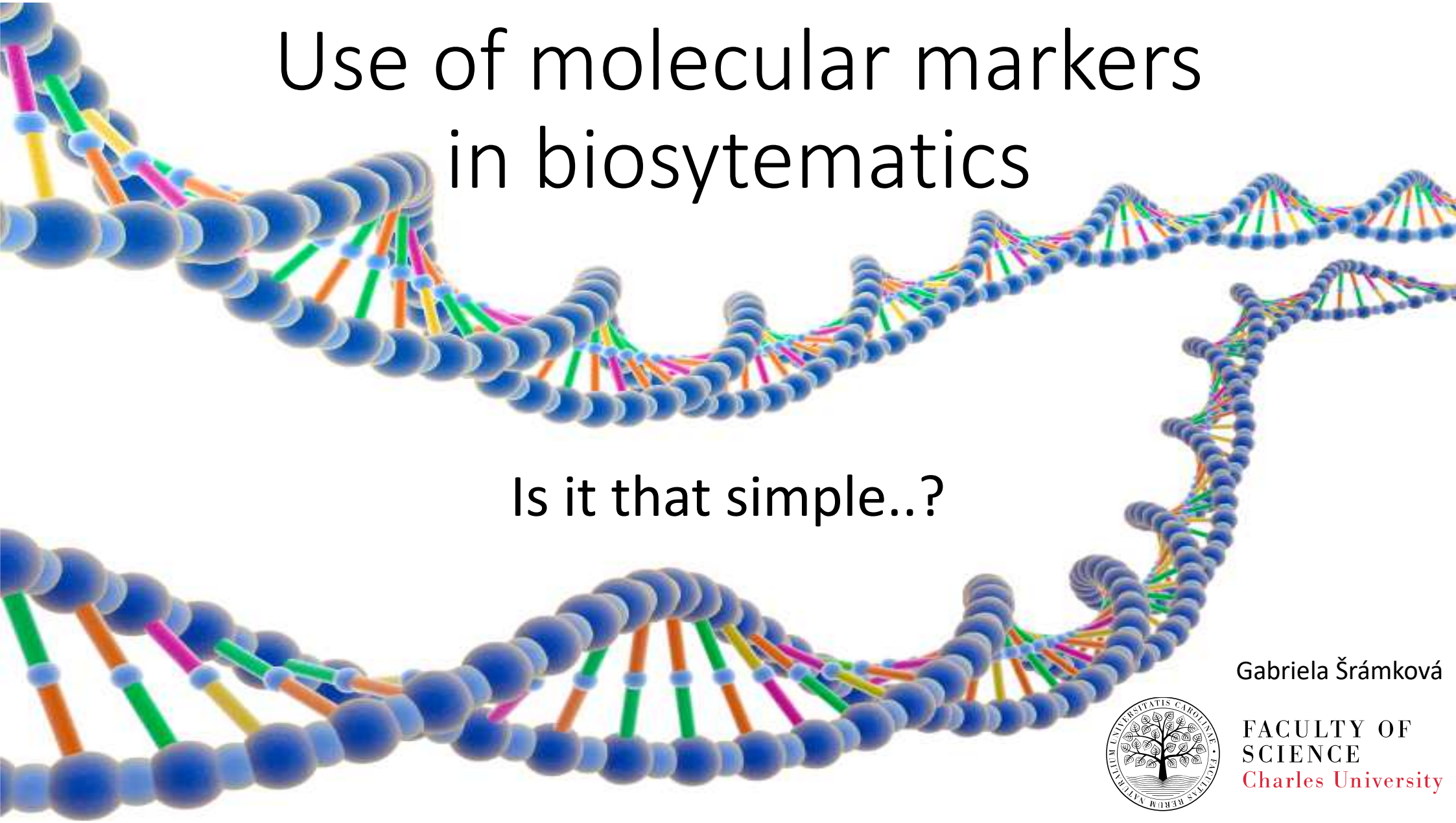
Use of molecular markers in biosytematics



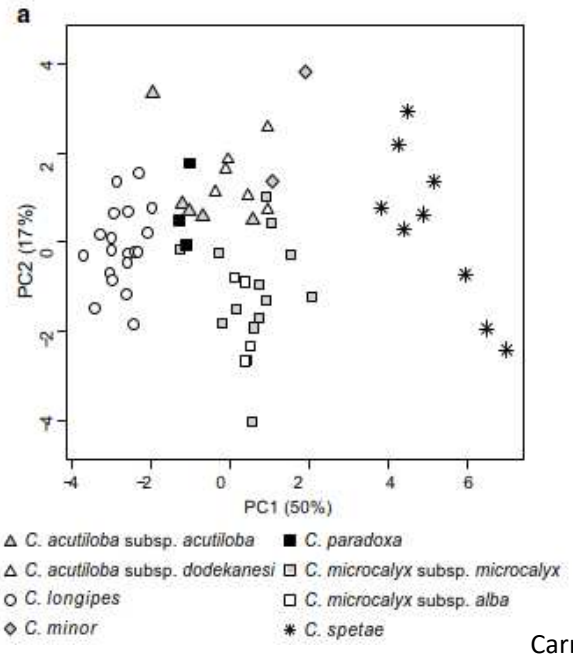
Is it that simple..?

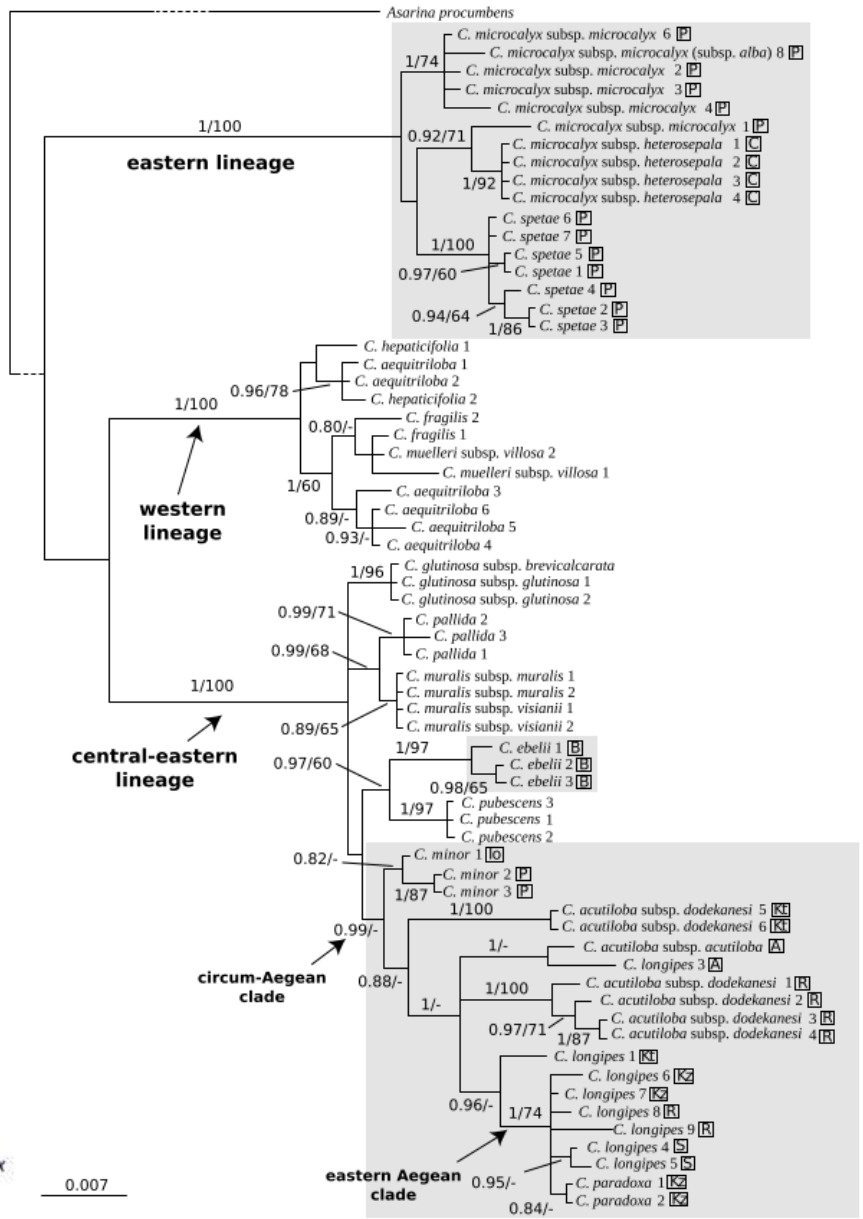
Gabriela Šrámková

FACULTY OF SCIENCE
Charles University

Biosystematics

- complex study of taxonomy of organisms based on comparisons of their ecological, cytological or genetic characteristics

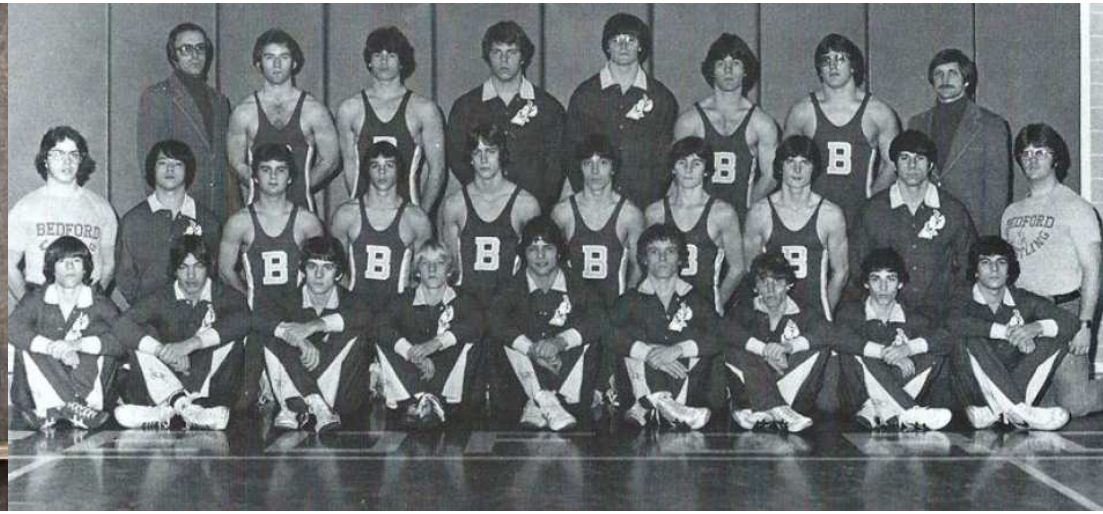




Carnicero et al. 2021 *Cymbalaria*

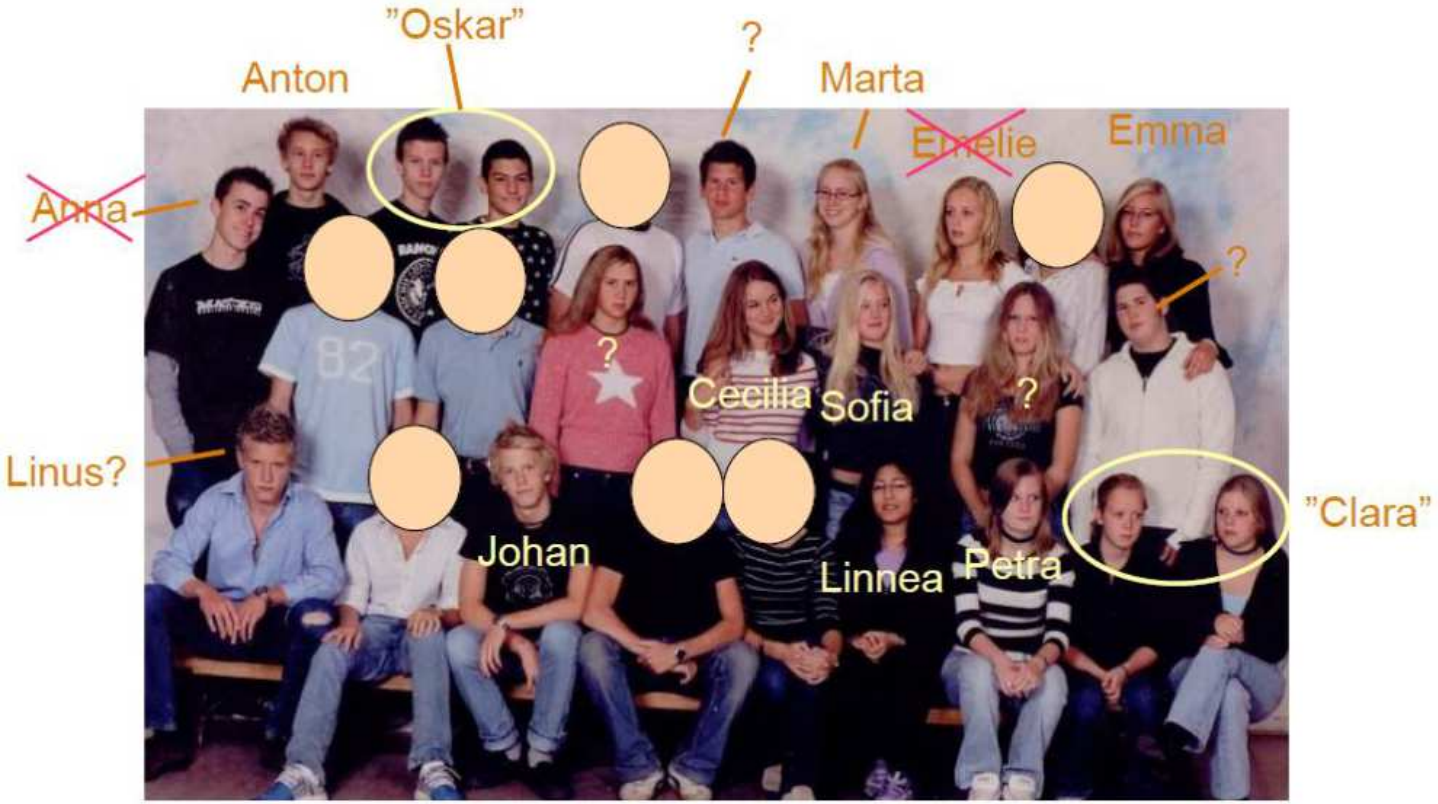
Biosystematics – what for?

- Imagine people or things do not have names…



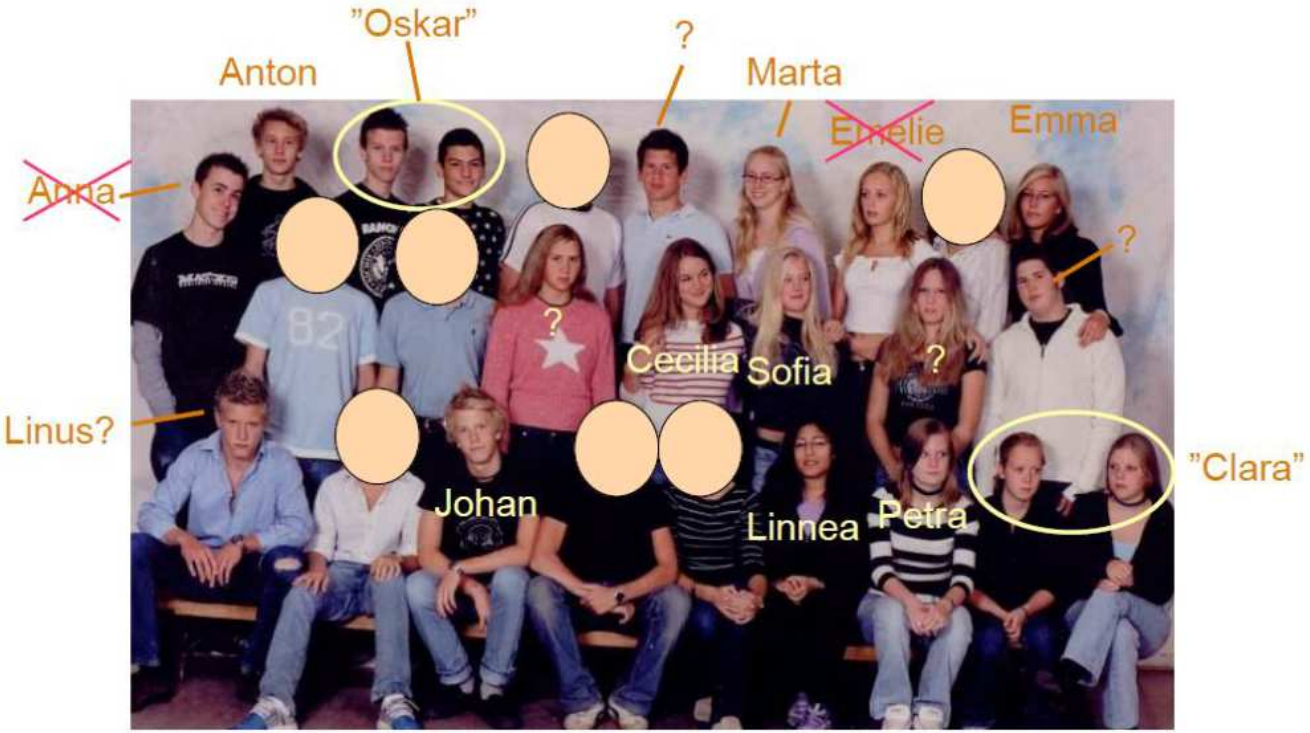
Biosystematics – what for?

- Or a lot of names would be wrong…

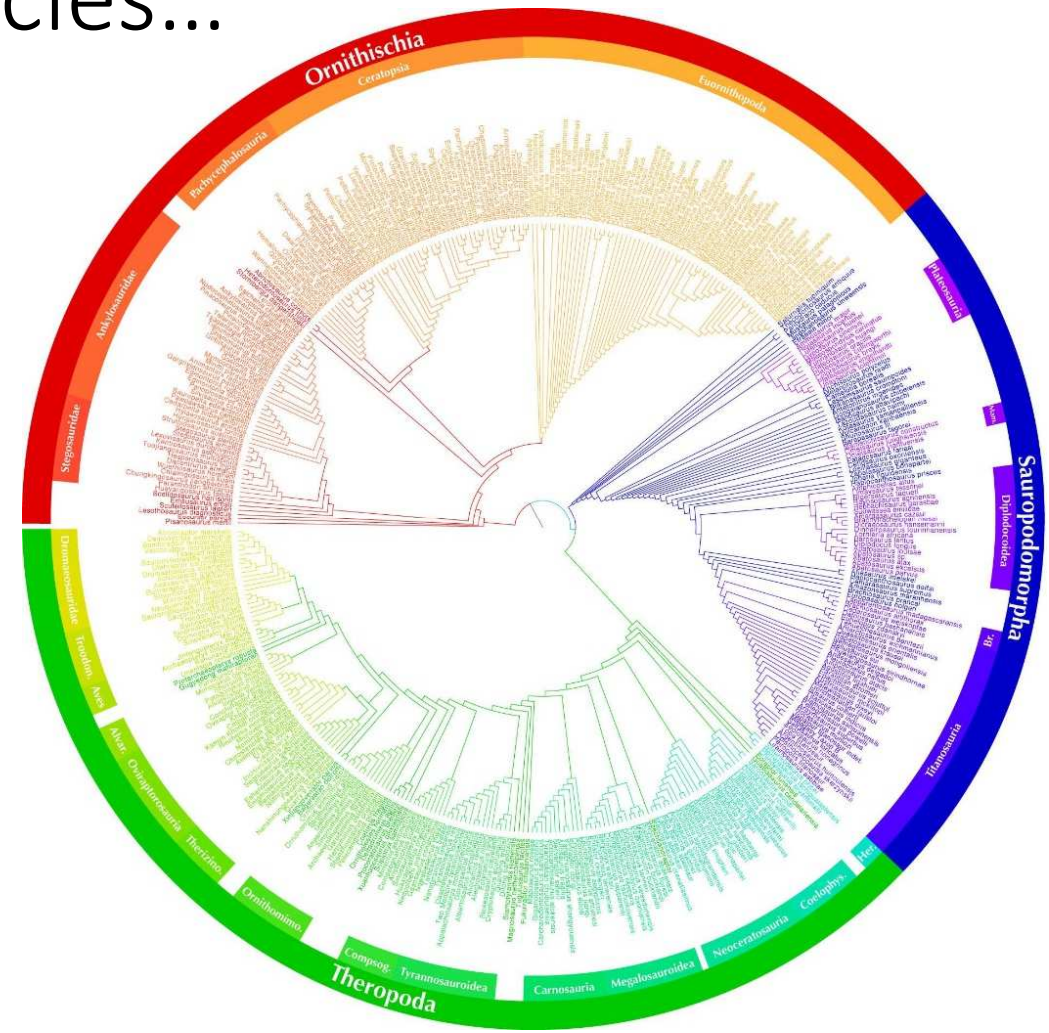


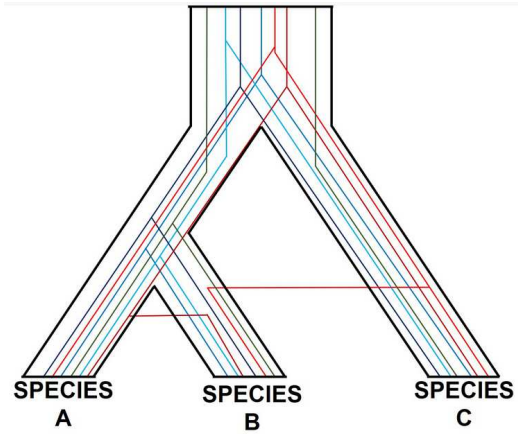
Biosystematics – what for?

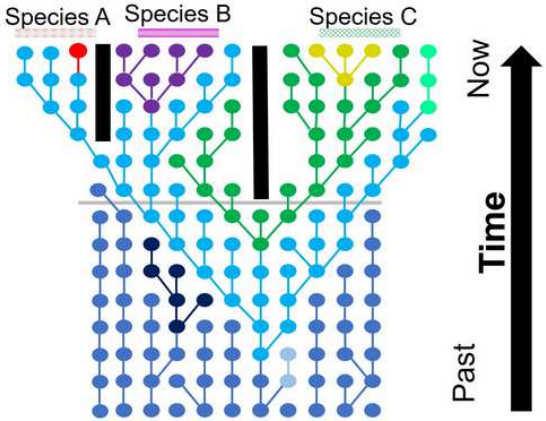
- Biosystematics creates "language" and interprets it. Without it, many other disciplines (such as protection, evolutionary biology etc.) could not exist...

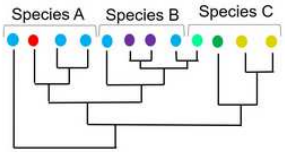


Species…


Ornithischia
Ceratopsia
Euornithopoda
Pachycephalosauria
Ankylosauridae
Stegosauridae
Thyreophora
Sauropodomorpha
Diplodocoidea
Titanosauria
Hadrosauridae
Neoceratosauria
Coelophysoidea
Carnosauria
Megalosauroidea
Tyrannosauroidea
Compsognathidae
Ornithomimosauria
Alvarezsauroidea
Oviraptorosauria
Therizinosauroidea
Troodontidae
Dromaeosauridae
Theropoda


SPECIES A
SPECIES B
SPECIES C


Species A
Species B
Species C
Now
Time
Past


Species A
Species B
Species C

Species…

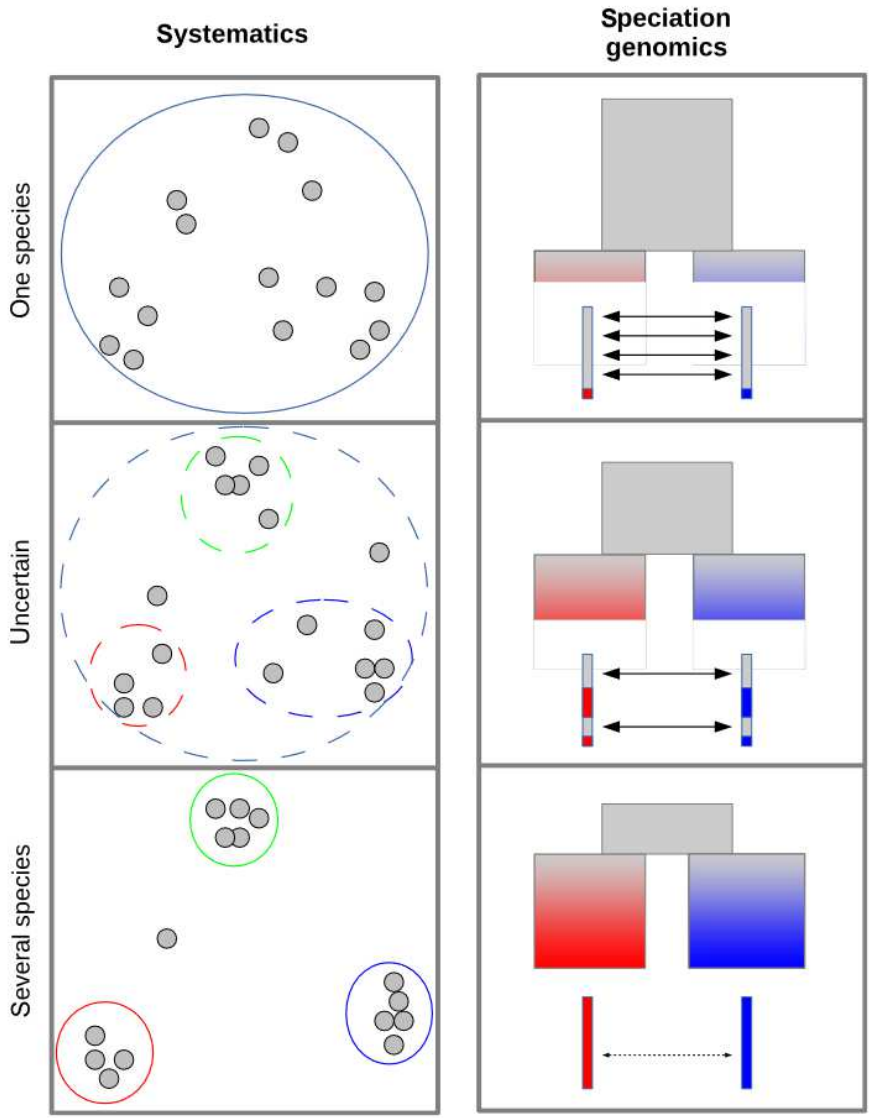


FIGURE 1 Two views on the continuum of speciation. Left: Species are defined as groups of organisms resembling each other according to an arbitrary set of variables. Right: Species are defined as entities sufficiently diverged such that gene flow (arrows) is very rare or inexistent. Top: unambiguous single-species situation. Bottom: unambiguous multiple-species situation. Intermediate: ambiguous situation. Ambiguous situations appear when groups can be identified but intermediate individuals are common (left) and when gene flow exists but is limited to a fraction of the genome (right)

Galtier 2018

Species...

Species concepts

- biological, morphological, phylogenetic...
- there are many and definitions vary
- the definition is based on breaking the continuum of variation
- the assumption of reproductive isolation

- the species is understood differently - by taxonomists, conservationists, in different countries, in
- different times, different groups of organisms
- (zoologists/botanists)

Where to begin?

Where to begin?

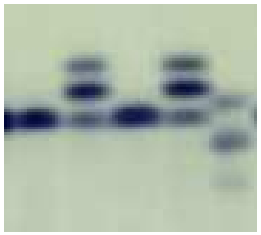
- Thorough study of the complex
- Sampling
- Phylogenetic relationships
- Ecological relationships
- Morphological analyses
- Taxonomic re-evaluation

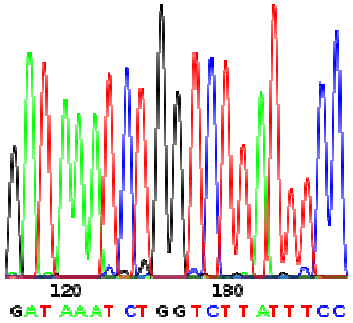
Where to begin?

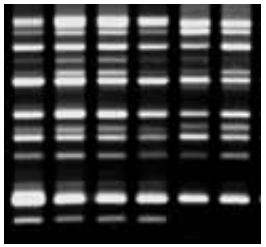
- Thorough study of the complex
- Sampling
- **Phylogenetic relationships**
- Ecological relationships
- Morphological analyses
- Taxonomic re-evaluation

Molecular markers

- information about an organism obtained from the analysis of its molecules - proteins, DNA, RNA
- marker - character, unit of information - a purposefully or randomly selected part of the total information
- markers tell about the genetic similarity (relatedness) of individuals, populations or species
- electrophoresis of macromolecules







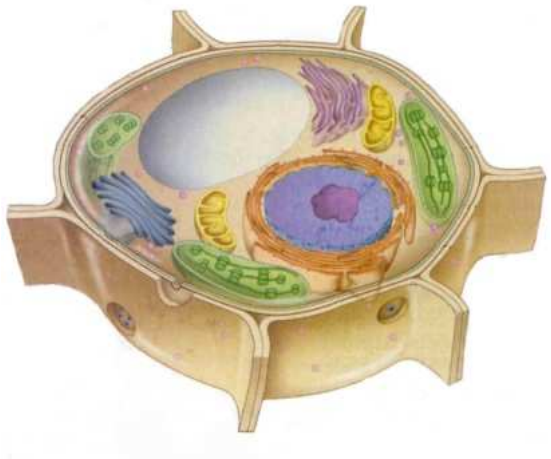
Molecular markers

The nature of molecular data:

- provide information about the genotype of an individual
- information independent of environmental conditions
- assumption of selective neutrality - no effect on fitness
- qualitative information - presence of fragment, allele, nucleotide
- unique information about the organism - clone identification
- changes during generative reproduction – recombination

Storage of genetic information:

- nucleus, plastids, mitochondria



Molecular markers

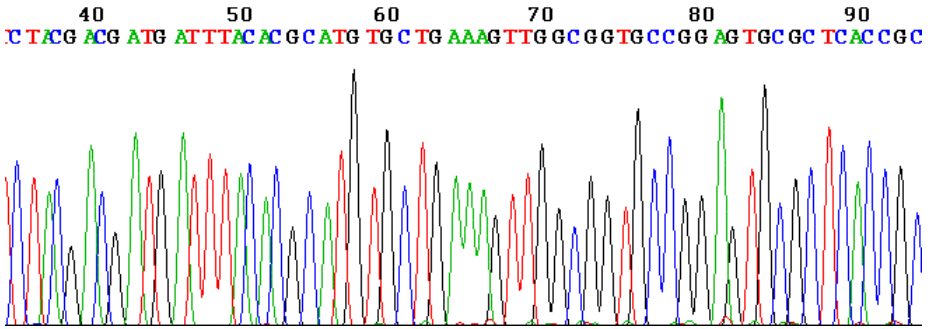
Types of questions:

- identification of clones
  - genetic diversity of clonal plants
- genetic structure of populations
  - intra-population genetic diversity
  - H-W equilibrium test
  - relationships between populations, distribution of genetic variability
  - gene flow
- study of plant migration
  - phylogeography
  - study of invasions
- type of reproductive systém
- systematic studies at all levels, reconstruction of phylogeny
- hybridization, polyploidization

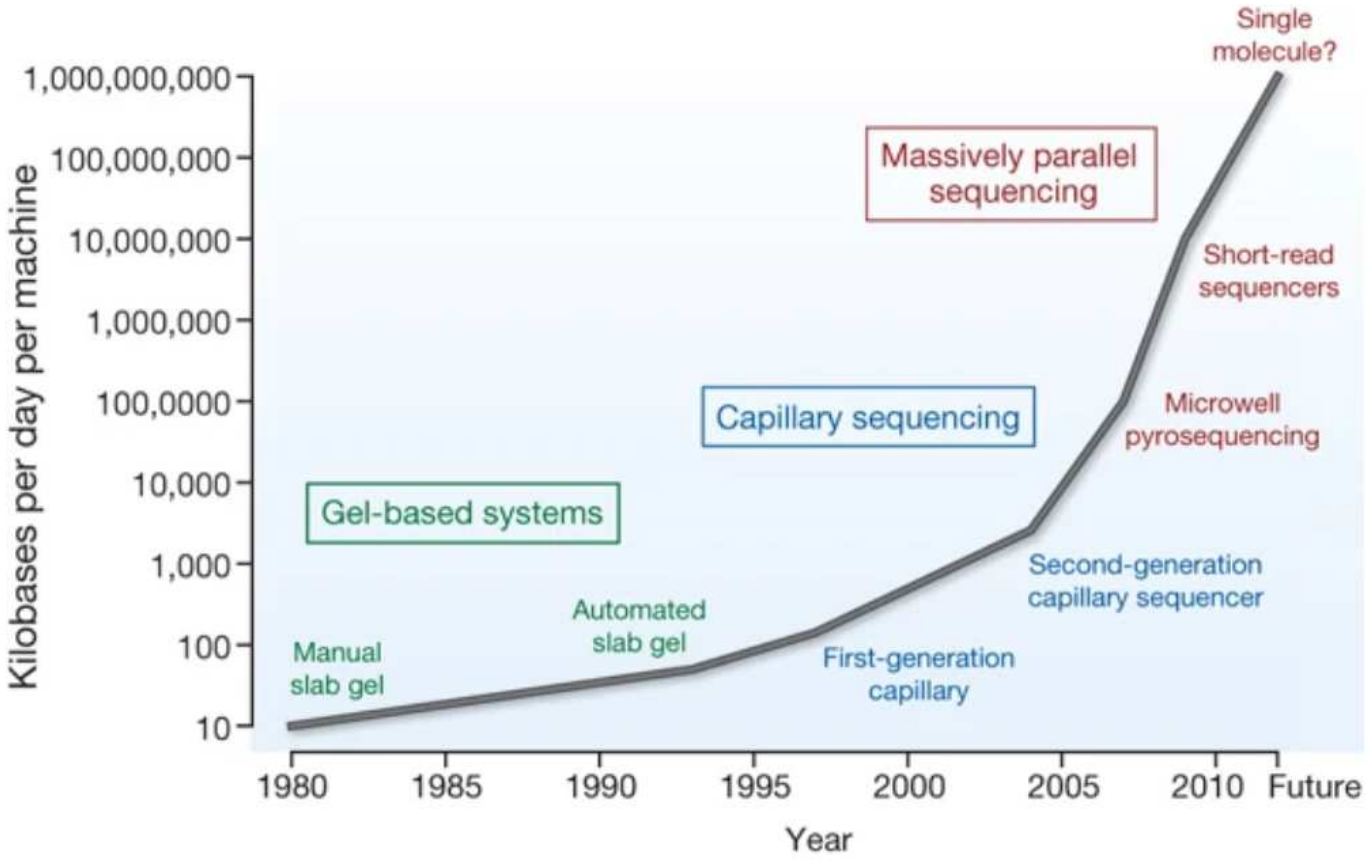
Sequencing

- finding the order of nucleotides in a DNA strand

  …ATATATAGGCAAGGAATCTCTATTATTAAATCATT…
- using information to determine the course and rate of evolution
- determining the similarity and relatedness of taxa
- capillary sequencing vs next generation sequencing

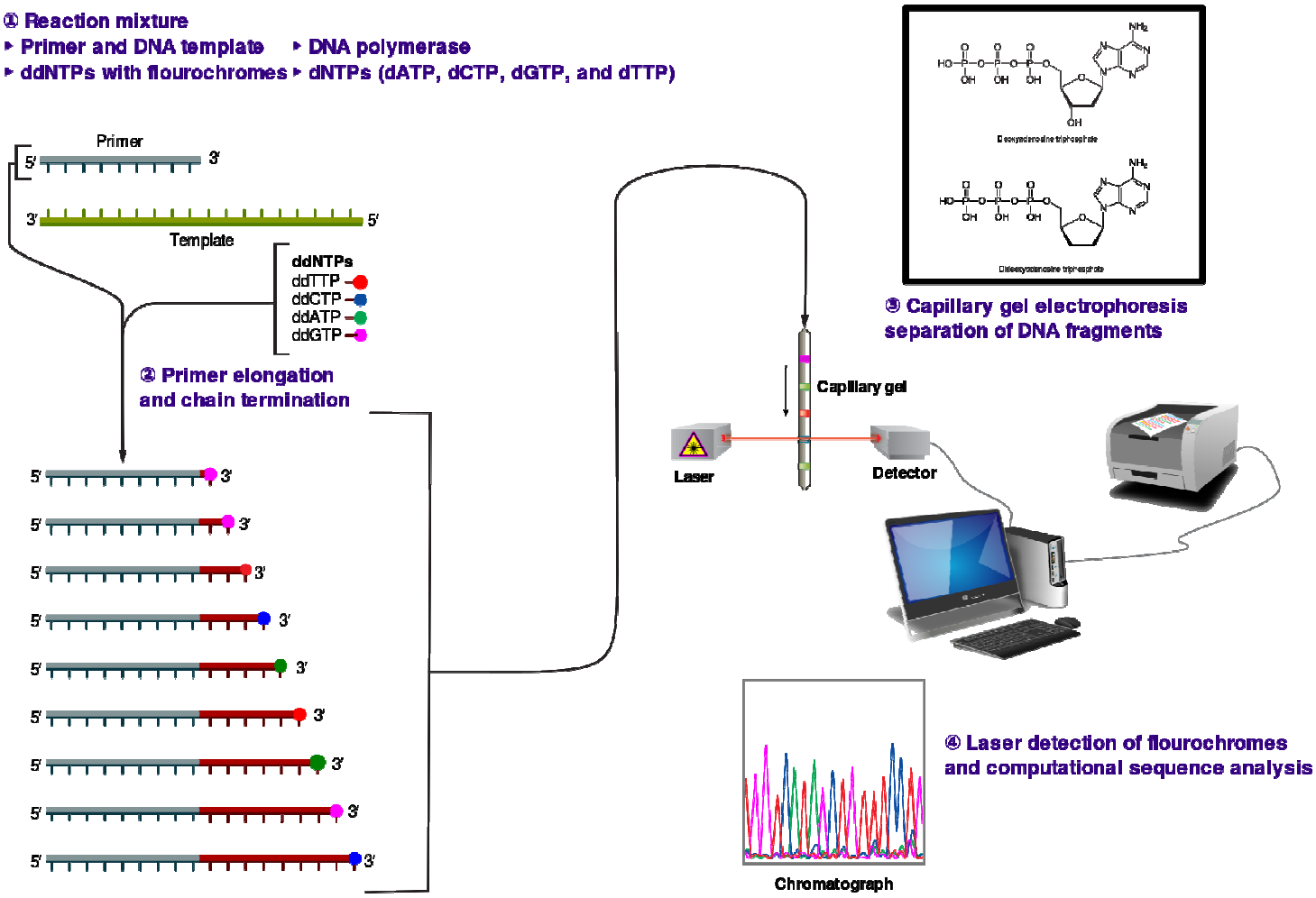


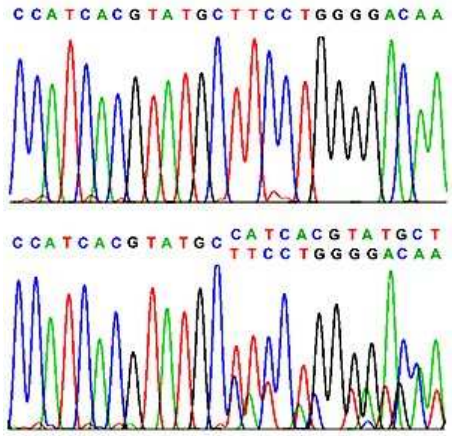
The evolution of sequencing



Stratton et al. , Nature 2009

Sanger sequencing

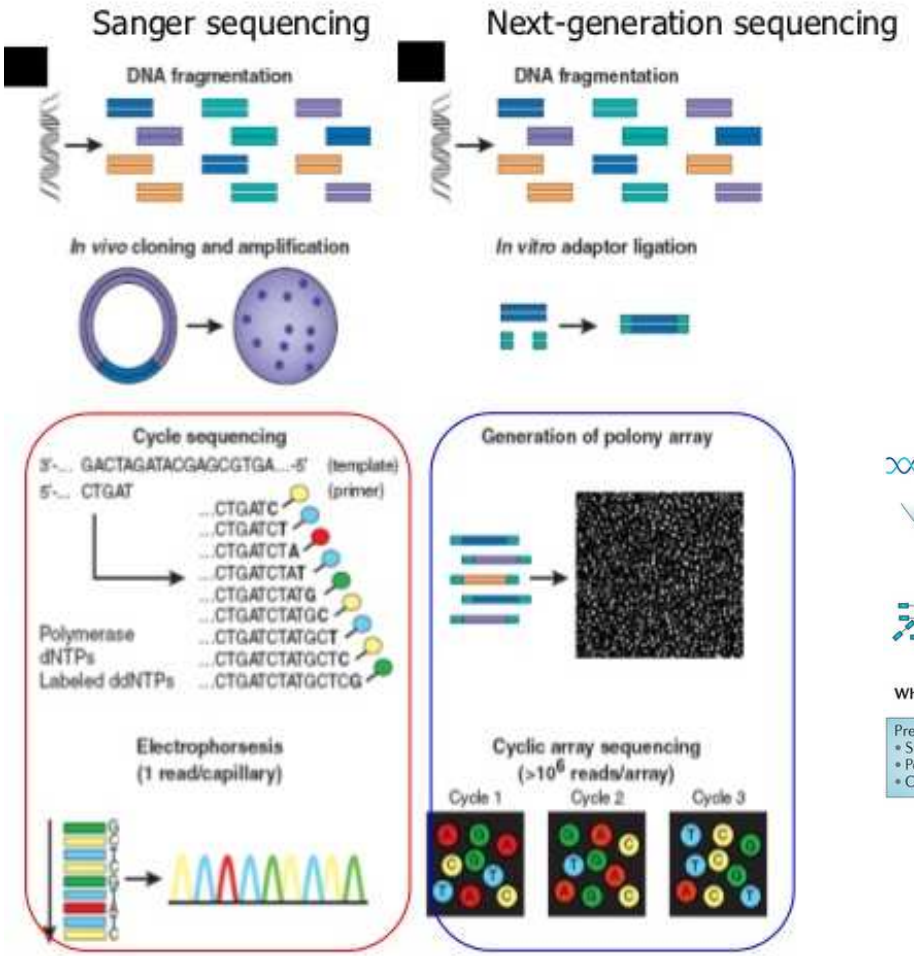




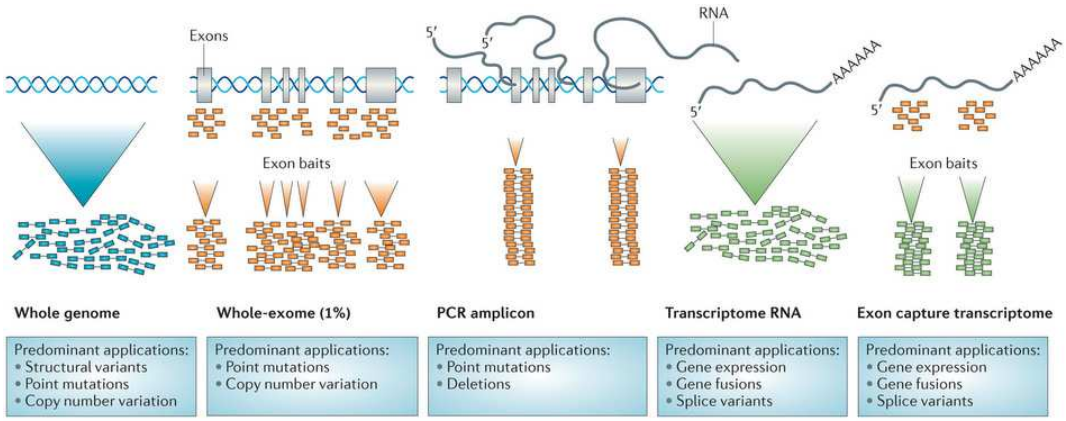
Disadvantages

- cloning
- low possibility of parallelisation

NGS (next generation sequencing)



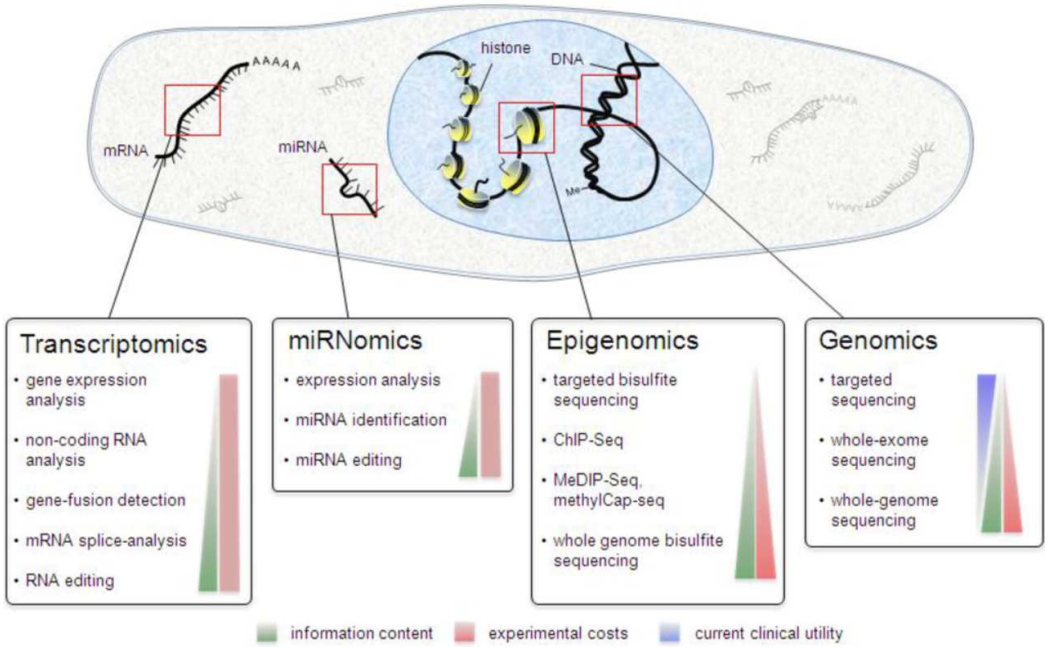
Massive parallelization



Simon et al. , Nat Rev Drug Discovery 2013

Methods

- Whole genome sequencing
- Libraries with reduced representation:
  - Transcriptomes
  - Exomes
  - Plastomes
  - Restriction-based methods
    - RadSeq, ddRadSeq
  - Target enrichment
    - HybSeq
  - Amplicons (PCR of multiple regions)

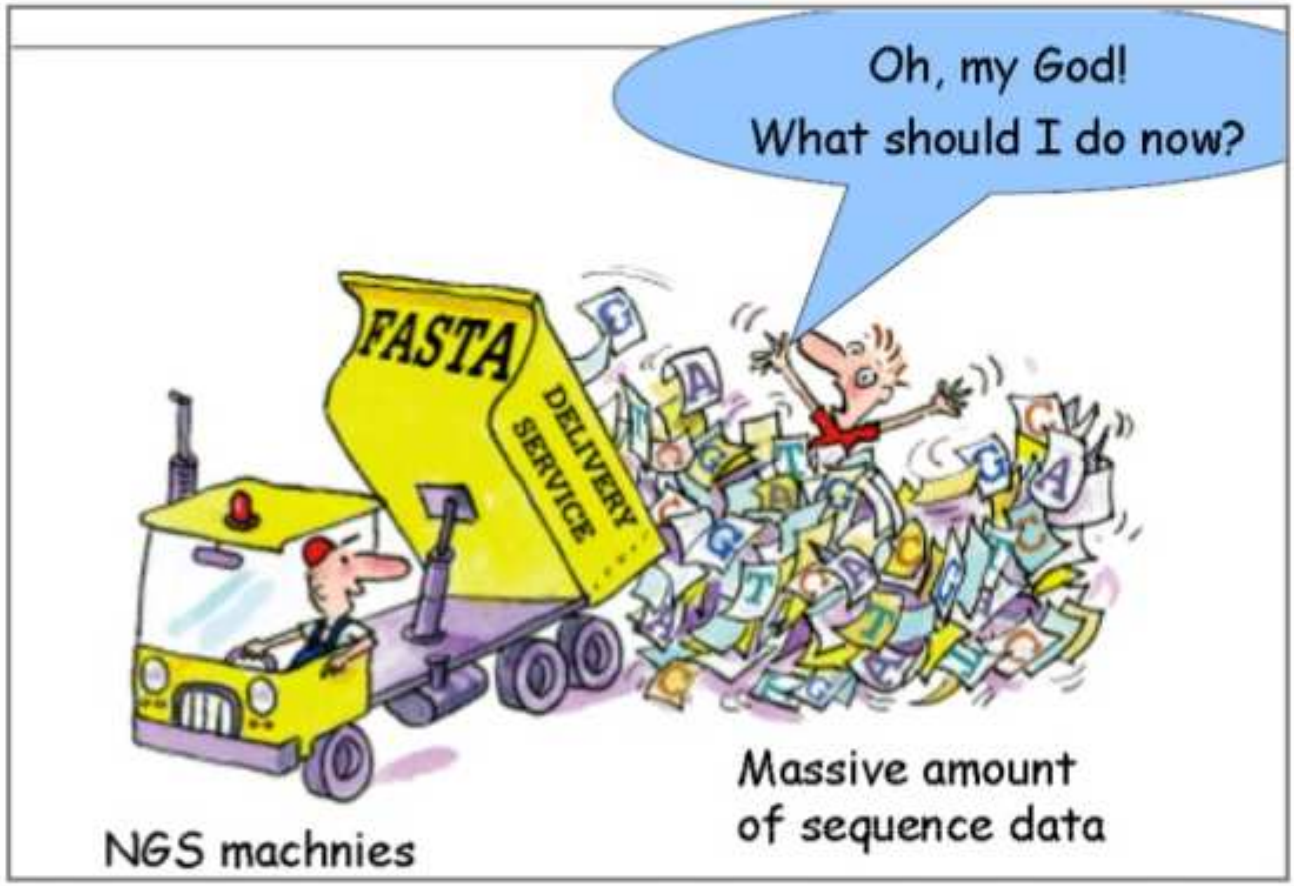


Bottle neck?

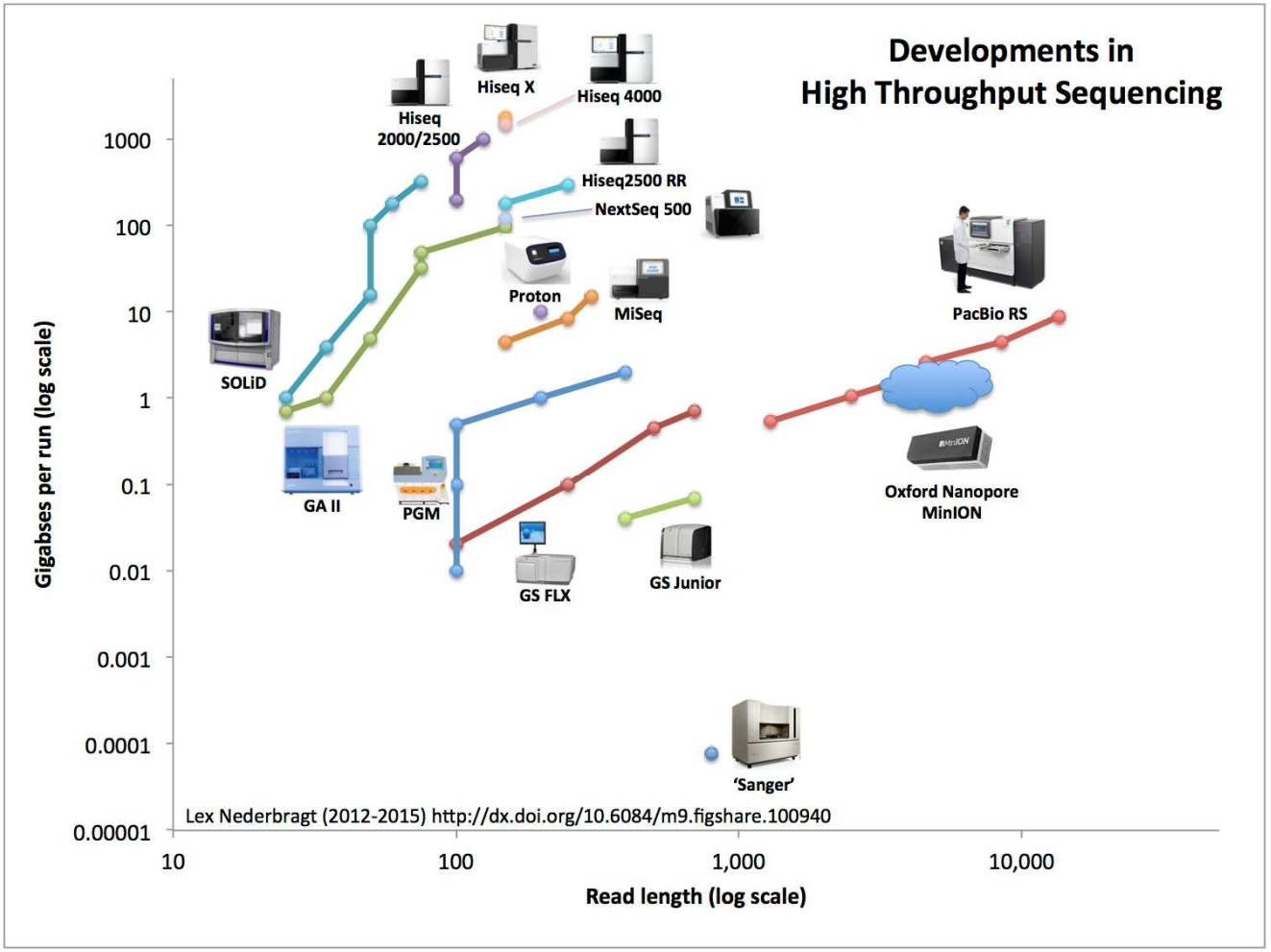
- Sanger
- NGS

- Sampling
- Wet-lab
- Sequencing
- Data procesing

Bottle neck


Oh, my God!
What should I do now?
FASTA
DELIVERY SERVICE
NGS machnies
Massive amount
of sequence data

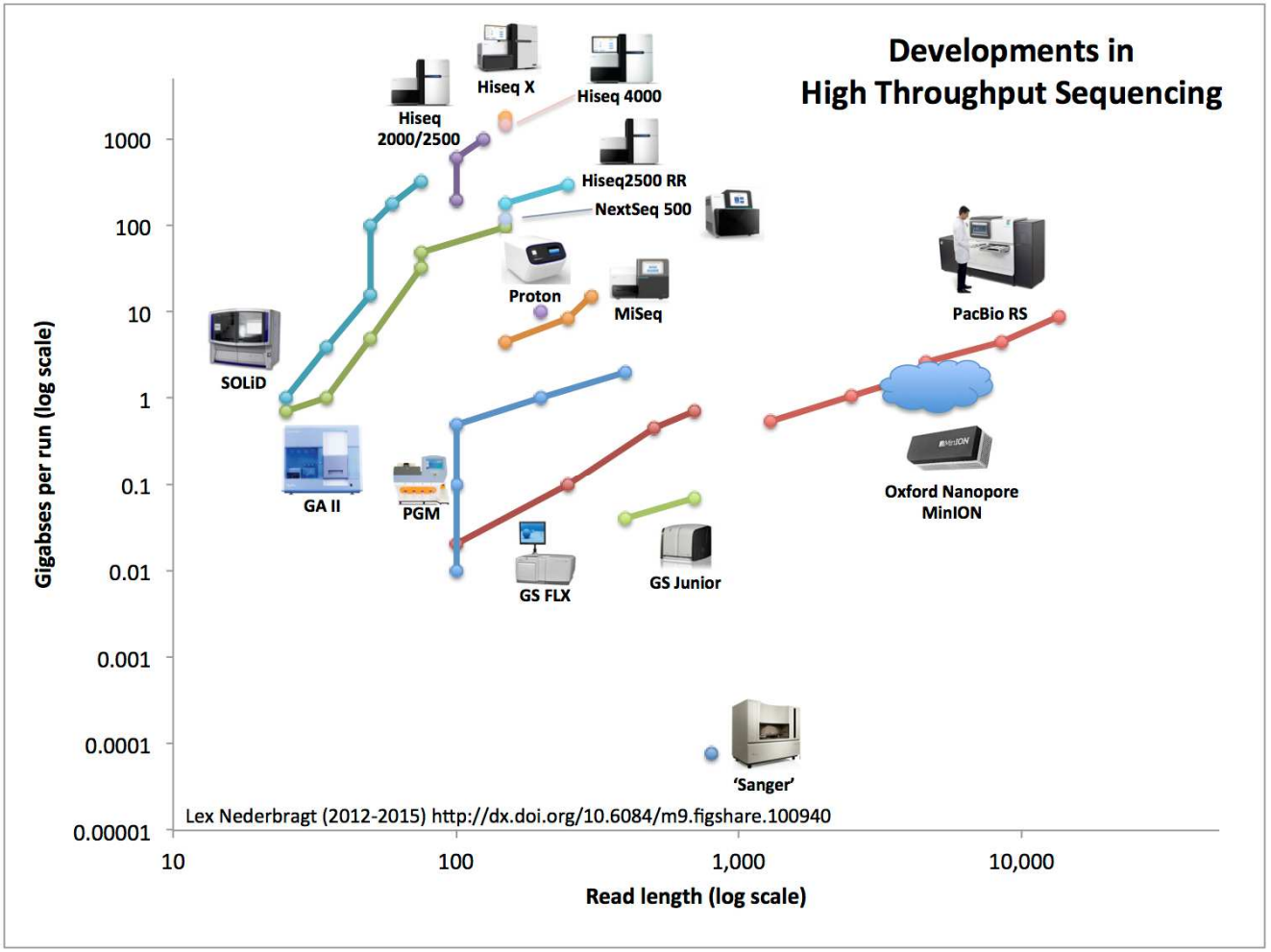
How to choose a method and procedure?



How to choose a method and procedure?

The question is #1

- related species, population, bulk segregant analyses, genomics, transcriptomics, phylogeography,...
- Equipment, protocol availability, data processing and data storage...



And what do you use?

- AFLP
- SSR
- cpDNA
- Single/low copy genes

- RNA

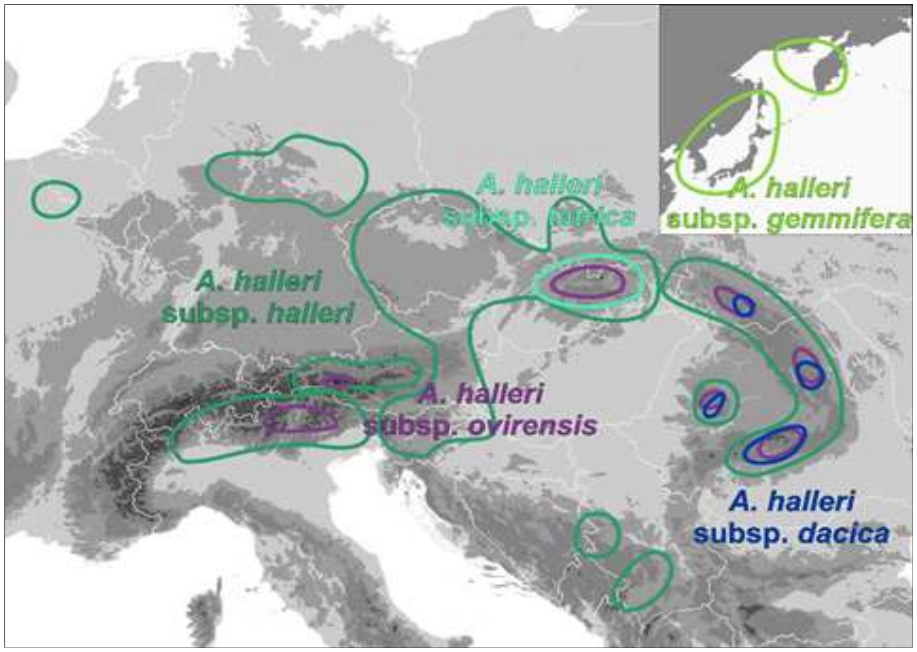
- RadSeq
- HybSeq
- WGS
- PoolSeq

…

And why?

genus *Arabidopsis*

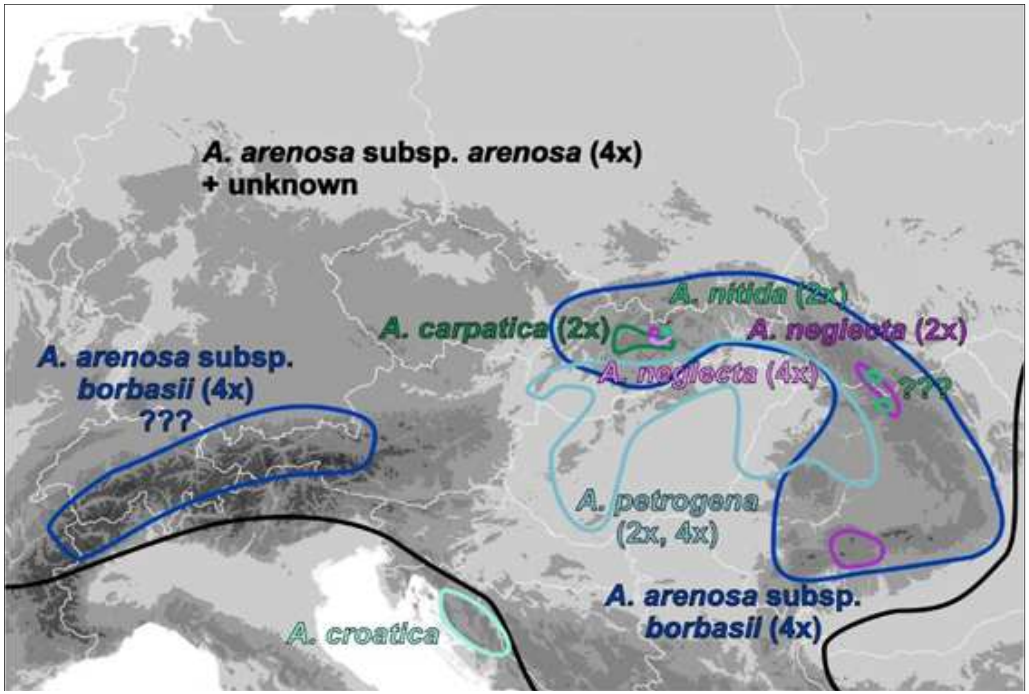
- ca 14 species worldwide – *A. thaliana* & “wild relatives“



Arabidopsis halleri
distribution of described and undescribed taxa

Based on Měsíček 1970, Kolník and Marhold 2006

- phytoremediation



Arabidopsis arenosa complex
distribution and ploidy levels of described and undescribed taxa

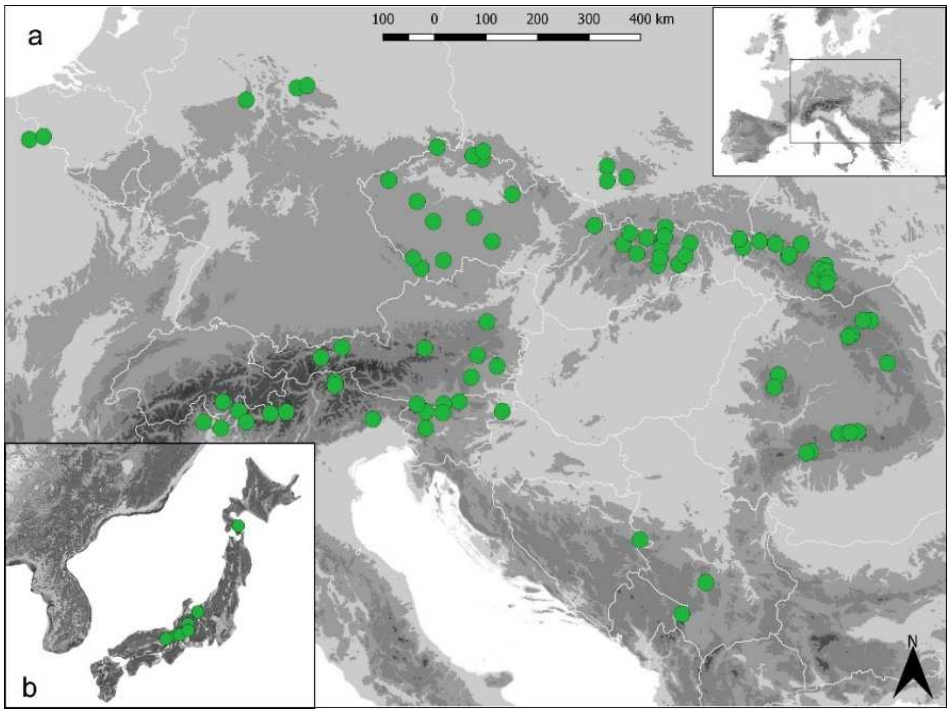
Based on Měsíček 1970

- polyploidy, parallel evolution

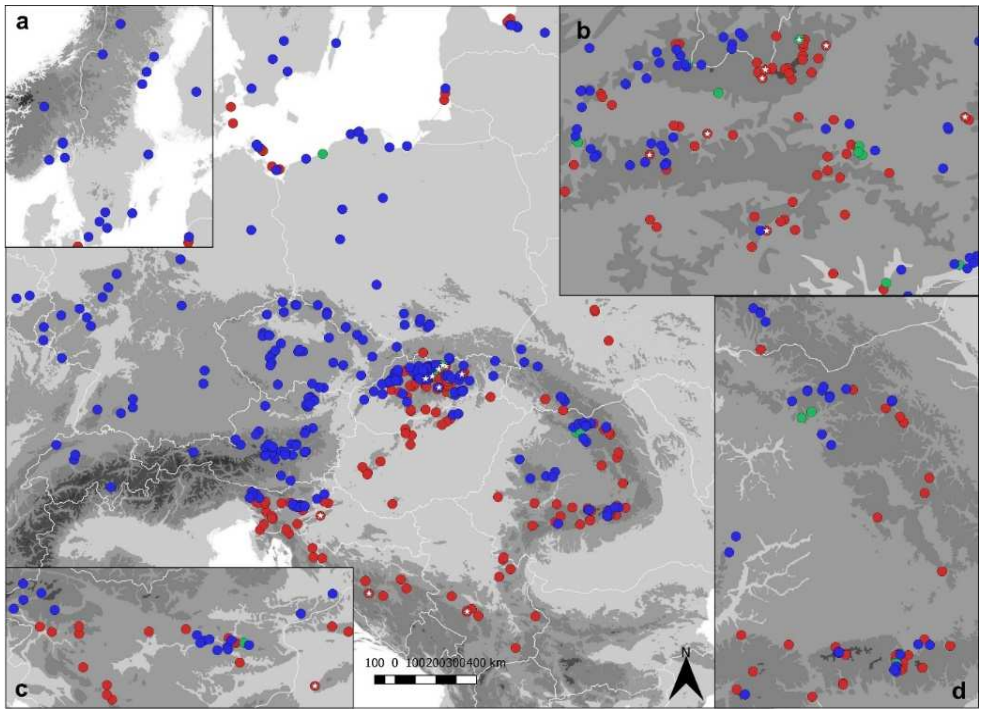
Materials and methods

- homogeneous sampling across the entire distributional range
- flow cytometry
- molecular methods: AFLP, SSR, cpDNA, single copy gene, dd RadSeq, WGS
- multivariate morphometrics

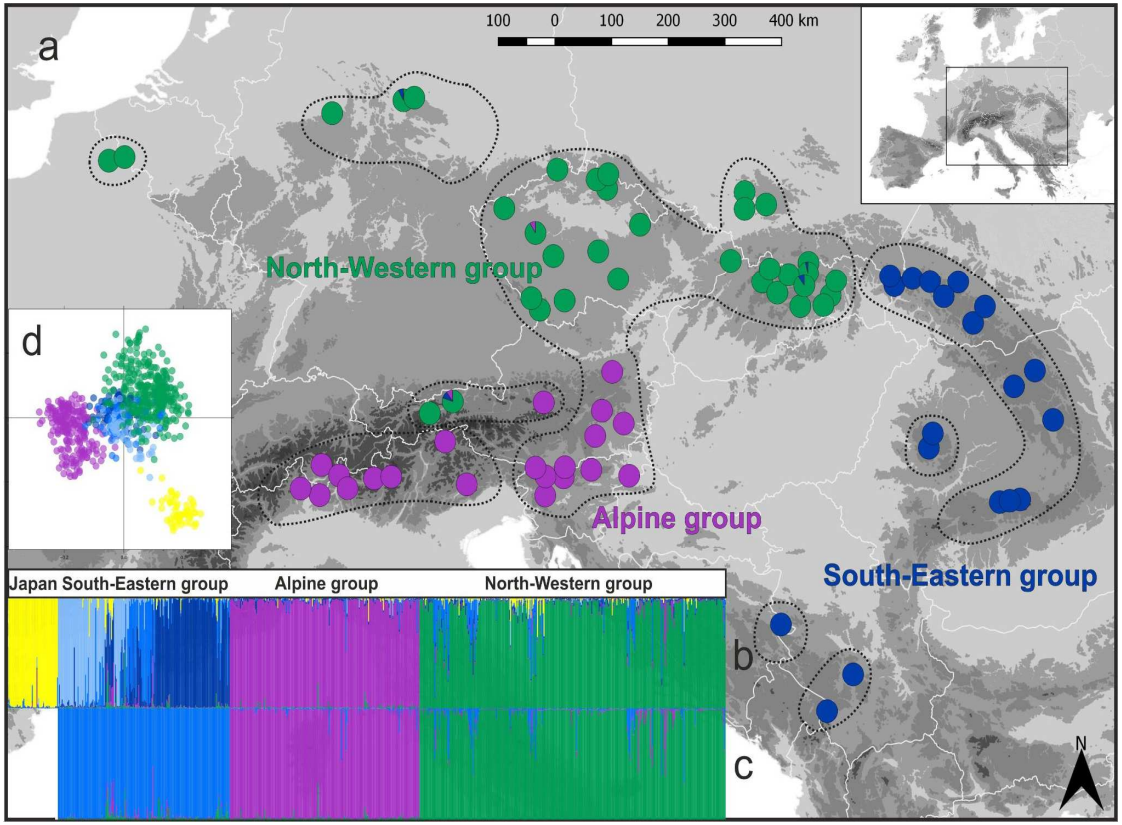
136 sampled populations – *A. halleri*



496 sampled populations – *A. arenosa*



Key results: *Arabidopsis halleri*



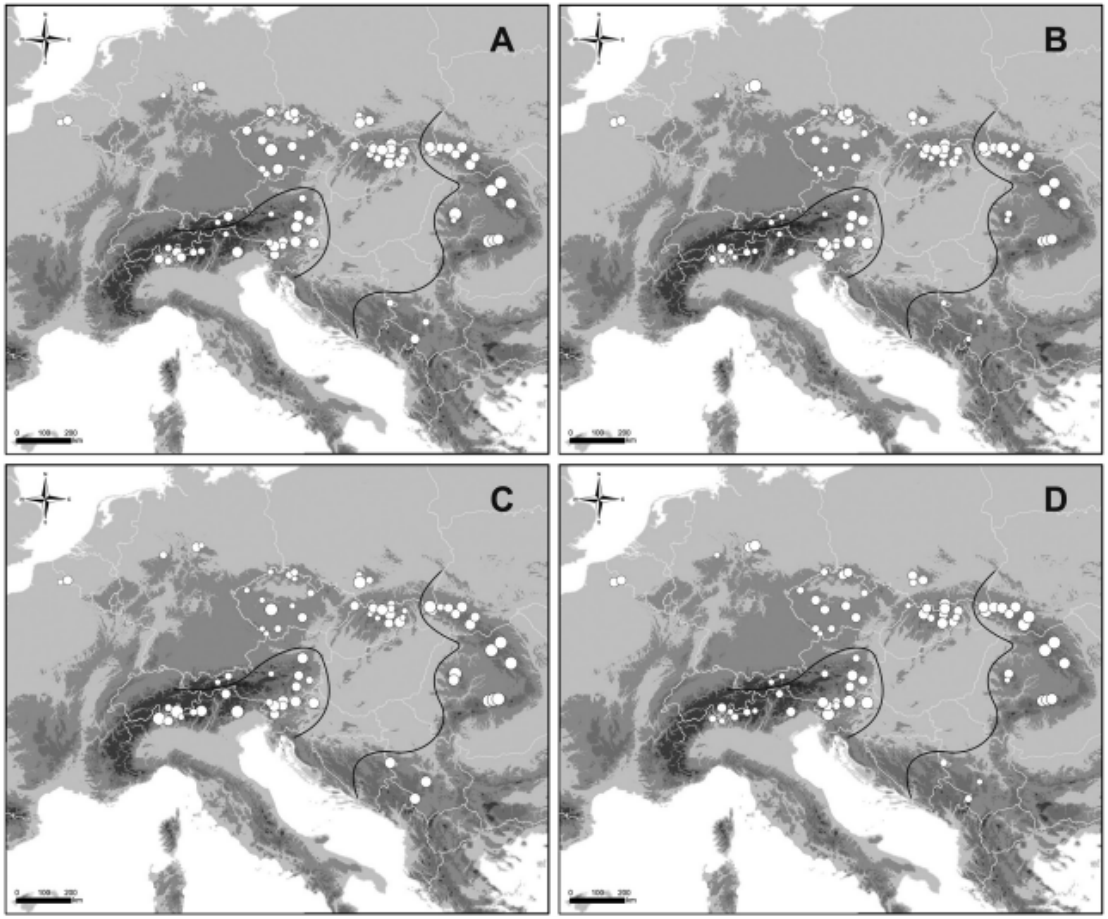
Genetic structure:

a geographic distribution (dotted line denotes borders of distribution)
b STRUCTURE of Eurasian dataset
c STRUCTURE of European dataset
d PCoA of Eurasian dataset

- northern hemisphere disjunction (Japanese populations)
- three main European lineages – Alpine, North-Western and South-Eastern

Šrámková-Fuxová, Záveská et al. 2017

Key results: *Arabidopsis halleri*



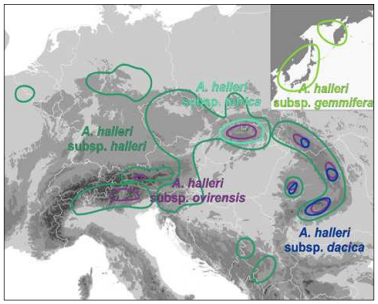
Population-level diversity:

a gene diversity AFLP
b expected heterozygosity SSR
c proportion of rare fragments AFLP
d allelic richness SSR

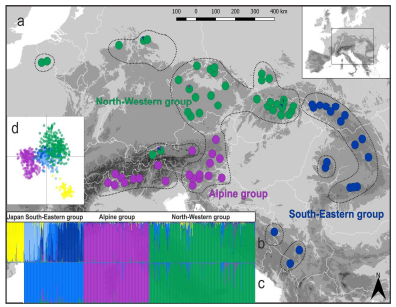
- the highest diversity and representation of rare alleles are in W and SE Carpathians

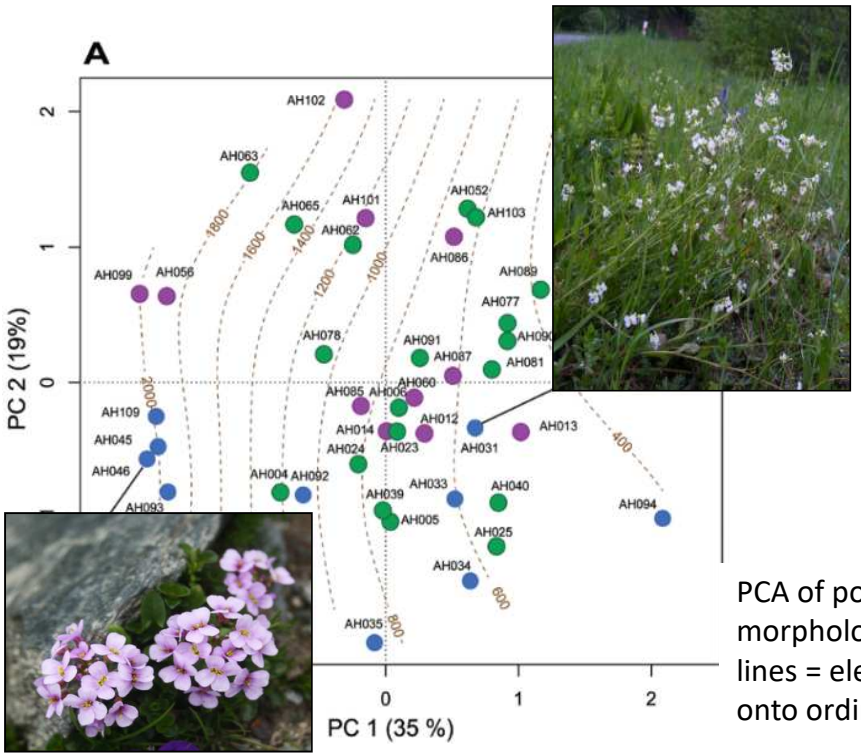
Šrámková-Fuxová, Záveská et al. 2017

Key results: *Arabidopsis halleri*



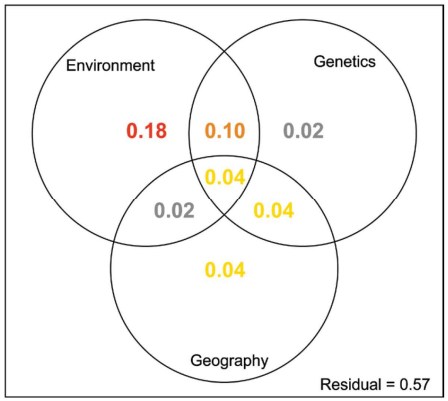
≠





PCA of population morphology (brown lines = elevation fitted onto ordination plot)

Šrámková-Fuxová, Záveská et al. 2017

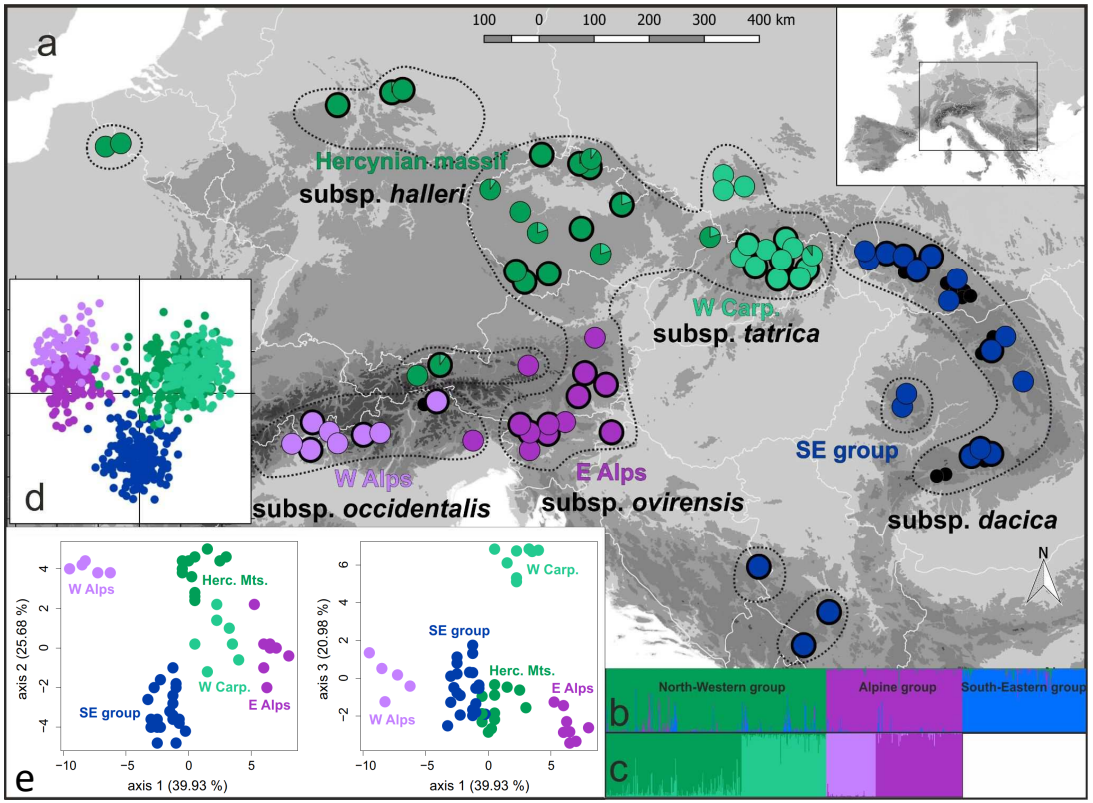


relative contributions of parameters to morphological variation (variation partitioning)

Genetic divergence is not a source of morphological variation…

Key results: *Arabidopsis halleri*

- detailed phylogeographic structure tested for morphological separation



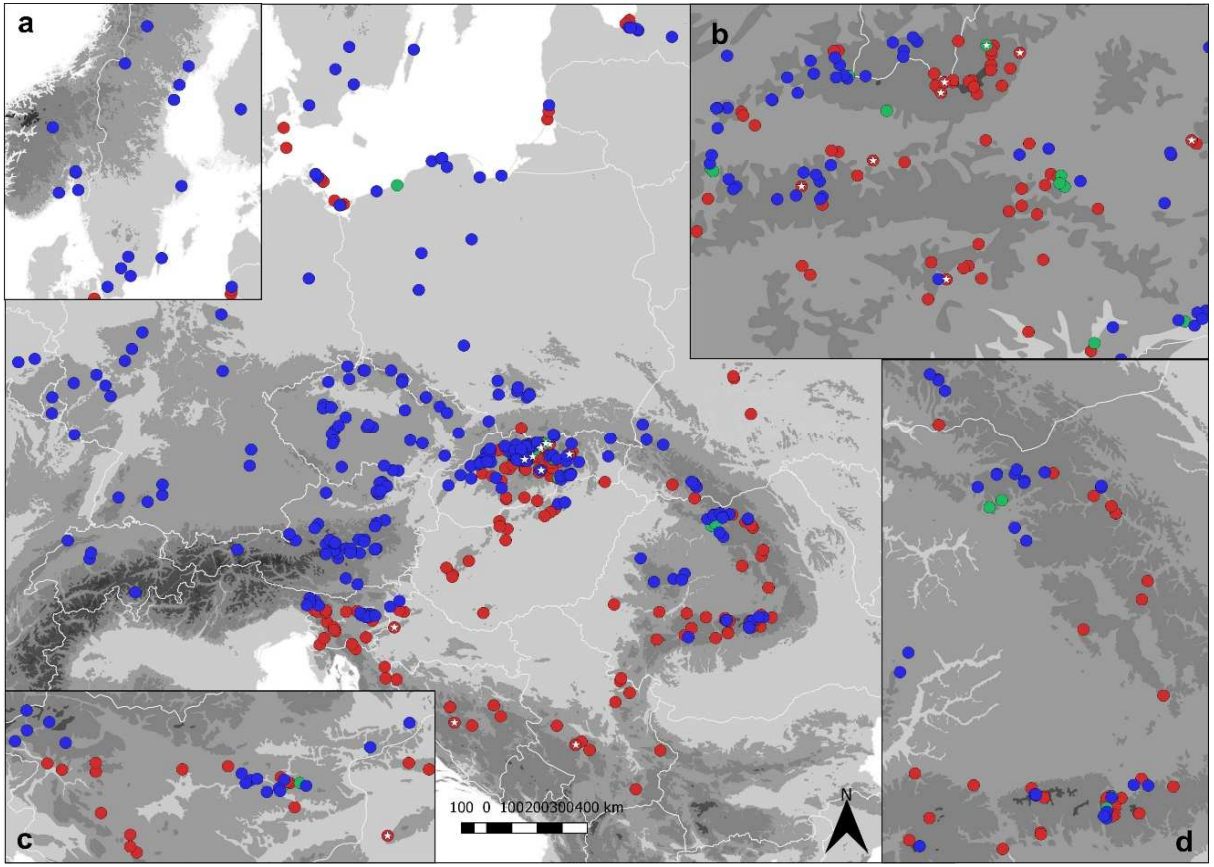
Taxonomic reassessment

a subgroups geographic distribution
b STRUCTURE EU dataset
c separate STRUCTURE for lineages
d PCoA of AFLP phenotypes
e morphological separation of 5 subgroups (canonical discriminant analyses)

- five subgroups morphologically separated → taxonomic re-evaluation

Šrámková et al. 2019

Key results: *Arabidopsis arenosa* complex



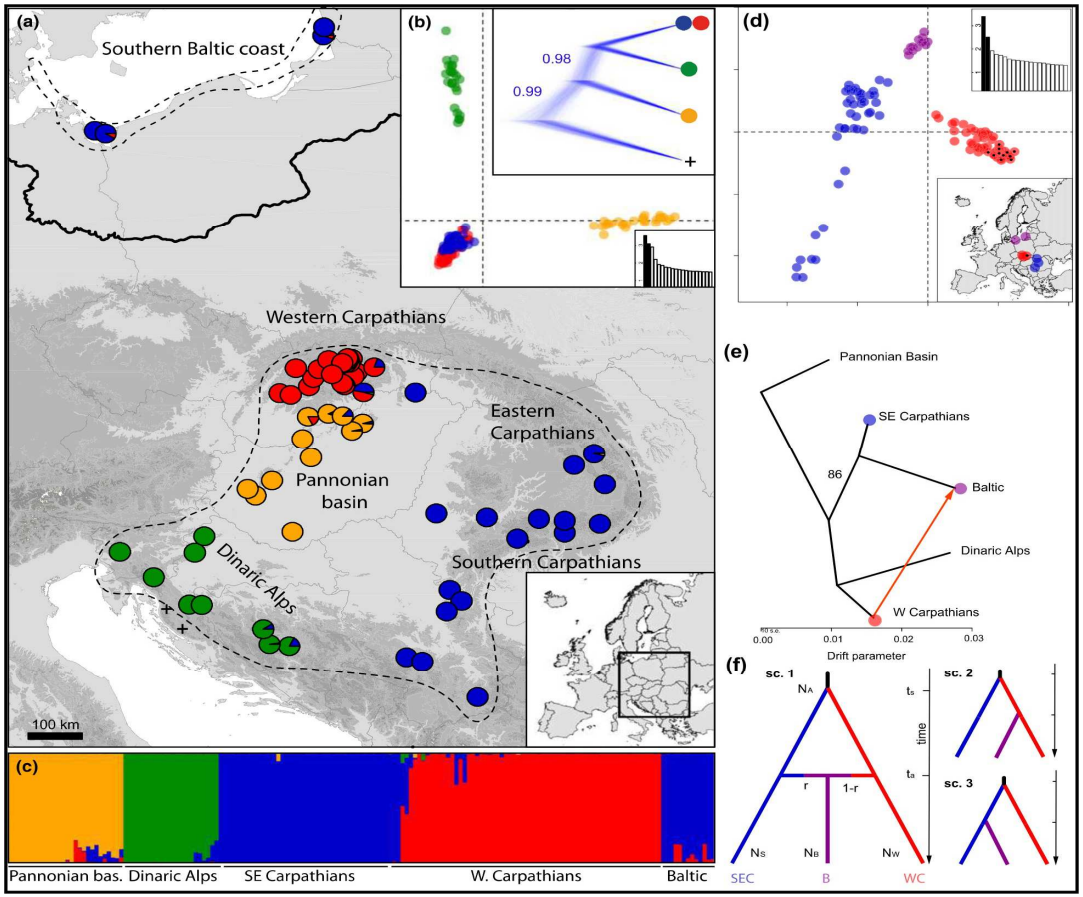
Distribution and ploidy level

red – diploid
blue – tetraploid
green – mixedploid
asterisk – triploid inds in population
a Scandinavia
b W Carp cont. zone
c Slovenia cont. zone
d SE cont. zone

- 3 different ploidy levels – 2x, 3x 4x
- distinct 2x populations on Baltic Sea coast
- 3 contact zones

Kolář et al. 2016 BiolJLinnSoc + unpub.

Key results: *Arabidopsis arenosa* 2x



Genetic structure:

a geographic distribution (dotted line denotes borders of 2x distribution)
b PCoA and species tree (rooted with *A. croatica*)
c STRUCTURE clustering

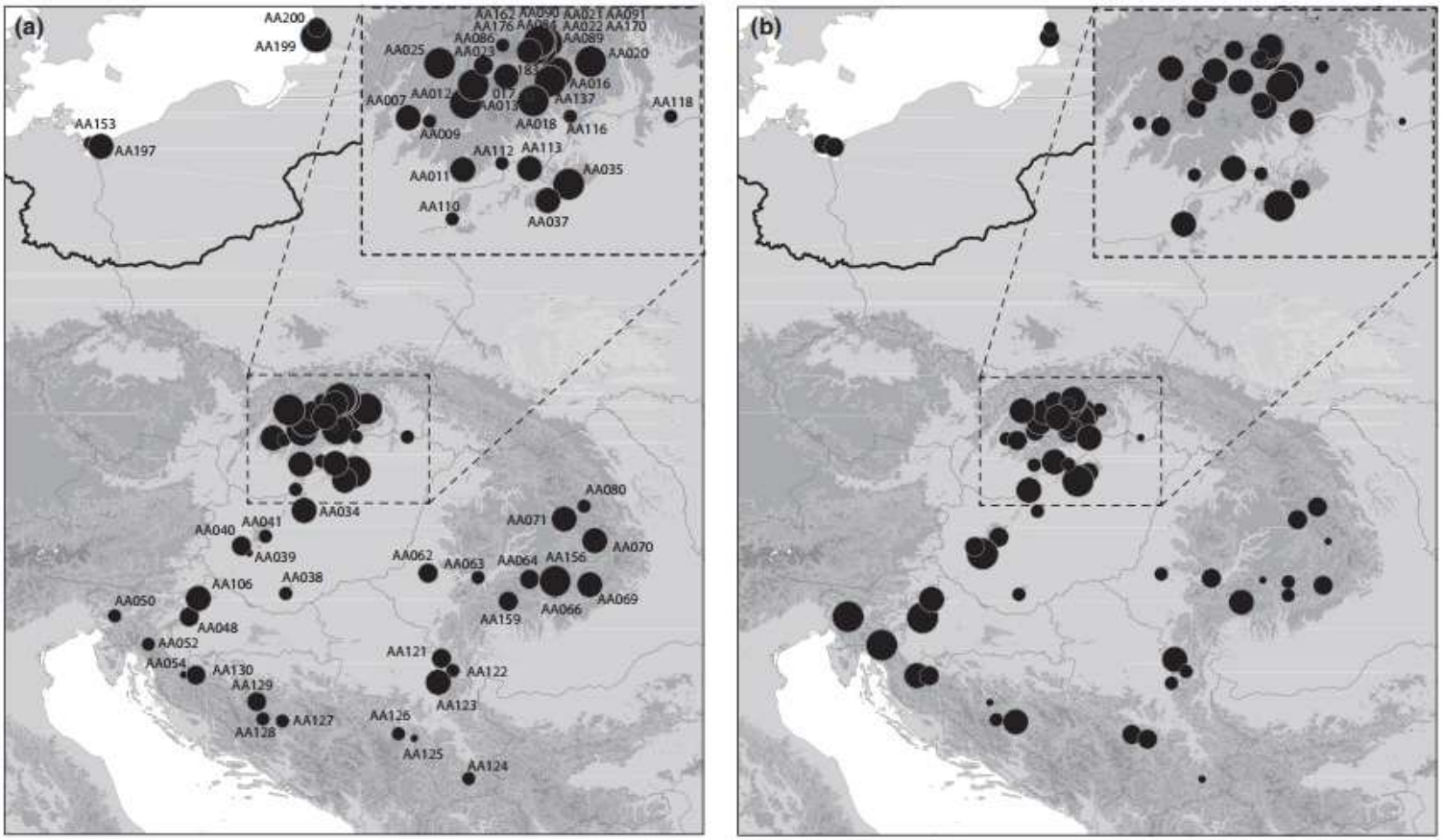
Reconstruction of Baltic-Carpathian relationships:

d PCoA of Baltic and Carpathian individuals
e graph of allele frequency covariance with admixture (Treemix)
f ABC modelling of Baltic origin

Kolář, Fuxová, Záveská, et al. 2016 MolEcol

- 4 divergent European lineages – Dinaric, Pannonian, W Carpathian, SE Carpathian + spatially isolated Baltics
- Baltics originated from SE and W Carpathian lineages

Key results: *Arabidopsis arenosa* 2x

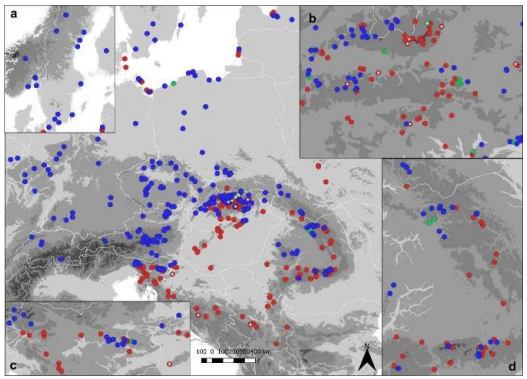


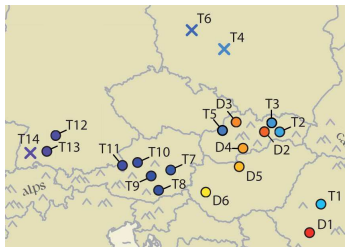
Population-level diversity (SSR): **a** expected heterozygosity **b** proportion of rare alleles (DW index)

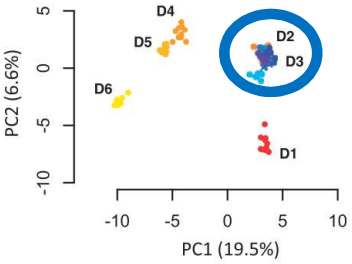
- the highest diversity and representation of rare alleles are in W Carpathians

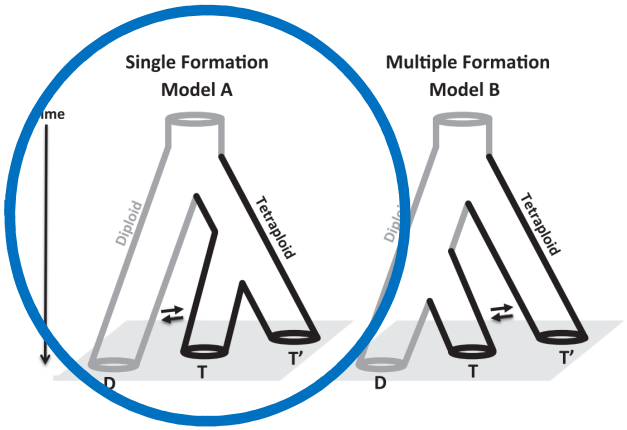
Kolář, Fuxová, Záveská, et al. 2016 MolEcol

Arabidopsis arenosa 4x

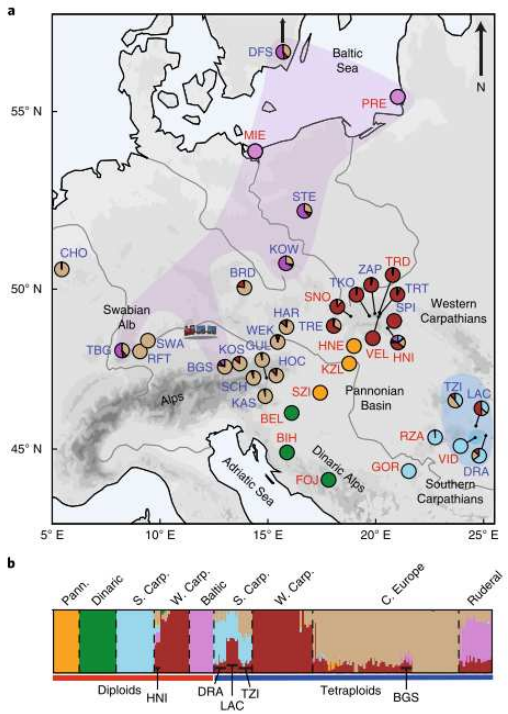


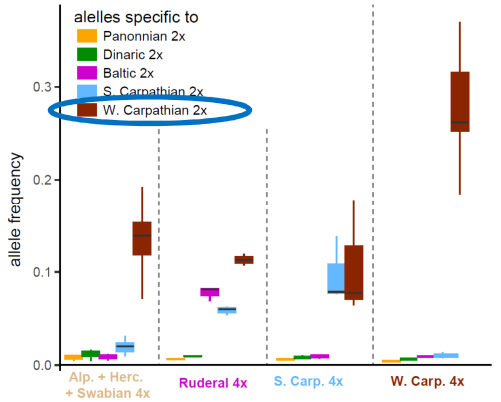






Arnold et al. 2015 Mol Biol Evol

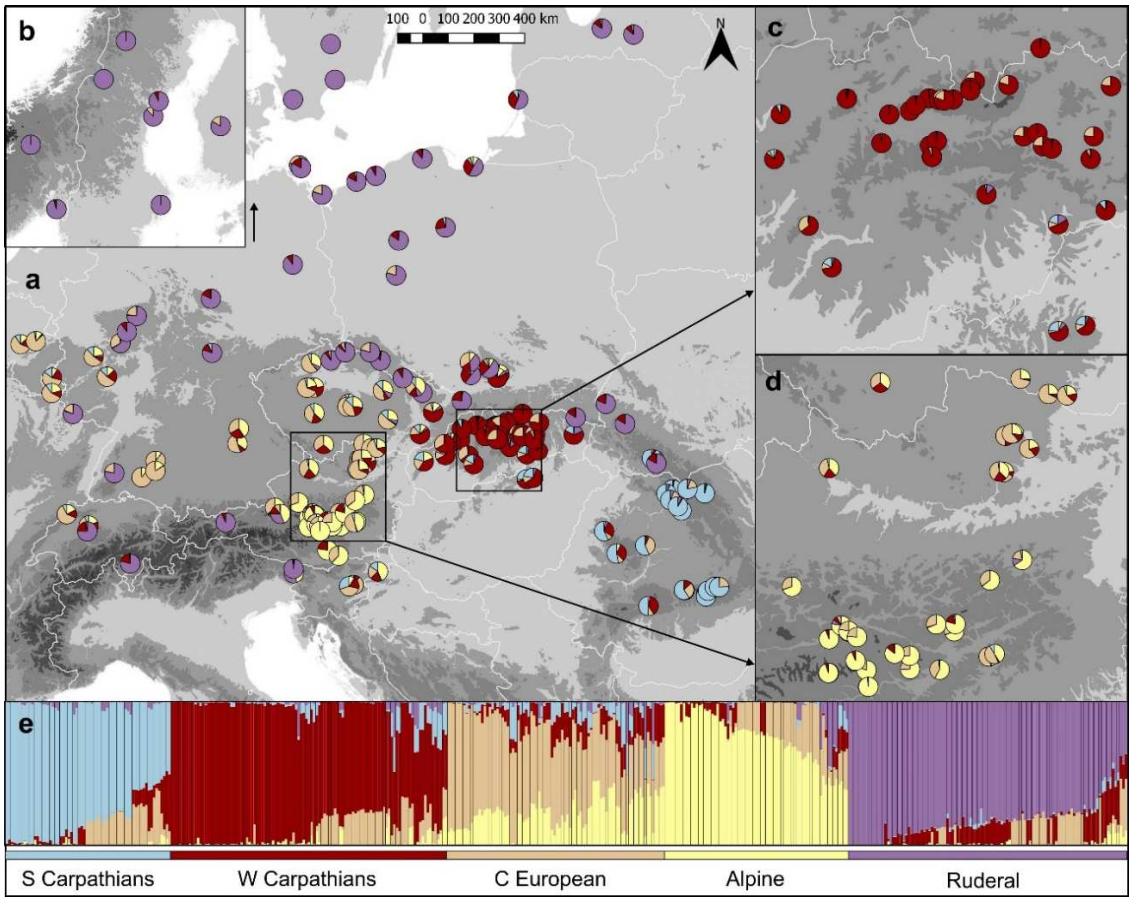




- single geographic origin of autotetraploid lineages

Monnahan et al. 2018 Nature Ecol Evo

Key results: *Arabidopsis arenosa* 4x



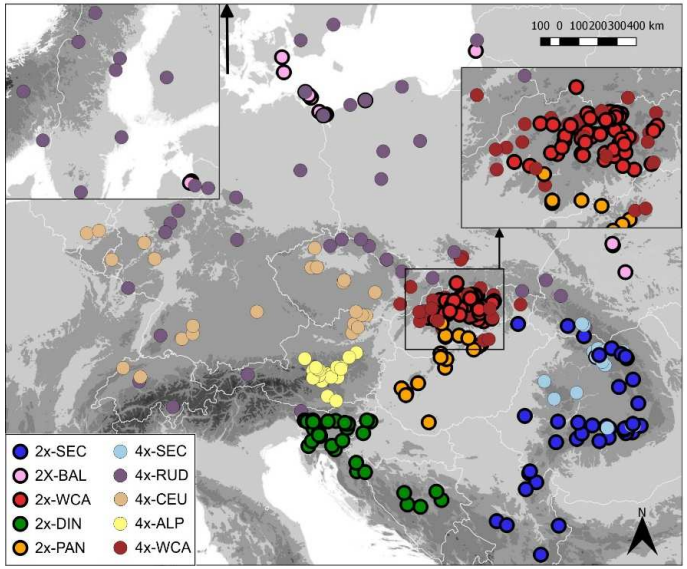
Geographic distribution of genetic structure of 4x populations

a sampled populations
b Scandinavia
c W Carpathians
d Austrian zone
e individual STRUCTURE assignment

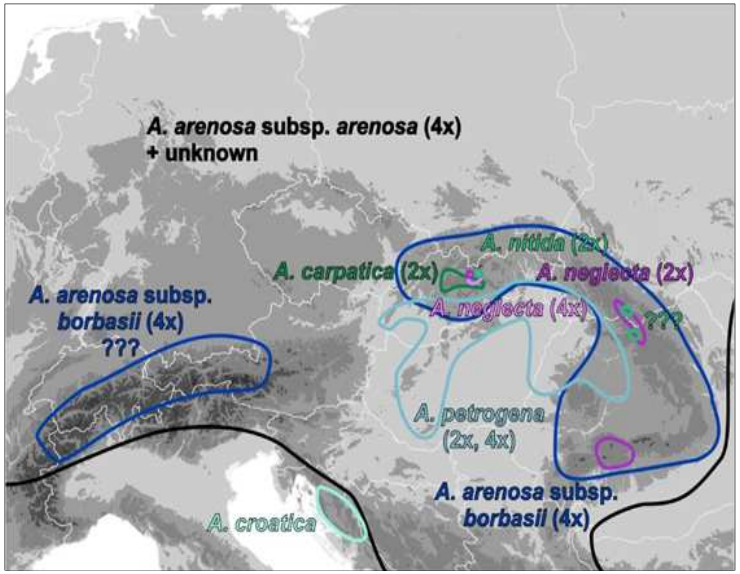
- five genetic clusters
- Ruderal lineage – anthropogenic niches, colonized higher latitudes

Padilla-García, Šrámková et al. Submitted J Biogeo

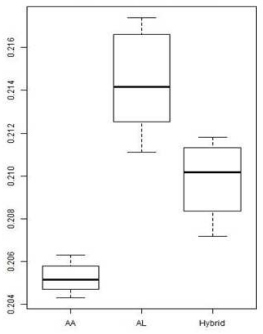
Future prospects

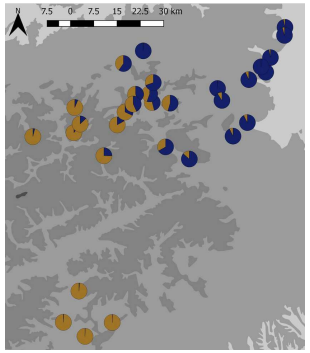


≠

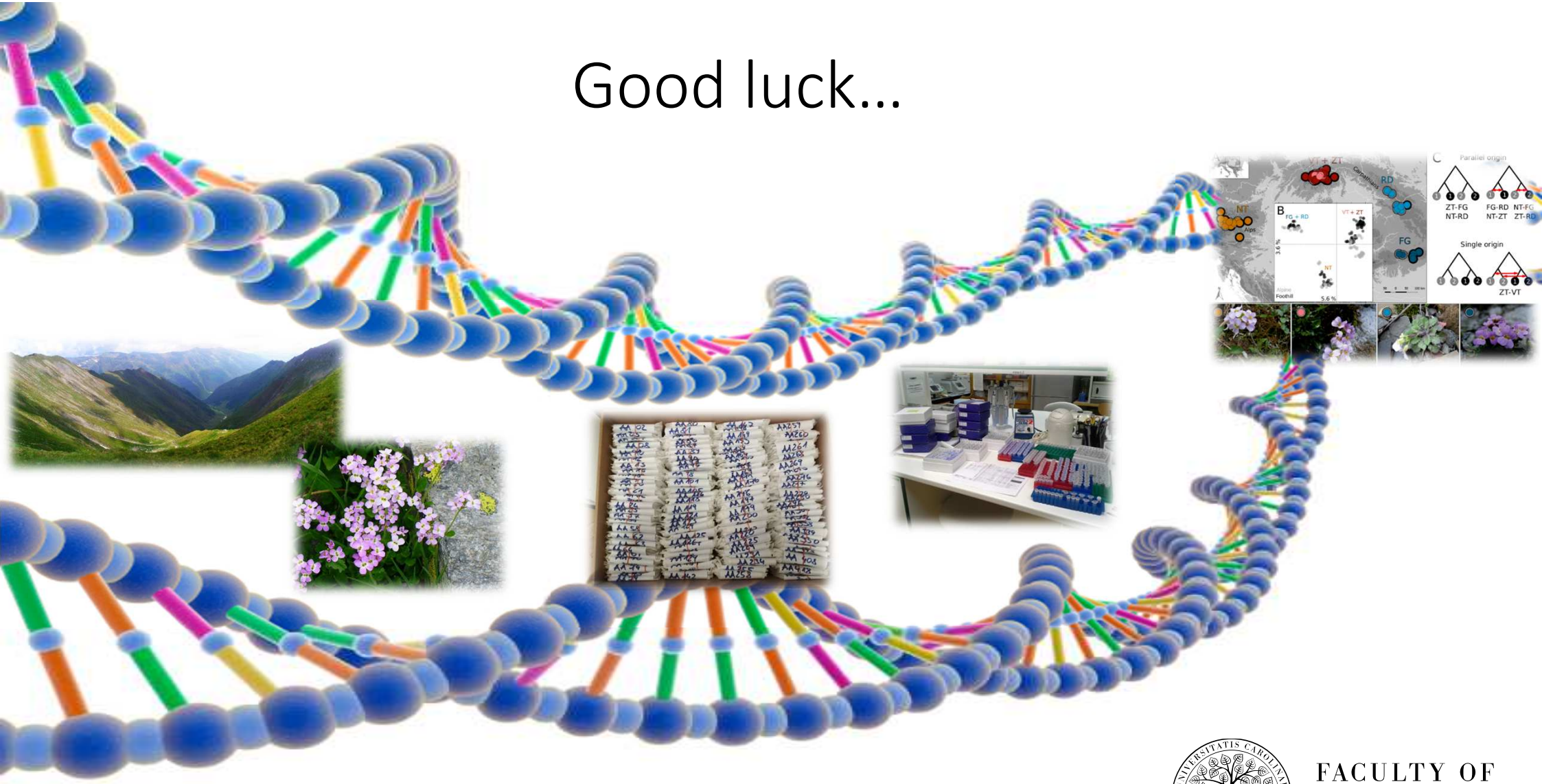


- taxonomical reassessment of *A. arenosa*
  - analyses of the whole dataset
  - morphometrics
- hybridization *A. arenosa* – *A. lyrata*
- …

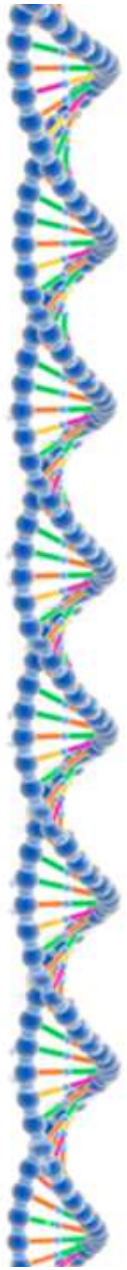




Good luck…



FACULTY OF SCIENCE
Charles University



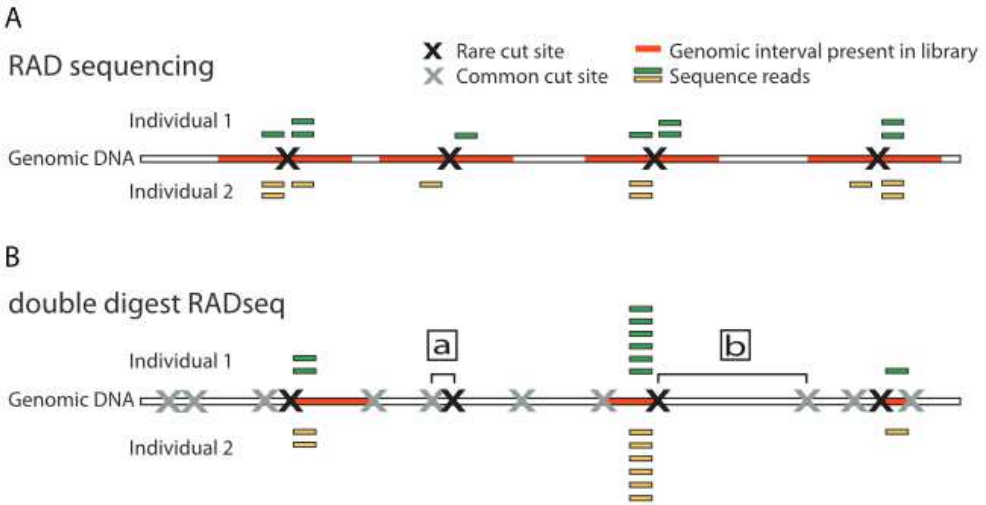
A few methods of library preparation...

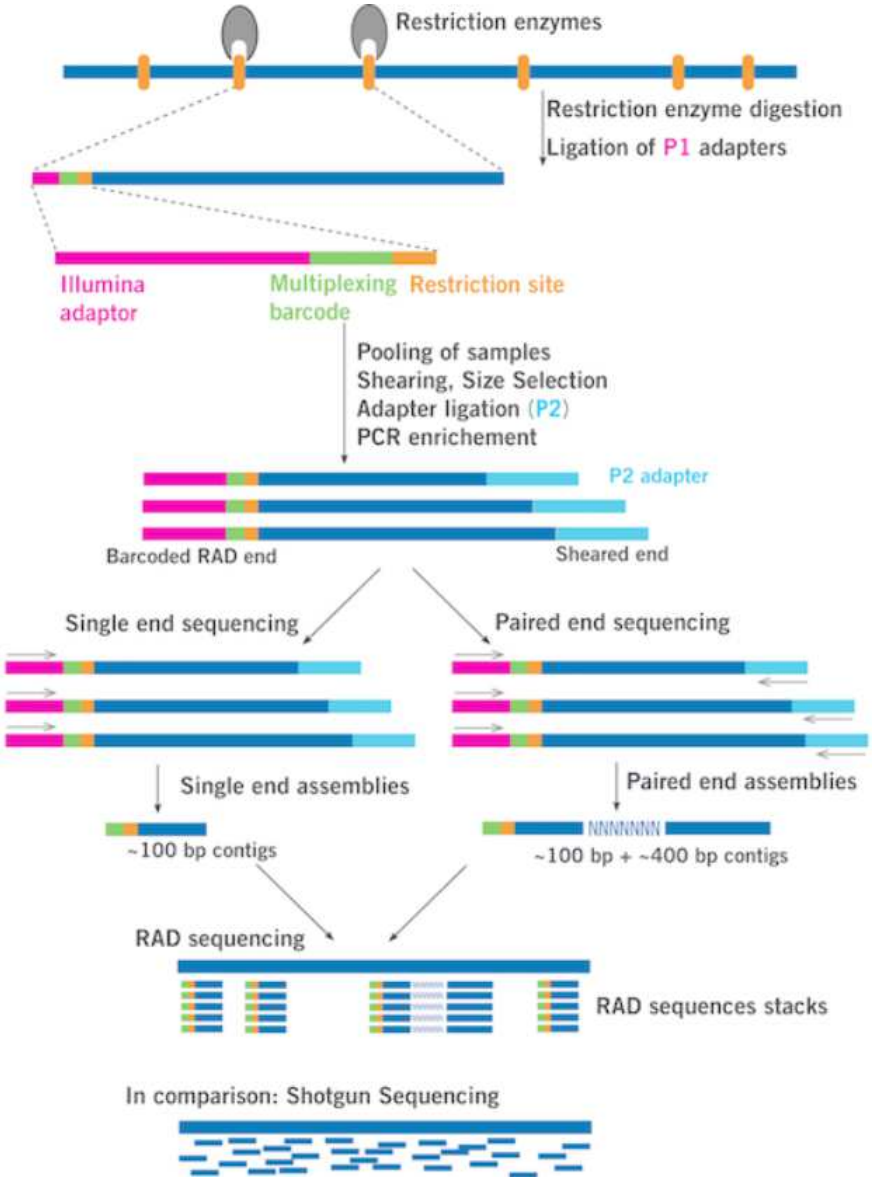
- RadSeq
- HybSeq
- WGS
- RNASeq
- HyRAD
- MigSeq
- SSR-GBS

RadSeq

Restriction (enzyme, sonication)

Radseq, ddRadseq





HybSeq

Custom-made proby (baits)

Genomická knihovna

Hybridizace

Sekvenace obohacených fragmentů (+ neobohacených částí genomu)

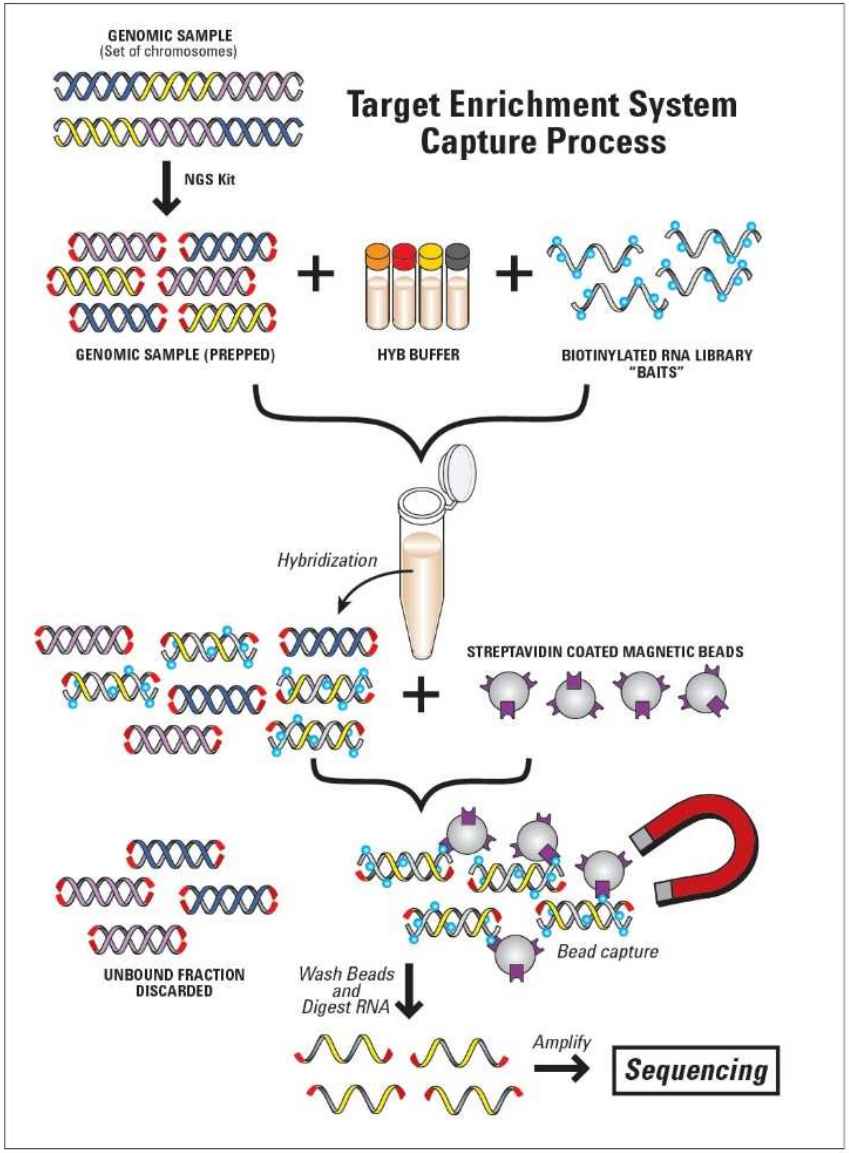
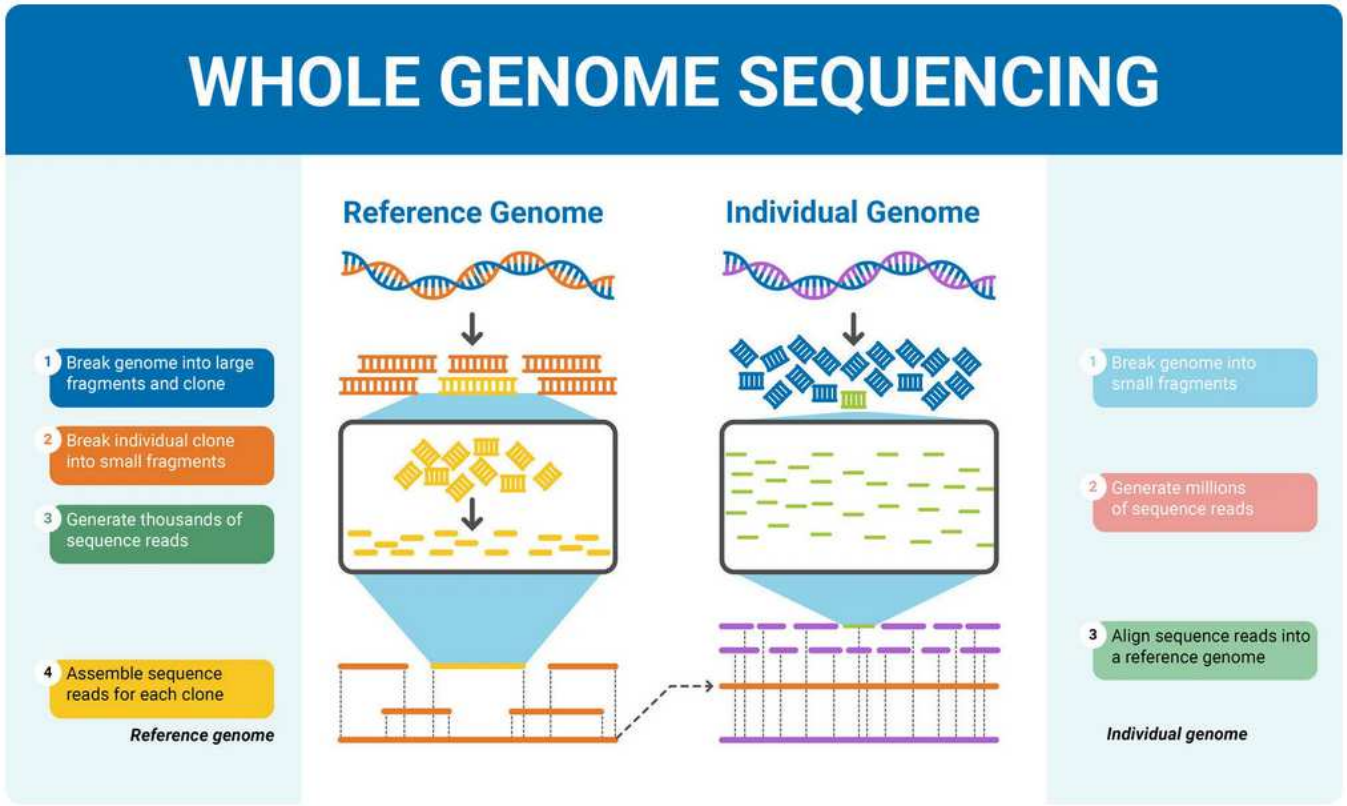


Figure 1 Target enrichment system workflow

https://www.ddw-online.com/bringing-cost-and-process-efficiency-to-next-generation-sequencing-731-200908/

Whole Genome Sequencing

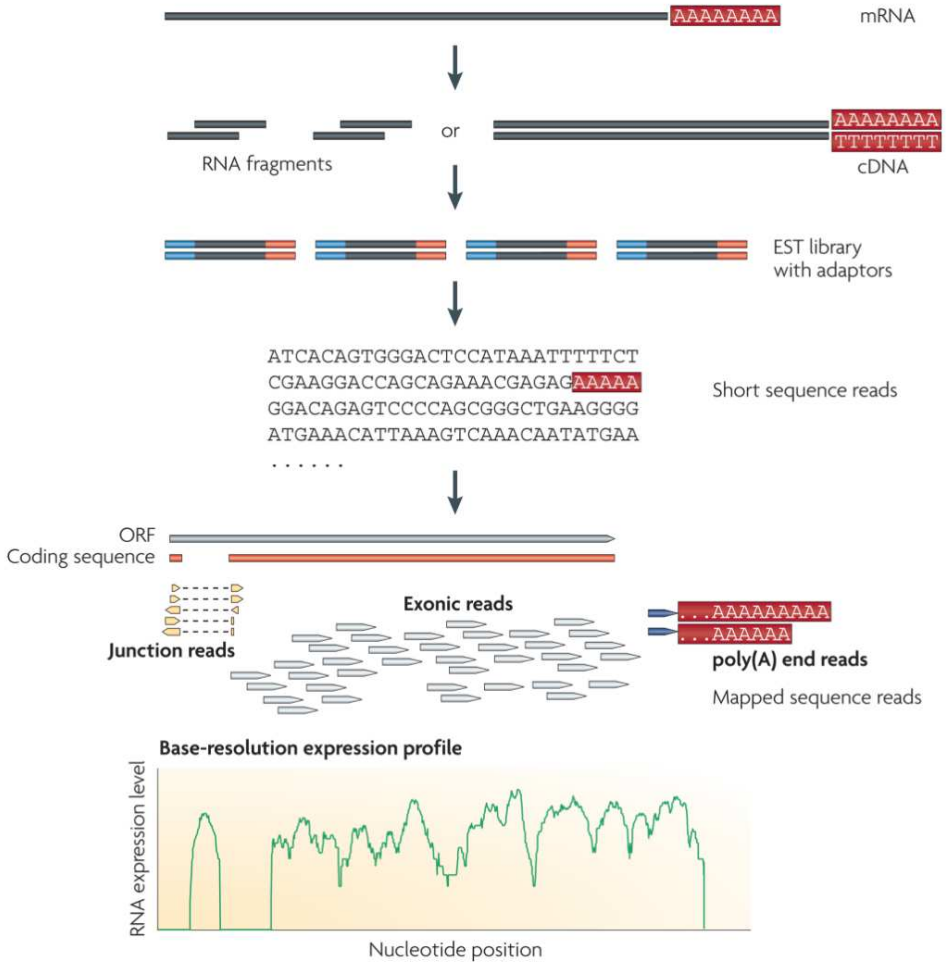
- Metodou služby (kity…)
- Home-made protokoly
  - LITE (Rowan et al. Genetics 2019) - transpozom



https://sequencing.com/education-center/whole-genome-sequencing

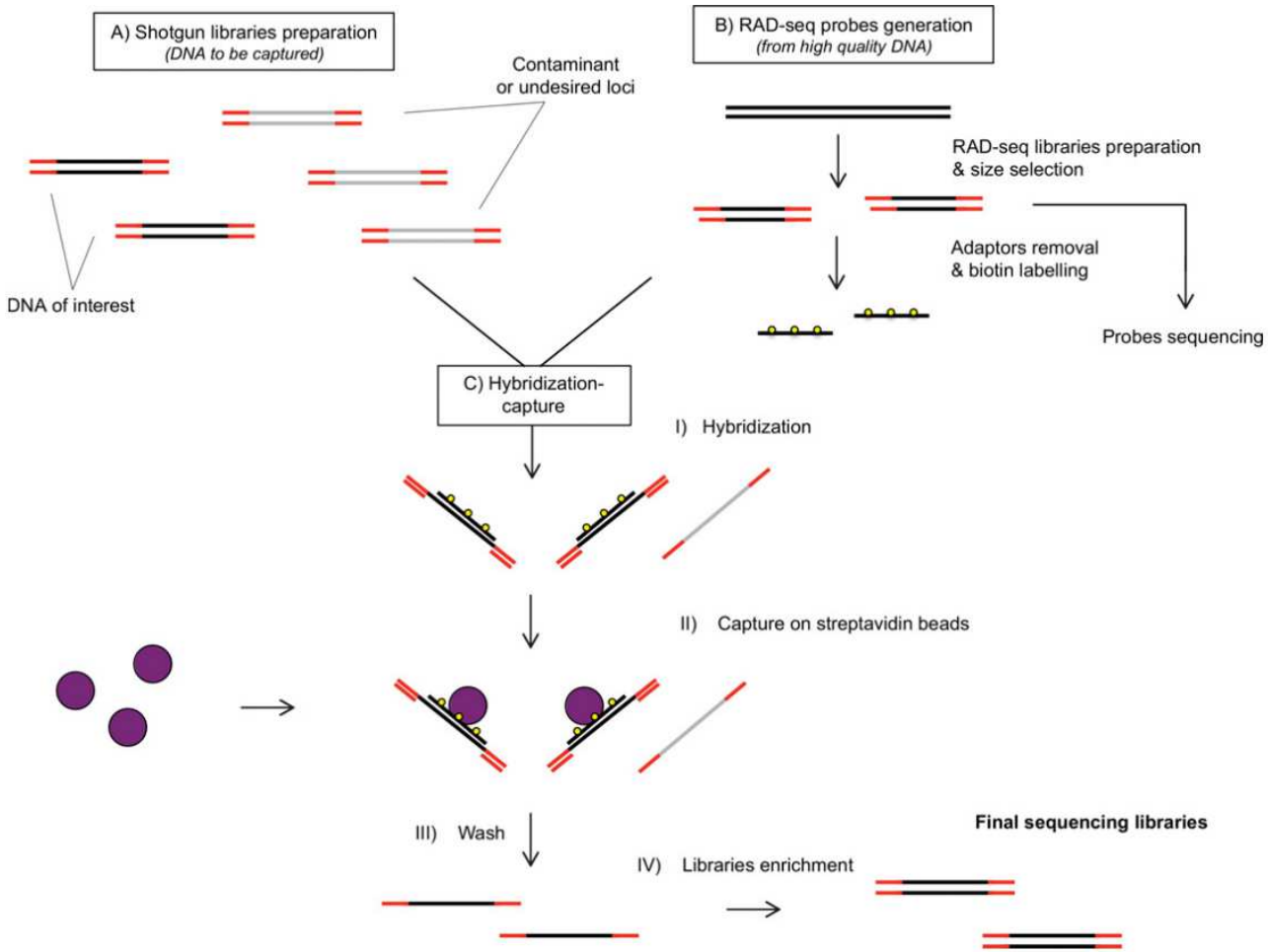
RNA Seq

Izolace RNA



HyRad

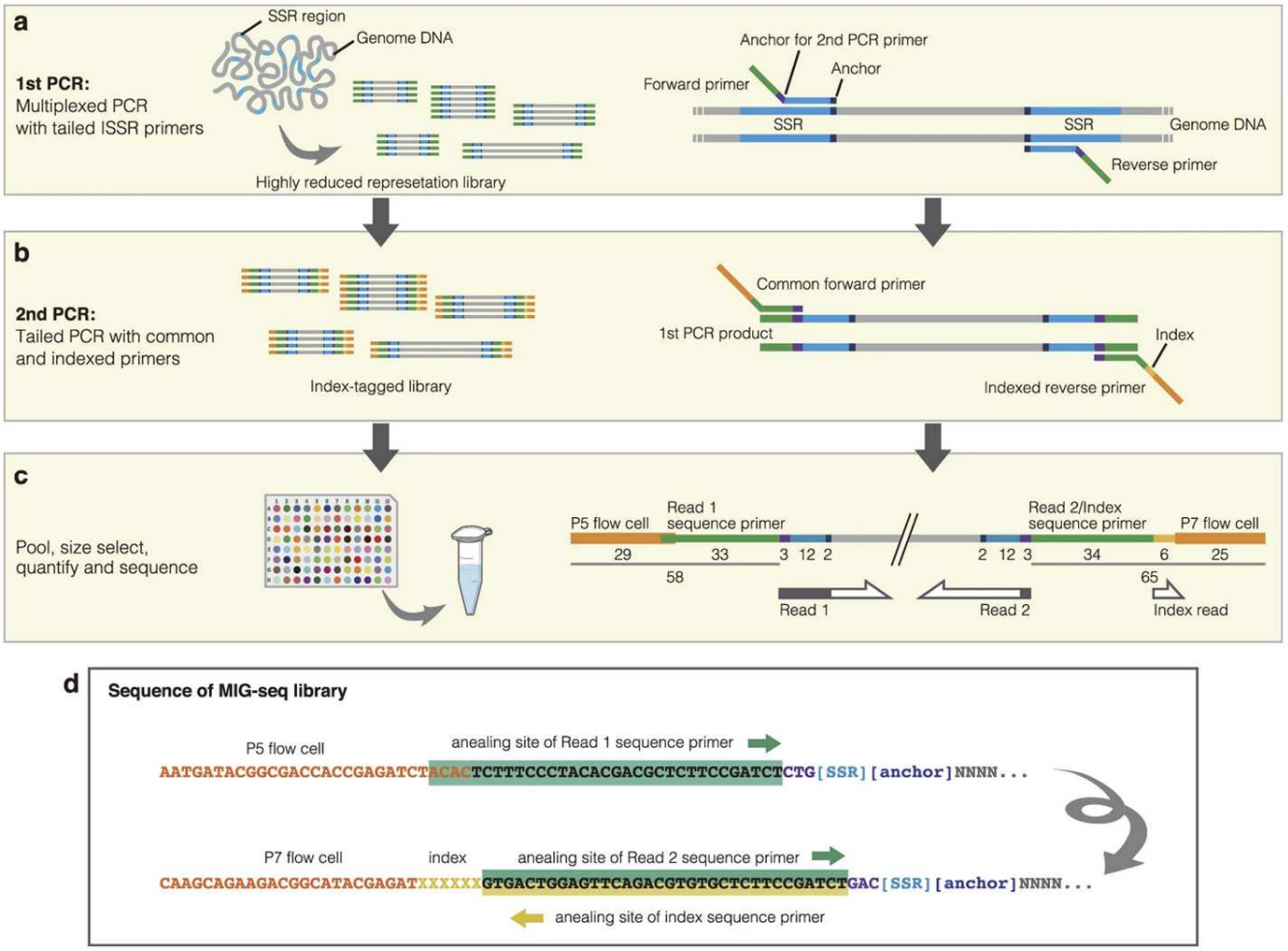
Kombinace RadSeq + target enrichment



Suchan et al. PLoS ONE 2016

MigSeq

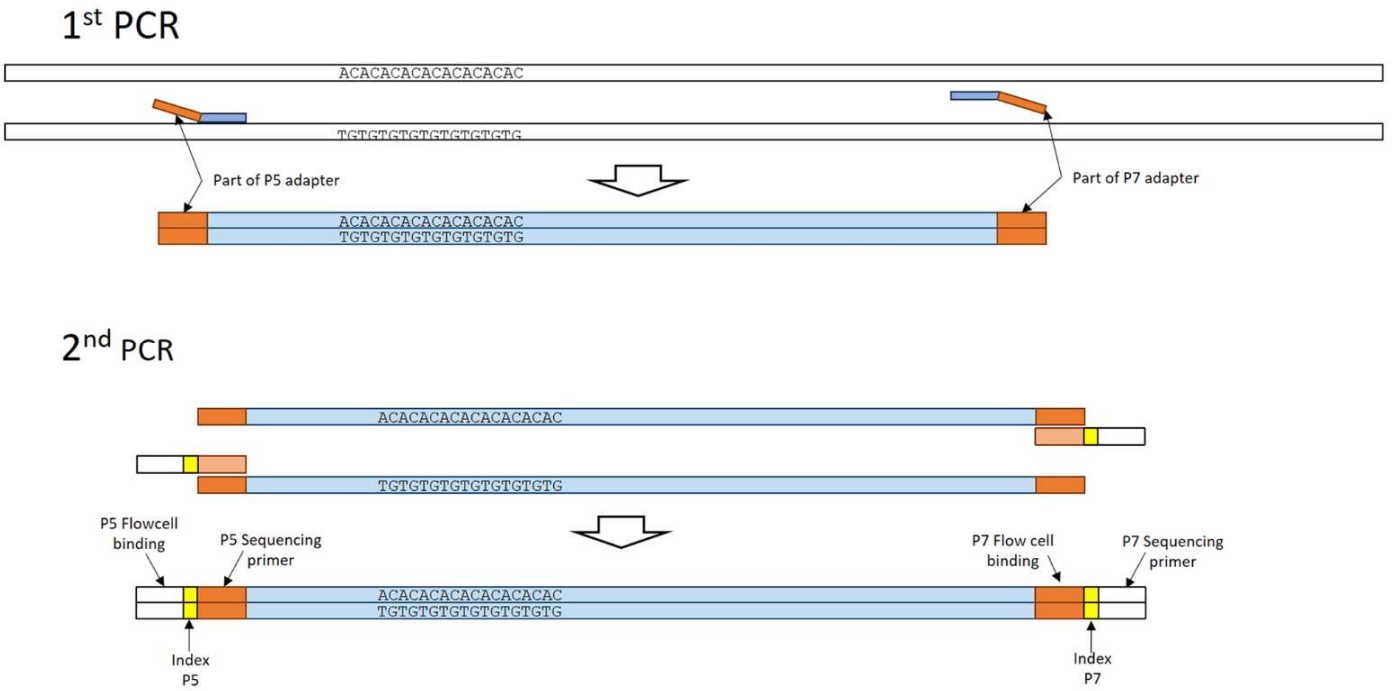
Restrikce (enzym - Inter-simple sequence repeats)



Suyama & Matsuki Scientific Reports 2015

SSR-GBS

4-primerová PCR



Tibihika et al. Conservation Genetics 2019