Využití molekulárních markerů v systematice a populační biologii rostlin

9. Sekvenování DNA II. – nrDNA, low-copy markery

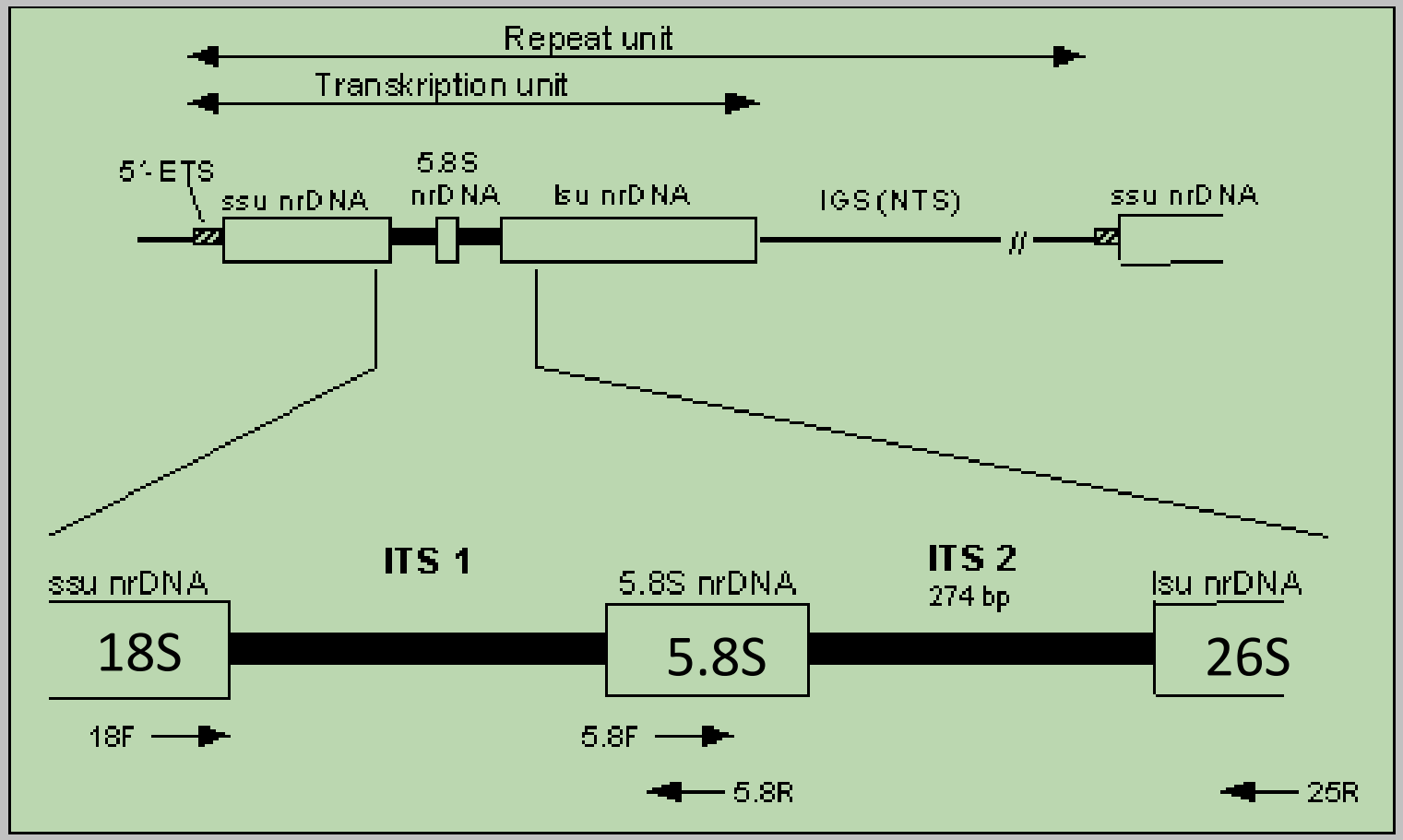
Jaderný genom

- mnoho genů je v mnoha kopiích (*multiple-copy*)
  - problém s homologií – nevíme, co vlastně sekvenujeme
  - např. geny pro rRNA
- *low-copy* nebo *single-copy* geny
  - problém s primery pro studovanou skupinu
  - geny pro konkrétní proteiny – *Adh*, *Tpi*, *Pgi*, phytochrome *c*, *waxy* (GBSSI)

rDNA

- geny pro rDNA – nejrozšířenější marker v systematice
- mnoho tisíc tandemových repetic (1-50 tis.)
- cca 10% celkové DNA
- v jednom nebo několika málo chromosomových lokusech
- přepisovaný úsek (ETS-18S-ITS1-5.8S-ITS2-26S) oddělený intergenickým spacerem (IGS)
- *concerted evolution* – zaručuje intragenomickou uniformitu opakujících se jednotek
  - pokud probíhá pomalu, existuje v rámci genomu více různých ITS sekvencí (paralogů – na různých chromozomech) → problematické určení fylogeneze, pozor na hybridizaci a polyploidizaci

Struktura rDNA

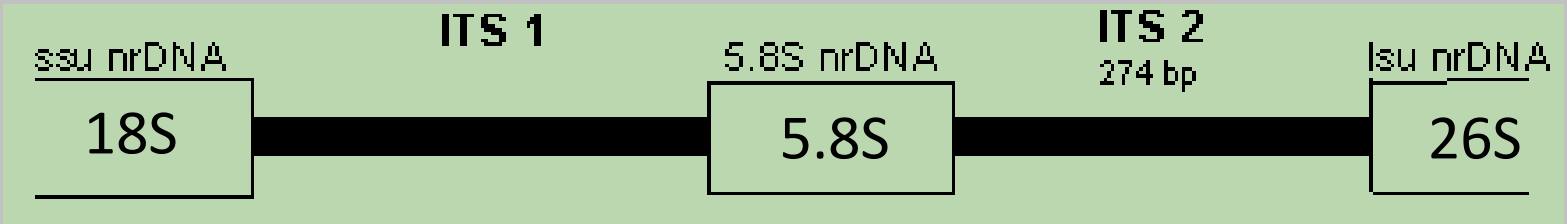


ETS – *external transcribed spacer*

ITS – *internal transcribed spacer*

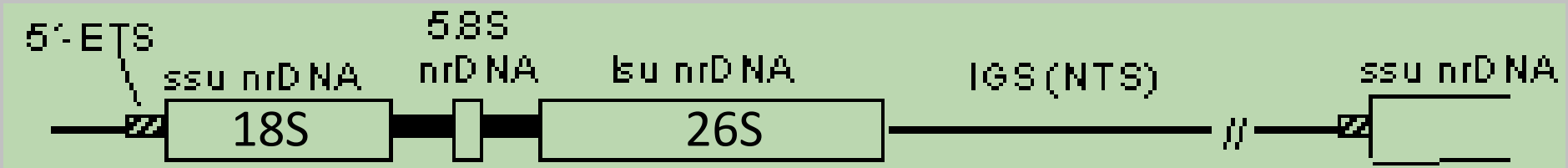
IGS – *intergenic spacer* (NTS – *non-transcribed spacer*)

ITS – internal transcribed spacer



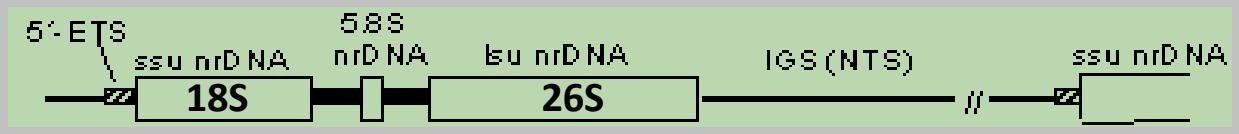
- ITS1 (200-300 bp) – vyšší délková variabilita než ITS2
- ITS2 (180-240 bp)
- velmi používán ke zjištění vztahů mezi blízce příbuznými rody i na druhové úrovni, někdy však málo variabilní
- mají určitou funkci při formování ribozomových podjednotek
- tj. existuje nějaké evoluční omezení ve struktuře a sekvenci
  - 40% ITS2 je konzervováno mezi všemi sekvenovanými krytosemennými
  - 50% ITS2 je možno alignovat na úrovni čeledí a vyšší
- mnohem delší u nahosemenných, výrazná délková variabilita (1550-3125 bp u *Pinaceae*)

ETS – external transcribed spacer



- vyvíjí se nejméně stejně rychle jako ITS
- 258-635 bp dlouhý
- problém se sekvenováním – chybí konzervovaný úsek na 5´ konci spaceru
- *long-distance* PCR k namnožení celého IGS nutná (využití univerzálních primerů z 18S a 26S DNA)
- po osekvenování produktu od 3´ konce (18S) je možno designovat interní primery

18S rDNA



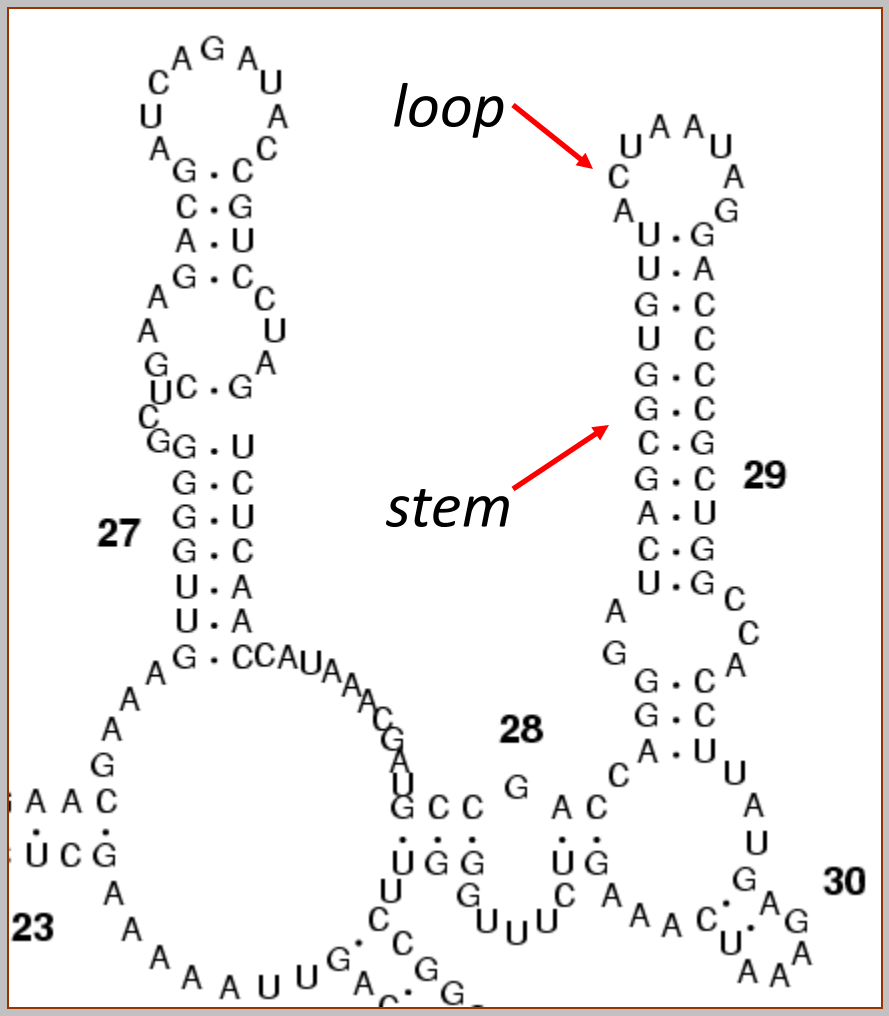
- asi 1800 bp dlouhá
- délkové mutace často jen 1 bp, v konkrétních místech
- tj. snadný alignment

26S rDNA

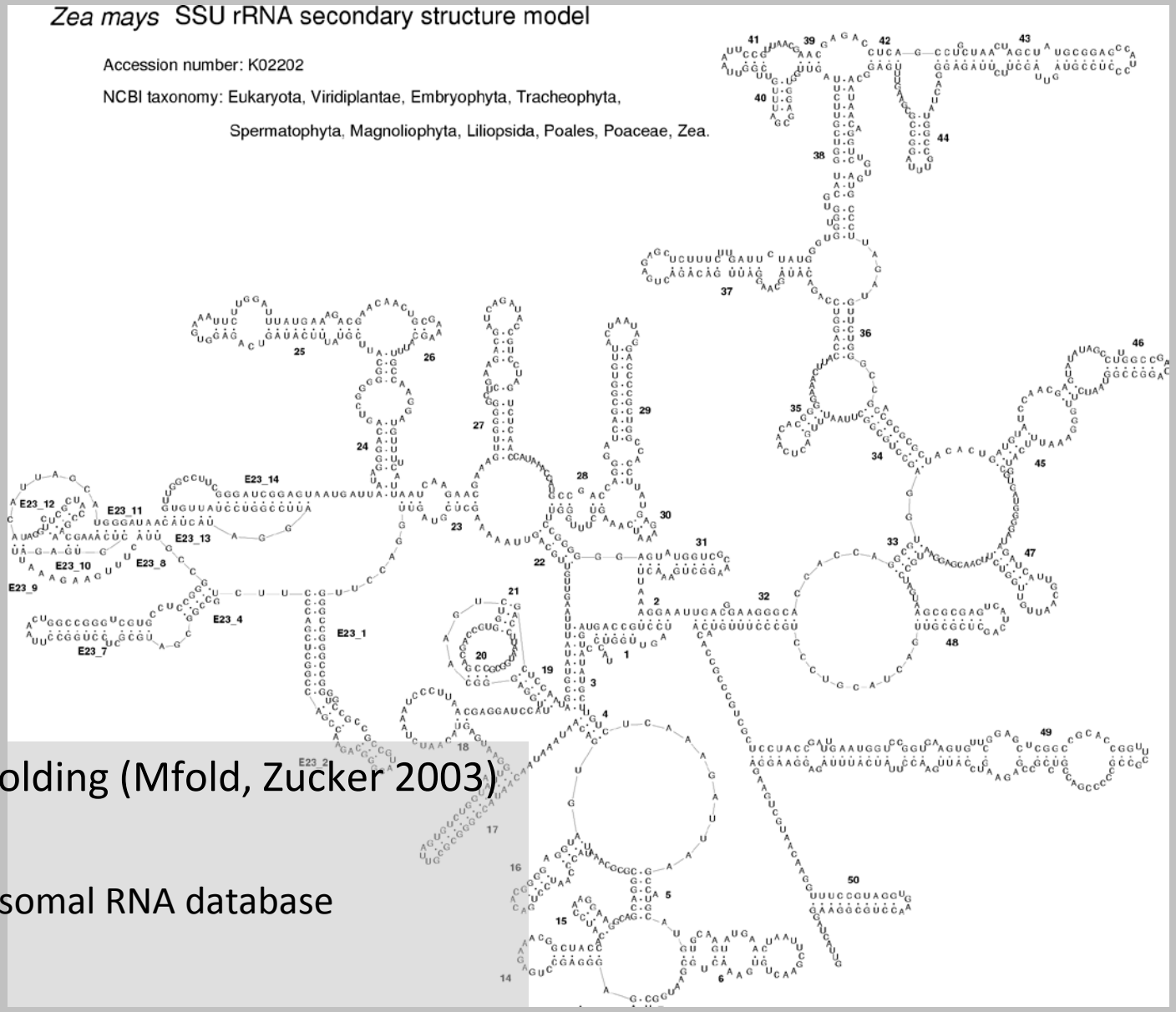
- délka mezi 3375 a 3393 bp
- velmi konzervovaná
- vyvíjí se 1,6-2,2x rychleji než 18S rDNA
- obsahuje konzervované úseky a expanzní segmenty
- konzervované úseky (*conserved core region*) – systematika na vyšších úrovních
- *expansion segments* – vyvíjí se až 10× rychleji

Model sekundární struktury rRNA

- *loop* – smyčka
  stem – stonek
- Watson-Crickovské párování A-U a G-C, ale často i G-U
- *CBC - compensatory base changes*
  (takové substituce, aby byla udržena *stem* struktura)
- tj. baze nejsou nezávislé – jiná váha než *loop*



Sekundární struktura RNA



- predikce – RNA folding (Mfold, Zucker 2003)
- databáze
  - European ribosomal RNA database
  - ITS2

Low-copy nukleární markery

- geny nacházející se v genomu pouze v několika málo kopiích
- × multiple copy – stovky až desetitisíce kopií (nrDNA…)
- vyšší variabilita než ITS i nekodující cpDNA
- problém homologie – paralogy × orthology × homeology

Low-copy markery

Výhody

- vyšší rychlost evoluce sekvencí než organelární genom
- přítomnost mnoha nezávislých (nesouvisejících) lokusů
- biparentální dědičnost

Nevýhody

- komplexnější genetická struktura (genová duplikace)
- obtížnost izolace a identifikace orthologních lokusů
- přítomnost variability vnitrodruhové, vnitropopulační a na úrovni jedince (heterozygosita)

Variabilita v rychlosti evoluce
(*rate variation*)

- synonymní substituce – 5× rychlejší než u cpDNA genů a 20× rychlejší než mtDNA
- např. vztahy v rodu *Gossypium* (Small et al. 1998)
  - 7000 bazí nekodující cpDNA poskytlo neúplné a slabě podpořené rozlišení
  - 1650 bazí *Adh*C – úplné a robustní rozlišení
  - velké rozdíly ve variabilitě (rychlosti mutací) mezi různými geny – až sedminásobné rozdíly
- tj. pro každou skupinu je potřeba testovat více markerů a vybrat ty variabilní

Struktura eukaryotních genů

- *5´ UTR (untranslated region)* – promotory pro genovou regulaci (konzervované), občas obsahují vysoce variabilní introny
- *exony* – více konzervované na nesynonymních místech (první a druhá pozice kodonu), na synonymních třetích pozicích kodonu podobné nekodujícím úsekům
- *introny* – méně funkčních omezení na úrovni sekvence, často omezená délka
- *3´ UTR (untranslated region)* – kontrolují zpracování mRNA a přidání poly-A signálu, ale jinak také často velmi variabilní
- různá funkce, různé evoluční omezení

Mnoho nezávislých lokusů

multiple unlinked loci

- použití pro nezávislou rekonstrukci evoluce
- markery na různých chromozomech (nebo dostatečně daleko od sebe na jednom chromozomu) – navzájem evolučně nezávislé
- nesoulad mezi různými markery – použití k detekci např. hybridizace, introgrese nebo *incomplete lineage sorting*

Biparentální dědičnost

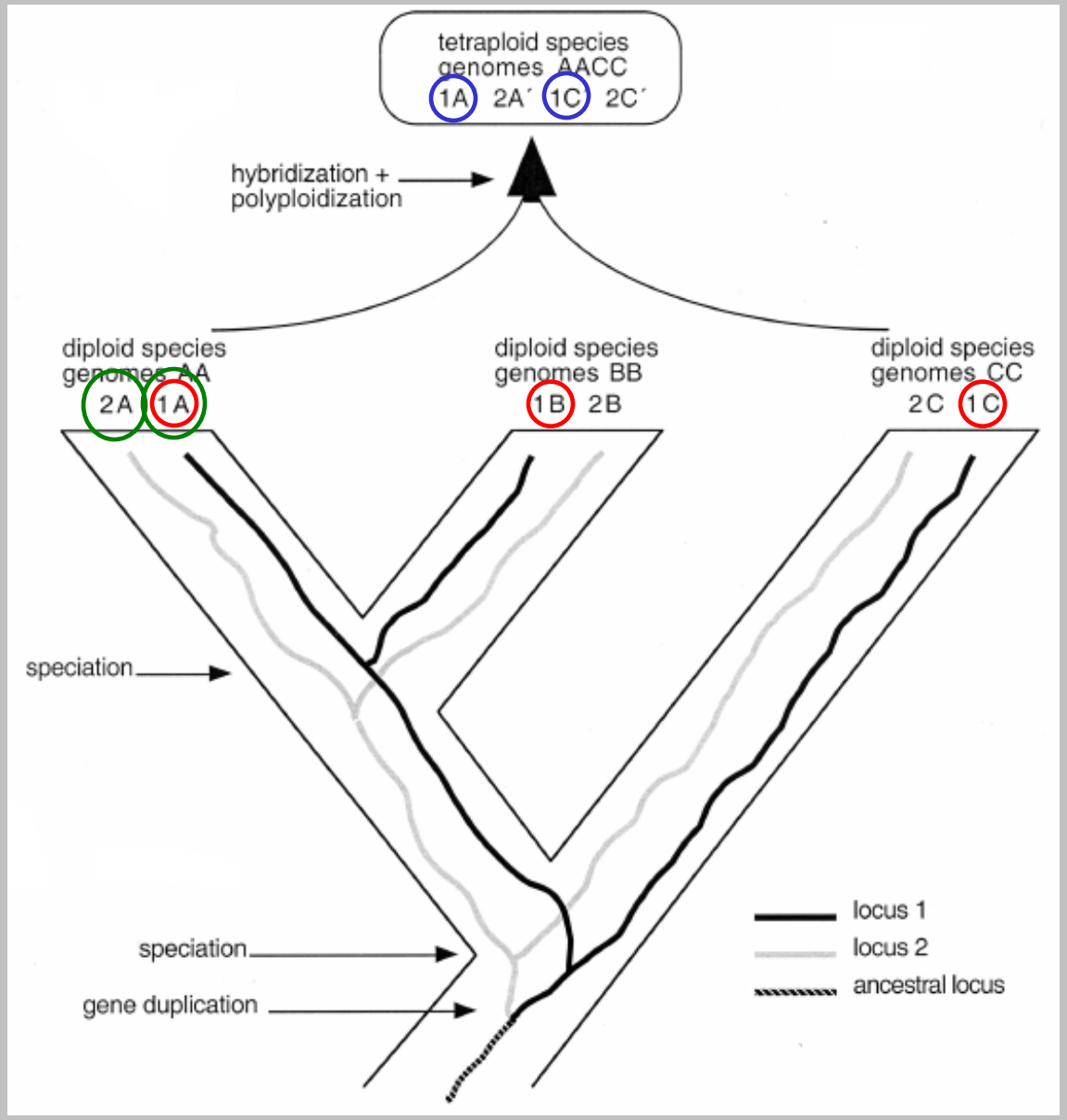
- low copy markery jsou méně často subjektem *concerted evolution*

- tj. jsou ideálními kandidáty pro identifikaci rodičovských kombinací předpokládaných ***hybridů*** nebo ***polyploidů*** (použito v rodě *Gossypium*, *Paeonia*, *Clarkia*, *Elymus, Silene, Cerastium, Bromus*…)

Genové rodiny

- mnohonásobné kopie homologních genů vznikající duplikací
- genové rodiny se velmi liší ve velikosti
  - single copy – GBSSI u diploidních *Poaceae*
  - stovky kopií – aktiny, small heat-shock proteiny
- genová a genomová duplikace (a následná ztráta genů) – dynamický a trvalý proces
- charakteristika genové rodiny – specifické pro taxon (skupinu)
- charakteristika genové rodiny u jedné skupiny nemusí být aplikovatelná na jinou
  - *Adh* obecně – 1 až 3 lokusy
  - u r. *Gossypium* nebo *Pinus* – až 7 lokusů
- špatná charakterizace genové rodiny vede k chybné fylogenetické rekonstrukci (vždy je nutné porovnávat orthologní kopie!)

Paralogní, orthologní a homeologní geny



orthologní geny

- vznik speciací

paralogní geny

-vznik genovou duplikací

homeologní geny

- vznik polyploidizací

orthologní geny

1A-1B-1C

paralogní geny

1A-2A, 1B-2B, 1C-2C

1A-2B, 1A-2C atd…

homeologní geny

1A´-1C´, 2A´-2C´

Studium orthologních sekvencí

1. design *universálních primerů* – produkují více PCR produktů různé délky → charakteristika genové rodiny (identifikace počtu lokusů)
2. vyvinutí *lokusově specifických primerů* – amplifikují pouze orthology

- evidence ortologie
  - celková sekvenční podobnost (orthology se navzájem více podobají než paralogy)
  - *expression pattern* – orthologní sekvence sdílejí stejná pattern
  - *Southern hybridisation analysis* – hybridizace lokusově specifických prob k restrikčnímu pattern genomické DNA – počet proužků = počet lokusů
- velké rozdíly ve variabilitě mezi jednotlivými geny a lokusy – nutná předběžná studie k určení dostatečné variability

Intraspecifická variabilita

- alelická variabilita v rámci a mezi populacemi
- *coalescence within species* - pokud se alely vyvinuly v rámci jednoho druhu nenarušuje to správné určení fylogeneze, tj. je to užitečná variabilita pro vnitrodruhové studie – populační, fylogeografické apod.
- *deep coalescence (incomplete lineage sorting)* - alelická variabilita přesahuje hranice druhu, tj. některé alely jsou příbuznější alelám z jiného druhu než některým alelám z toho samého druhu – pravděpodobnější u druhů s vysokými počty jedinců v populacích
- lokusy podrobené balancující selekci (udržuje vysokou alelickou variabilitu) – nevhodné pro fylogenetické rekonstrukce
  - např. self-incompatibility geny u *Solanaceae* – alelická variabilita přesahuje druhové a dokonce rodové hranice
- také díky hybridizaci a introgresi

Rekombinace

1. *alelická rekombinace*
   - rekombinace na úrovni individuálního lokusu
   - generování alelické variability
   – narušuje předpoklad bifurkátních vztahů mezi alelami
   – vnáší retikulátní evoluci
   – neporušuje možnost správné rekonstrukce fylogeneze, pokud jsou alely v rámci druhu monofyletické
2. *nehomologní rekombinace*
   - rekombinace mezi paralogními lokusy
   - může být pouze sporadická nebo být specifická pro příslušný gen

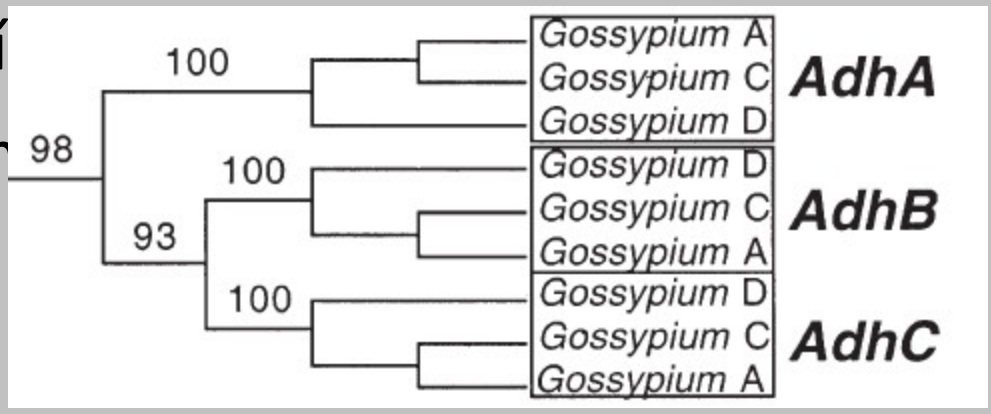
Concerted evolution

- běžná u vysoce repetitivních lokusů (nrDNA)
- vyskytuje se i u low copy markerů - *nehomologní rekombinace*

- *nevyskytuje se* → sekvenování všech genů genové rodiny dává tzv. orthology-paralogy tree (OP-tree)
- *úplná* (předpokládá se u nrDNA) → sekvenování jakéhokoliv genu genové rodiny poskytne správný fylogenetický strom
- *neúplná* → směs ortologních a neúplně homogenisovaných paralogních sekvencí, tj. správná rekonstrukce fylogeneze je prakticky nemožná

Concerted evolution

- běžná u vysoce repetitivní
- vyskytuje se i u low copy n
  rekombinace



- *nevyskytuje se* → sekvenování všech genů genové rodiny dává tzv. orthology-paralogy tree (OP-tree)
- *úplná* (předpokládá se u nrDNA) → sekvenování jakéhokoliv genu genové rodiny poskytne správný fylogenetický strom
- *neúplná* → směs ortologních a neúplně homogenisovaných paralogních sekvencí, tj. správná rekonstrukce fylogeneze je prakticky nemožná

PCR-mediated recombination

- *in vitro* nehomologní rekombinace
- zdrojem jsou
  - 1. výměna templátu během PCR
  - 2. nekompletně prodloužené kopie jednoho lokusu, které slouží jako primery v následné extenzi z paralogního lokusu
- závisí na
  - stupni podobnosti sekvencí mezi paralogními lokusy
  - univerzalitě/specificitě primerů
  - podmínkách PCR (nutná optimalizace teploty annealingu, délky produktu a doby extenze)

Procedura determinace vhodných nukleárních markerů pro fylogenetické analýzy

1. selekce kandidátních genů (a representativních taxonů) pro předběžnou studii
2. isolace kandidátních genů z representativních taxonů
3. určení ortologie mezi izolovanými sekvencemi
4. určení relativní rychlosti evoluce sekvencí – vybrání vhodného lokusu
5. generování sekvencí ze studovaných taxonů v daném lokusu

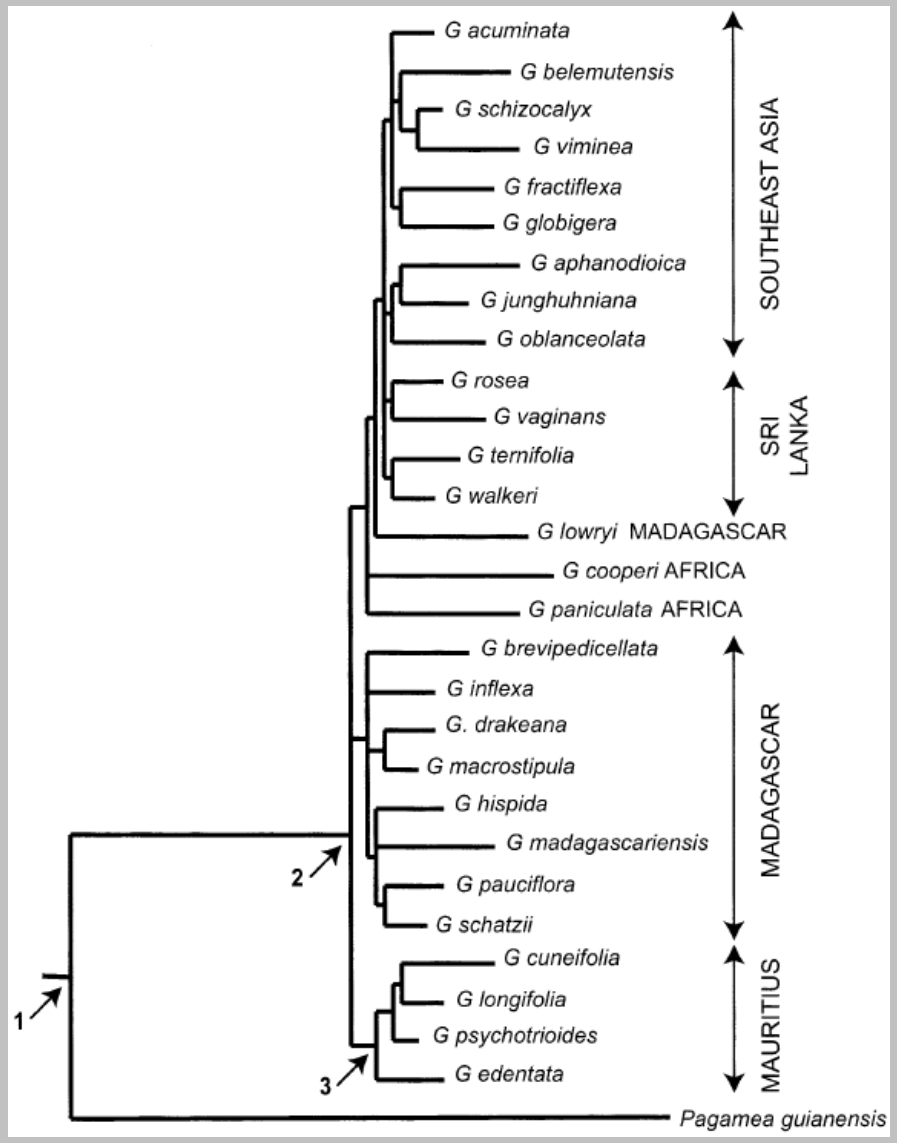
Selekce kandidátních genů

- není žádný důvod předpokládat, že nějaký konkrétní gen bude univerzálně použitelný
- není nutné používat často sekvenované geny (*Adh*…)
- lze použít málo prozkoumané geny nebo dokonce anonymní nukleární lokusy
- kde tedy začít s hledáním?
  - využití předchozích studií v dané skupině
  - prohledání literatury ohledně využitelnosti genu na dané taxonomické úrovni v různých skupinách
  - GenBank, EMBL… - kombinace hledání podle taxonu a názvu genu
  - BLAST search – hledání podobných sekvencí $\rightarrow$ využitelné pro design primerů, identifikaci genové struktury (exony, introny)…

Využití single-copy genů

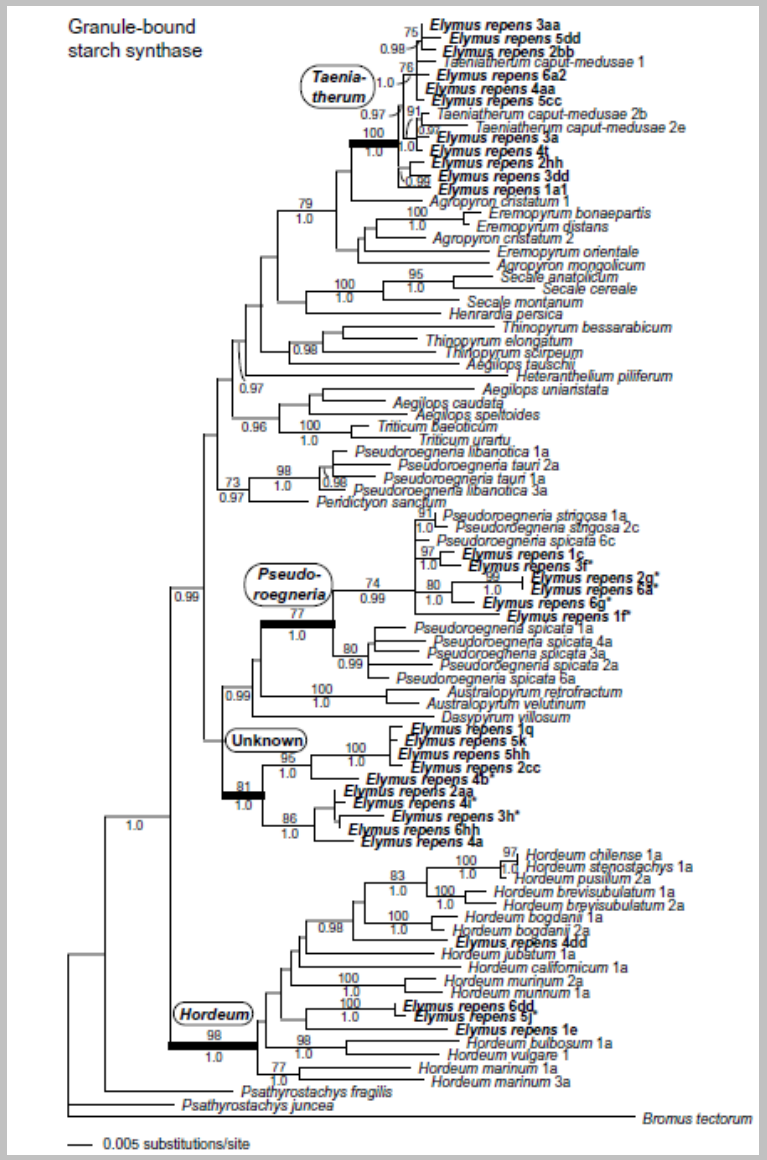
- fylogenetické studie – mohou poskytnout dostatek variability pro úplné rozlišení vztahů na nižší úrovni (např. blízce příbuzné druhy)
- studium polyploidů – „vyklonování“ jednotlivých rodičovských sekvencí a identifikace komlexního pattern vzniku allopolyploidů
- fylogeografie

Vztahy mezi blízce příbuznými druhy



Gaertnera
PepC, Tpi
Malcomber 2002

Původ allohexaploida



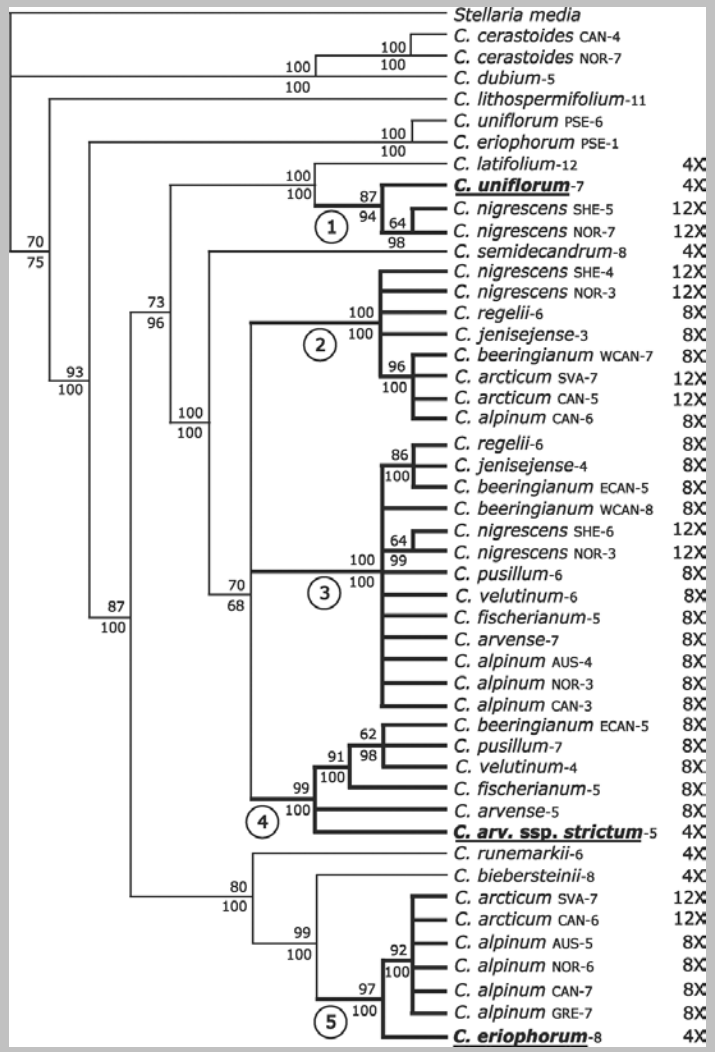
Taeniatherum

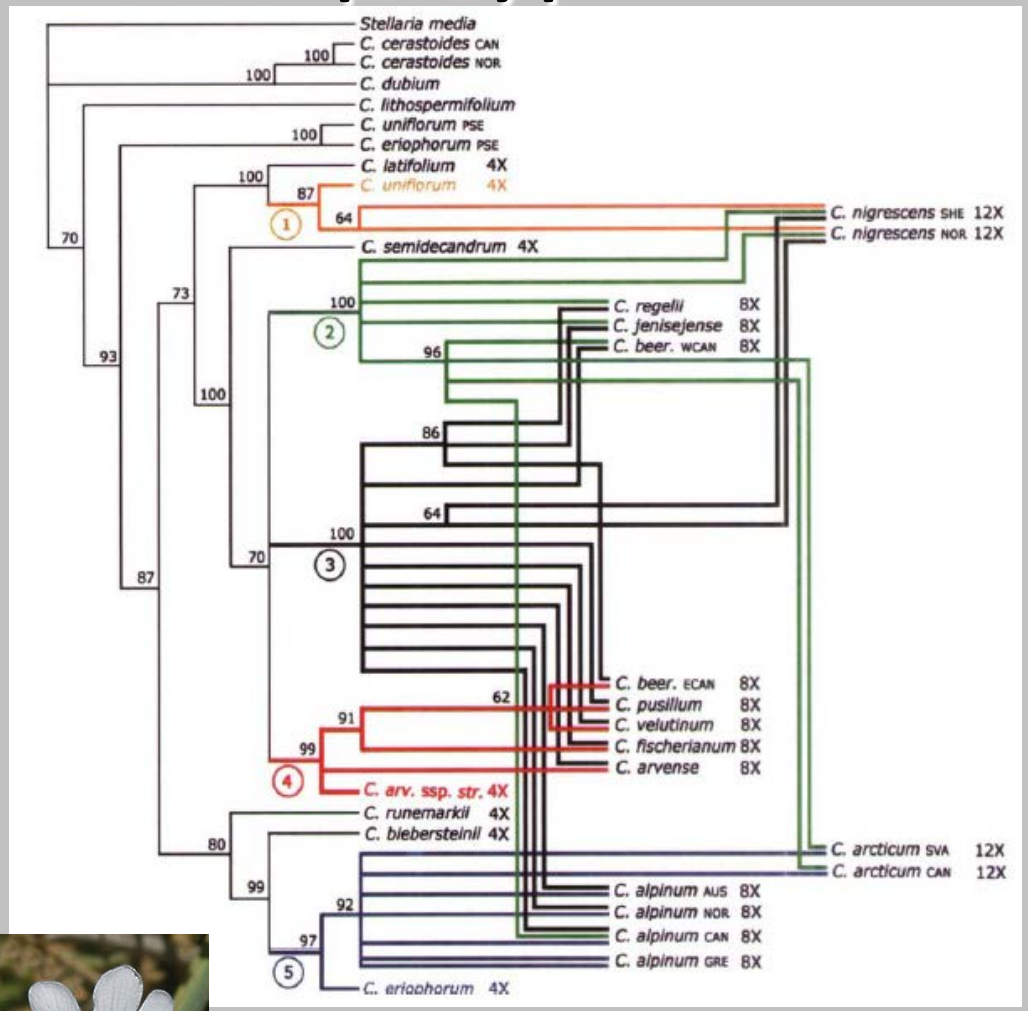
Pseudoroegneria

Hordeum

Elymus repens
GBSSI (single-copy)
Mason-Gamer 2008

Hledání rodičů allopolyploidů





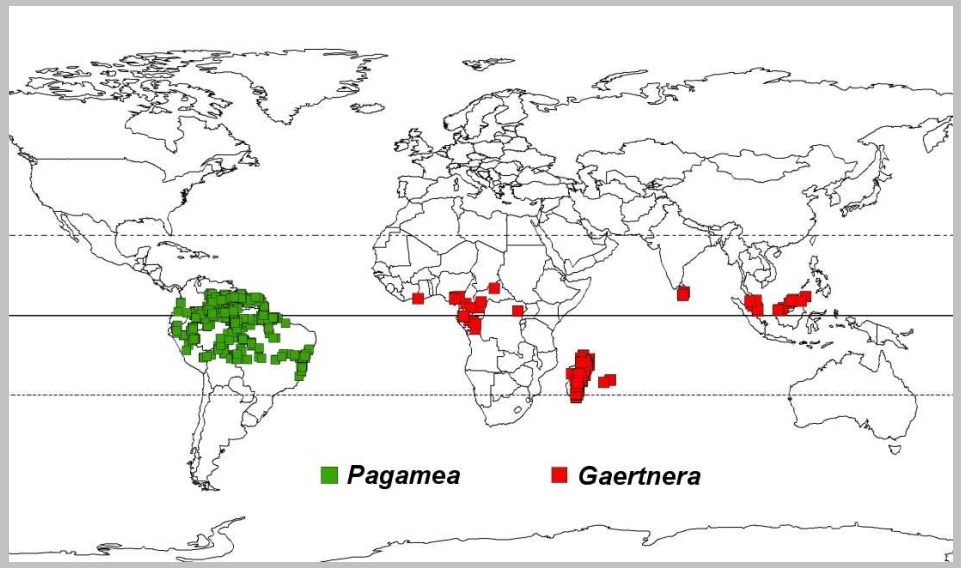
Cerastium

4x, 8x, 12x

RPB2 strom, network

Brysting et al. 2007

Malcomber S.T. (2000): Phylogeny of *Gaertnera* Lam. (Rubiaceae) based on multiple DNA markers: evidence of a rapid radiation in a widespread, morphologically diverse genus. *Evolution* 56(1):42-57



Literatura

- Small R.L., Cronn R.C. & Wendel J.F. (2004): *Use of nuclear genes for phylogeny reconstruction*. Australian Systematic Botany 17: 145-170
- Hughes C.E., Eastwood R.J. & Bailey C.D. (2006): *From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction*. Phil. Trans. R. Soc. B 361: 211-225
- Wu F. et al. (2006): *Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid plant clade*. Genetics 174: 1407-1420
- Li M. et al. (2008): *Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species*. Cladistics 24: 1-19.
- Alvarez I. & Wendel J.F. (2003): *Ribosomal ITS sequences and plant phylogenetic inference*. Molecular Phylogenetics and Evolution 29: 417–434