++	?	+++	+++
Polyploidy	+++	-	(+)	+	++	++	-	-	+++	++	+++
Hybridization	++	++	++	+	++	++	+	+	+++	+++	+++
Phylogeny	(+)	-	++	(+)	+++	+++	(+)	+++	+++	++	+++
Individual genotyping	(+)	+++	+++	+++	+++	-	-	-	?	+++	+++
Phylogeography	(+)	-	++	-	(+)	+++	(+)	+++	(+)	+++	+++
Selection	(+)	(+)	(+)	+	++	-	-	-	++	++	+++
Diversification	?	?	(+)	-	++	++	?	++	+++	+++	+++

+++	excellent	(+)	has been used
++	good	-	unlikely to be usefull or useless
+	OK	?	uncertain or not used

first part based on Lowe et al. 2004

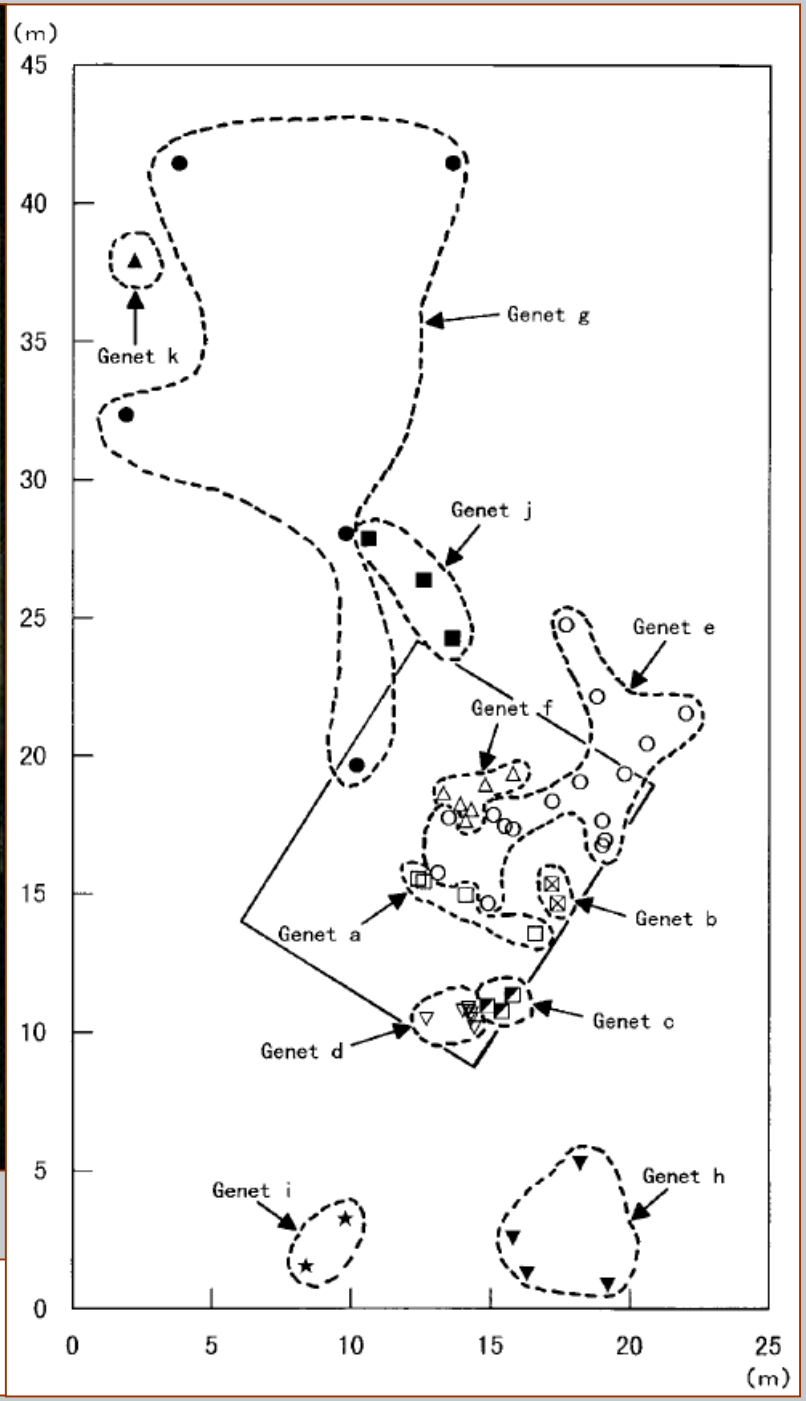
Types of questions

…to be solved with molecular methods

- identification of clones
  - genetic diversity of clonal plants
- reproduction systems
  - molecular evidence
  - mating system evolution
- population genetic structure
  - intrapopulation genetic diversity
  - Hardy-Weinberg equilibrium testing
  - relationship among populations, pattern of genetic variation
  - gene flow
- plant migration
  - phylogeography
  - invasions
- phylogenetic studies, evolution reconstruction
- hybridization/introgression, polyploidization (auto/allo)
- selection, adaptation, diversification, trait evolution, molecular dating

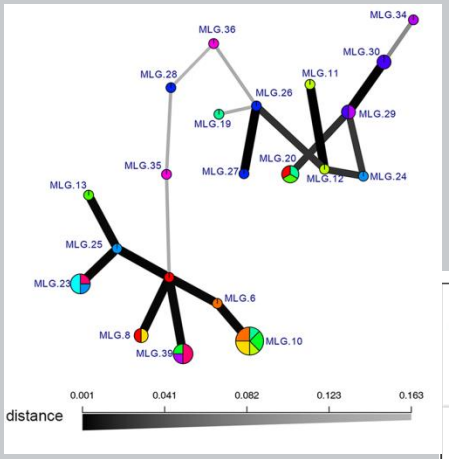
Identification of clones

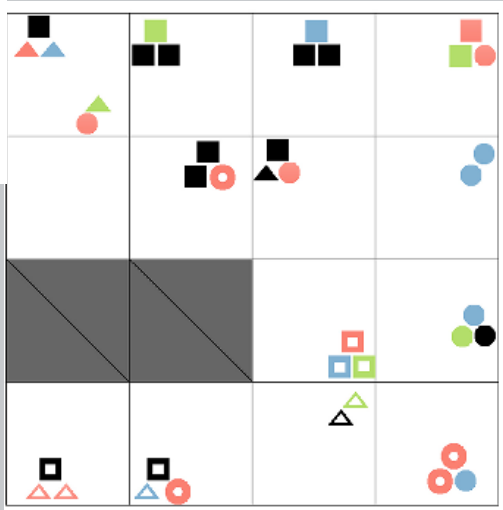
- number of genotypes in population and spatial pattern of genets and ramets
- ratio between vegetative and generative reproduction
- intensity of generative reproduction
  - RSR – *repeated seedling recruitment*
  - ISR – *initial seedling recruitment*

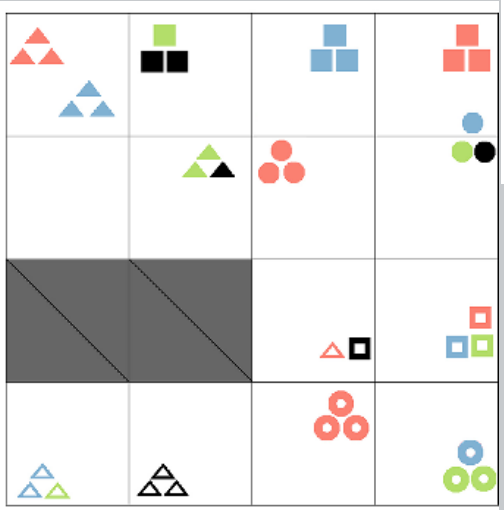


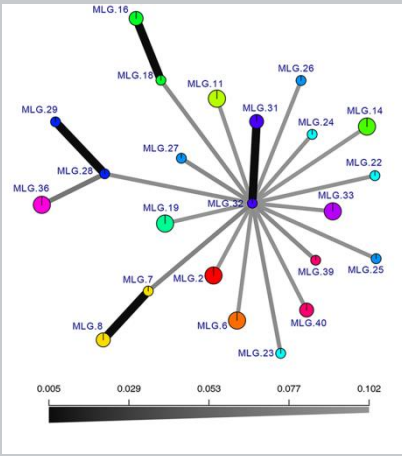
Miwa et al. (2001): Analysis of clonal structure of *Melaleuca cajuputi* (Myrtaceae) at a barren sandy site in Thailand using microsatellite polymorphism. Trees 15:242-248

Clonal structure









Blysmus compressus *Kobresia alpina*

Ning et al. (2018): Contrasting fine-scale genetic structure of two sympatric clonal plants in an alpine swampy meadow featured by tussocks. PLoS ONE15:242-248

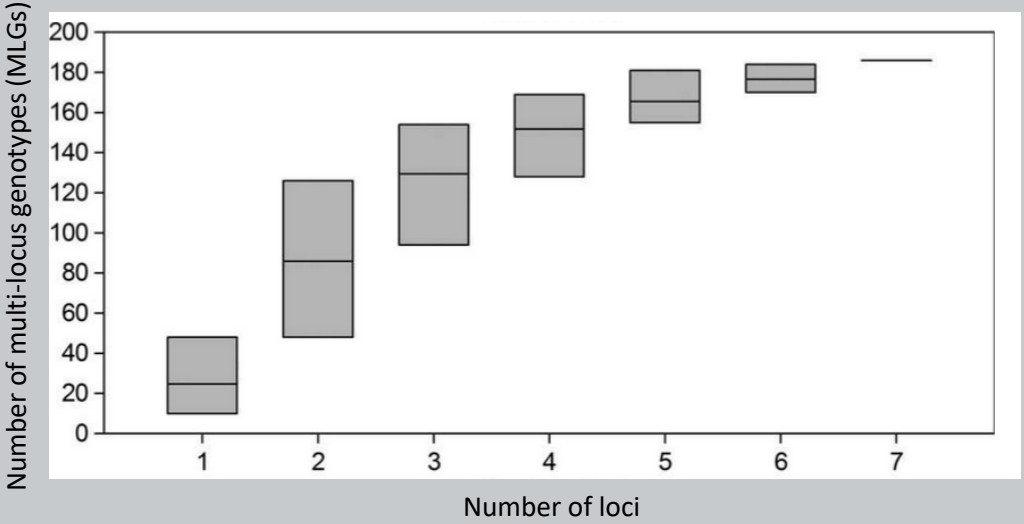
> 5,000 SNPs identified by 2b-RAD

Markers to identify clones

- isozymes – sometimes insufficient variation
- RAPD
- AFLP
- microsatellites
- RADseq

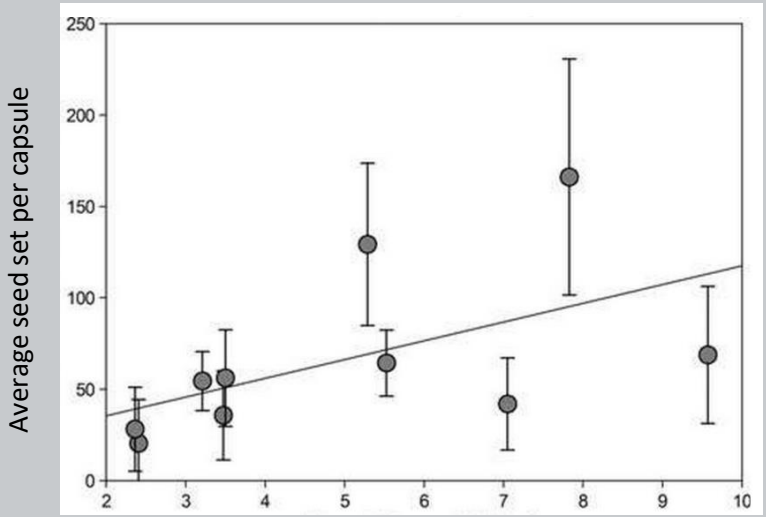
- overestimation of clonal variation
  - ramets of the same genet are detected as different genotypes
  - methodological artifact – *error rate* calculation required (repeated analyses)
  - somatic mutations?
- underestimation of clonal variation
  - genetically independent individuals are detected as a clone
  - insufficient variability of marker
  - low number of markers
  - marker strength calculation (probability that two individuals with the same genotype originated from an independent sexual process)

Clones identification vs. number of markers



Linaria vulgaris

Number of clones vs. seed production



Bartlewicz J. et al. 2015: Population genetic diversity of the clonal self-incompatible herbaceous plant *Linaria vulgaris* along an urbanization gradient. Biol. J. Linn. Soc. 116:603–613.

Type of reproduction system

- Hardy-Weinberg equilibrium – assume *random mating* – frequently not fulfilled:

H-W equilibrium expectations:

- random mating
- infinite population size (no drift)
- no mutations (no new alleles appear)
- no migration
- no selection
- (diploids, sexual reproduction, non-overlapping generations)

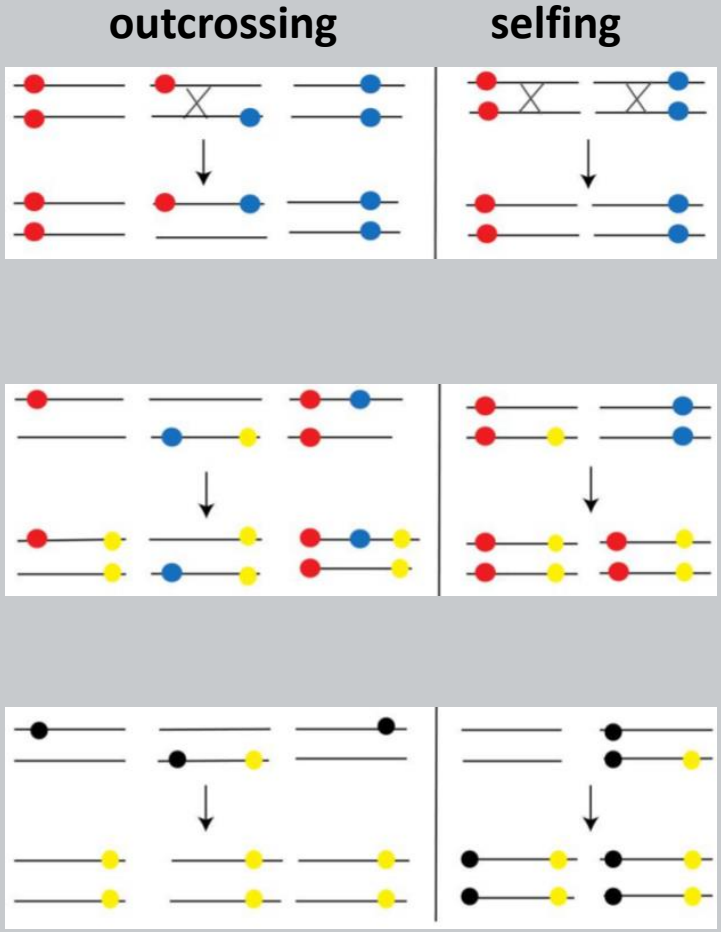
Type of reproduction system

- Hardy-Weinberg equilibrium – assume *random mating* – frequently not fulfilled:
  - *positive assortative mating*
  - *inbreeding* – pollination of plants from neighbourhood
  - regular *inbreeding* – autogamy
- autogamy (selfing)
  - low variation within population – 2 pure homozygote lineages
  - high variation among populations – locally adapted populations
- allogamy (outcrossing)
  - high variation within population – continuous formation of heterozygotes
  - low variation among populations
- BUT– pattern of variation depends also on *gene flow*

Mating system

- asexual reproduction
  - vegetative (clonal)
  - apomixis
- sexual reproduction
  - outcrossing
  - selfing
  - mixed system

nonrandom association of alleles of different loci



Effect on linkage disequilibrium (LD)
- heterozygosity
- breaking up LD (crossing over polymorphic sites)
- selfing – maintaining LD due to low heterozygosity level

Effect on diversity
- outcrossing – uncoupling advantageous mutations from linked variation
- selfing – fixation of linked neutral variation (hitchhiking)

Effect on deleterious mutations
- outcrossing – deleterious mutations eliminated by selection
- selfing – fixation of advantageous mutation can be linked with deleterious

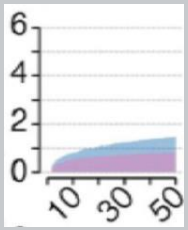
Wright S. et al. 2008: Genomic consequences of outcrossing and selfing in plants. Int. J. Pl. Sci. 169:105–118.

Test of reproduction system

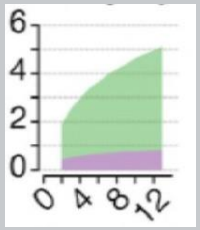
- *inbreeding intensity* – inbreeding coefficient
  - F_{IS} – heterozygote deficiency
- comparison of genetic information of mother plant and its progeny – rate of autogamy
- codominant markers
  - isozymes
  - microsatellites
  - SNPs – contiguous runs of homozygosity (ROH)

Selfing vs. outcrossing

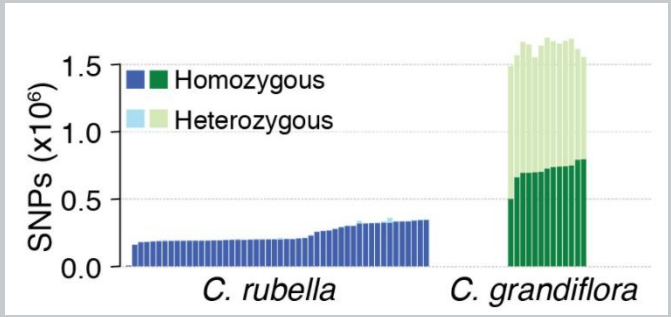
Capsella rubella
selfer



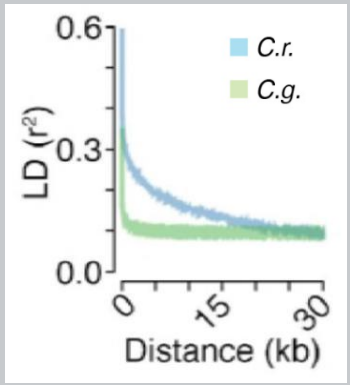
number of segregating (variable) sites
less in selfer



Capsella grandiflora
outcrosser



number of heterozygous (light colors) and homozygous SNP calls (dark colors)
nearly no heterozygotes in selfer



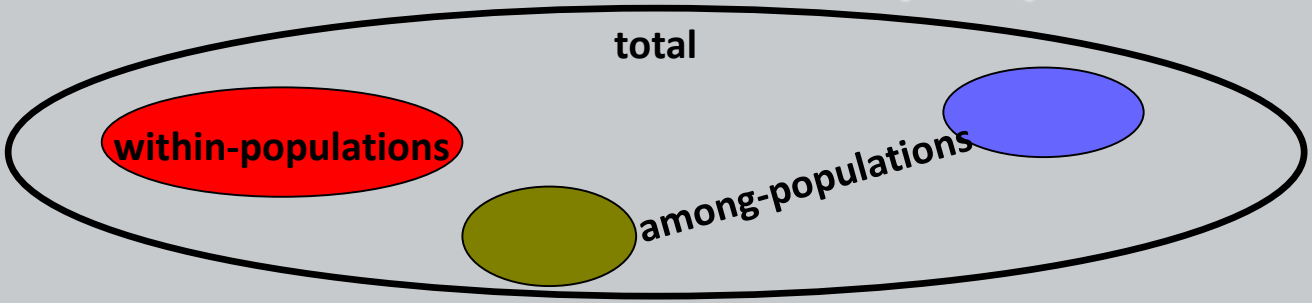
average decay of linkage disequilibrium
higher LD in selfer

Koenig D. et al. 2019: Long-term balancing selection drives evolution of immunity genes in *Capsella*. eLife 8:e43606.

Genetic diversity

- heterozygosity
  - observed – H_o
  - expected – H_e
    - = *gene diversity* – probability that two randomly chosen alleles are identical
- various diversity coefficient, e.g. Shannon
- number of alleles
- *allelic richness*
  - number of alleles standardized according to sample size
- number of rare or private alleles
  - DW-index…
- molecular diversity
  - nucleotide diversity (π) – average number of pair-wise differences/number of nucleotides + correction for multiple mutations…
  - number of polymorfic sites (S)

Genetic structure of populations



- genetic diversity – index of diversity (*H*)
  - can be correlated with population size, geographic location etc.
- pattern of genetic variability
  - *within populations*
  - *among populations*
- total diversity - H_T
- within-population diversity - H_S
- among-population diversity - $D_{ST} = H_T - H_S$

$$G_{ST} = \frac{H_T - H_S}{H_T}$$

coefficient of genetic differentiation – extent of **subpopulation differentiation**

Gene flow

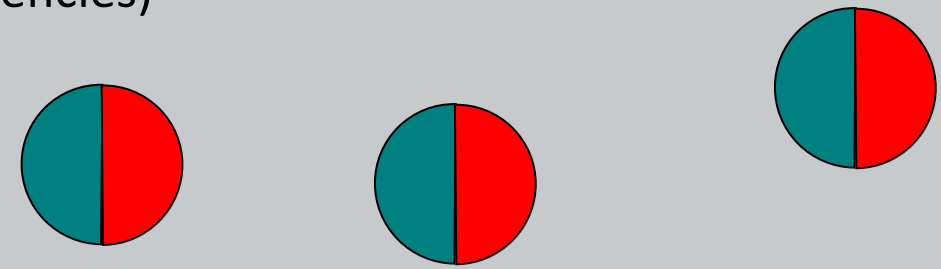
- are populations connected or isolated ?
- what is the intensity of *gene flow* among populations ?

  gene flow = intensity of communication, i.e. seed and/or pollen dispersal
- how far are seeds dispersed ?
- over which distance is pollen transfered ?
- spatial autocorrelation analysis
  - correlation of genetic and geographic distances

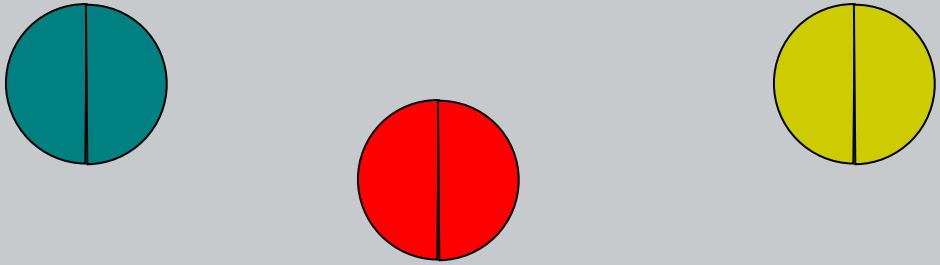
- **indirect** determination of *gene flow*
  - from the pattern of genetic diversity among populations – G_{ST}/F_{ST}
- **direct**
  - *parentage analysis*

Gene flow estimation

- interpretation of G_{ST}
  - degree of subpopulation differentiation
    - 0 – no population genetic structure (all populations with the same allele frequencies)



    - 1 – maximum genetic structure (each population fixed for different allele)



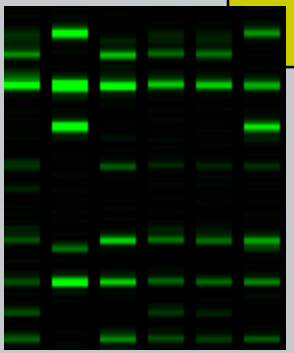
  - according to population-genetic models equals to the number of migrants per generation

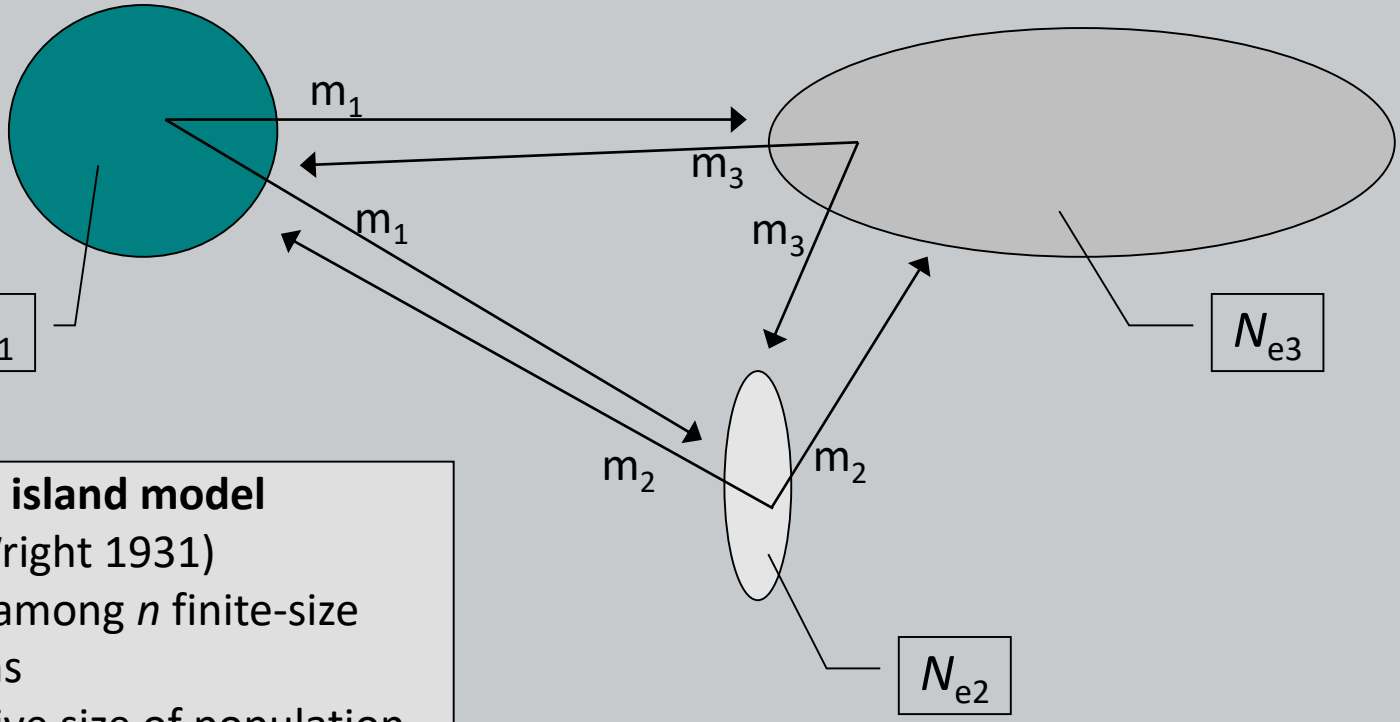
Interpretation of G_{ST}

migration among populations

genetic marker distribution

demografic model





finite island model
(Wright 1931)

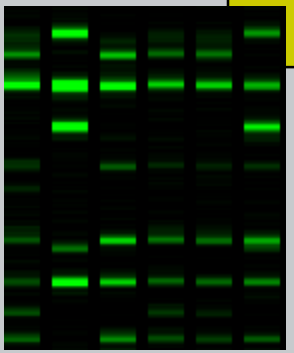
- migration among *n* finite-size populations
- N_e – effective size of population
- each population gives $N_e m$ migrants to all other populations

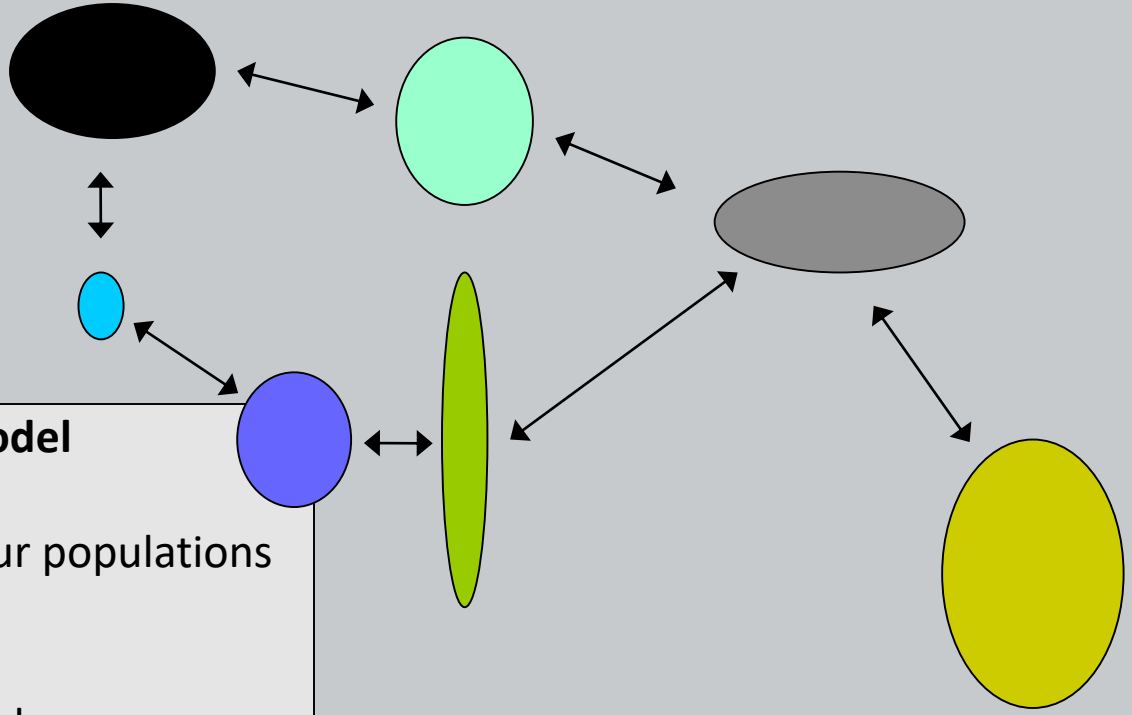
Interpretation of G_{ST}

migration among populations

genetic marker distribution

demografic model





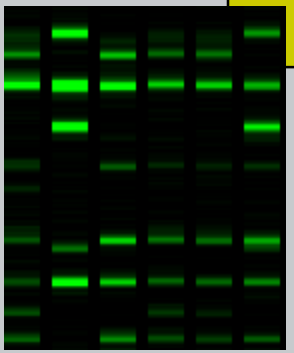
stepping stone model
(Kimura 1953)

- migration among neighbour populations only
- one- or two-dimensional
- dispersal among non-neighbour populations is achieved with several steps

Interpretation of G_{ST}

migration among populations

genetic marker distribution



demografic model

population-genetic model

$$G_{ST} = \frac{1}{1 + 4N_e m}$$

estimation of genetic population differentiation

indirect estimate of **gene flow**

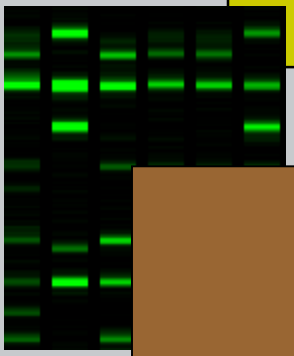
which number of migrants (N_em) caused observed pattern of genetic diversity

Interpretation of G_{ST}

migration among populations

genetic marker distribution

demografic model



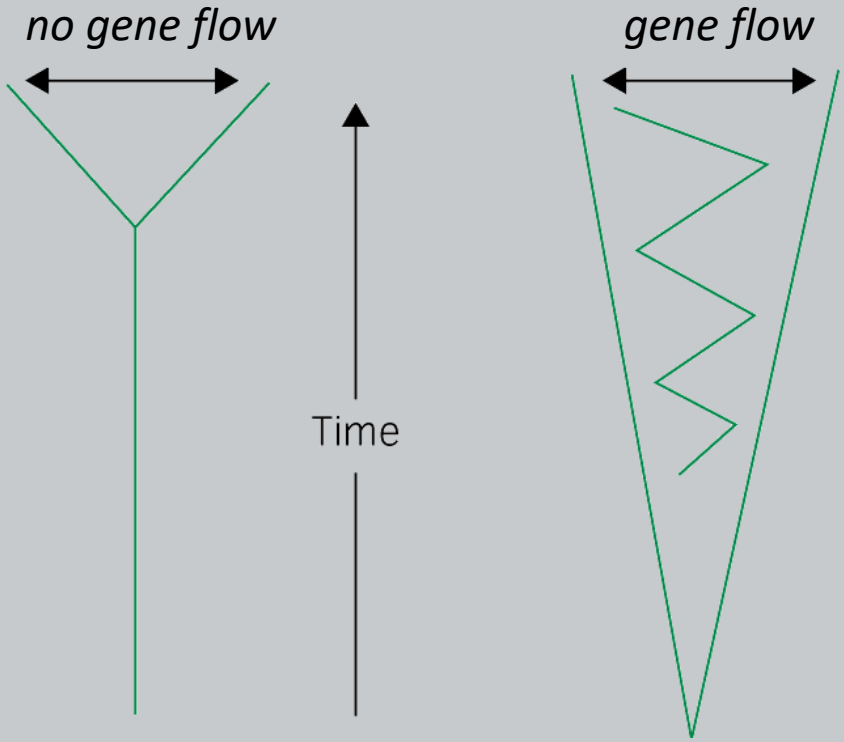
assumptions for the interpretation

- equilibrium among gene flow and drift
- gene flow is constant during time
- mutations are much lower than migration

indirect estimate of **gene flow**

which number of migrants (N_em) caused observed pattern of genetic diversity

Historical and contemporary gene flow



non-equilibrium

recent population separation

no gene flow

equilibrium

long-term population separation

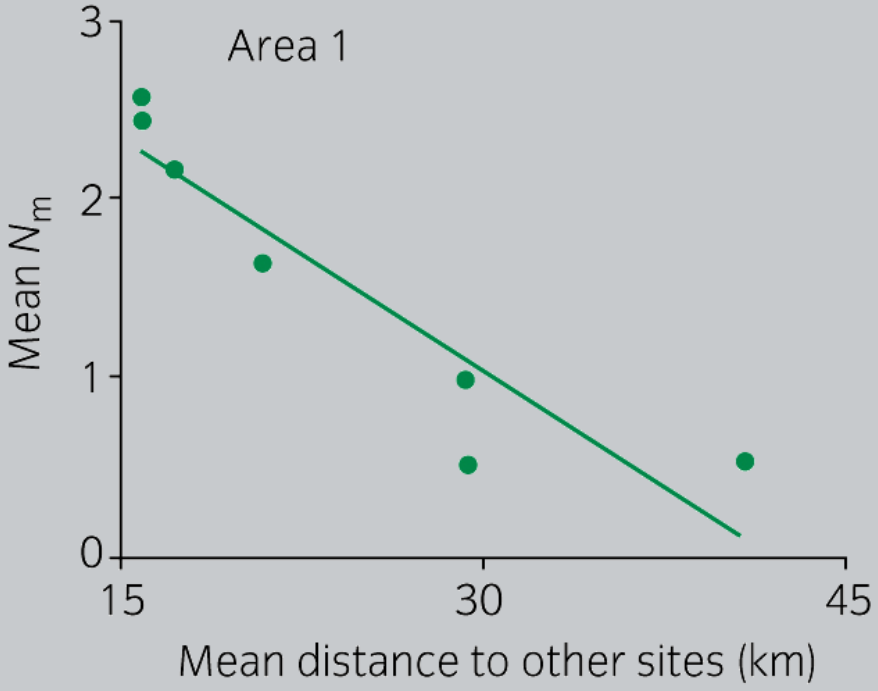
continuous gene flow

it is not easy to distinguish among continuous gene flow from the similarity due to common origin just from the allelic frequencies

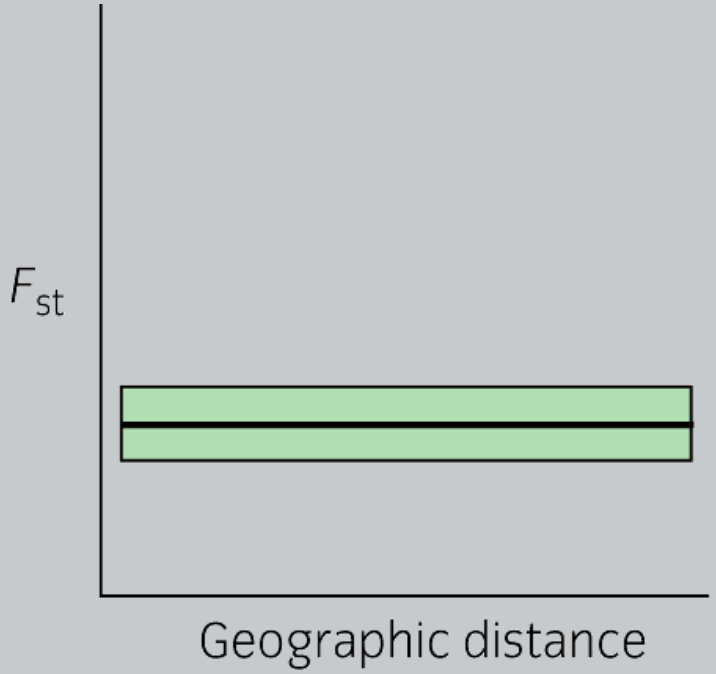
→ model-based approach (e.g. using joint site frequency spectrum, jSFS)

Relationship between genetical and geographical structure

- *isolation-by-distance*



- intense *gene flow*

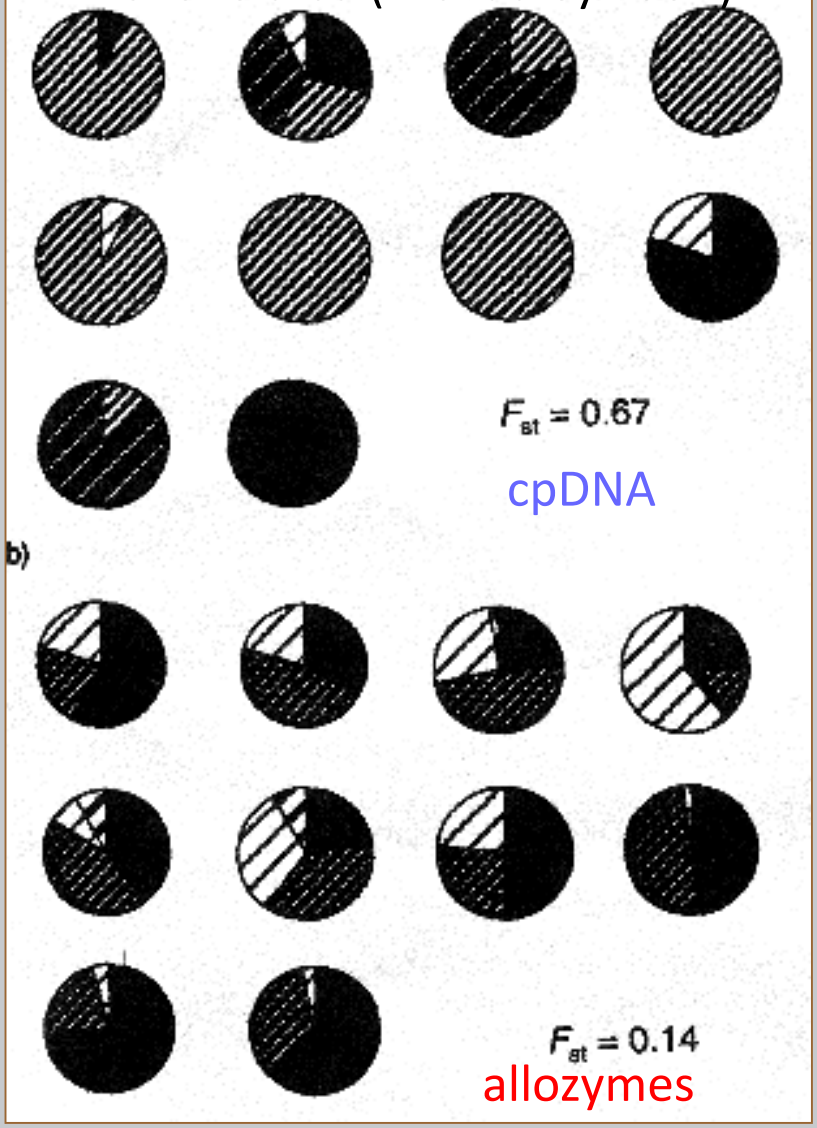


Gene flow = pollen flow + seed flow

- **pollen** - haploid nuclear DNA
- **seeds** - diploid nuclear DNA
  - cpDNA

$$\frac{\text{pollen migration}}{\text{seed migration}} \approx \frac{\left(\frac{1}{F_{STb}} - 1\right) - 2\left(\frac{1}{F_{STm}} - 1\right)}{\left(\frac{1}{F_{STm}} - 1\right)}$$

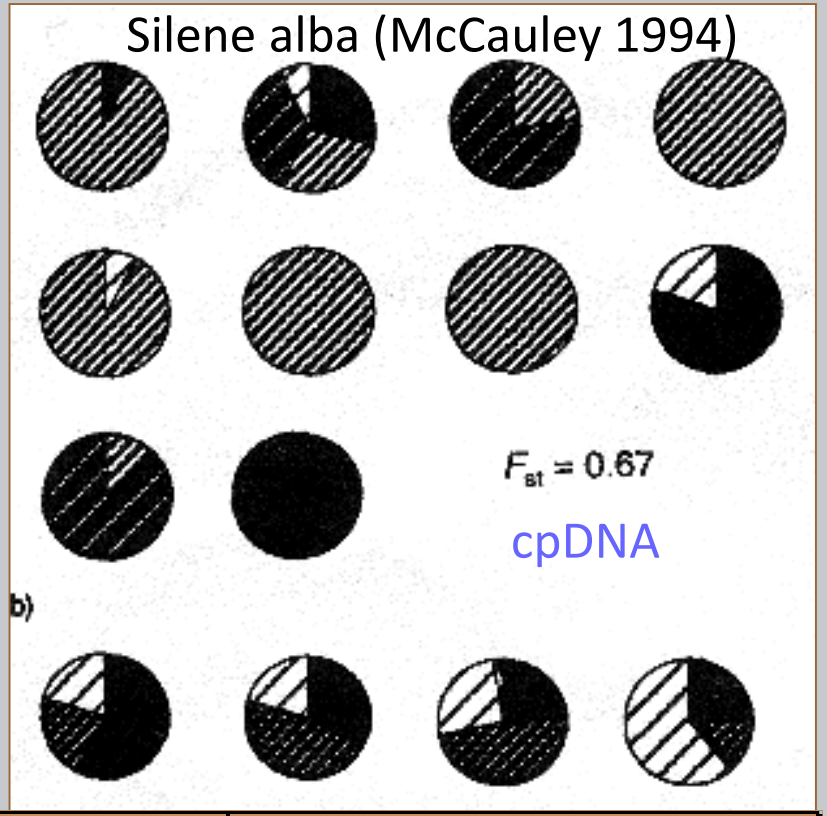
Silene alba (McCauley 1994)



Gene flow = pollen flow + seed flow

- **pollen** - haploid nuclear DNA
- **seeds** - diploid nuclear DNA
  - cpDNA

$$\frac{\text{pollen migration}}{\text{seed migration}} \approx \frac{\left(\frac{1}{F_{STb}} - 1\right) - 2\left(\frac{1}{F_{STm}} - 1\right)}{\left(\frac{1}{F_{STm}} - 1\right)}$$



Species	Pollen dispersal	Seed dispersal	Pollen/seed ratio	Reference
Quercus sp.	wind	birds	196	Kremer et al. (1991)
Pinus contorta	wind	wind	28	Dong and Wagner (1993)
Argania spinosa	insects	ruminants	2.5	El Mousadik and Petit (1996)
Pinus sylvestris (Scotland)	wind	wind	18	Sinclair et al. (1998)
Pinus sylvestris (Spain)	wind	wind	105	Sinclair et al. (1999)

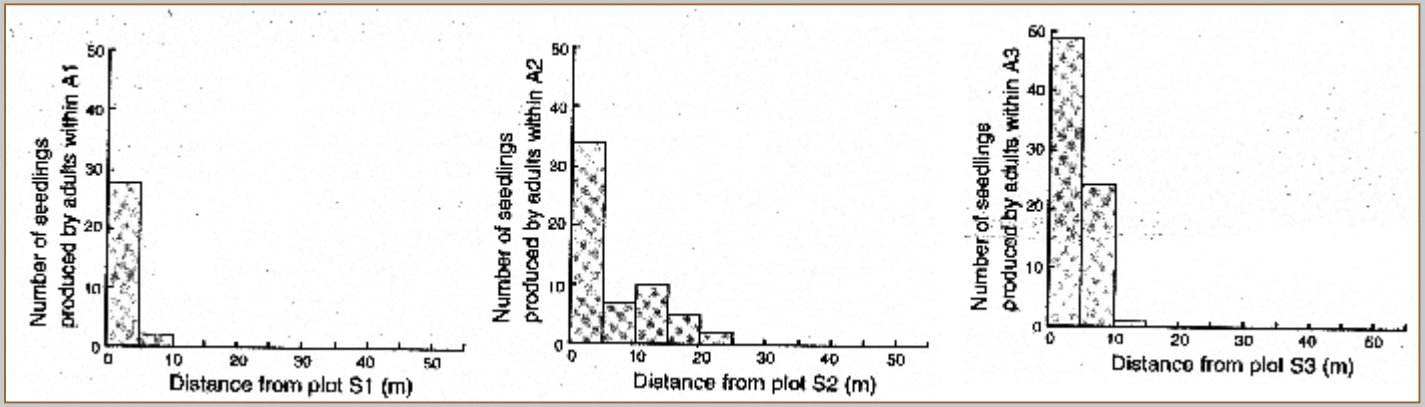
Markers for population studies

- nuclear (ot total) DNA
  - isozymes, microsatellites
  - RAPD, AFLP
  - sequencing (ITS…), *low-copy* markers
  - genome-wide SNPs (NGS – RADseq, resequencing)

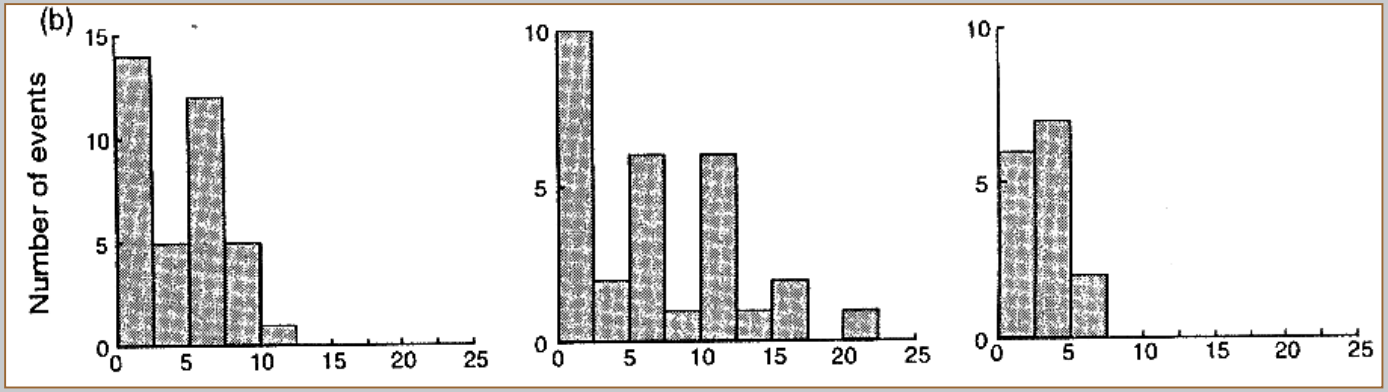
- chloroplast DNA
  - RFLP, PCR-RFLP
  - cpDNA microsatellites
  - sequencing (*trn*L-*trn*F…)
  - whole plastome

Parentage analysis

- *gene flow* in subpopulation
  - calculated from distances between parents and seeds



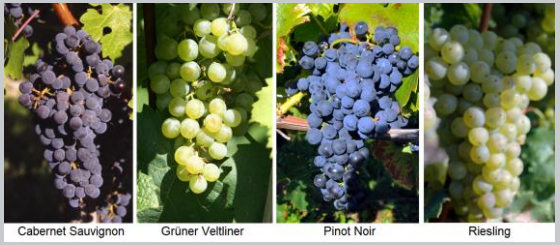
- *pollen flow* in subpopulation
  - pollen travelled from one parent to the other

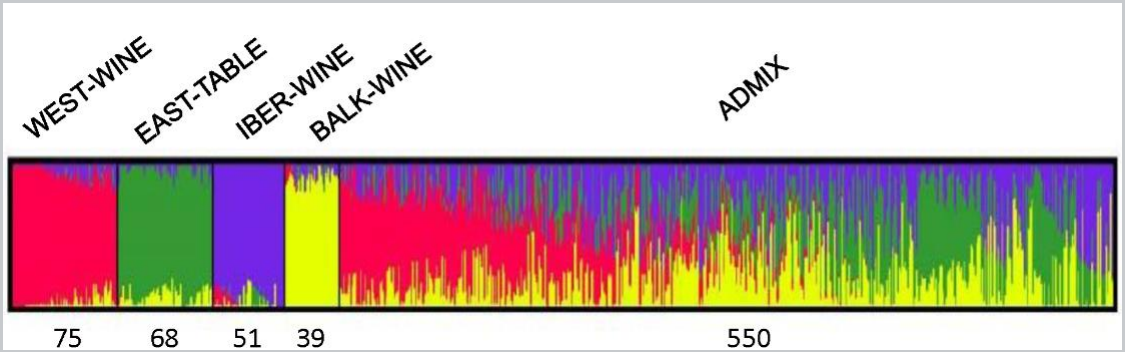


Parentage analysis with SNPs

Vitis vinifera (wine)

- 18k SNP genotyping array
- 10,207 SNPs and 783 different genotypes
- 14 SNPs sufficient to identify each genotype
- 118 full parentages and 490 parent-offspring duos



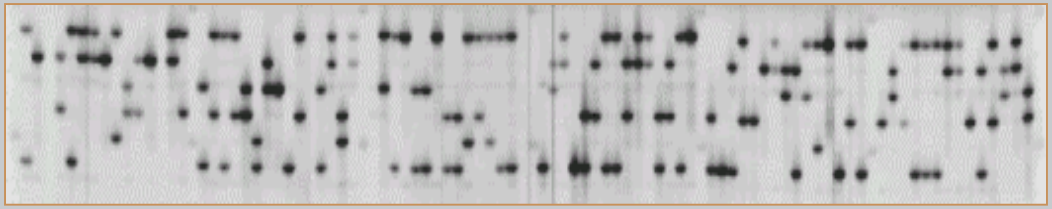


Pinotage = Pinot noir × Cinsaut
Chardonnay = Gouais blanc × Pinot noir
Merlot = Cabernet franc × Magdeleine noire des Charentes

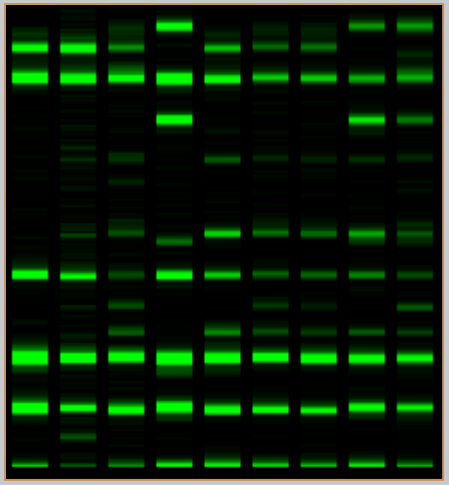
Laucou V. et al. 2018: Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide SNPs. PLOS ONE 8:e43606.

Markers for *parentage analysis*

- microsatellites



- AFLP
  - high variability
  - reliability



- SNPs
  - high power

Study of migration

- *phylogeography* – study of postglacial migration
  - identification of migration routes
  - use of cpDNA
    - maternally inherited (i.e., dispersal through seeds in angiosperms)
    - haploid
    - absence of recombinations
- relationships among species distribution and their evolution
- recent migration of species in a landscape
  - plant invasions
  - …

Phylogeography

influence of historical factors (e.g., glaciation) on geographic distribution of genealogical lineages

- maximum extent of glaciation – 20,000-18,000 BP
- maximum (concentrated) variation in Mediterranean region
- 3 basic refugia – Iberian Penninsula, Appenines, the Balkans
- only a small part of variability migrated back to the Central Europe
- recolonization started ca. 13,000 BP
- molecular methods – use of cpDNA
- lineages definition (haplotypes) and correlation of their genealogy (relationships) with geographic distribution

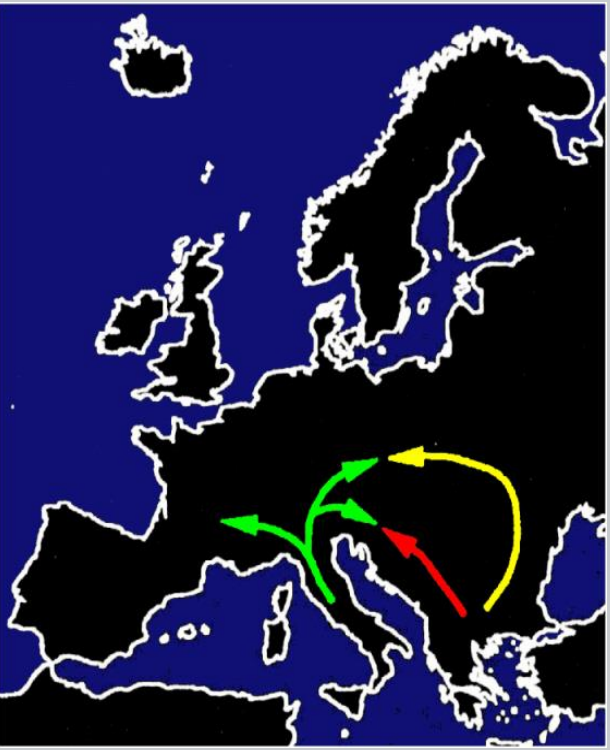
maximum extent of glaciation during last ice age

permafrost

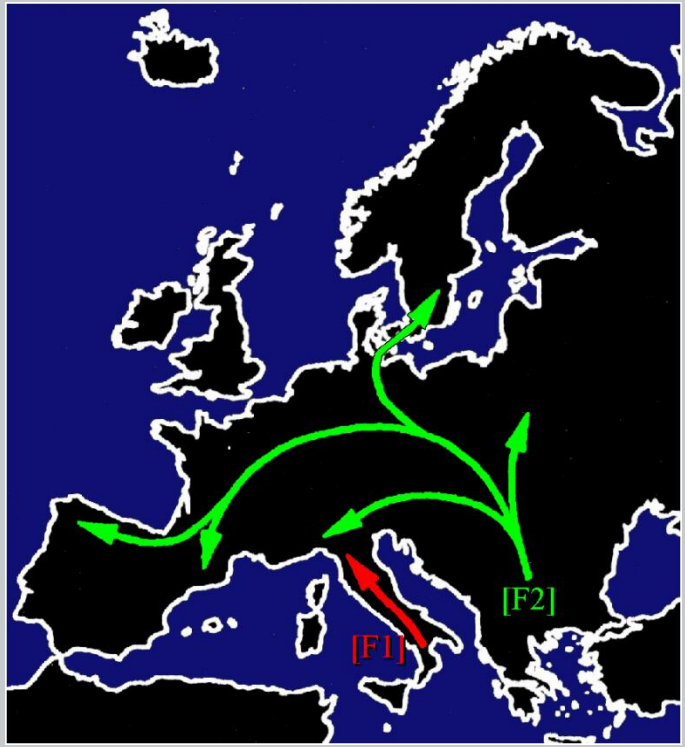
R1, R2, R3 – basic refugia

Postglacial recolonizace of Europe

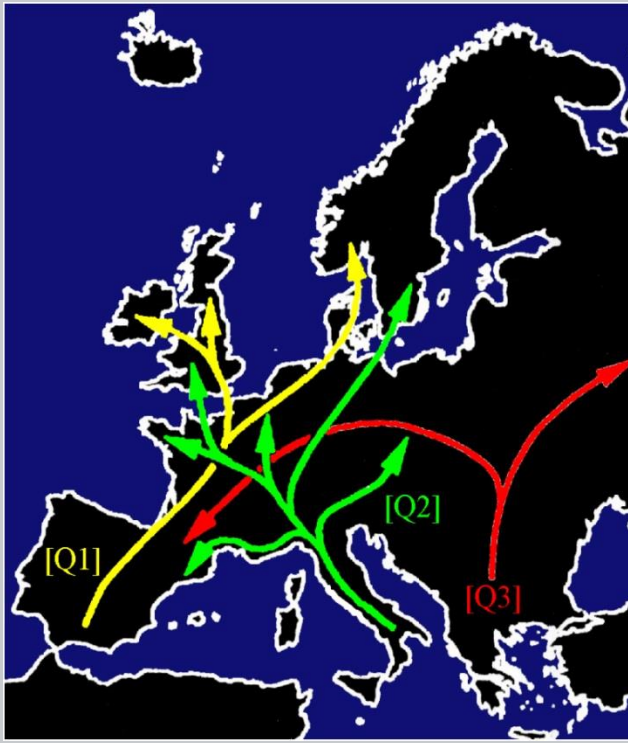
traditionally estimated migration routes of trees



fir (*Abies*)



beech (*Fagus*)



oak (*Quercus*)

Systematics studies
at all taxonomic levels

- evolution of (higher) plants
  - group monophyly, relationships among families
  - primitive, derived families
- relationships among genera within a family, among species within a genus
- microspecies identification and their origin
- origin and spread of new taxa
- relationships among taxa with different ploidy levels, origin of polyploids
- hybridization, identification of parental taxa…
- introgression
- selection
- diversification
- trait evolution
- molecular dating

Markers in systematics

phylogenetic reconstruction

- sequences
  - ITS, low-copy, cpDNA, mtDNA
  - coding × non-coding
  - NGS
    - whole plastomes
    - transcriptomes
    - hundreds of orthologous genes (Hyb-Seq)
    - SNPs (re-sequencing, RADseq…)
- RFLP of chloroplast DNA
- AFLP – for closely related and recently evolving taxa
- (cpDNA) microsatellites – at the species level
- isozymes – differentiation of very closely related taxa

hybrid detection

- PCR-RFLP – e.g., ITS region
- AFLP – band-sharing analysis, Bayesian-based clustering…
- microsatellites – F1 and advanced hybrids
- SNPs

Sequencing

1. coding genes (exons) – *conserved*
   - for the family, genus level (*rbc*L*)*
   - but some genes quite variable
2. spacers, introns – *more variable regions*
   - for the genus, species and lower levels (*trn*L-F, ITS…)

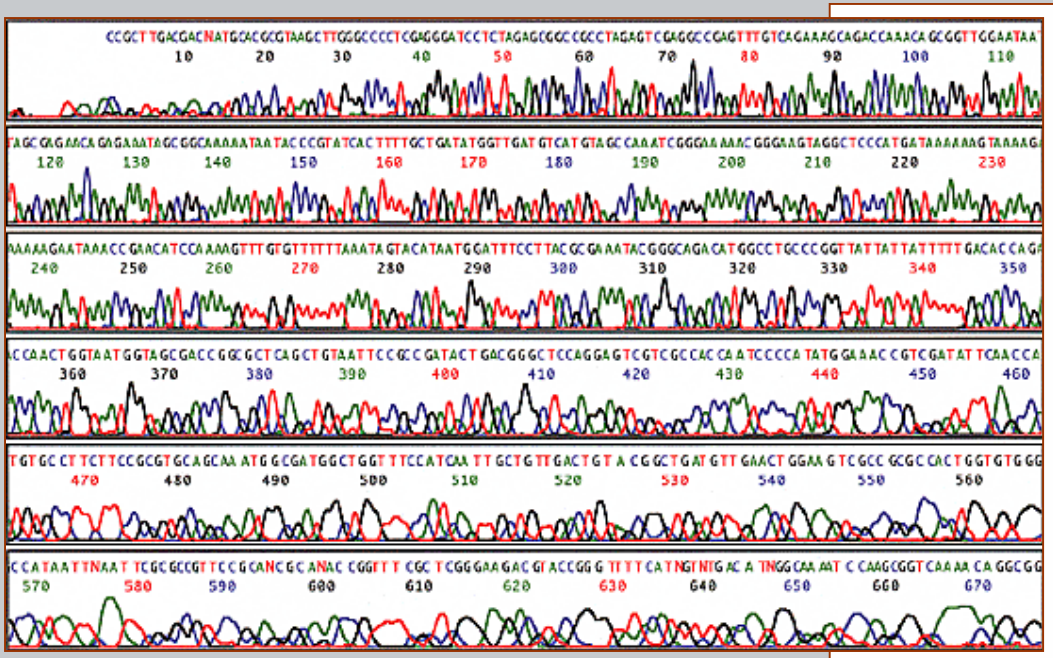
chloroplast

- *rbc*L
- *mat*K
- *trn*L-*trn*F…

nuclear

- ITS
- 26S rDNA
- low-copy genes …

- orthology × paralogy
- gene trees × species tree



ACER ACGTAA-GAACCGAAG

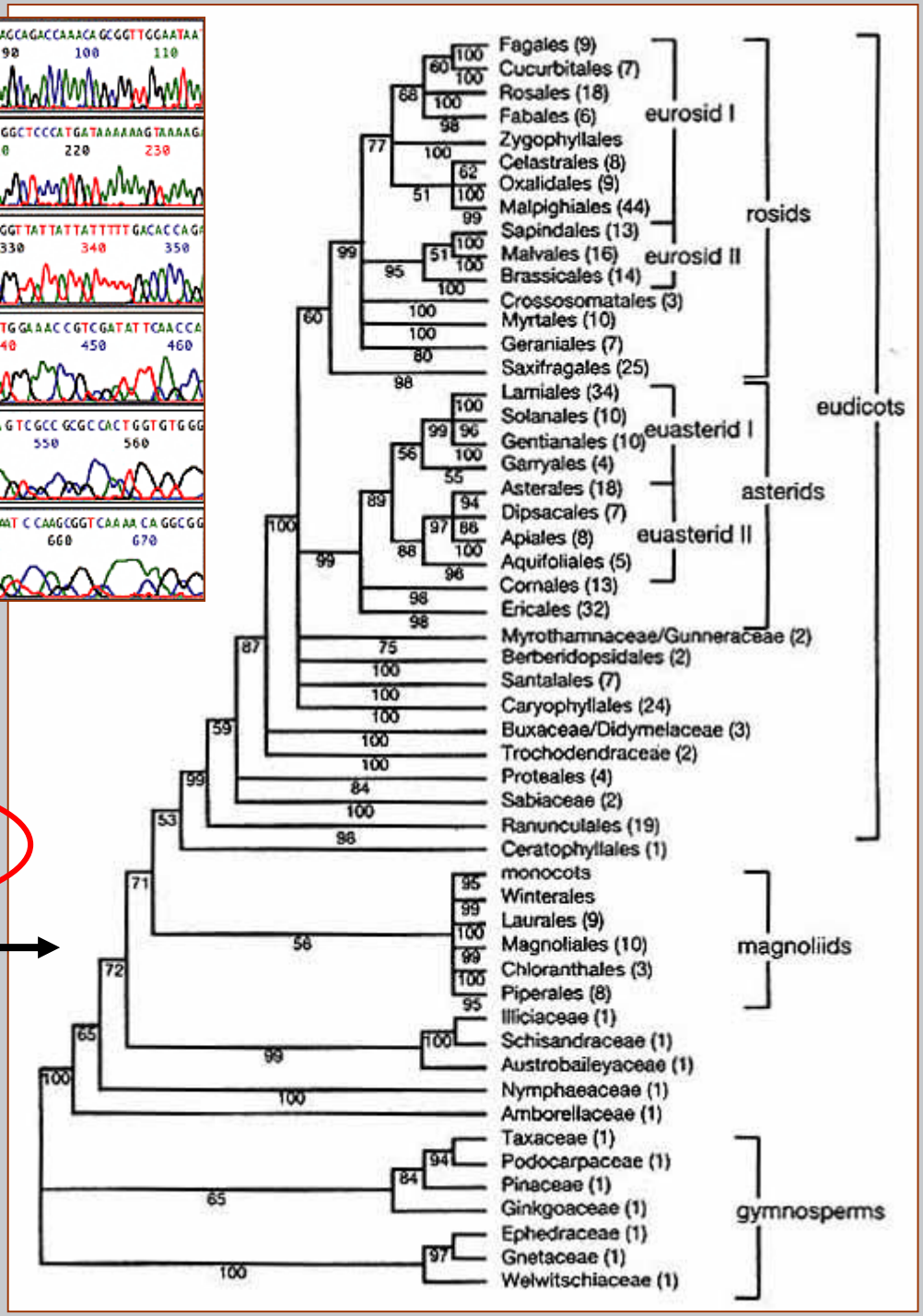
QUERCUS ACGTTATGA---GAAG

Substitution

Insertion

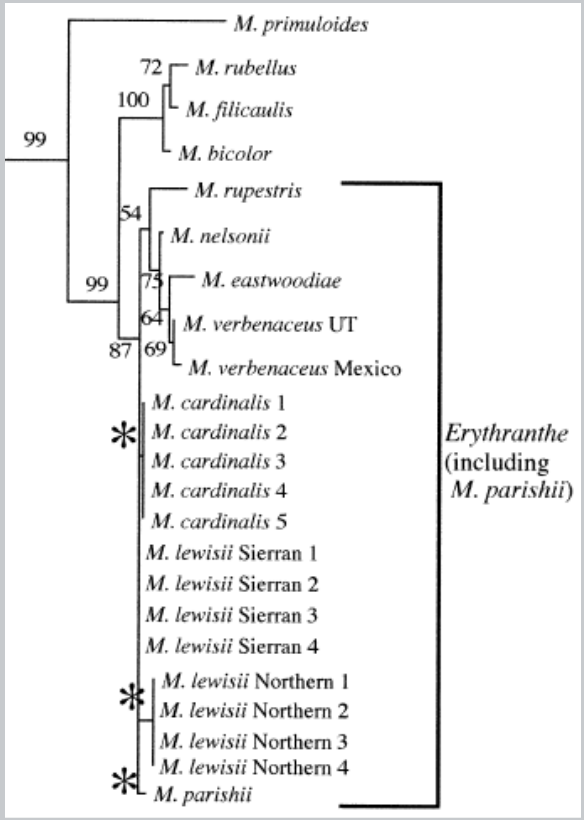
Deletion

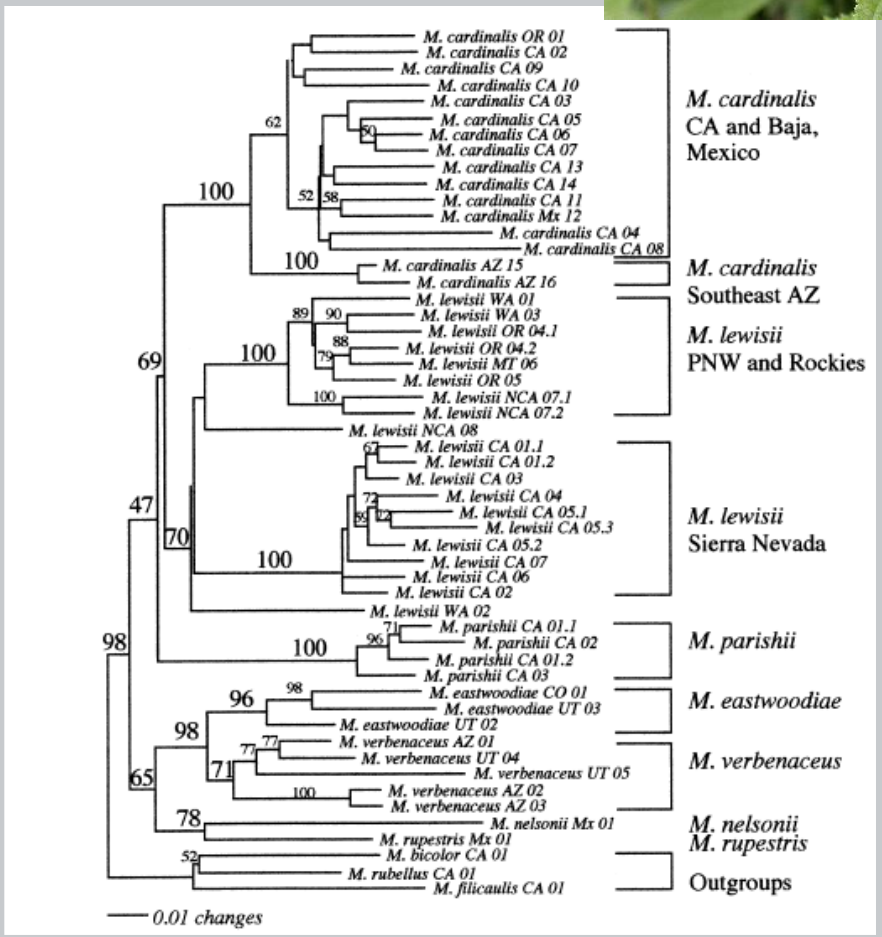
Sequence alignment



Phylogeny ITS vs. AFLP

ITS





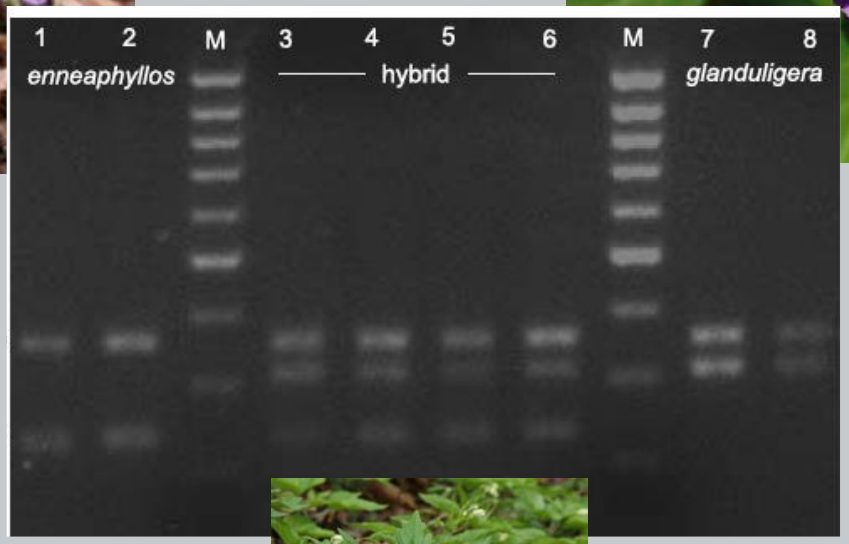
AFLP

species complex – *Mimulus* sect. *Erythranthe*

Beardsley et al. 2003

Hybridization

Dentaria enneaphylos



Dentaria glandulosa

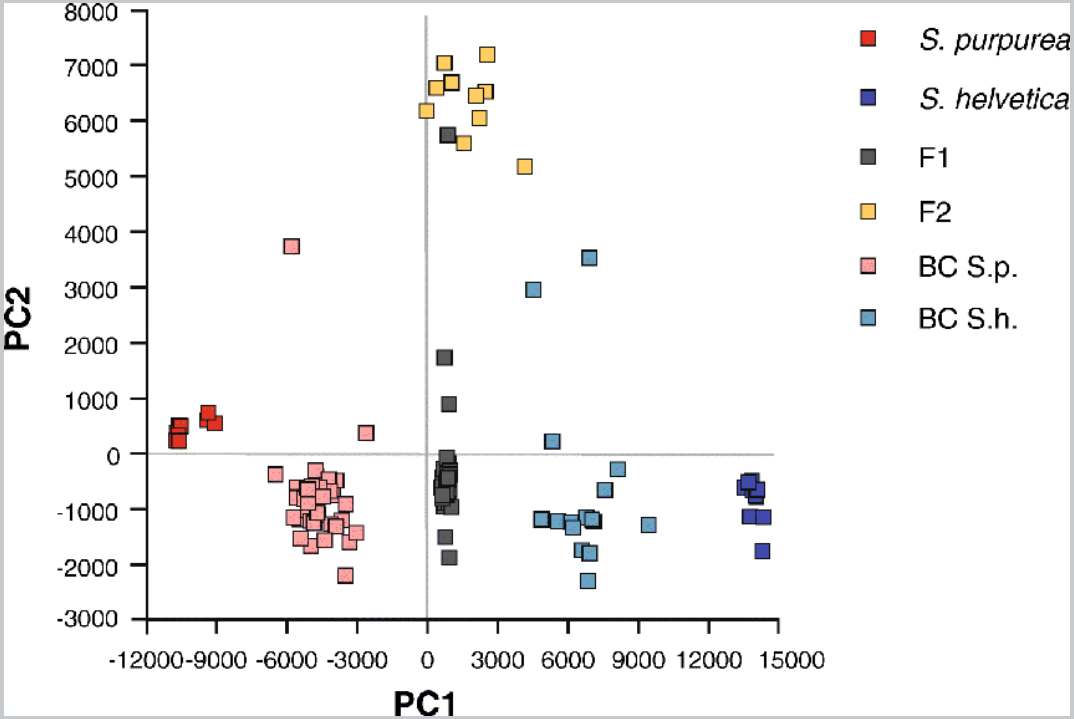
Dentaria x *paxiana*

PCR-RFLP of ITS region

Lihová et al. 2007

Hybridization

Salix purpurea



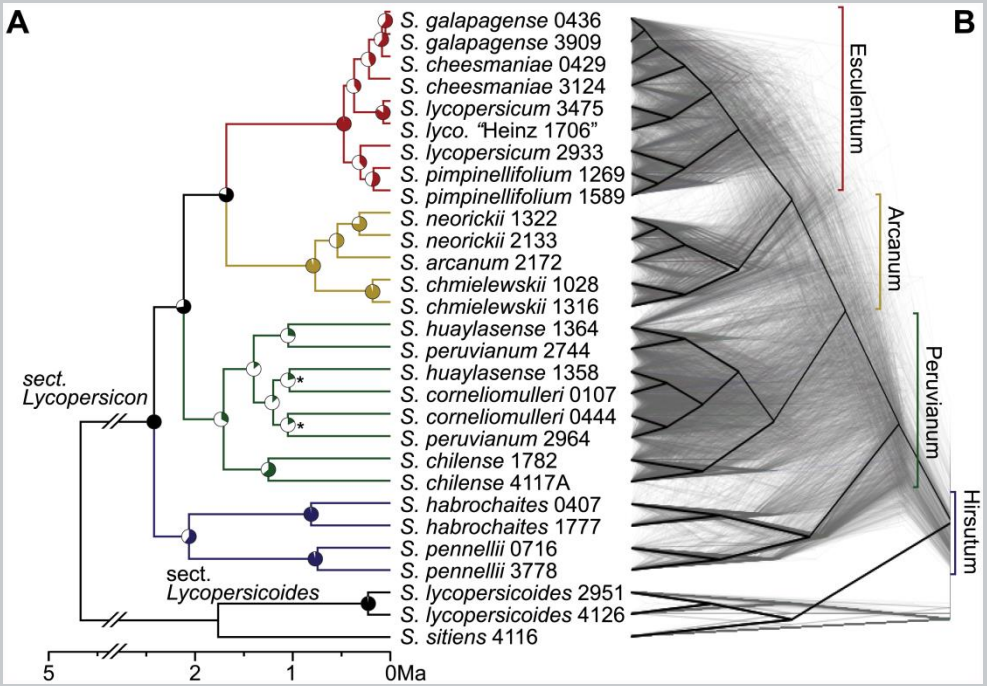
Salix helvetica

5,758 RADseq loci

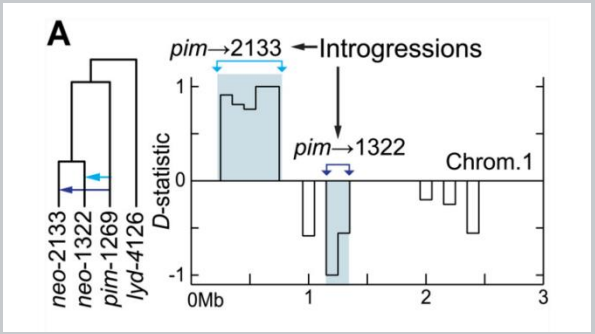
Gramlich et al. 2018

Introgression

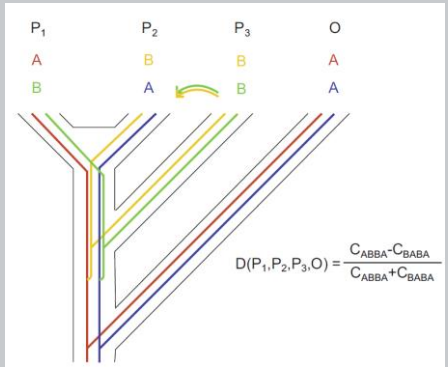
Solanum sect. *Lycopersicon*



A) A **whole-transcriptome** concatenated molecular clock phylogeny. Pie charts: majority rule extended bipartition support scores (out of 100). All nodes have 100 BS, "*" denotes BS 68.
(B) A "cloudogram" of 2,745 trees (grey) inferred from nonoverlapping 100-kb genomic windows, the consensus phylogeny is shown in black.



A) The *D*-statistic for **testing introgression** for 100-kb windows on the short arm of chromosome 1 using the tree shown. Shaded regions indicate windows where introgression is significantly detected ($p \leq 1.45 \times 10^{-4}$).

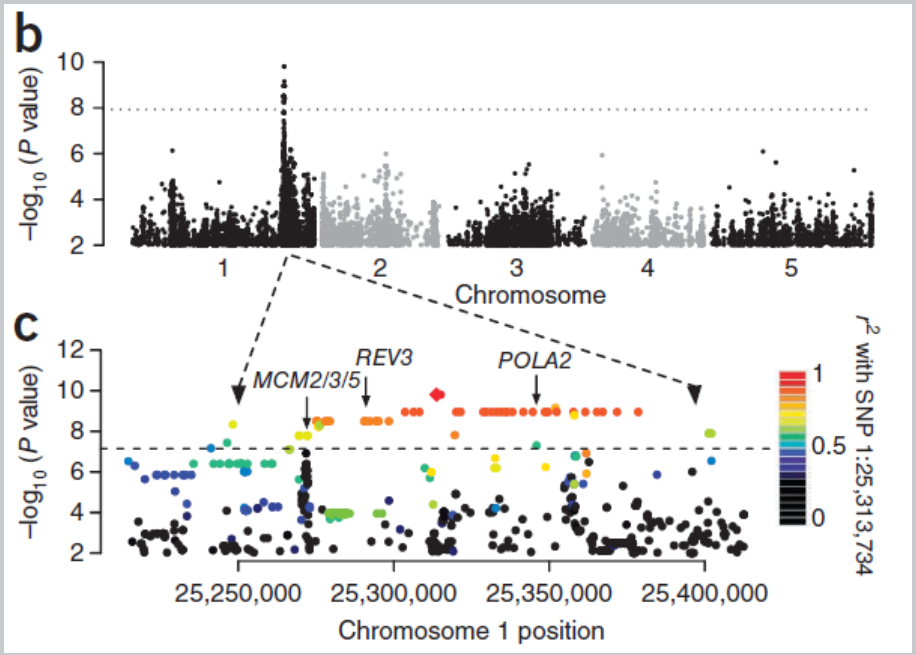


ABBA-BABA *D*-statistics

Pease et al. (2013): Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* 14(2): e1002379.

Selection tests

Arabidopsis

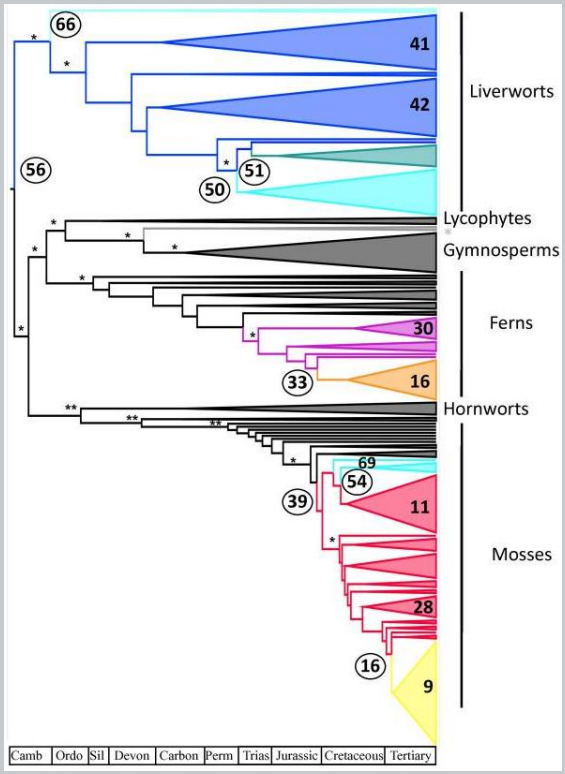


genome-wide association studies (GWAS)

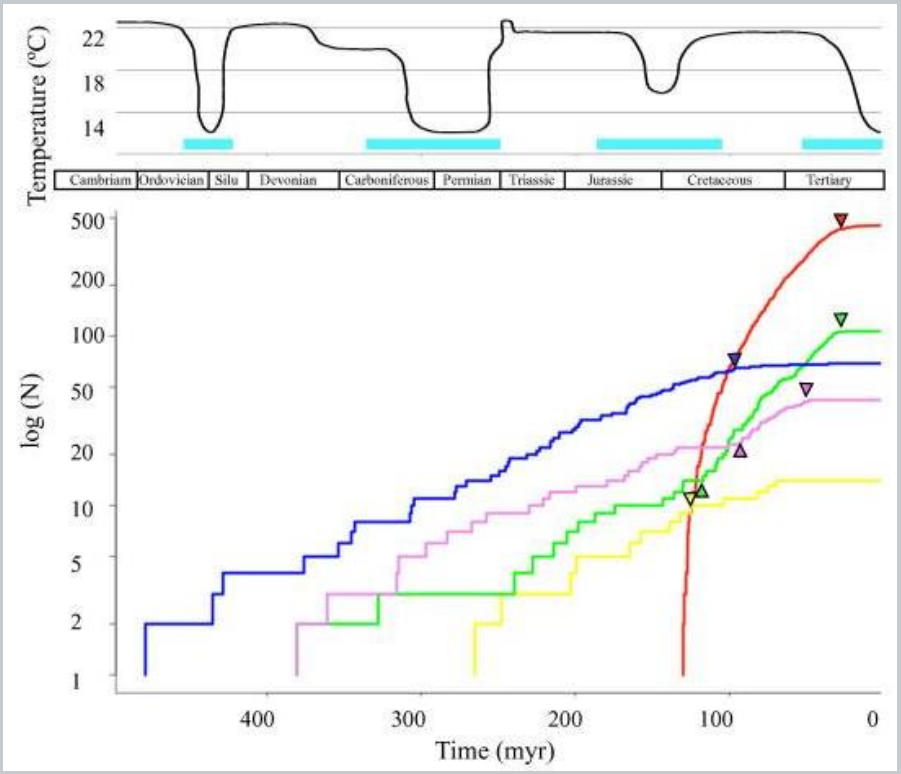
identification of loci associated with, e.g., particular phenotype/trait

Long et al. (2013): Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics 45*(8): 884–891.

Diversification



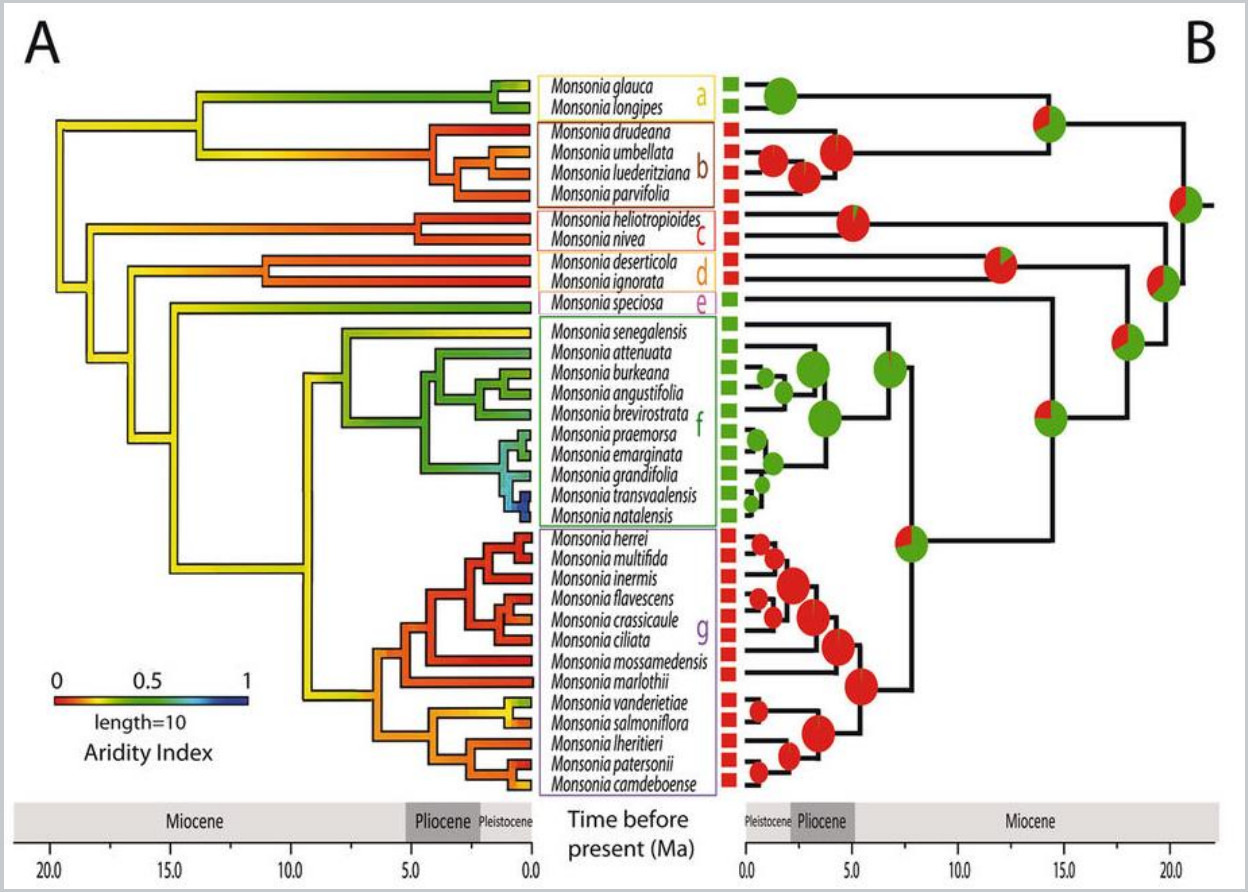
Diversification chronogram with **rate shifts** located for different groups of land plants. Numbers correspond to the rate shifts. Different colours indicate different **net diversification rates** found in the tree.



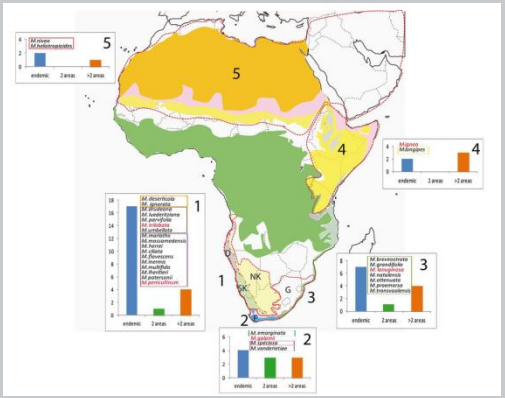
Lineage Through Time (LTT) plot for liverworts (blue), mosses (green), ferns (purple), gymnosperms (yellow) and angiosperms (red) with indication of average global temperature and cool climate modes (blue bars). Triangles pointing up or down indicate **diversification rate shifts.**

Fiz-Palacios O. et al. (2011): Diversification of land plants: insights from a family-level phylogenetic analysis. *BMC Evol Biol.* 11: 341.

Trait evolution



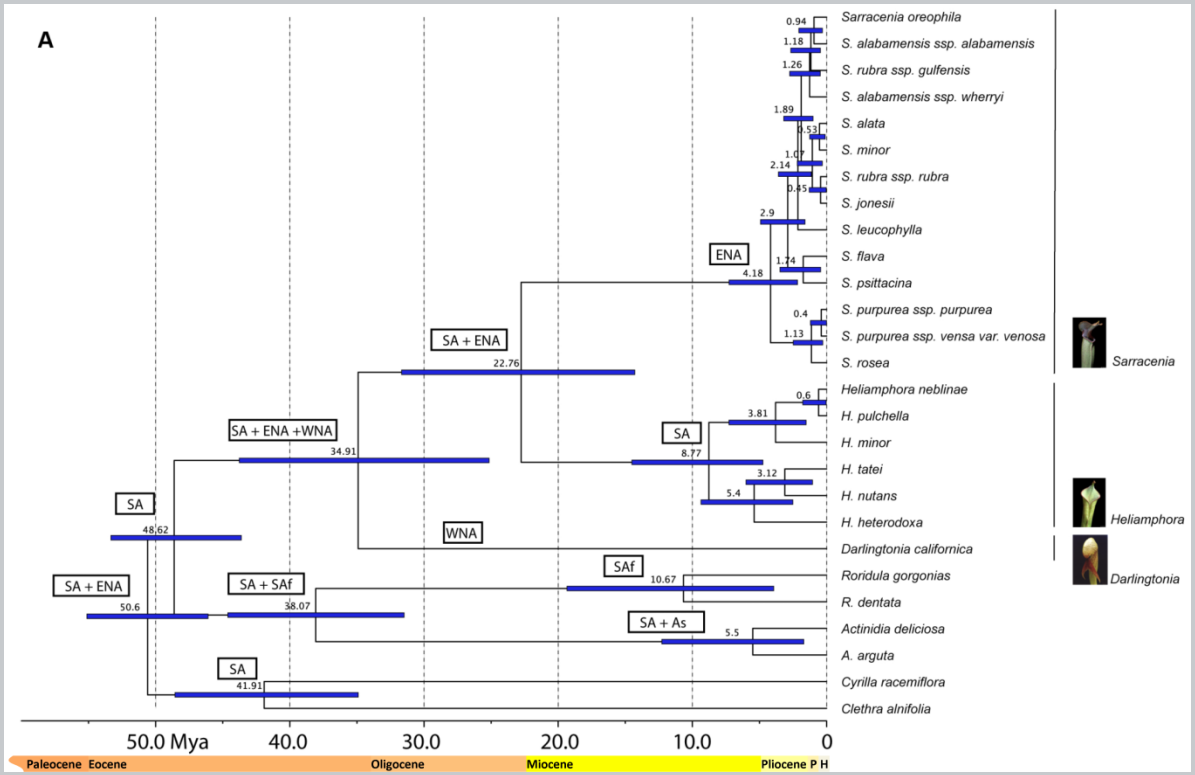
Monsonia



(A) Shifts between arid and semiarid-wet habitats inferred using aridity values.
(B) Reconstruction of fruit type evolution.

García-Aloy et al. (2017): Opposite trends in the genus *Monsonia* (Geraniaceae): specialization in the African deserts and range expansions throughout eastern Africa. *Scientific Reports* 7, Article number: 9872.

Molecular dating & biogeography



Sarraceniaceae

(A) Mean **divergence times estimates** (95% posterior probability distribution shown with blue lines). **Ancestral areas reconstructions** in boxes. SA = South America; ENA = Eastern North America; WNA = Western North America; SAf = South Africa; and As = Asia.

Ellison et al. (2012): Phylogeny and biogeography of the carnivorous plant family Sarraceniaceae. *PLoS ONE7(6)*: e39291.

Literature

Henry R.J. (2012): Molecular Markers in Plants.

Besse P. (2014): Molecular Plant Taxonomy. Methods and protocols.

Lemmon E.M. & Lemmon A.R. (2013): High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Avise J.C. (2004): Molecular markers, natural history and evolution.

Baker A.J. (2000): Molecular methods in ecology.

Beebee T. & Rowe G. (2004): An introduction to molecular ecology.

Henry R.J. (2001): Plant genotyping. The DNA fingerprinting of plants.

Karp A. et al. (1998): Molecular tools for screening biodiversity.

Lowe A., Harris S. & Ashton P. (2004): Ecological Genetics: Design, Analysis, and Application.

Weising K. et al. (2005): DNA fingerprinting in plants. Principles, methods, and applications.

Karp A. et al. (1996): Molecular techniques in the assesment of botanical diversity. *Annals of Botany* 78:143-149

Ouborg N.J. et al. (1999): Population genetics, molecular markers and the study of dispersal in plants. *J. Ecol.* 87:551-568.

Parker G.P. et al. (1998): What molecules can tell us about populations: Choosing and using a molecular marker. *Ecology* 79: 361-382