Molecular markers in plant systematics and population biology

10. RADseq, population genomics

Tomáš Fér

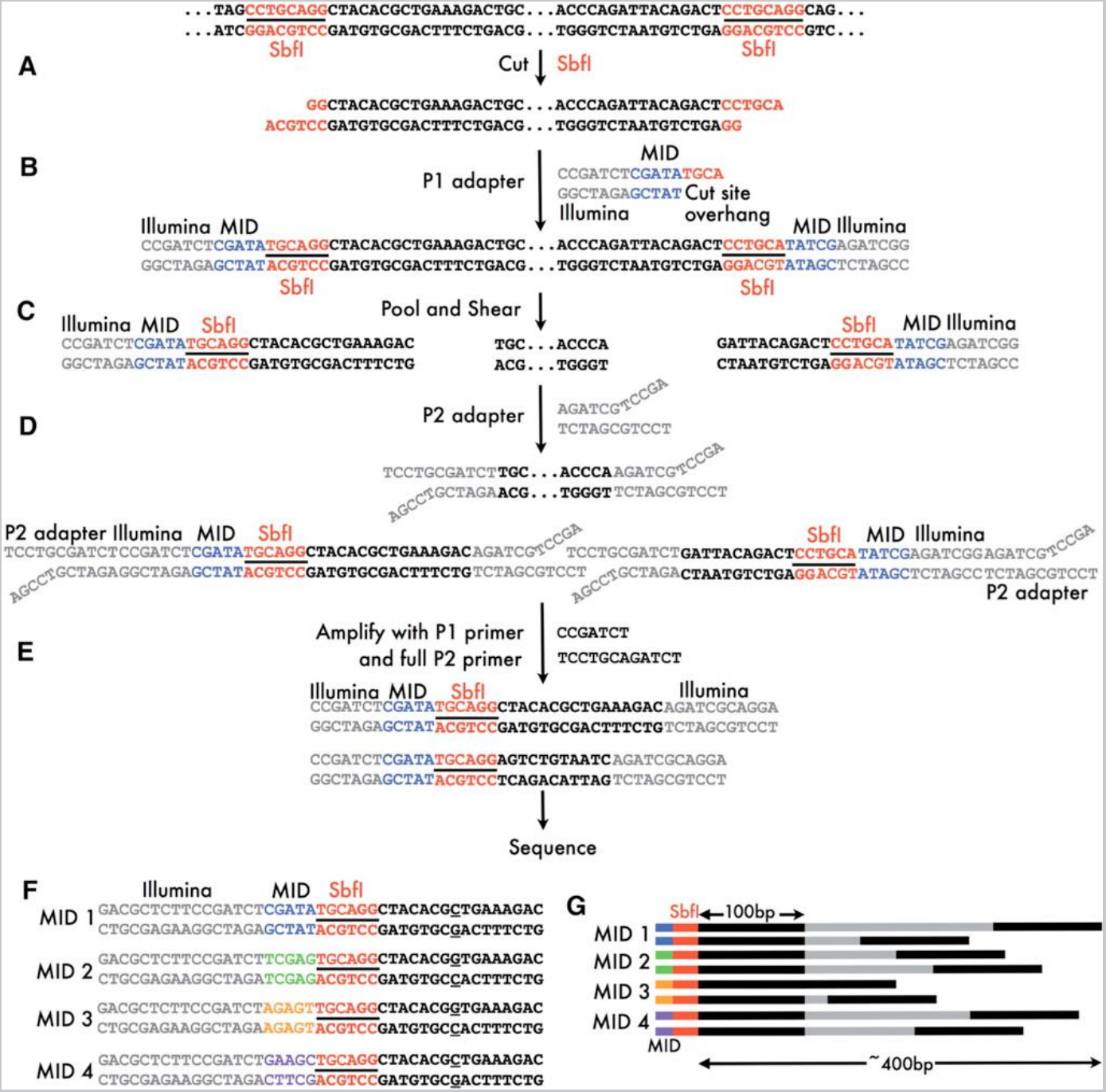
tomas.fer@natur.cuni.cz

RADseq

Restriction site-associated DNA sequencing

- genome complexity reduced by DNA cutting by restriction endonuclease(s)
- only sequences associated with restriction sites are sequenced
- many modifications of the basic protocol

Original RADseq



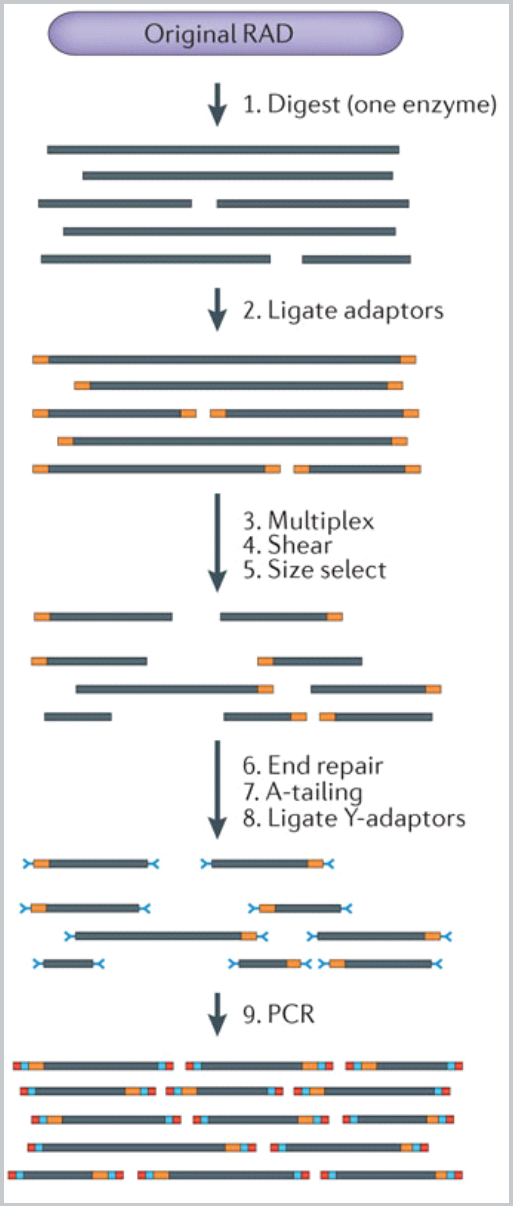
1. digestion with one enyzme
2. adapter/barcode ligation
3. pooling
4. mechanical shearing
5. size selection
6. adapter ligation
7. PCR amplification with adapter specific primers
8. (PE) sequencing

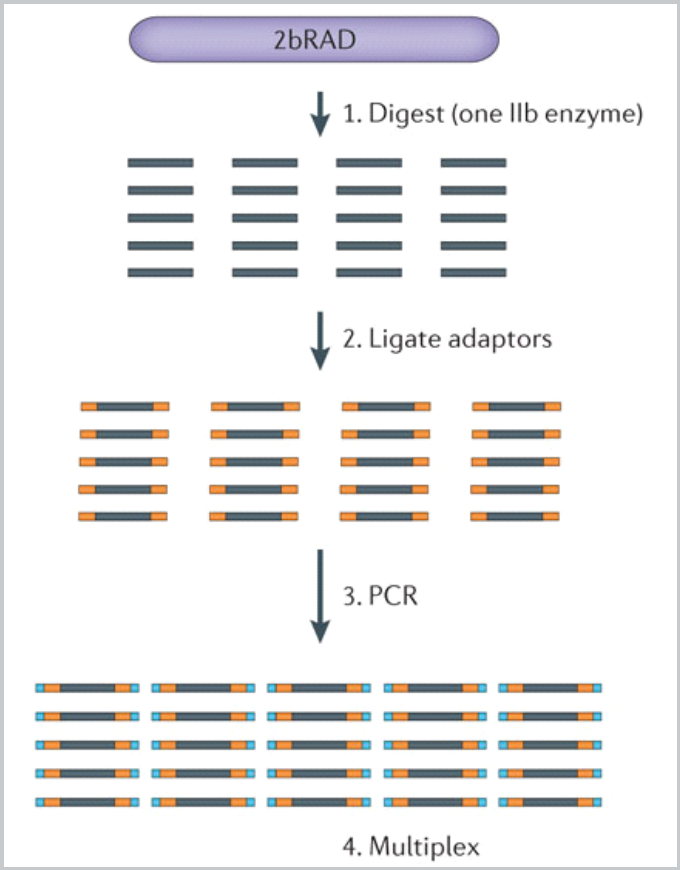
Davey J.W. & Blaxter M.L. (2011): *RADSeq: next-generation population genetics*. Briefings in Functional Genomics 9: 416-423.

RADseq modifications

- sequencing of fragments adjacent to single restriction enzyme cut sites
  - original **RADseq** (Baird et al. 2008)
  - **2bRAD** (Wang et al. 2012)
- sequencing of fragments flanked by two restriction enzyme cut sites
  - single enzyme, indirect size selection
    - **GBS** – genotyping by sequencing (Elshire et al. 2011)
    - **SBG** – sequence-based genotyping (Truong et al. 2012)
  - double enzyme, indirect size selection
    - **CRoPS** – complexity reduction of polymorphic sequences (Orsouw et al. 2007)
  - single enzyme, direct size selection
    - **RRLs** – reduced representation libraries (van Tassel et al. 2008)
    - **MSG** – multiplexed shotgun genotyping (Andolfatto et al. 2011)
    - **ezRAD** (Toonen et al. 2013)
  - double enzyme, direct size selection
    - **ddRAD** – double-digest RAD (Peterson et al. 2012)

Sequence next to RE cut site





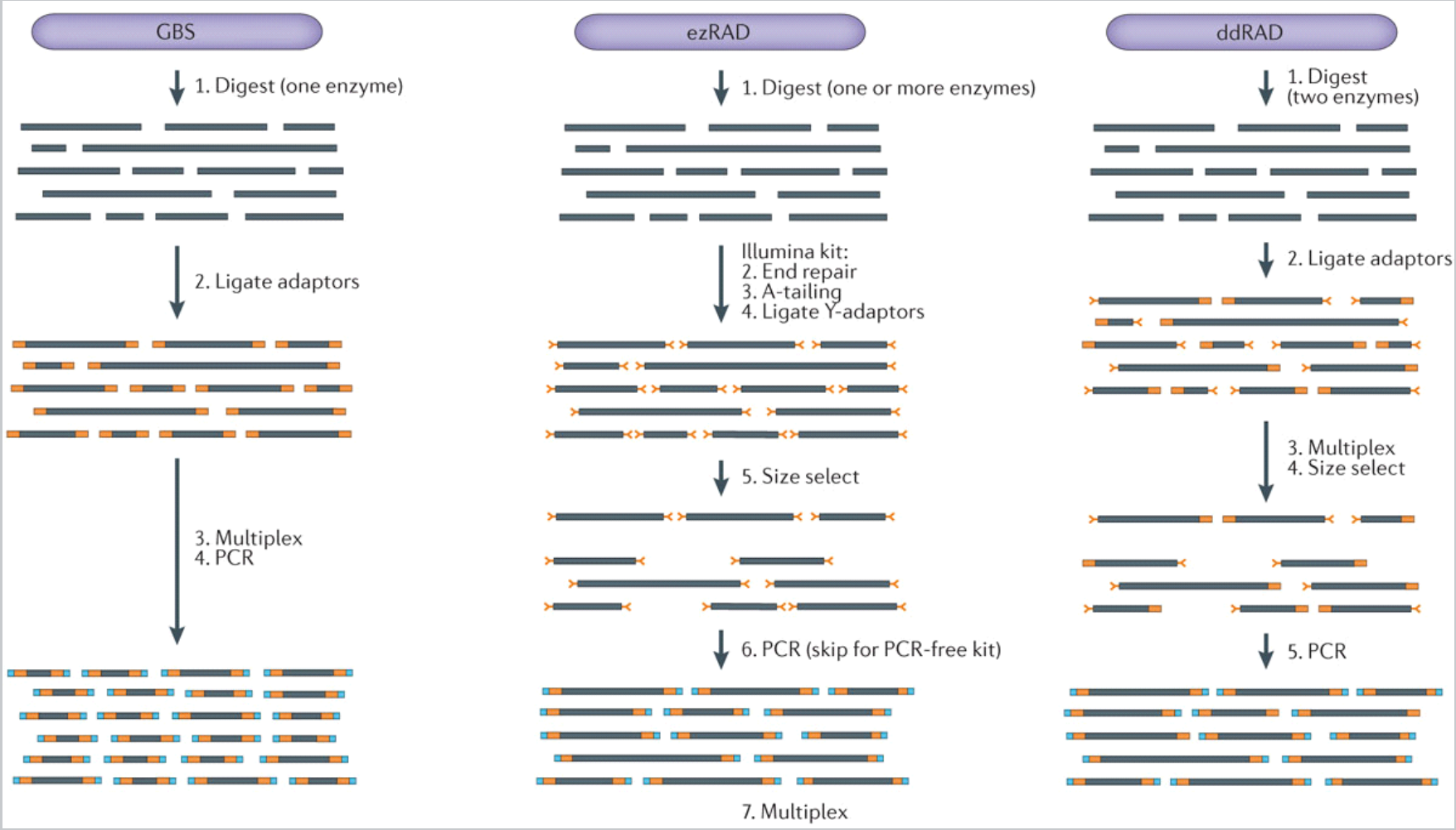
- uses IIB restriction enzymes – cleave DNA upstream and downstream of the recognition site
- results in short fragments of uniform length

Sequence flanked by two RE cut sites

single enzyme, indirect selection

single enzyme, direct selection

double enzyme, direct selection



- common-cutter enzyme
- PCR size selection (shorter fragments preferentially amplified)

- common-cutter enzyme(s)
- proprietary kit for Illumina library preparation

- two enzymes
- size selection by automated gel cut

RADseq methods comparison

Table 1 | **Summary of trade-offs among five RADseq methods**

	Original RAD	2bRAD	GBS	ddRAD	ezRAD
Options for tailoring number of loci	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme or size selection window	Change restriction enzyme or size selection window
Number of loci per 1 Mb of genome size*	30–500	50–1,000	5–40	0.3–200	10–800
Length of loci	≤1kb if building contigs; otherwise ≤300 bp‡	33–36 bp	<300 bp‡	≤300 bp‡	≤300 bp‡
Cost per barcoded or indexed sample	Low	Low	Low	Low	High
Effort per barcoded or indexed sample§	Medium	Low	Low	Low	High
Use of proprietary kit	No	No	No	No	Yes
Identification of PCR duplicates	With paired-end sequencing	No	With degenerate barcodes	With degenerate barcodes	No
Specialized equipment needed	Sonicator	None	None	Pippin Prep‖	Pippin Prep‖
Suitability for large or complex genomes¶	Good	Poor	Moderate	Good	Good
Suitability for *de novo* locus identification (no reference genome)#	Good	Poor	Moderate	Moderate	Moderate
Available from commercial companies	Yes	No	Yes	Yes	No

Andrews et al. (2016): Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Review Genetics* 17: 81-92.

RADseq bioinformatics

- demultiplexing, trimming barcodes
- filtering reads
  - presence of expected restriction site
  - quality
- PCR duplicate removal

- reference genome existing
  - align reads to the genome
  - call SNPs – define genotypes/haplotypes

- reference genome missing
  - de-novo assembly of reads
  - call SNPs – define genotypes/haplotypes

- software – Stacks, pyRAD, AftrRAD, dDOCENT

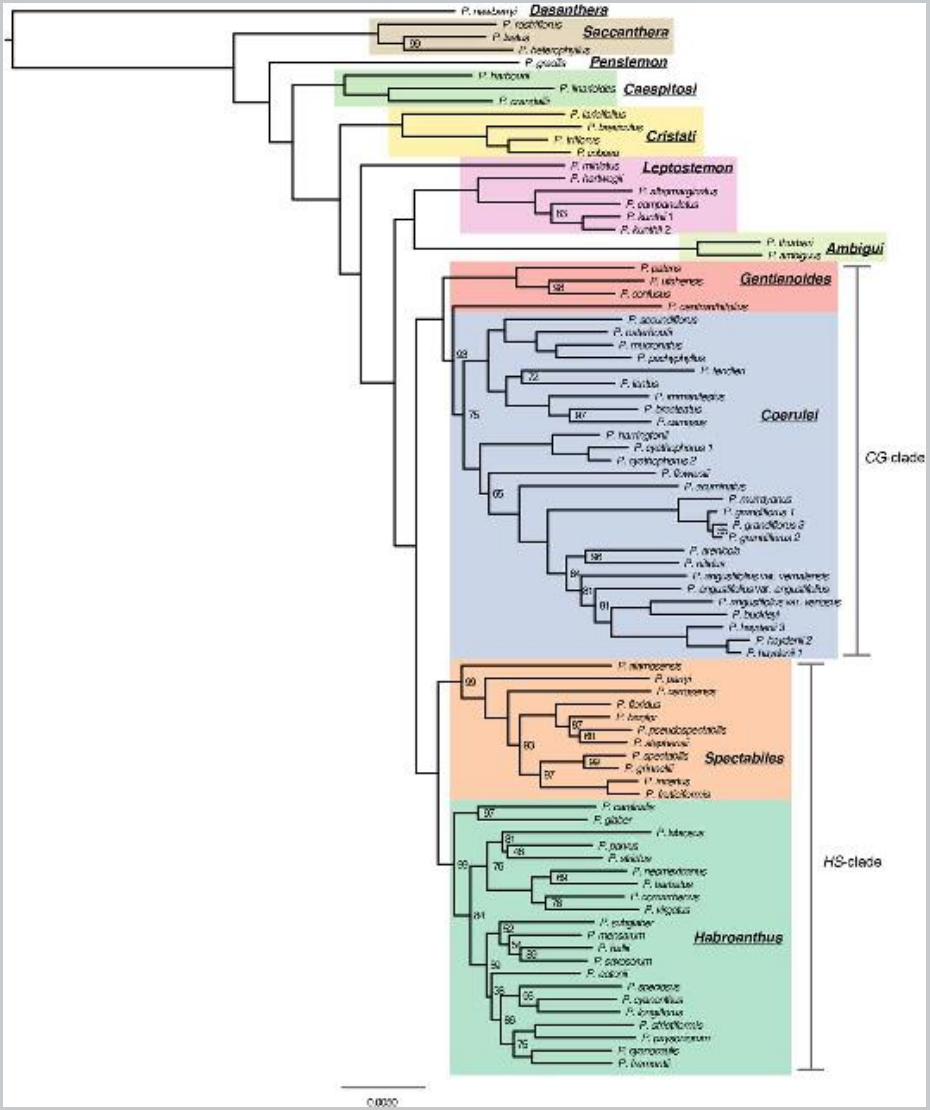
RADseq data properties

- relatively short loci
- wide genomic distribution
- allelic dropout/null alleles
- large proportion of missing data
- orthology/paralogy – bioinformatic assessment

RADseq application

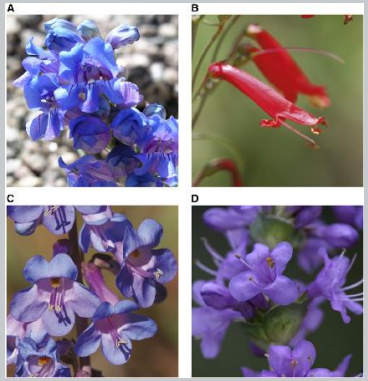
- phylogenomics
- population structure, phylogeography
- population genomics
- evolution of recently radiated groups
- hybridization, introgression
- …

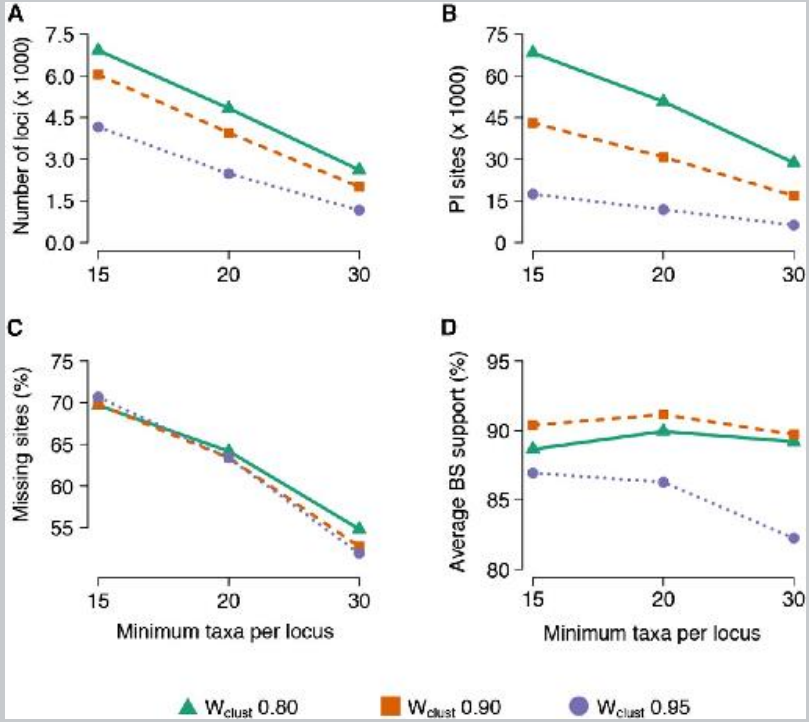
RADseq phylogenomics



Penstemon

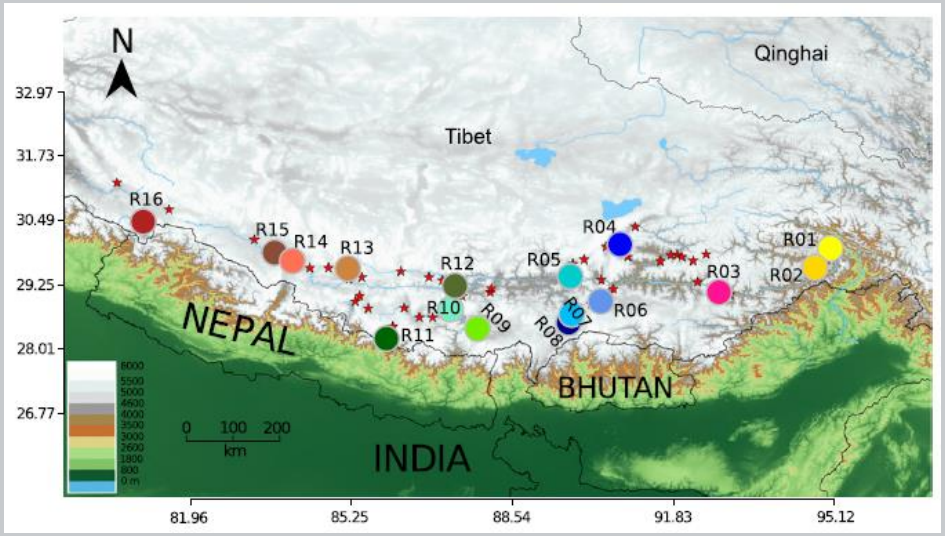
- 75 species
- 13 sections





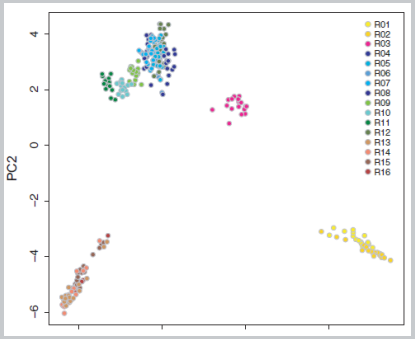
Wessinger et al. (2016): Multiplexed shotgun genotyping resolves species relationships within the North American genus *Penstemon*. Am. J. Bot. 103(5): 912-922.

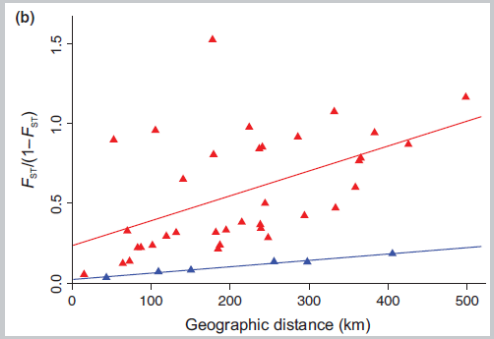
RADseq population structure

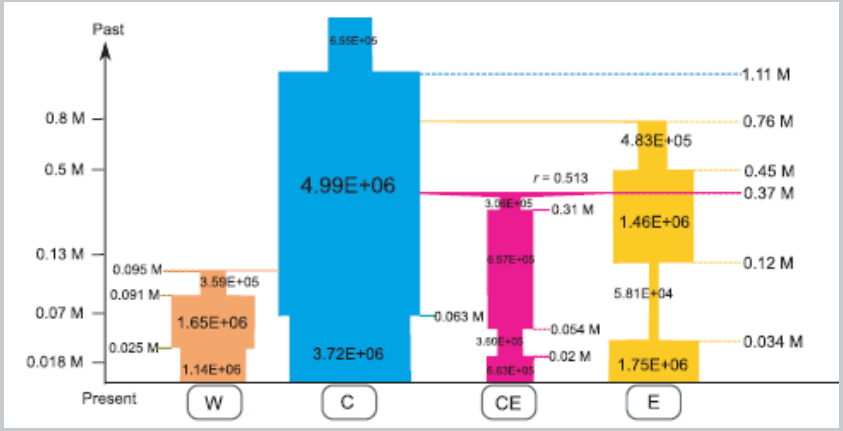


Primula tibetica

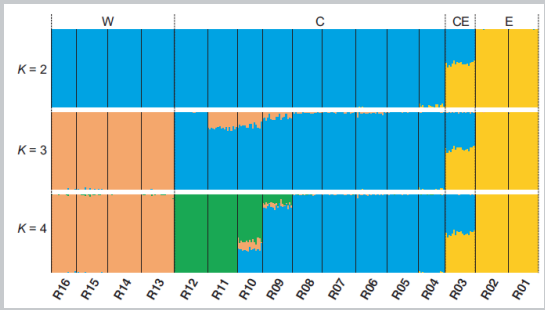
- 293 samples
- 61 populations
- 4 groups







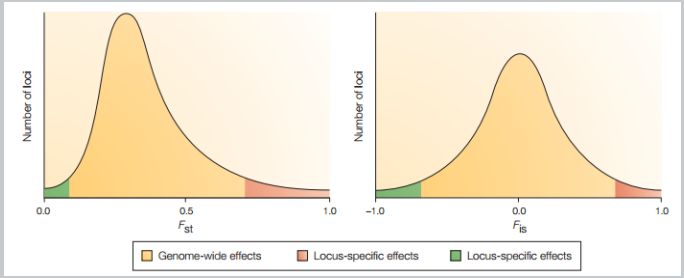
inferred demographic histories



Ren et al. (2017): Genetic consequences of Quaternary climatic oscillations in the Himalayas: *Primula tibetica* as a case study based on restriction site-associated DNA sequencing. New Phytol. 213(3): 1500-1512.

Population genomics

- simultaneous study of numerous loci to better understand the roles of evolutionary processes (such as mutation, random genetic drift, gene flow and natural selection) that influence variation across genomes and populations
- **neutral loci** – will be similarly affected by demography and the evolutionary history of populations
- **loci under selection** (adaptive) – often behave differently and reveal 'outlier' patterns of variation
- identification of outlier loci (high or low F_{ST} between populations)

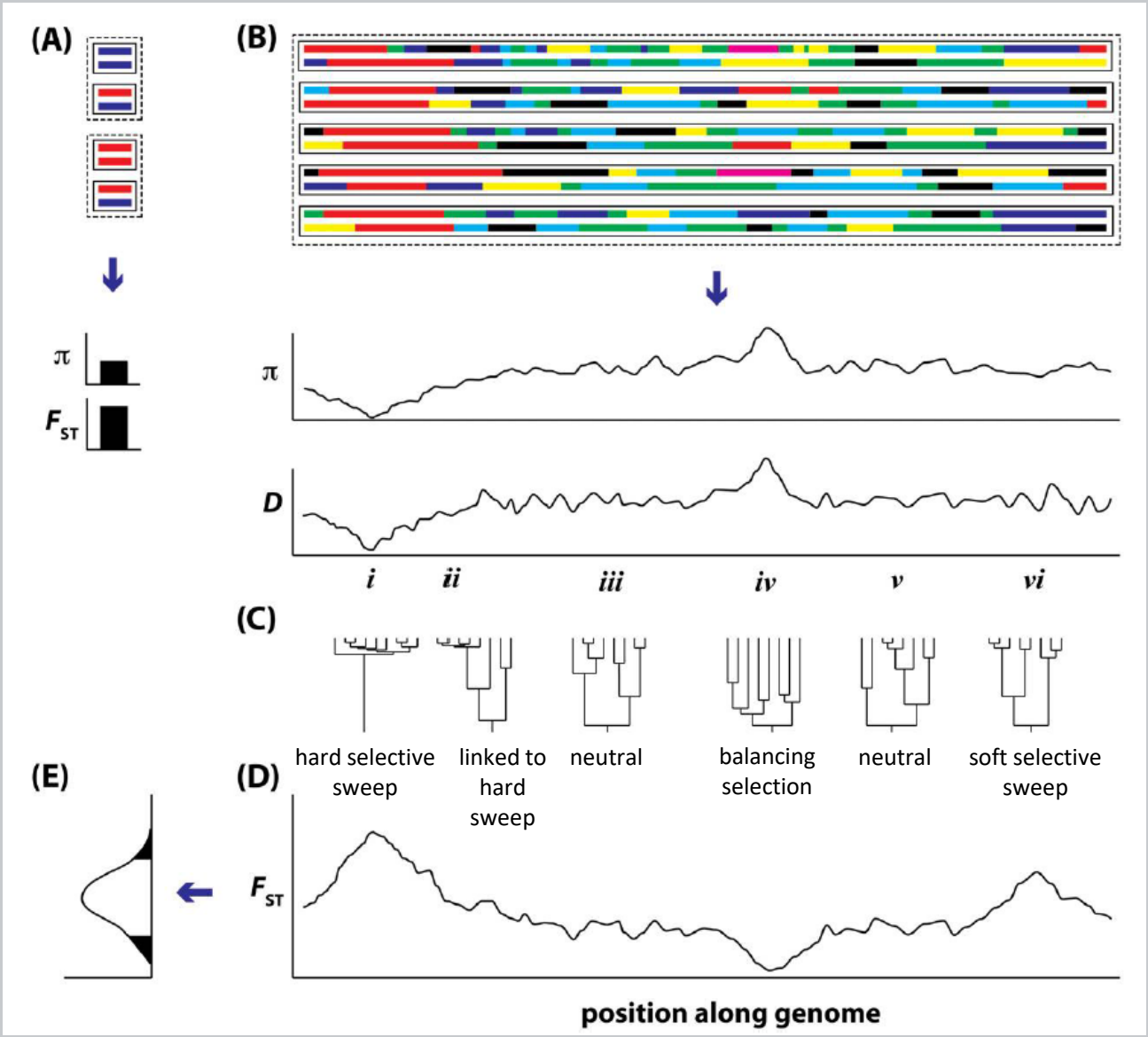


Luikart et al. (2003): The power and promise of population genomics: from genotyping to genome typing. *Nature Review Genetics* 4: 981-994.

Population genomics perspective

population genetics

population genomics



nucleotide diversity

allele frequency spectrum – Tajima's *D*

coalescent structure of ancestral relationships among alleles

genetic variation across populations

estimate of genome wide average

Hohenlohe et al. (2010): Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *Int. J. Plant Sci*. 171(9): 1059–1071.

Site (allele) frequency spectrum

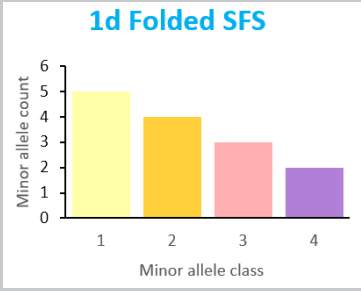
distribution of allele frequencies

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ind #1 allele 1	A	A	T	A	G	C	G	G	A	T	G	C	A	T
Ind #1 allele 2	A	A	A	C	G	C	C	C	A	T	G	C	A	T
Ind #2 allele 1	A	T	A	C	T	C	C	C	A	A	G	G	A	T
Ind #2 allele 2	A	A	A	C	T	C	G	C	A	A	C	C	T	T
Ind #3 allele 1	T	A	T	C	T	C	G	G	T	T	G	C	A	A
Ind #3 allele 2	T	A	T	C	G	G	G	G	T	T	G	C	A	T
Ind #4 allele 1	T	A	T	C	G	G	G	G	A	A	G	C	A	T
Ind #4 allele 2	T	A	T	C	G	C	G	G	A	A	C	C	A	T

Folded SFS

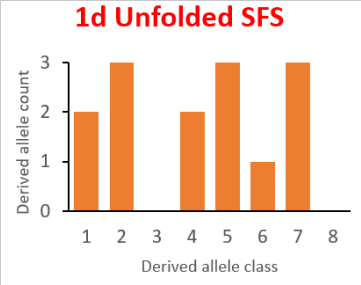
minor allele	T	T	A	A	T	G	C	C	T	A	C	G	T	A
minor allele count	4	1	3	1	3	2	2	3	2	4	2	1	1	1

singletons	5
doubletons	4
tripletons	3
freq. 4	2

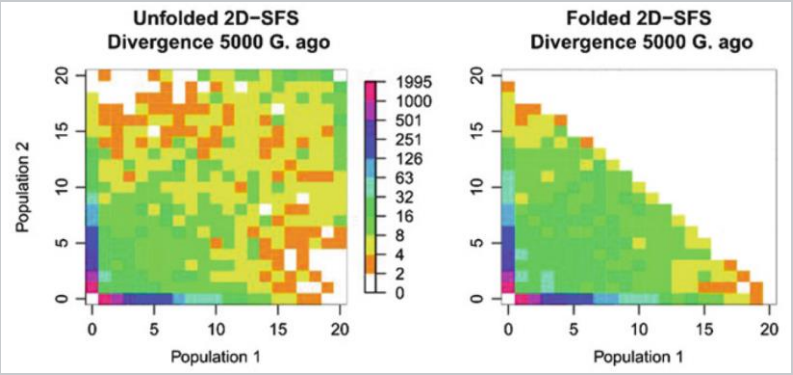


Unfolded SFS

ancestral allele	A	A	A	A	T	C	G	C	T	T	G	G	T	T
derived allele count	4	1	5	7	5	2	2	5	6	4	2	7	7	1



- 1dSFS - for single population
- 2dSFS - for two populations
- jointSFS - for more populations
- known ancestral allele - unfolded SFS
- unknown ancestral allele - folded SFS

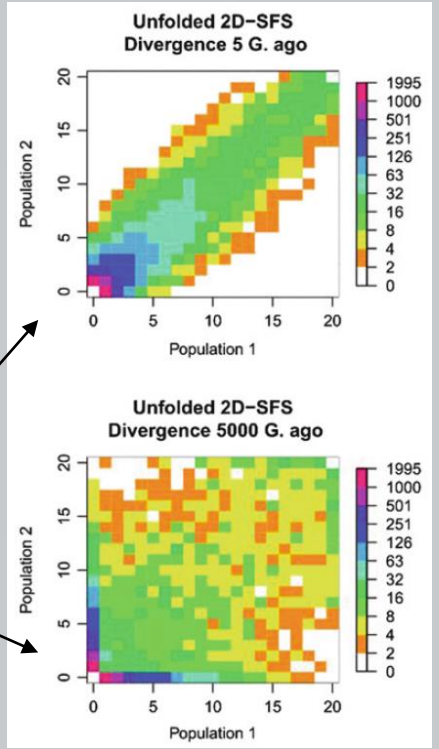


Salmona et al. 2019

SFS and population history

- changes in N_e distort gene genealogies and impact freq. distribution across SFS
  - inferred from fitting SFS under a particular demographic model against observed SFS
- expanding population – more singletons in 1dSFS than population of constant size
- declining population – deficit of singletons

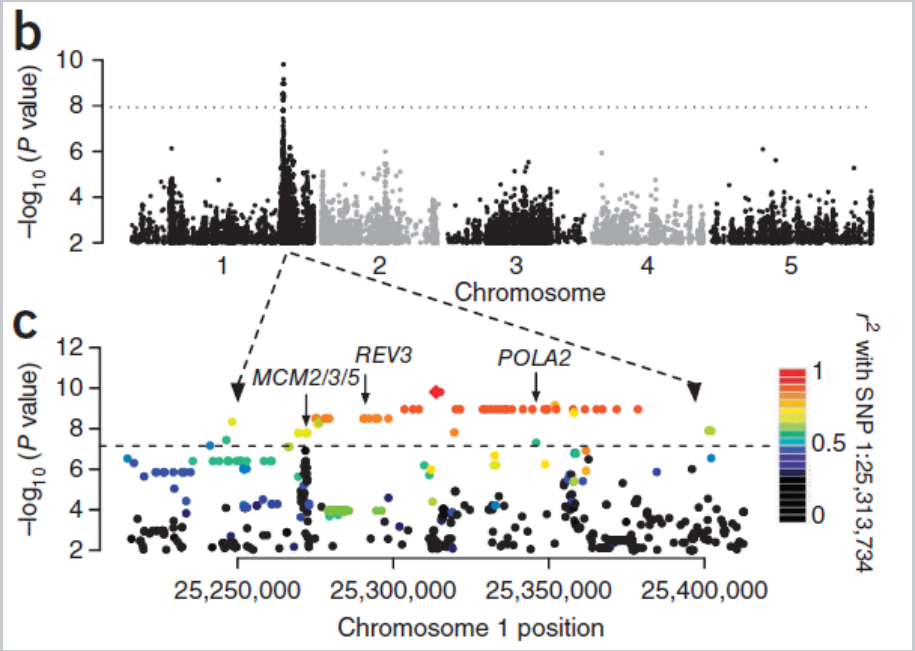
- 2dSFS – total number of segregating sites in which the allele is observed in each population
  - recently diverged pops – density concentrated along the diagonal, i.e. shared history
  - highly divergent pops – density along axes (most alleles private)
- model-based approaches (e.g. fastsimcoal2)
  - allow to infer size changes, splits, divergence, migration…



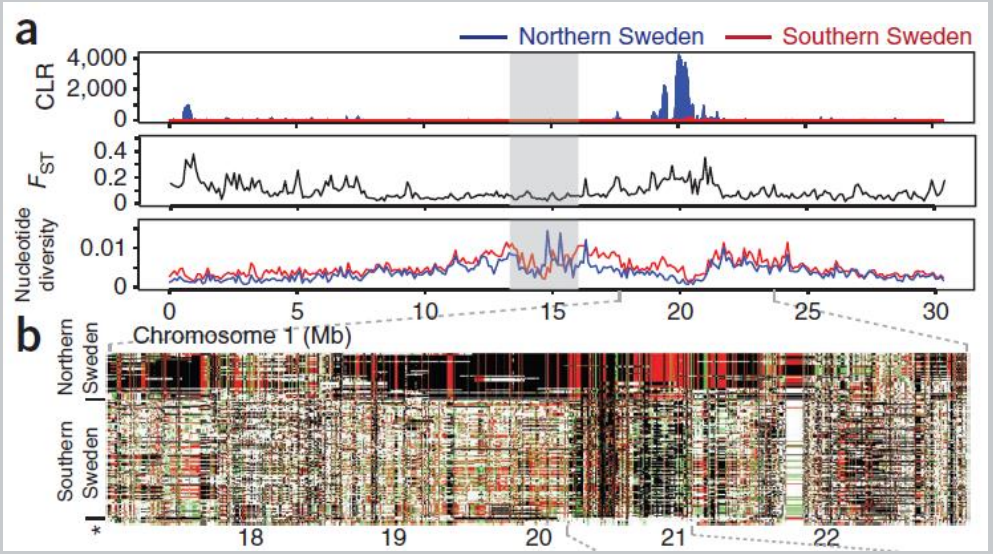
Salmona et al. 2019

Selection in *Arabidopsis thaliana*

- 180 lines from S and N Sweden
- massive variation in genome size due to 45S rDNA copy number variation
- massive global selective sweep (700-kb transposition)



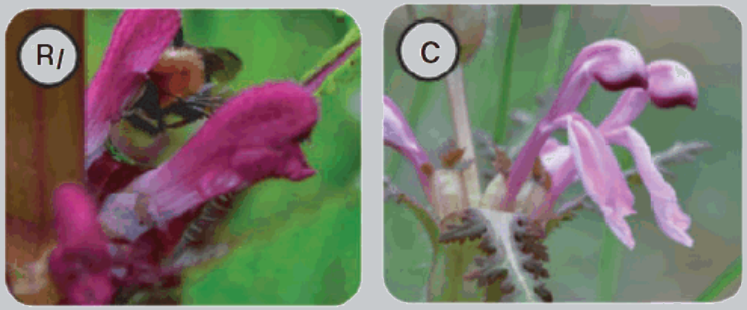
genome-wide association studies (GWAS)
identification of loci associated with, e.g., particular phenotype/trait

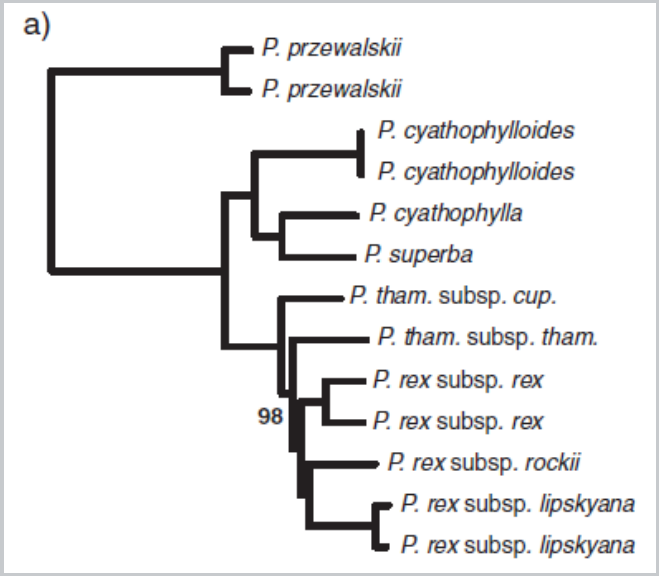


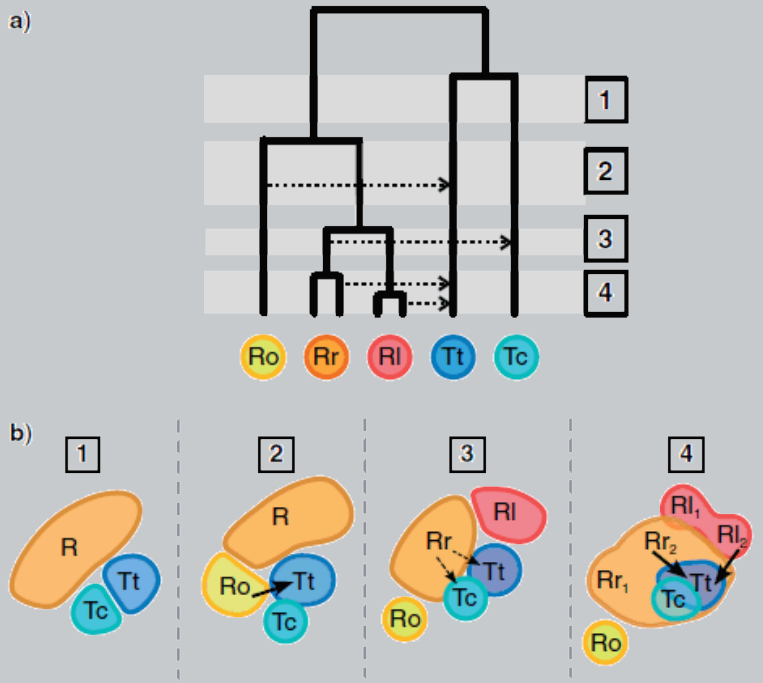
Long et al. (2013): Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics 45*(8): 884–891.

RADseq in recently diversified group

- recently diversified group – closely related species
- phylogeny and detection of ancestral hybridization
- 40,000 loci

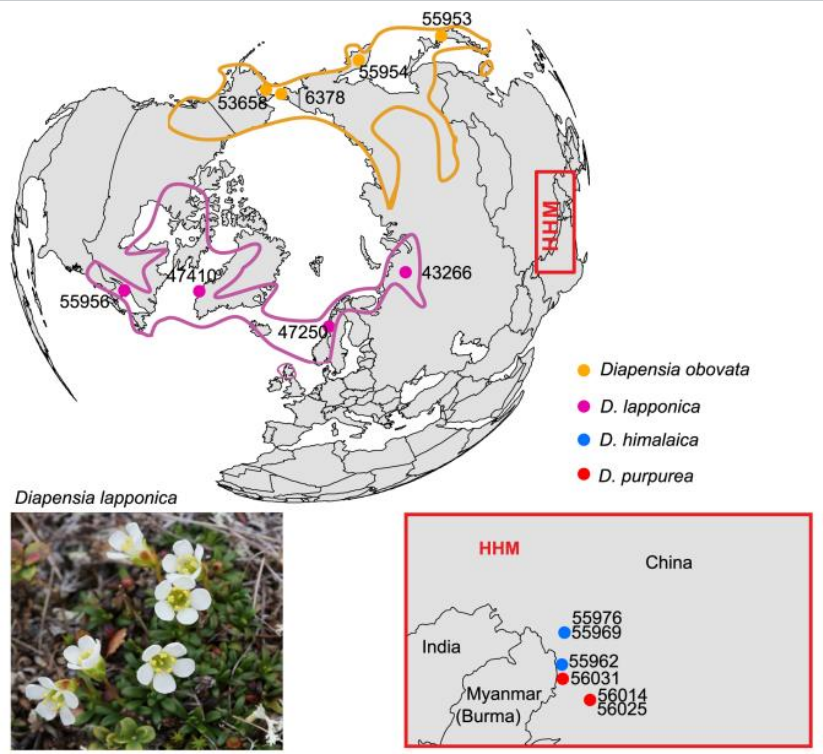






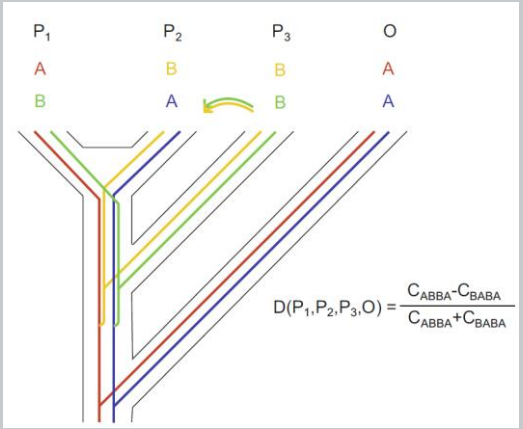
Eaton & Ree (2013): Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst. Biol*. 62(5):689–706.

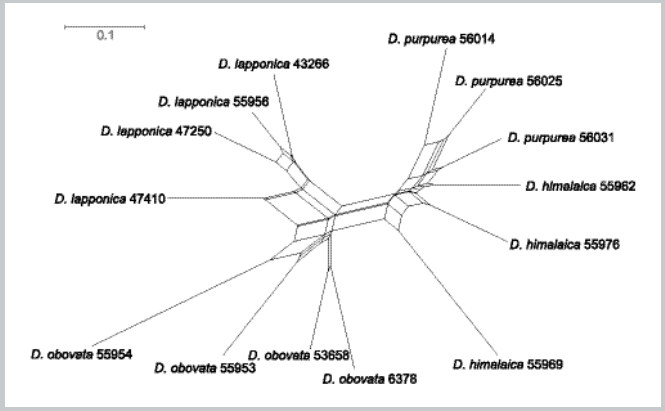
Testing for admixture



four-taxon D-statistic test

- two incongruent patterns of two biallelic SNPs (ABBA, BABA)
- these should be equally present under a scenario of ILS without gene flow
- excess of ABBA or BABA patterns is indicative of gene flow





testing admixture between *Diapensia purpurea* and *D. himalaica*

- 9 out of 18 tests detected significant signal
- congruent with reticulation in network

Hou et al. (2015): Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus *Diapensia* (Diapensiaceae). *PLoS ONE* 10(10): e0140175.

Comparison of RADseq and target enrichment

Category	RAD-Seq	Sequence capture
Marker distribution and genomic context	Pro: Widely dispersed across genome	Pro: Can be tailored using new genomic information
	Con: Anonymous, evolutionary processes largely unknown	Con: Purifying selection impacts allele frequencies
Practical considerations	Pro: Less expensive, faster	Pro: Works with low-quality and highly contaminated samples
Assembly and orthology identification	Pro: Deep coverage, high read overlap	Pro: Over-splitting less problematic
Variant-calling and genotyping	Pro: Fewer rare alleles may make errors easier to distinguish, phasing more straightforward	Pro: Fewer low-coverage rare alleles, no allele dropout
Information content	Pro: More overall information	Pro: More information per locus
Applications	Genome scans, rapid and inexpensive analyses, analyses using species in clades without genomic information, extremely shallow divergences and otherwise intractable relationships.	Comparisons across species, calibrating parameter estimates, targeting loci of known utility or interest, studies using poor-quality samples, studies requiring resolved gene trees, deeper phylogenetic studies.

Harvey et al. (2016): Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Syst. Biol*. 65(5):910–924

Literature

Davey J.W. & Blaxter M.L. (2011): RADSeq: next-generation population genetics. *Briefings in Functional Genomics* 9: 416-423.

Davey J.W. et al. (2011): Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews* 12: 499-510.

Peterson B.K. (2012): Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7(5): e37135.

Andrews et al. (2016): Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Review Genetics* 17: 81-92.

Rubin B.E.R. et al. (2012): Inferring Phylogenies from RAD Sequence Data. *PLoS ONE* 7(4): e33394.

Ree R.H. & Hipp A.L. (2015): Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl E. & Appelhans M.S. (eds.): Next-generation sequencing in plant systematics. IAPT

Harvey et al. (2016): Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Syst. Biol*. 65(5):910-924.

Hohenlohe P.A. et al. (2010): Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *Int. J. Plant Sci*. 171(9): 1059-1071.

Ellegreen H. (2014): Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol*. 29: 51-63.

Rajora O.P., ed. (2019): Population Genomics. Concepts, Approaches and Applications. Springer.