Molecular markers in plant systematics and population biology

7. DNA sequencing (cpDNA)

Tomáš Fér

tomas.fer@natur.cuni.cz

DNA sequencing

- detection the order of nucleotides in a DNA strand

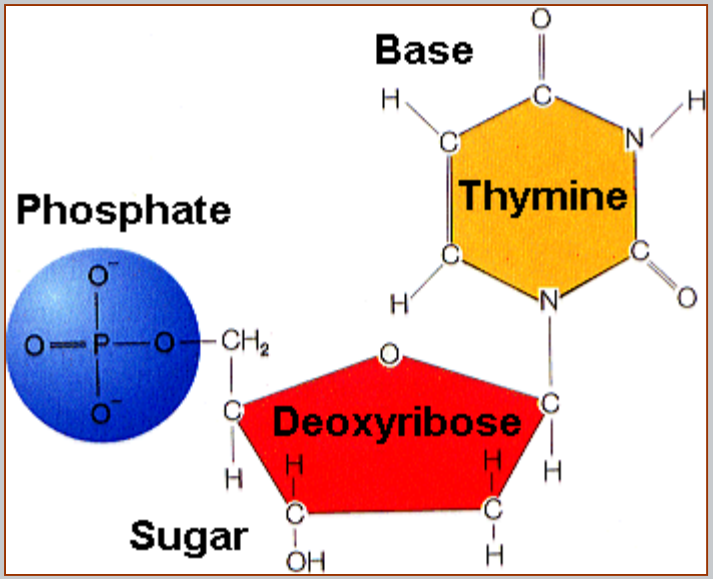
  …ATATATAGGCAAGGAATCTCTATTATTAAATCATT…

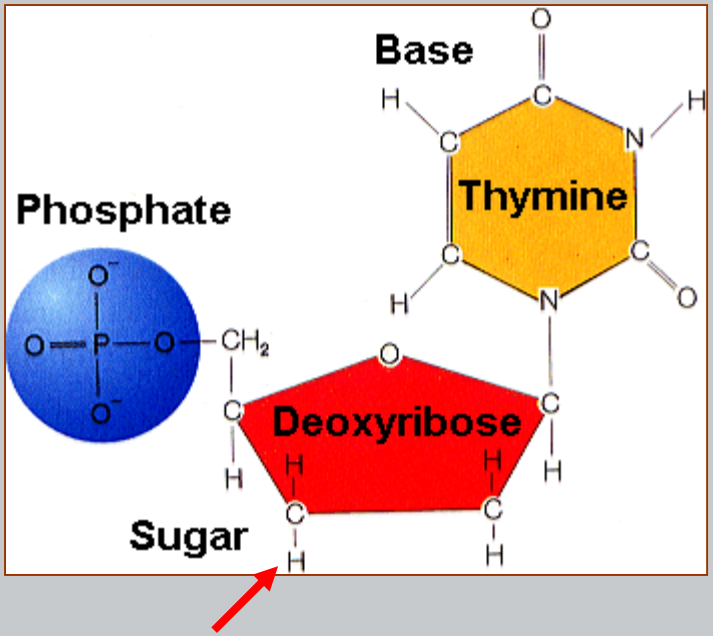
- use the information to model evolutionary and population genetic processes
- make hypothesis about similarity and relationships among taxa

Sequencing principle

- PCR with a primer pair
  - amplification of the target region
- cycle sequencing (dideoxy, Sanger)
  - use of one primer only
  - dNTP as well as ddNTP are present in the mixture
  - produce fragments differing exactly by one base
- electrophoretic separation of fragments in the gel
  - automated sequencer

2´, 3´- dideoxy NTPs





dTTP

ddTTP

3´-CTGGACTGCA-5´
5´-GACCT

Cycle sequencing

3´-TACG-5´ primer

5´-ATGCATGC-3´ template

ddGTP

GTACG
ATGCATGC

ddCTP

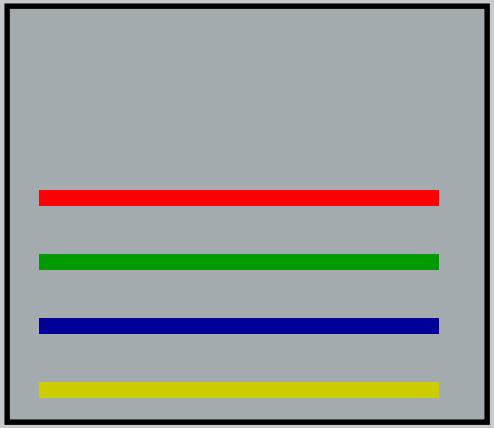
CGTACG
ATGCATGC

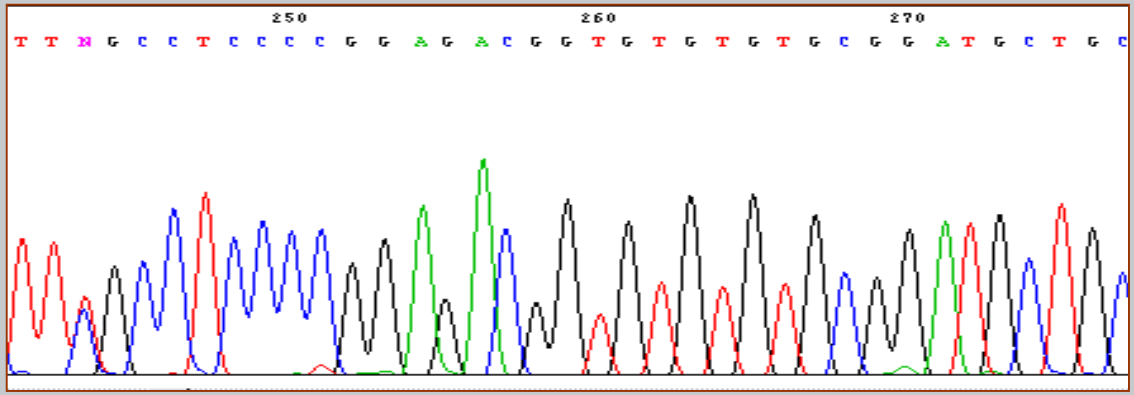
ddATP

ACGTACG
ATGCATGC

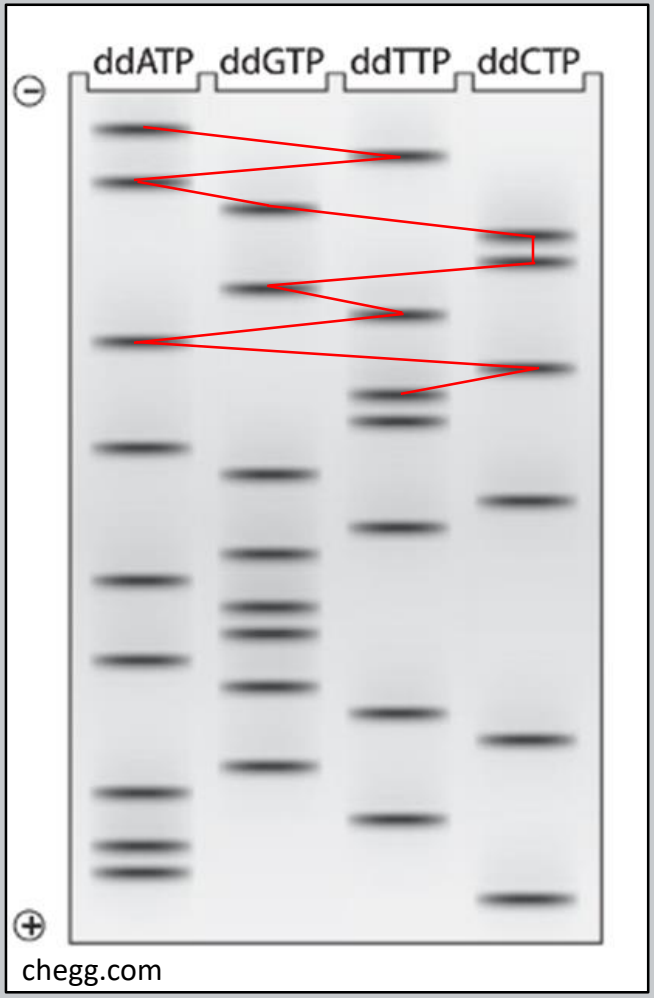
ddTTP

TACGTACG
ATGCATGC

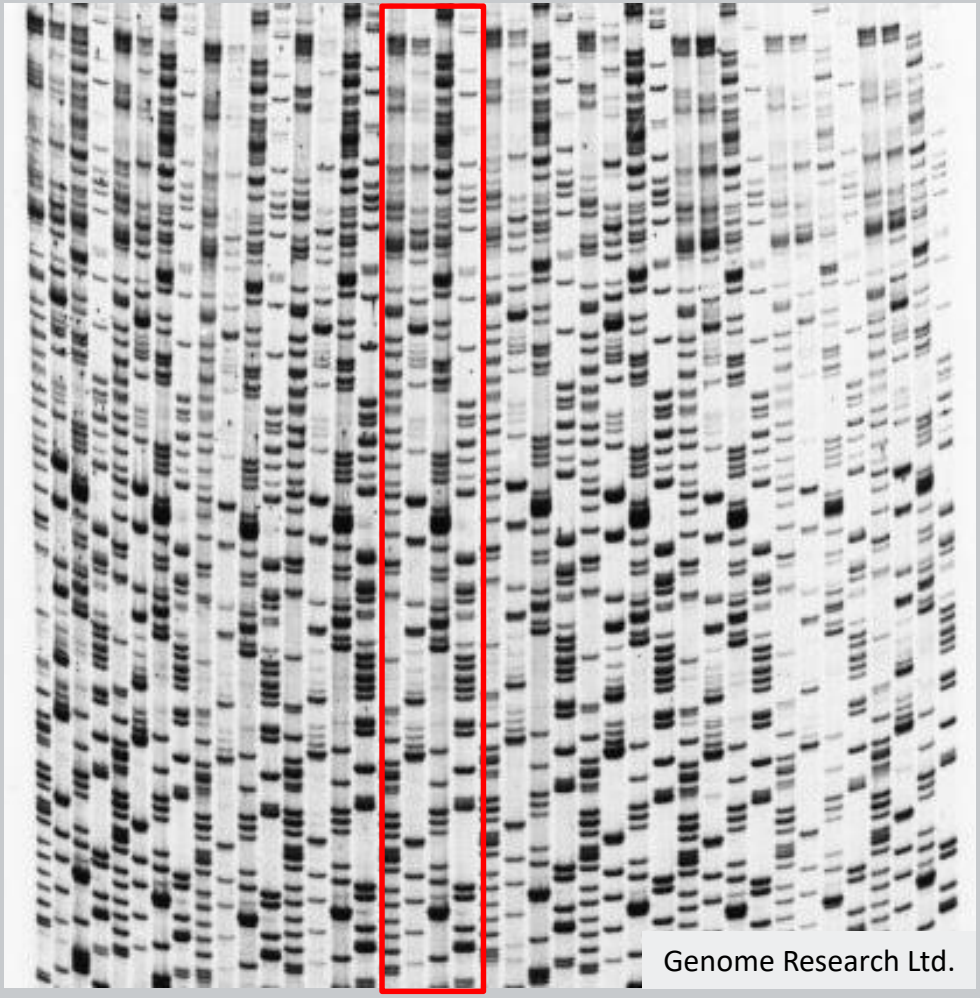




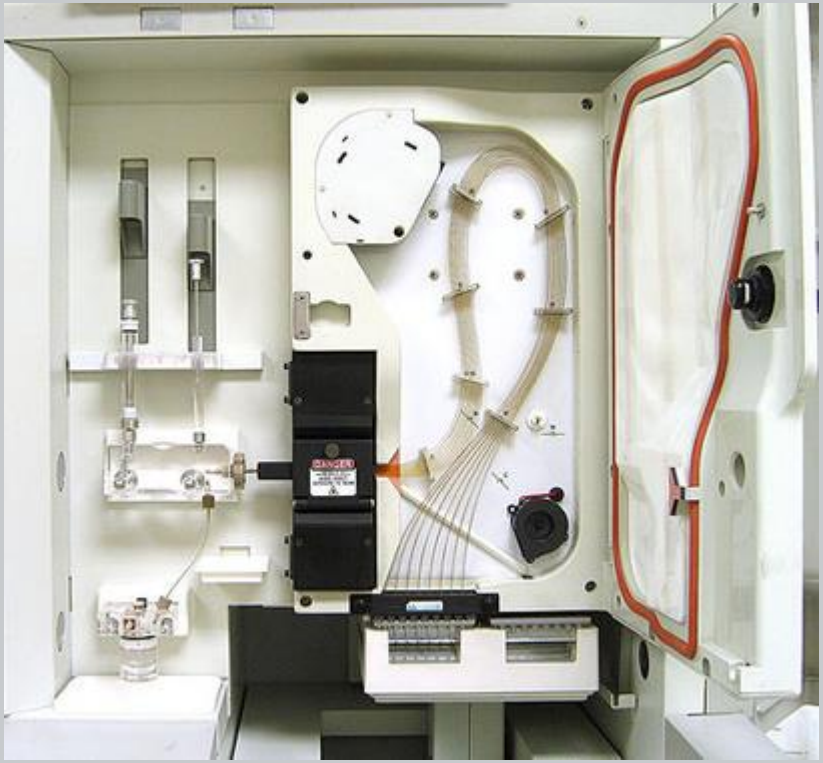
Classical sequencing gel



chegg.com

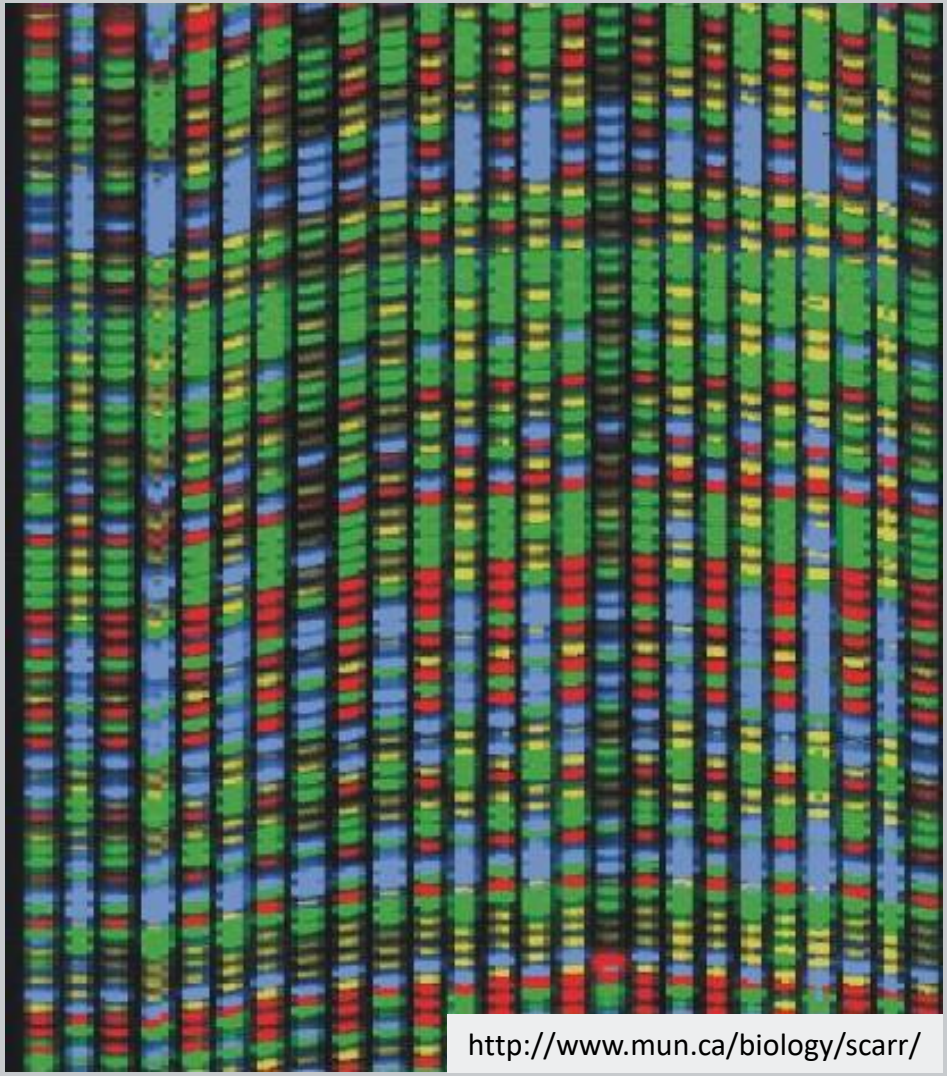

Genome Research Ltd.

Automated sequencer



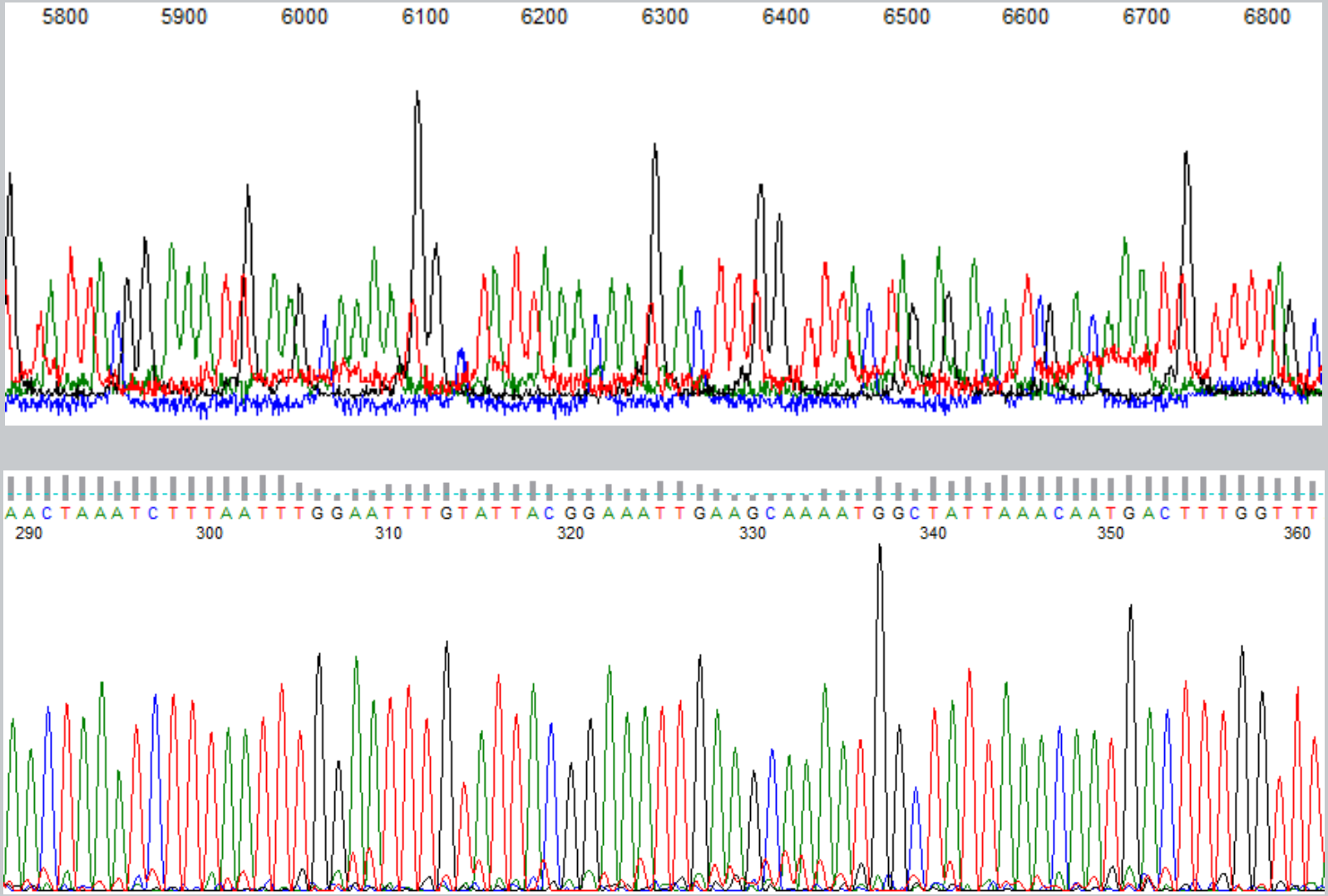
ABI (Applied Biosystems) – slab gel & capillary systems (up to 96)

Gel sequencer raw data



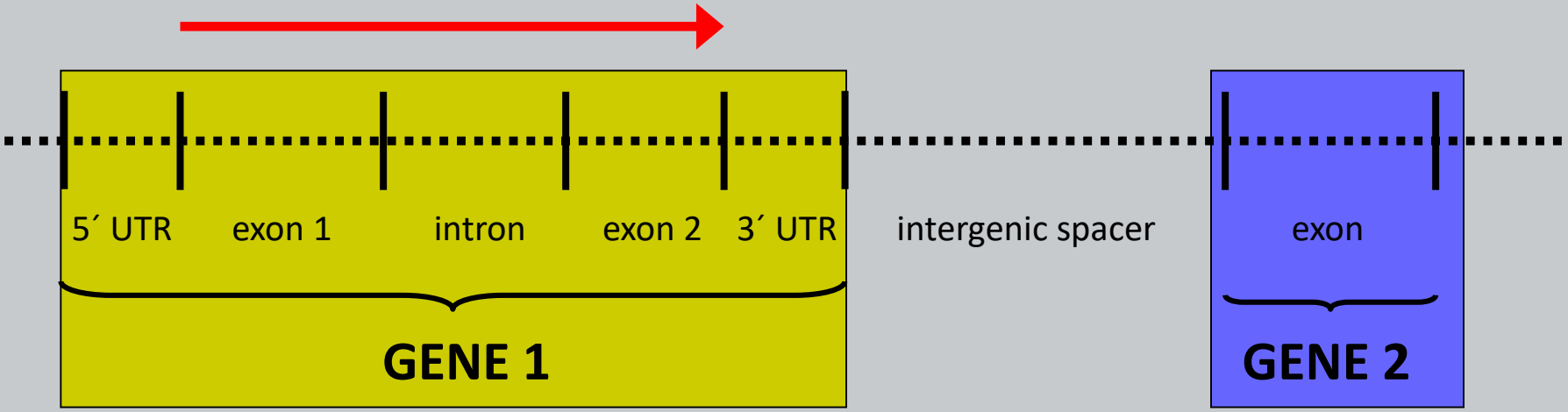
http://www.mun.ca/biology/scarr/

Capillary sequencer raw data



Genome structure

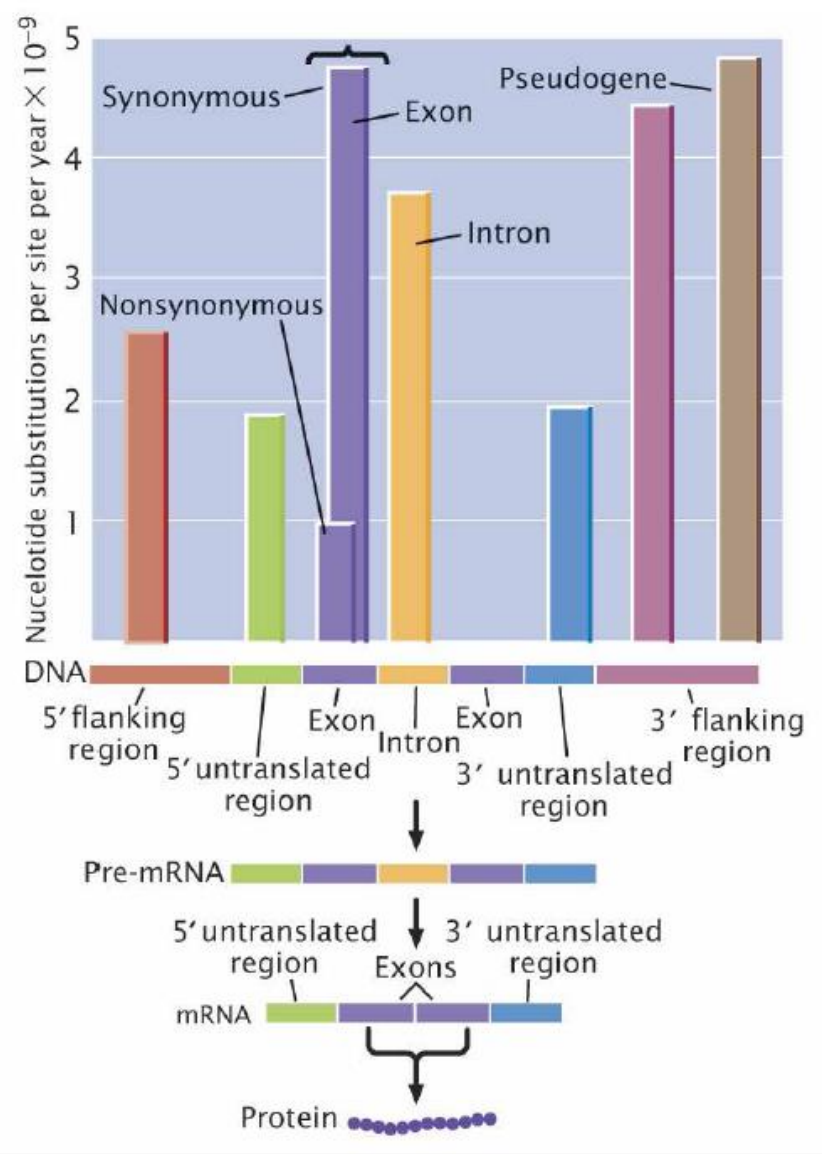
- genetic information – order of nucleotides (ACGT)



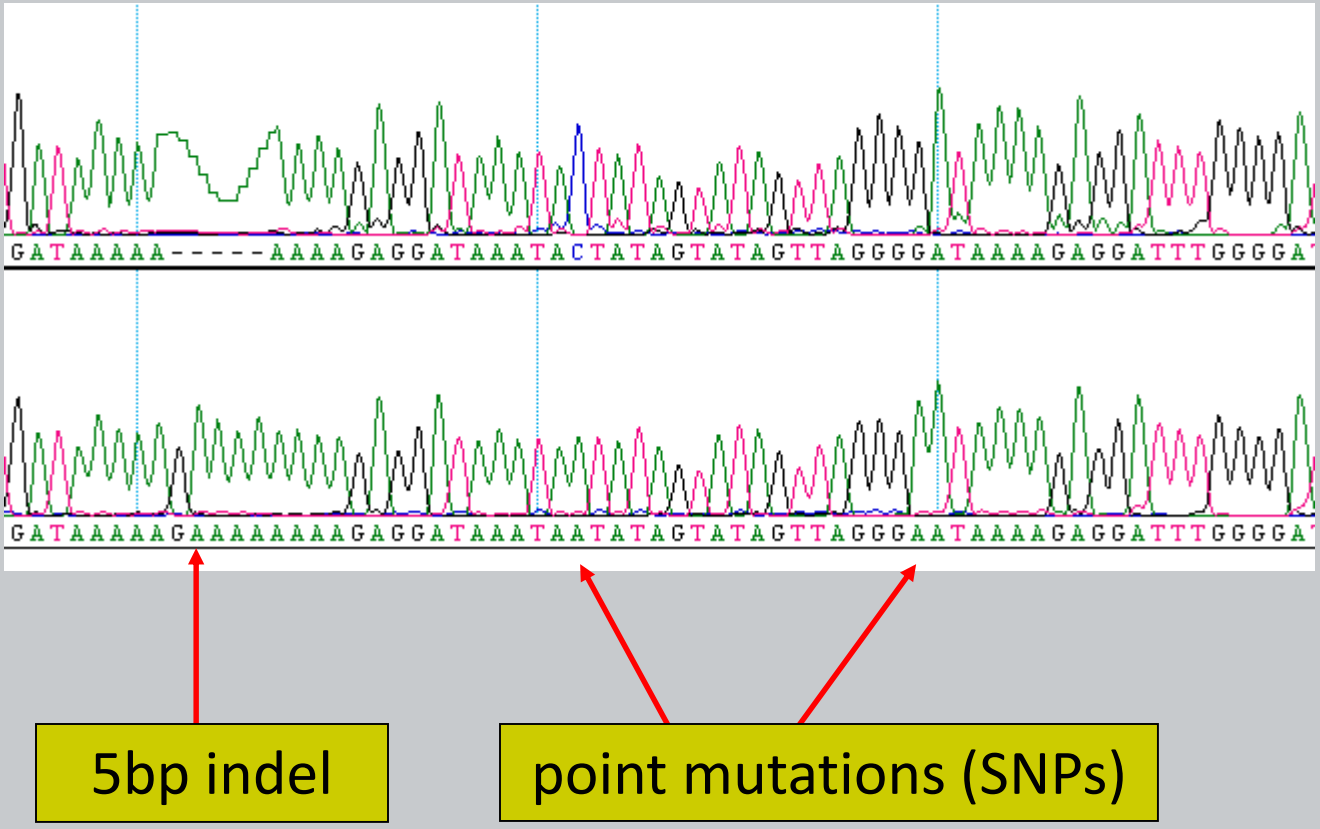
- coding regions – exons – c*onserved*
- non-coding regions – introns, spacers – *variable*
- UTR – untranslated regions

- nuclear, chloroplast and mitochondrial genome

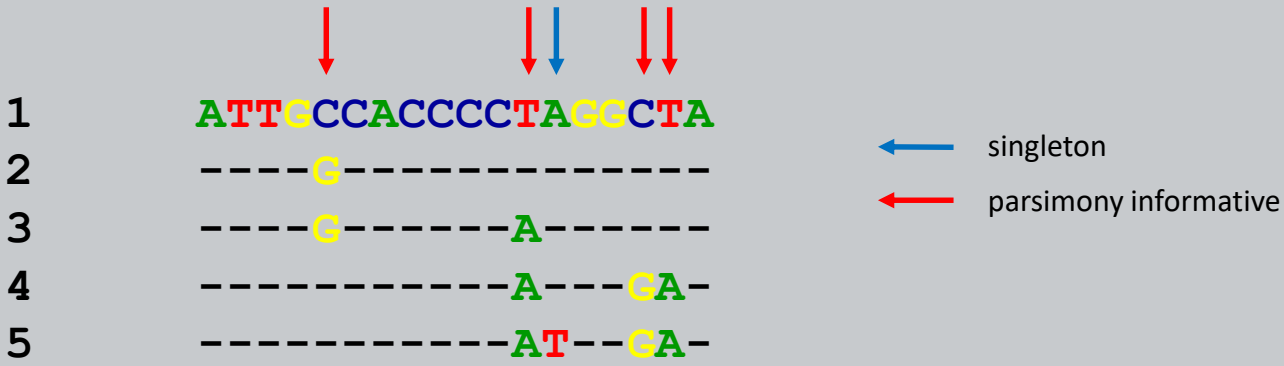
Sequence evolutionary rate



Types of variability in DNA sequences

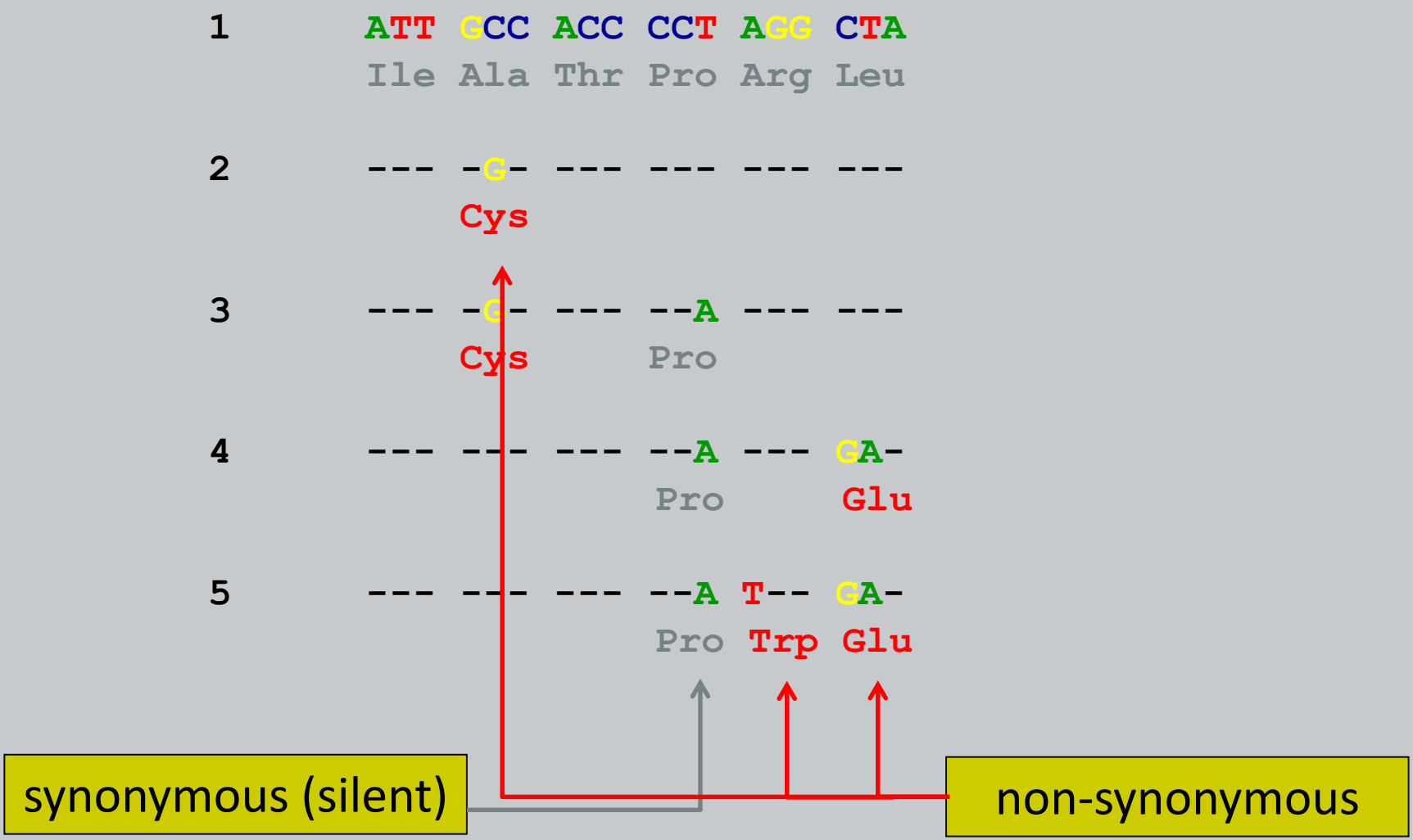


Descriptive statistics for DNA sequences



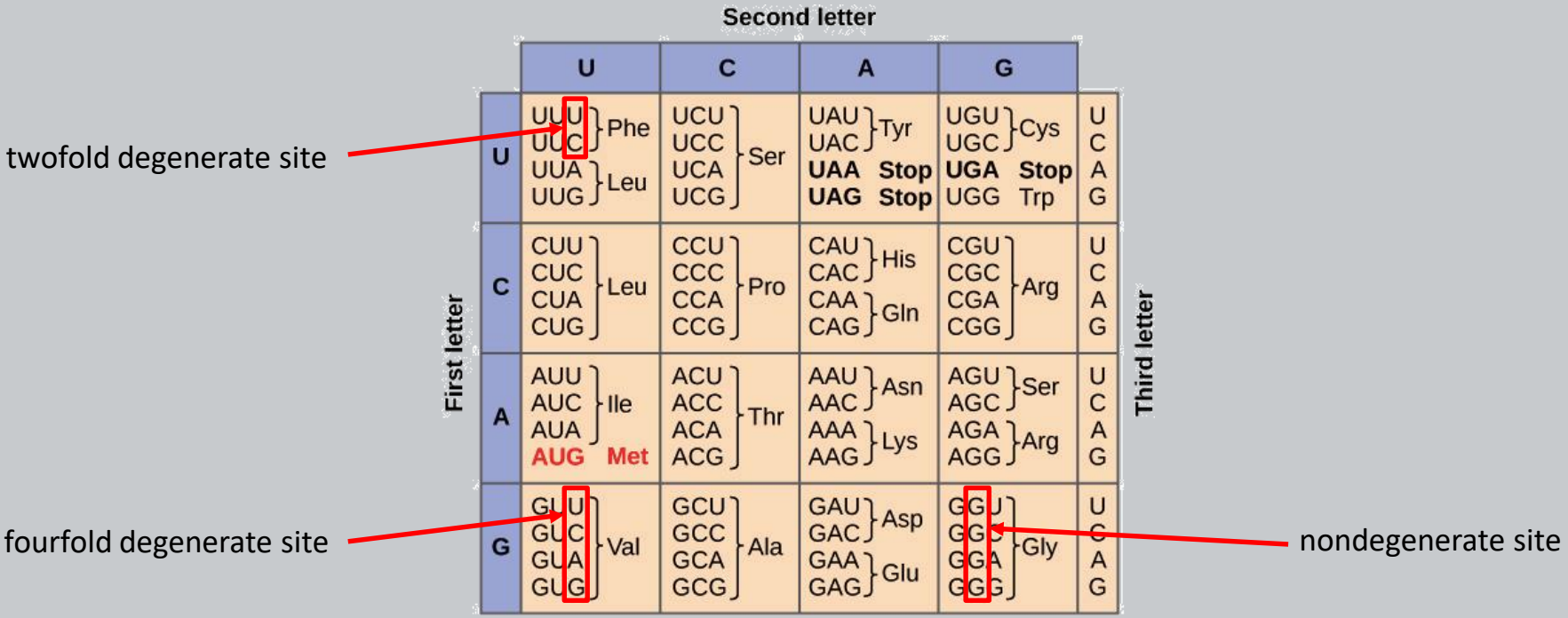
L	sequence length (nr. sites)	18
Π	average number of nucleotide difference	2.8
S	number of segregating (polymorphic) sites	5
π	nucleotide diversity (Π per site)	0.155
s	number of segregating sites per site	0.277

Synonymous vs non-synonymous mutations



Degenaration of the genetic code

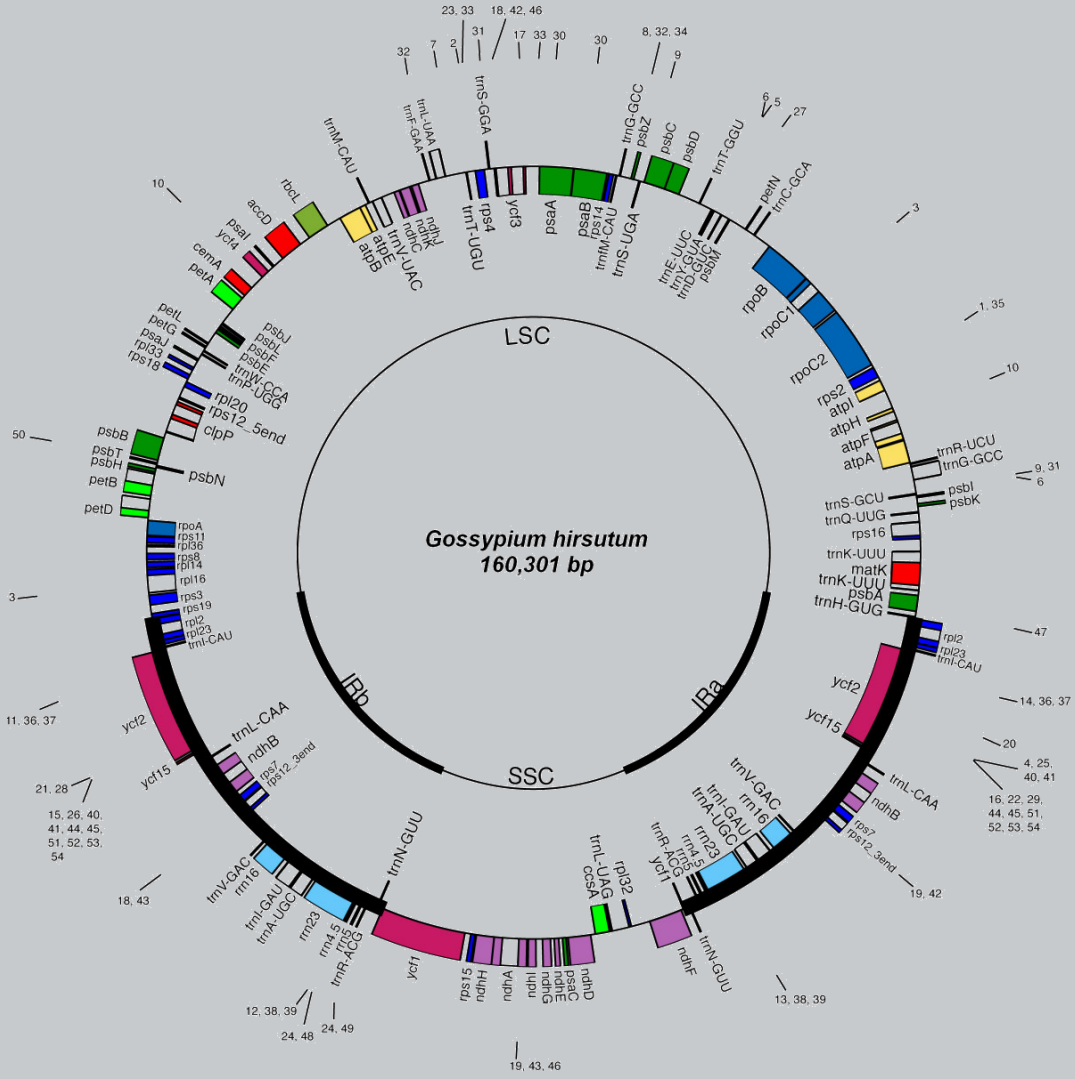
- multiple codons are coding the same aminoacid
  - fourfold degenerate site – any change do not change the aminoacid
  - nondegenerate site – any change also change the aminoacid



Chloroplast genome

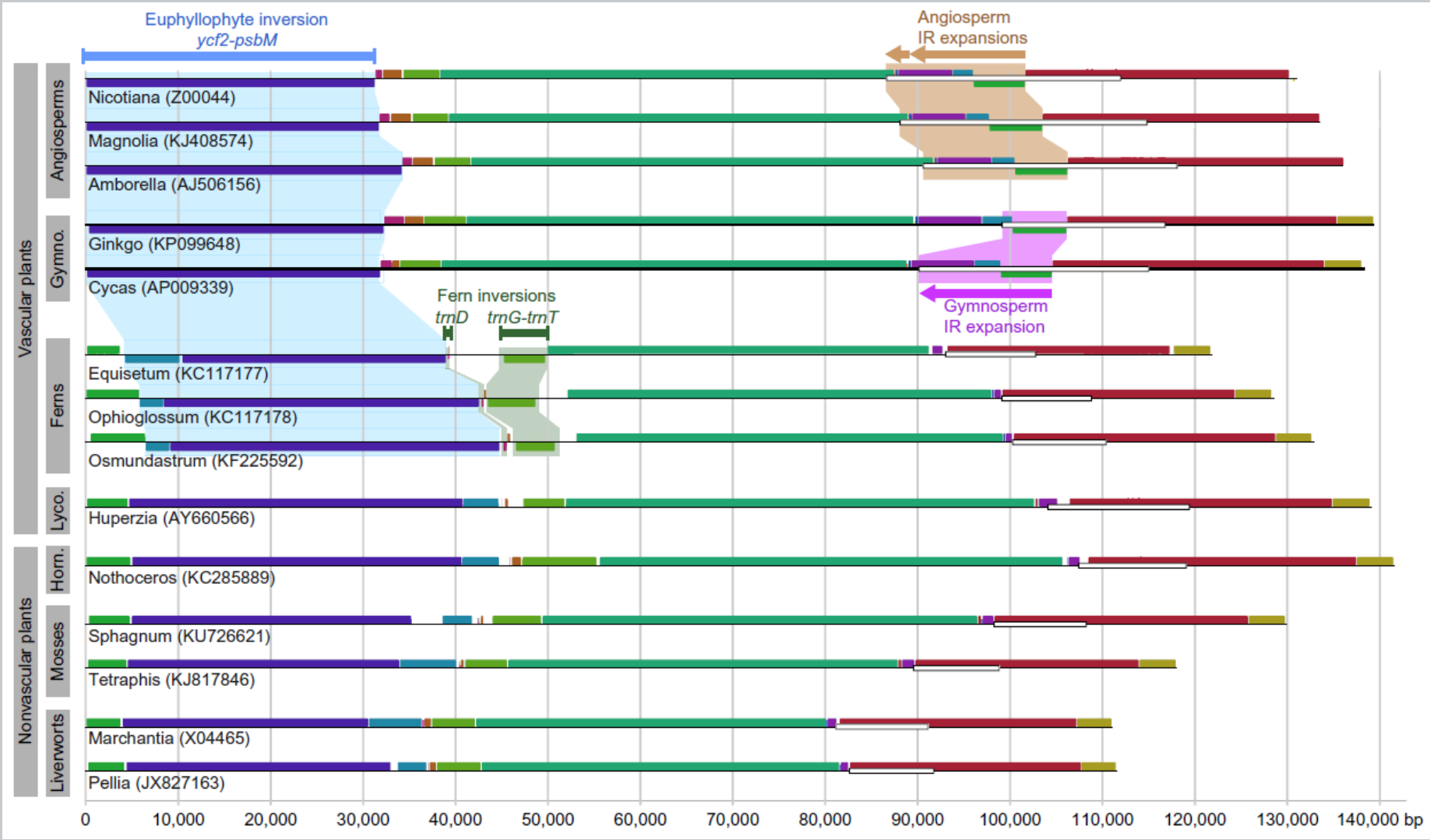
- many genes are *single-copy* (only 1 copy in the whole genome)
- conserved evolution of the chloroplast genome
  - disadvantage when studying intraspecific or population variability
  - many conserved regions can be used as *priming sites*
- structural rearrangements of chloroplast genome
  - mainly on larger evolotionary scale
  - inversion – e.g., 30kb inversion differentiates bryophytes and higher plants
  - extensive deletions
  - loss of specific genes and introns
- *chloroplast capture*
  - chloroplast transfer from one species to another by introgression
  - can influence phylogeny in a wrong way (when not recognized)

Chloroplast genome



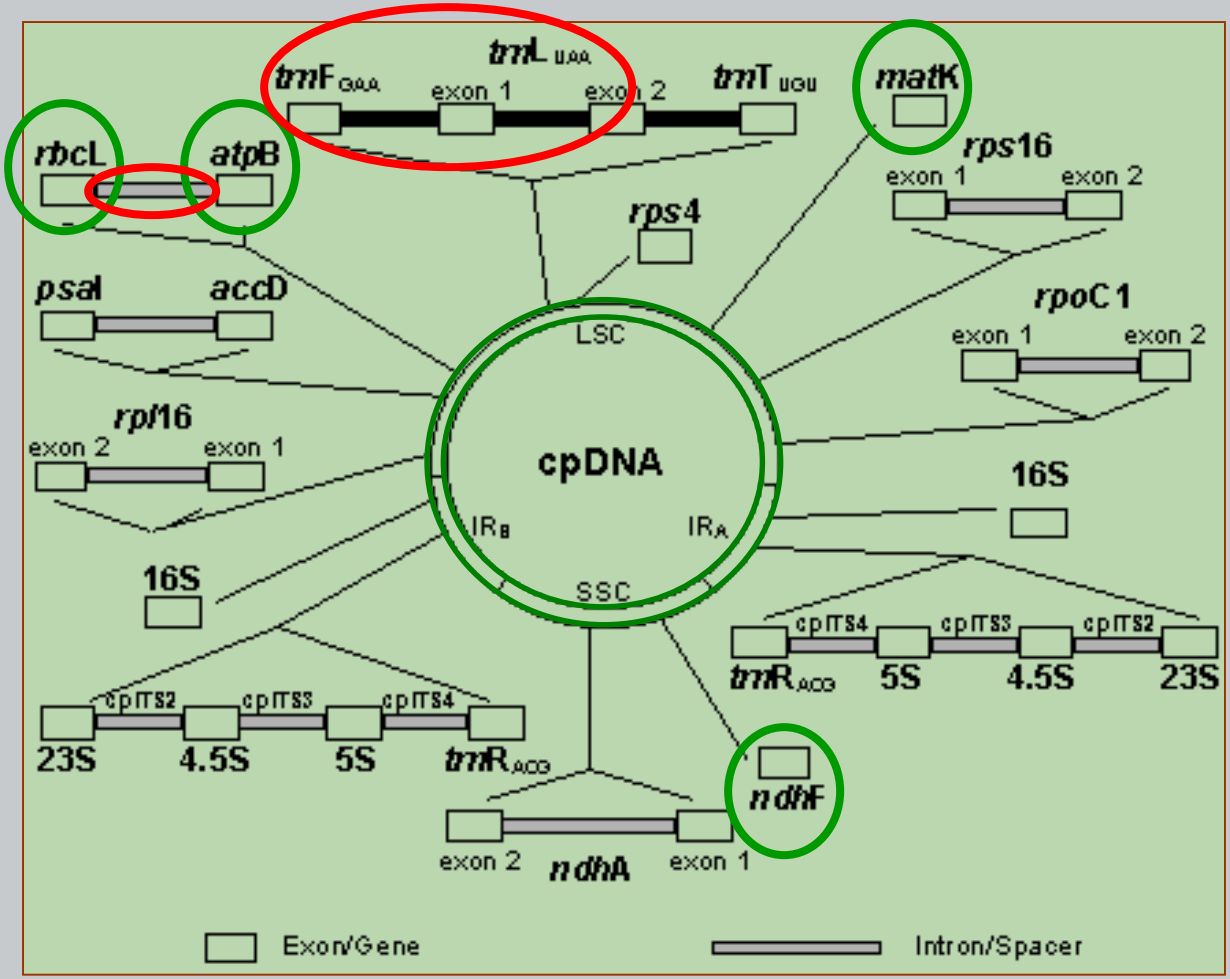
- 4 rRNAs
- 30-11 tRNAs
- 21 ribosomal proteins (*rps*)
- 4 RNA polymerase subunits (*rpo*)
- 28 thylakoid proteins (*ps*)
- *rbc*L (large RuBisCO subunit)
- 11 proteins similar to NADH (*ndh*)

Chloroplast genome



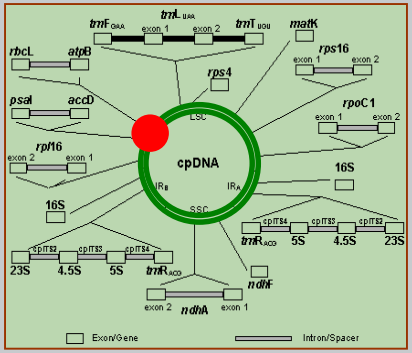
Genome alignment highlighting diagnostic changes among land plant plastomes. Coloured boxes – genome homology segments, horizontal white box – a copy of IR.

Frequently sequenced cpDNA regions



+ many others…

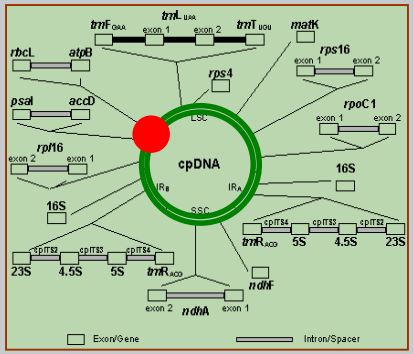
*rbc*L



- gene for large subunit of ribuloso-1,5-bisphosphate-carboxylase/oxygenase (RUBISCO)
- 1,428, 1,431 or 1,434 bp in length – indels are extremely rare
- one of the first sequenced genes
- very conserved, systematics at family or generic level, in some groups at species level

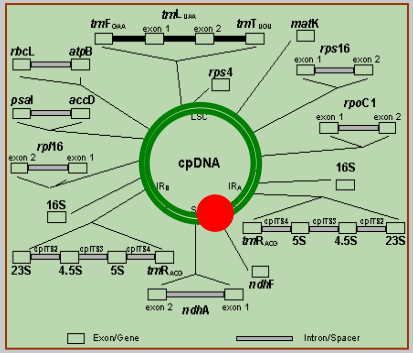
*atp*B

- gene coding beta subunit of ATP synthase
- 1,497 bp in length, indels not found
- similar use as *rbc*L

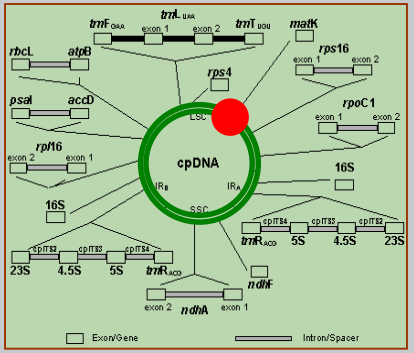


*ndh*F

- codes a subunit of chloroplast NADH-dehydrogenase
- 2,233 bp in length (tobacco)
- about 2x more substitutions then *rbc*L
- for generic level

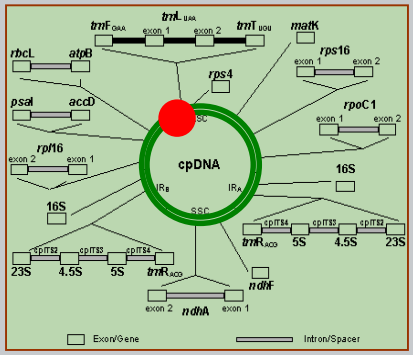


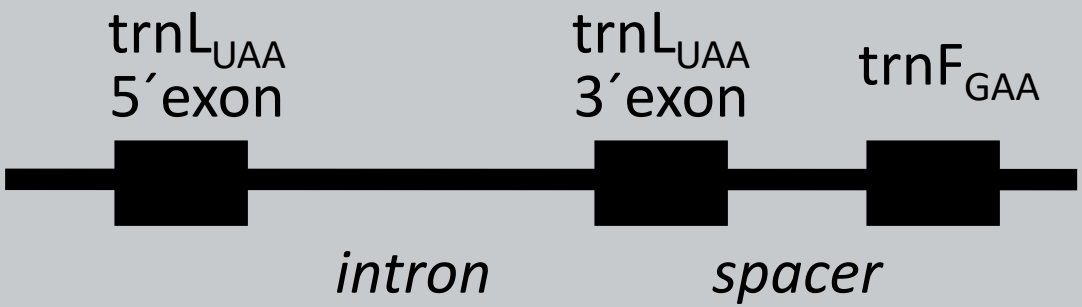
*mat*K



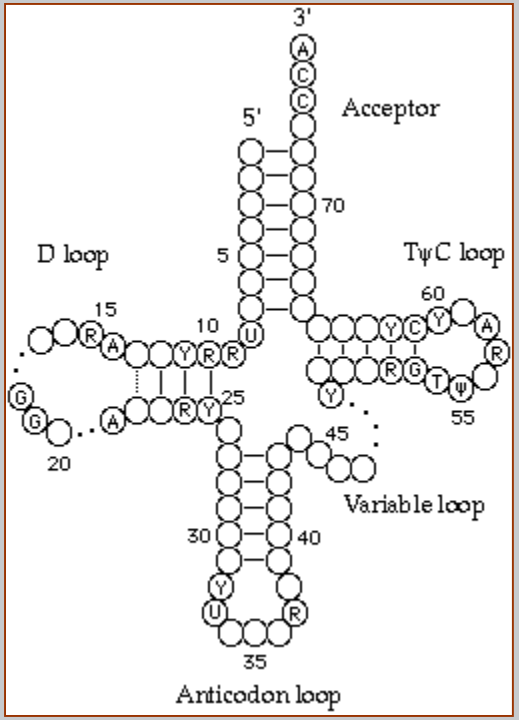
- gene coding maturase (splicing of plastid genes)
- about 1,550 bp in length – low number of indels
- systematics at family and generic level

*trn*L intron and spacer between *trn*L and *trn*F

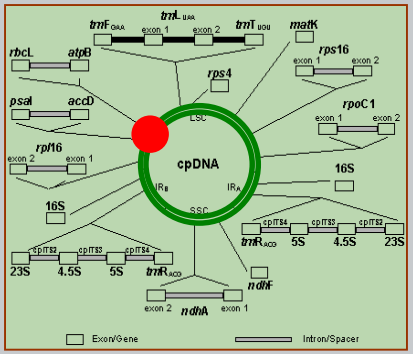




- tRNA genes – secondary structure
- accumulation of insertions/deletions with the same rate as nucleotide substitutions
- alignment problems, especially in distant organisms (sometimes already at family level)
- suitable for systematics of (closely) related species



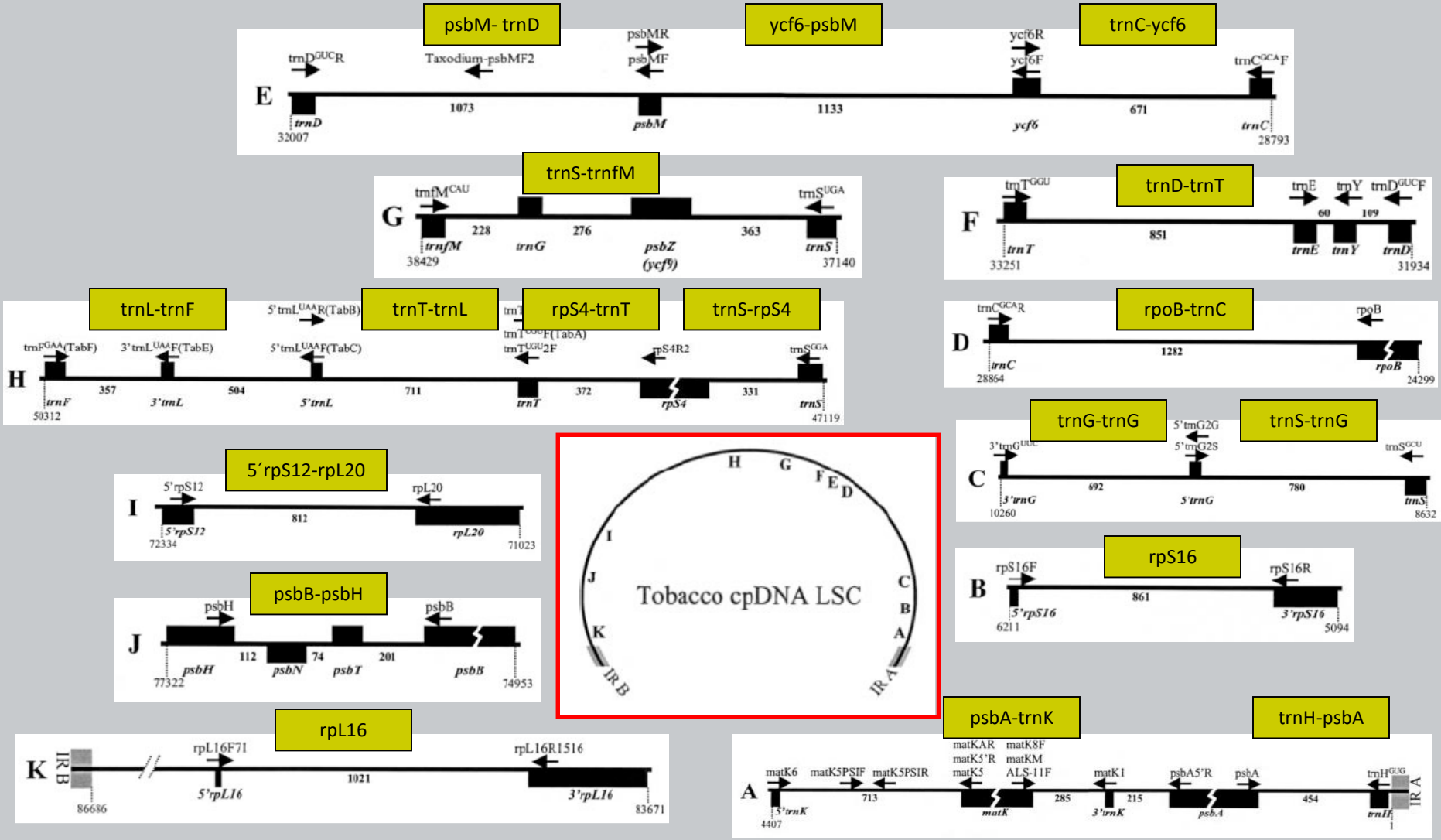
*atp*B-*rbc*L



- spacer of about 900-1,000 bp in length
- systematics at family and generic level

Variable non-coding cpDNA regions

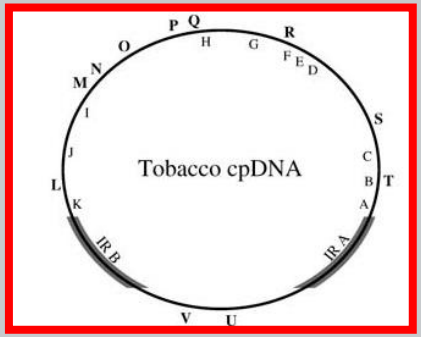
Shaw et al. (2005): The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* **92**: 142-166



Variable non-coding cpDNA regions

Shaw et al. (2007): Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.* **94**: 275–288.

- another 13 regions



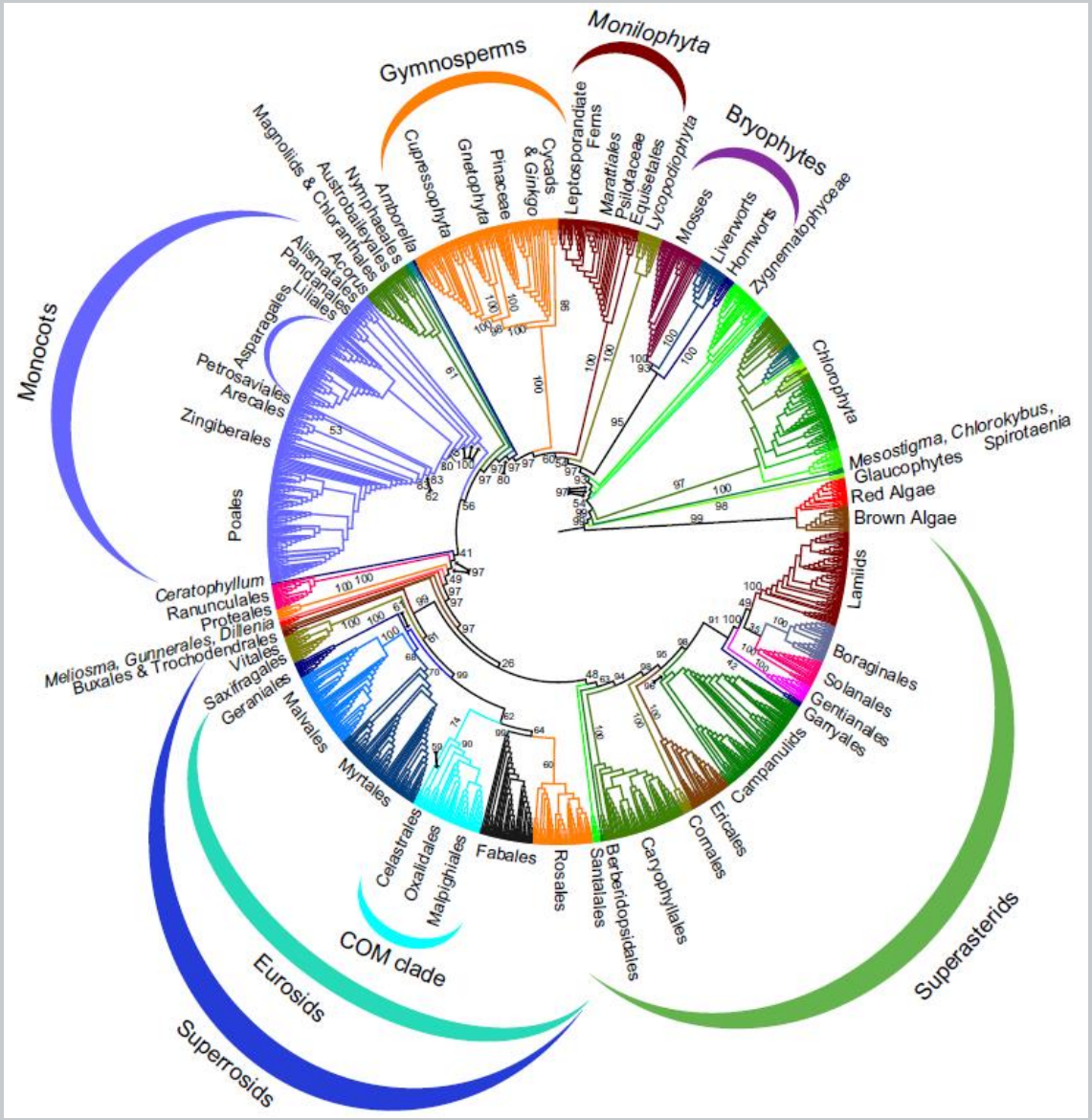
Shaw et al. (2014): Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *Am. J. Bot.* **101**: 1987–2004.

- top 13 regions within each major evolutionary lineage

Use of chloroplast sequences

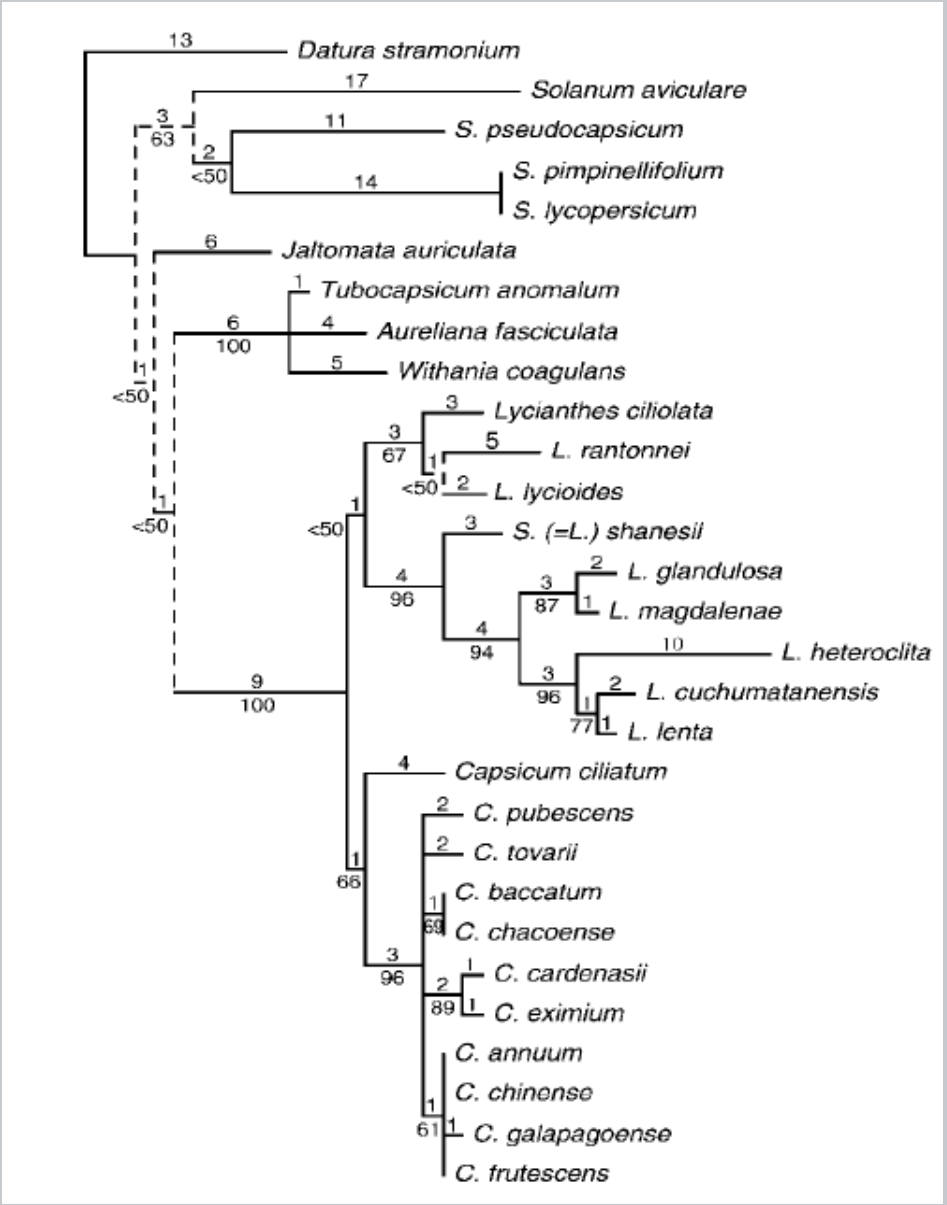
- phylogeny of large groups
- among-species relationships within a genus
- within-species phylogeography (haplotype definition)
- hybridization – inference of the maternal taxon (individual) – cpDNA maternaly inherited in angiosperms

Viridiplantae plastid phylogeny



Gitzendanner et al. (2018)
78 coding plastid genes
1827 taxa + 52 outgroups

Relationships among species



Capsicum
*atp*B-*rbc*L spacer
Walsh & Hoot (2001)

Phylogeography

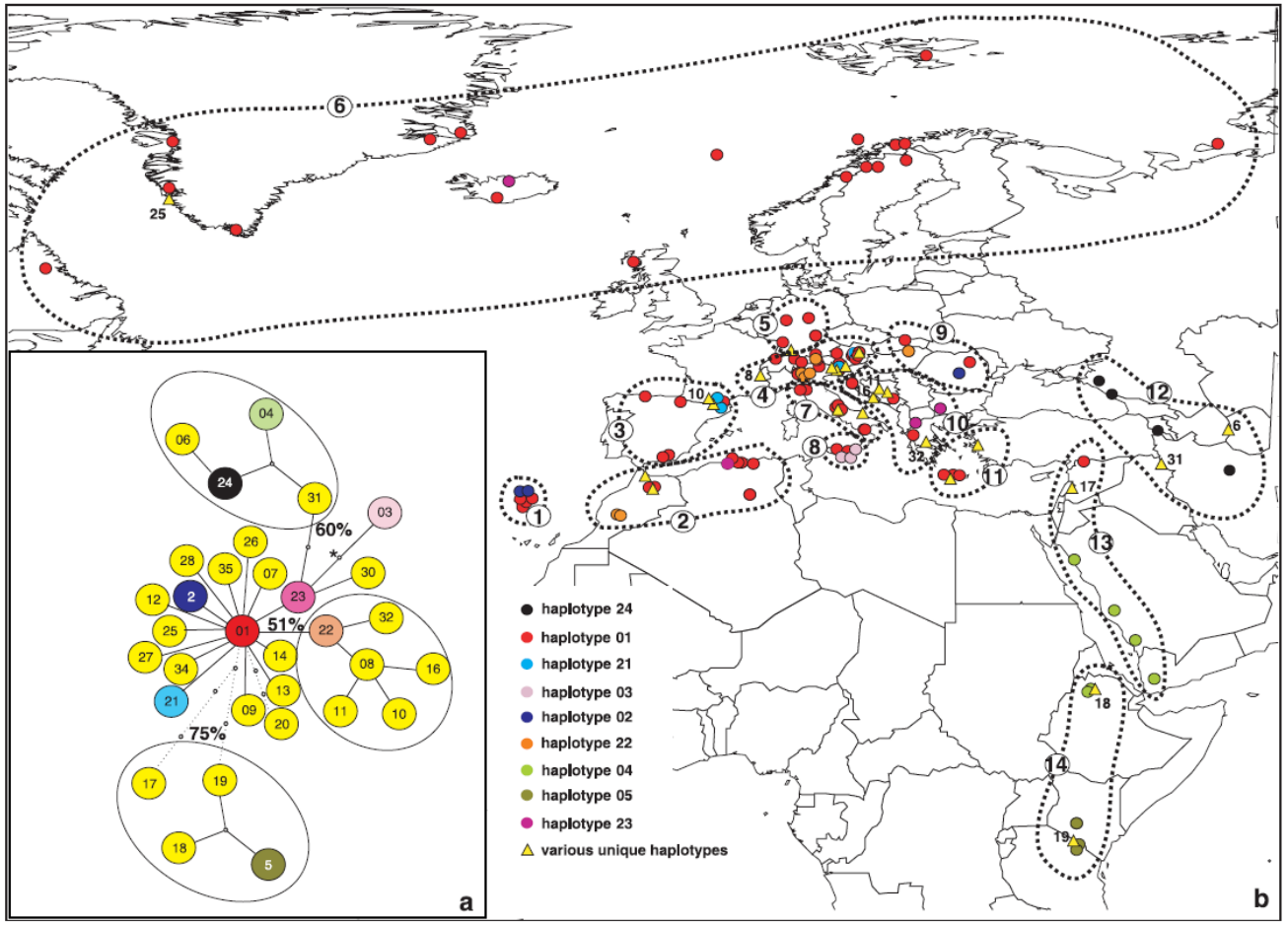
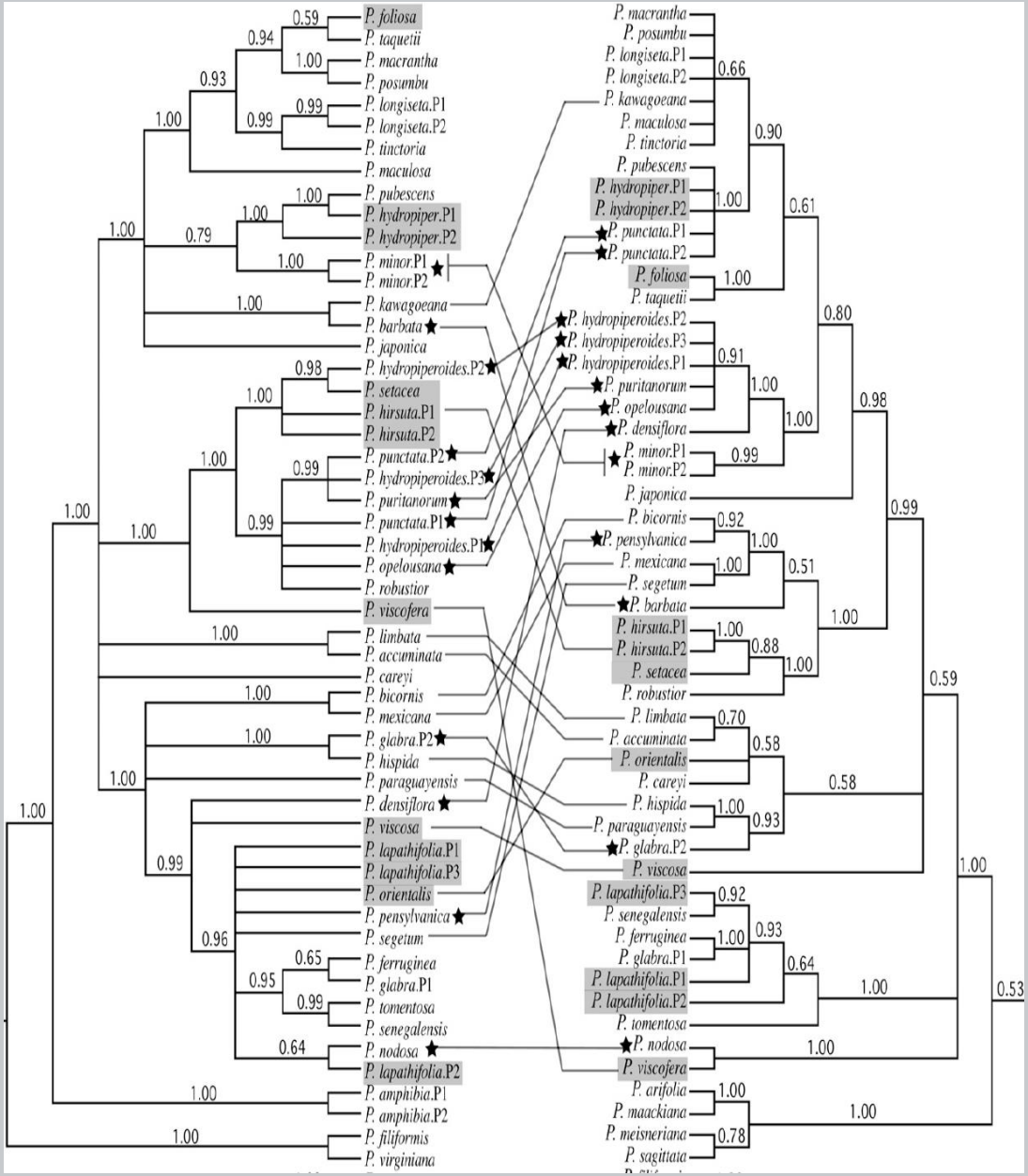


Fig. 1 (a) *Arabis alpina* cpDNA haplotype network as inferred from TCS (see Material and methods section for details). The position of the root as revealed from parsimony and ML analyses is indicated (*). Groups of haplotypes circled correspond to clades that have been also recognized by parsimony and ML analyses (bootstrap values redrawn from parsimony analysis). Colour code and haplotype designation follow Fig. 1 (b), with yellow symbols showing unique haplotypes. (b) Distribution of *A. alpina* cpDNA haplotypes. The distribution range of *A. alpina* is represented by the accessions. Regional subdivision of the accessions into 14 areas is indicated (refer to Table 1 and Table S1, Supplementary materials).

Arabis alpina
*trn*L-*trn*F
Koch et al. 2006

Inter-specific hybridization



incongruence between cpDNA and nDNA

Persicaria
*mat*K, *psb*A-*trn*H, *trn*L-*trn*F
versus ITS
Kim & Donoghue 2008

Data analysis

- *multiple alignment*

```
S206          ATATATATATAGGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
S207          ATATATATA--GGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
S208          ATATATATA--GGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
S209          ATATATATA--GGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
S210          ATATATATA--GGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
S0G3          ATATATATA--GGCAAGGAATCTCTATTATTAAATCATTTAGAATCCATA
TL            ATATATATATAGGCAAGGAATCTCTATTATTAAATCATTCATAATTCATA
```

- construction of phylogenetic tree
  - distance methods
  - maximum parsimony (MP)
  - maximum likelihood (ML)
  - Bayesian inference (BI)

Maximum parsimony (MP)

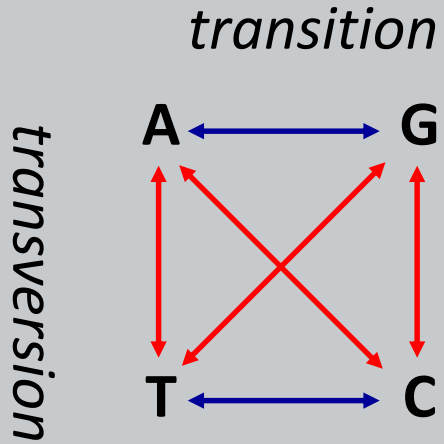
- cladistic method
- search for the simplest tree (*most parsimonious tree*)
- i.e., tree in which the evolution is explained by minimum number of substitutions
- software
  - PAUP *
    Phylogenetic **A**nalysis **U**sing **P**arsimony
    (* and other methods)
  - TNT
    Tree Analysis Using **N**ew **T**echnology

Maximum likelihood (ML)

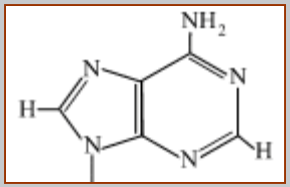
- search for tree with the highest probability (likelihood – L)
- probability that observed sequences evolved under given tree topology (and under given evolutionary model)
- software GARLI, PhyML, RAxML, RAxML-ng, PAML…

Evolutionary models for DNA sequences

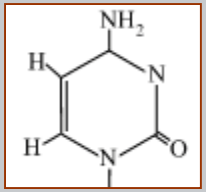
- models for sequence changes



purines



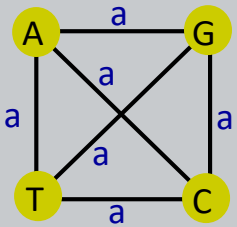
pyrimidines



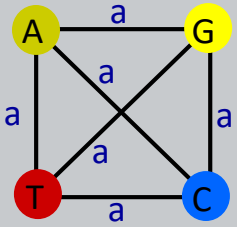
- parameters
  - base frequences
  - substitution types (transitions, transversions)
  - heterogeneity in substitution rates (G)
  - proportion of invariant sites (I)

Substitution models

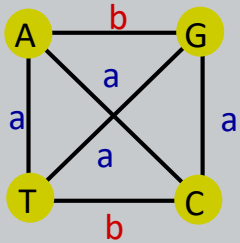
Increasing number of parameters



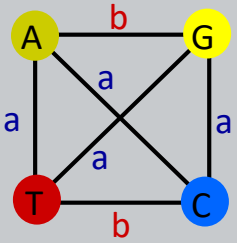
JC (Jukes-Cantor 1969)
– same substitution rates
– same base frequencies



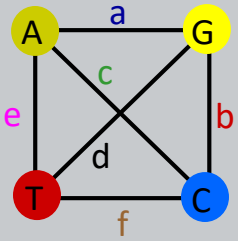
F81 (Felsenstein 1981)
– same substitution rates
– different base frequencies



K2P (Kimura 2 parameter 1980)
– two different substitution rates
– same base frequencies



HKY (Hasegawa, Kishino & Yano 1985)
– two different substitution rates
– different base frequencies



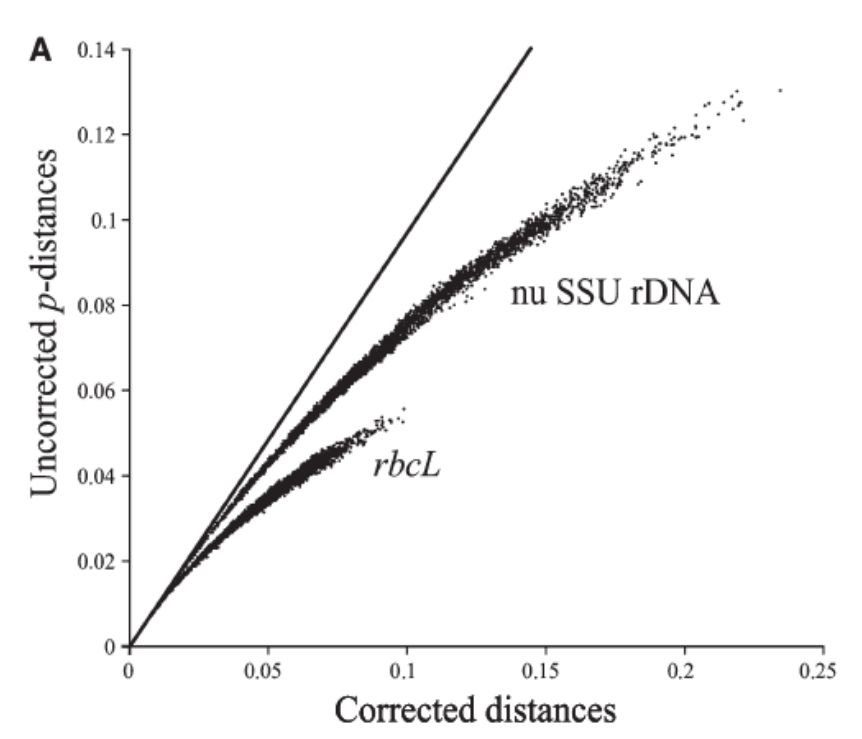
GTR (General time-reversible model)
(Tavaré et al. 1986)
– six different substitution rates
– different base frequencies

Which model to select?

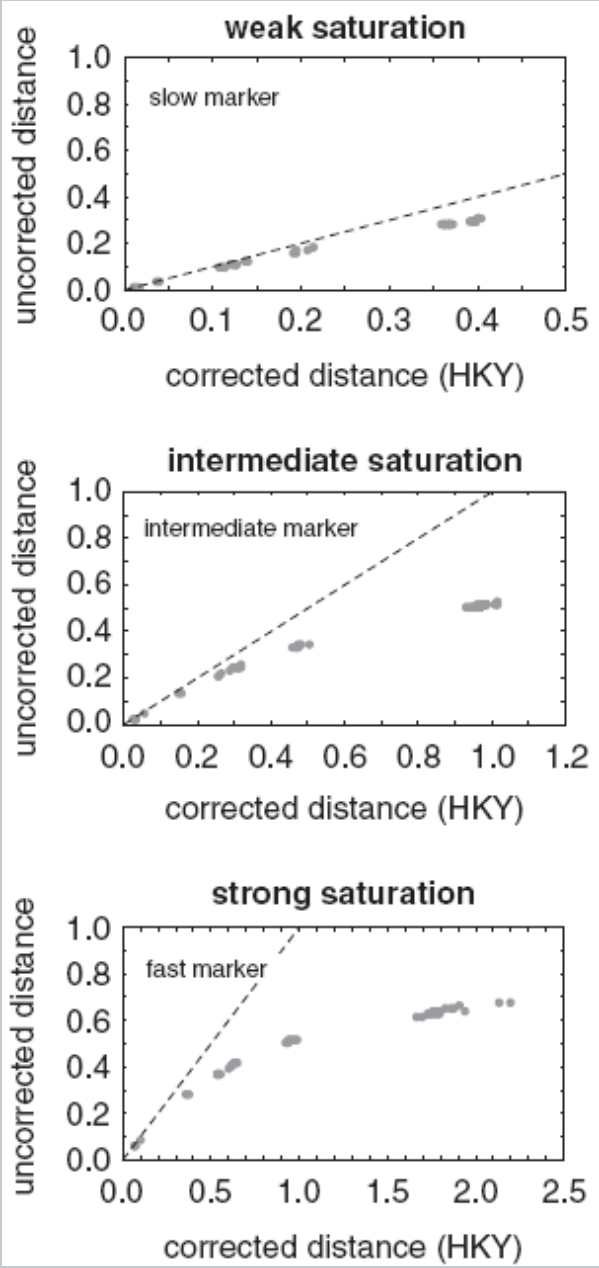
- MODELTEST: A tool to select the best-fit model of nucleotide substitution (Posada et al.)
- testing different models – selecting the simpliest that sufficiently explain the data using
  - hierarchical likelihood ratio tests (hLRTs)
  - Akaike information criterion (AIC)
- jModelTest2 (https://code.google.com/p/jmodeltest2/)
- ModelTest-NG (https://github.com/ddarriba/modeltest/)

Saturation

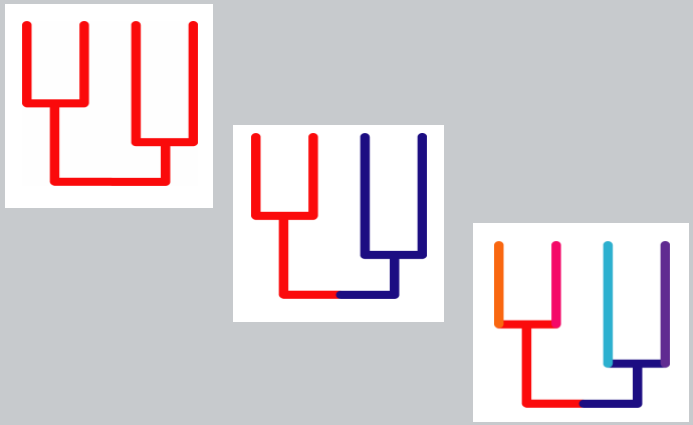
- signal and noise in the data
- corrected versus uncorrected distance
- skewness (g_1-statistics), I_{SS}



Gontcharov & Melkonian 2008



Molecular clock

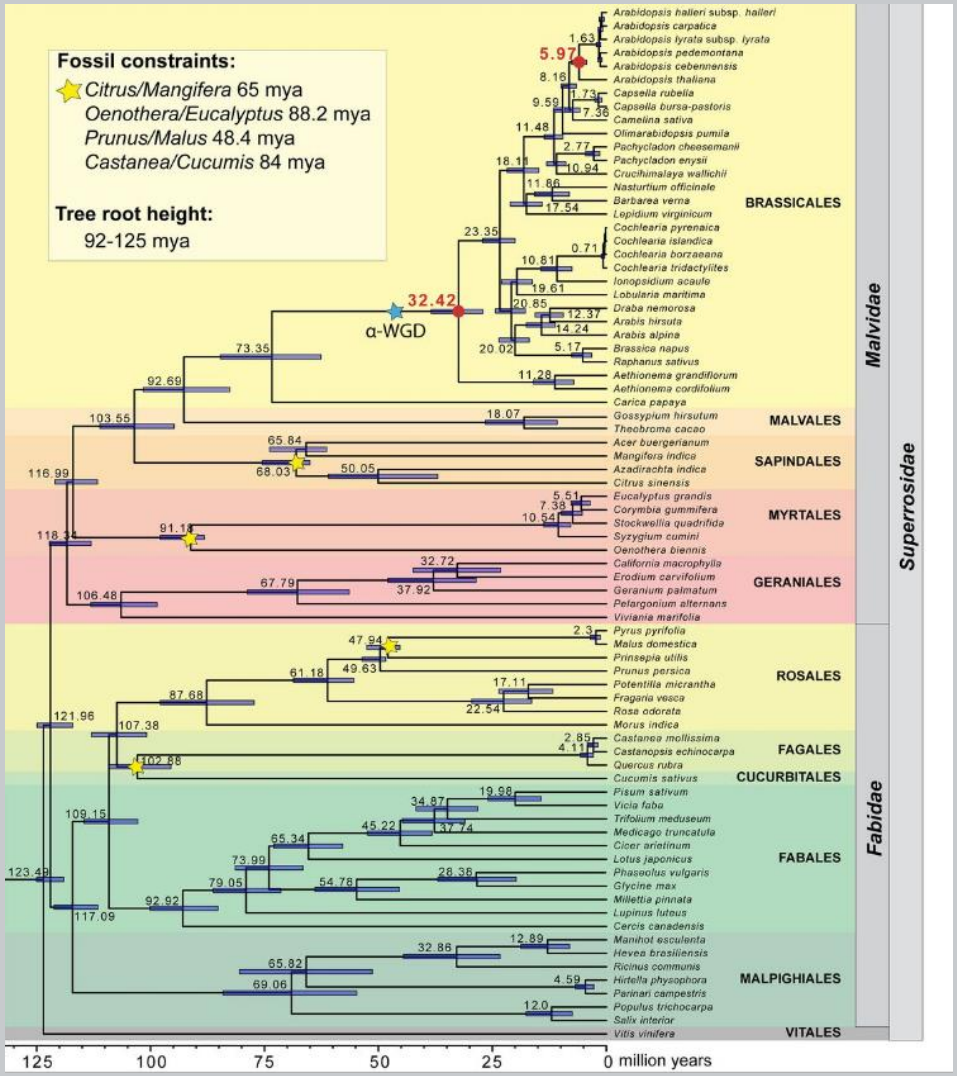


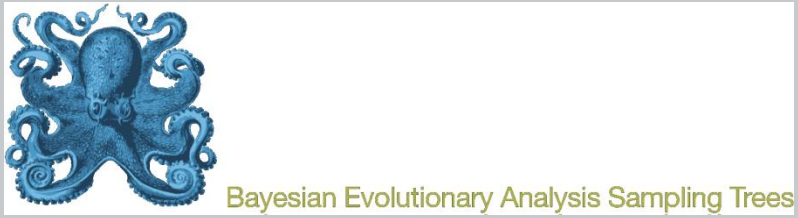
- strict (global)
  - *clocklike evolution*
- local
- *relaxed clocks*
  - autocorrelated (closely related taxa have similar mutation rates)
  - uncorrelated (lognormal, exponential)
- calibration
  - substitution rates from another study or generally assumed rate (e.g., for cpDNA)
  - fossils
  - biogeography
- software
  - BEAST (Bayesian), r8s (non-parametric rate smoothing, penalized likelihood), …

http://sydney.edu.au/science/biology/meep/documents/Workshop_Lecture4.pdf

Estimates of divergence times

(BEAST – Bayesian Evolutionary Analysis Sampling Trees)





Hohmann N. et al. (2015): A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell* **27**: 2770–2784.

Gene banks – databases of sequences

- **GenBank**
  National Centre for Biotechnology Information (NCBI)
  http://www.ncbi.nlm.nih.gov/
- **EMBL**
  European Bioinformatics Institute (EBI)
  http://www.ebi.ac.uk/embl/

GenBank example

```
LOCUS       JQ409881                562 bp    DNA     linear   PLN 31-DEC-2012
DEFINITION  Curcuma ecomata voucher JLS 73353 18S ribosomal RNA gene, partial
            sequence; internal transcribed spacer 1 and 5.8S ribosomal RNA
            gene, complete sequence; and internal transcribed spacer 2, partial
            sequence.
ACCESSION   JQ409881
VERSION     JQ409881.1
KEYWORDS    .
SOURCE      Curcuma ecomata
  ORGANISM  Curcuma ecomata
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Zingiberales;
            Zingiberaceae; Curcuma.
REFERENCE   1  (bases 1 to 562)
  AUTHORS   Zaveska,E., Fer,T., Sida,O., Krak,K., Ma
            Leong-Skornickova,J.
  TITLE     Conquest of ginger paradise: first insig
            based on plastid and nuclear sequences
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 562)
  AUTHORS   Zaveska,E., Fer,T., Sida,O., Krak,K., Ma
            Leong-Skornickova,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-JAN-2012) Department of Bo
            Prague, Faculty of Science, Benatska 2,
            Czech Republic
```

```
FEATURES             Location/Qualifiers
     source          1..562
                     /organism="Curcuma ecomata"
                     /mol_type="genomic DNA"
                     /specimen_voucher="JLS 73353"
                     /db_xref="taxon:252240"
                     /note="authority: Curcuma ecomata Craib"
     rRNA            <1..8
                     /product="18S ribosomal RNA"
     misc_RNA        9..179
                     /product="internal transcribed spacer 1"
                     /note="ITS 1"
     rRNA            180..344
                     /product="5.8S ribosomal RNA"
     misc_RNA        345..>562
                     /product="internal transcribed spacer 2"
                     /note="ITS 2"
ORIGIN
        1 cattgttgag agagcataga atgatggatg attgtgaatg tgtgaacgcg acccttttcgt
       61 tagcccacgt tggtgggcga ttgactacgg tgcgatcggc actaaggaac aatgaactcg
      121 gaagcakagg gccccttgct gtgagcgggg agcccaatgc atcgaagatt cctcggaatc
      181 aaatgactct cggcaatgga tatctcggct cttgcatcga tgaagaacgt agtgaaatgc
      241 gatacttggt gtgaattgca gaatctcgtg aaccattgag tctttgaacg caagttgtgc
      301 ccgaggcctt gtggtcgagg gcacgcctgc ttgggtgtca tggcatygtc gcttttgctc
      361 catgcttcgt tagcattgag cgcggaaatt ggccccgtgt gccctcgggc acagtcggtc
      421 gaagagtggg tagtcggtat tcgtcgagca cgatggatgt tggtcgtcgc gaacgggaac
      481 tgaacgtcgt cctcgtcgtt tcgggatgag tcctcaagag accctgtgtg attgcggagt
      541 cggttgaaag tgccgtgtca at
```

Population study

Sanz M. et al. (2014): Southern isolation and northern long-distance dispersal shaped the phylogeography of the widespread, but highly disjunct, European high mountain plant *Artemisia eriantha* (Asteraceae). *Botanical Journal of the Linnean Society* 174: 214–226.

Systematic study

Renner S.S. (2004): A chloroplast phylogeny of *Arisaema* (Araceae) illustrates Tertiary floristic links between Asia, North America, and East Africa. *American Journal of Botany* 91(6): 881–888

Literature

Giani A. M. (2020): *Long walk to genomics: History and current approaches to genome sequencing and assembly*. Comp. Struct. Biotech. J. 18: 9-19.

Soltis D.E. & al. [eds.] (1998): *Molecular systematics of plants.II. DNA sequencing*.

Hollingsworth & al. [eds.] (1999): *Molecular systematics and plant evolution.*

Hall B.G. (2001): *Phylogenetic trees made easy*.

Felsenstein J. (2004): *Inferring phylogenies*.

Lemey P. & al. [eds.] (2009): *The phylogenetic handbook*. 2nd ed.

Wiley E.O. & Lieberman B.S. (2011): *Phylogenetics. Theory and Practice of Phylogenetic Systematics*. 2nd ed.

Wheeler W. C. (2012): *Systematics*. *A course of lectures*.

Drummond et al. (2009): *Relaxed phylogenetics and dating with confidence*. PLoS Biol 4(5): e88