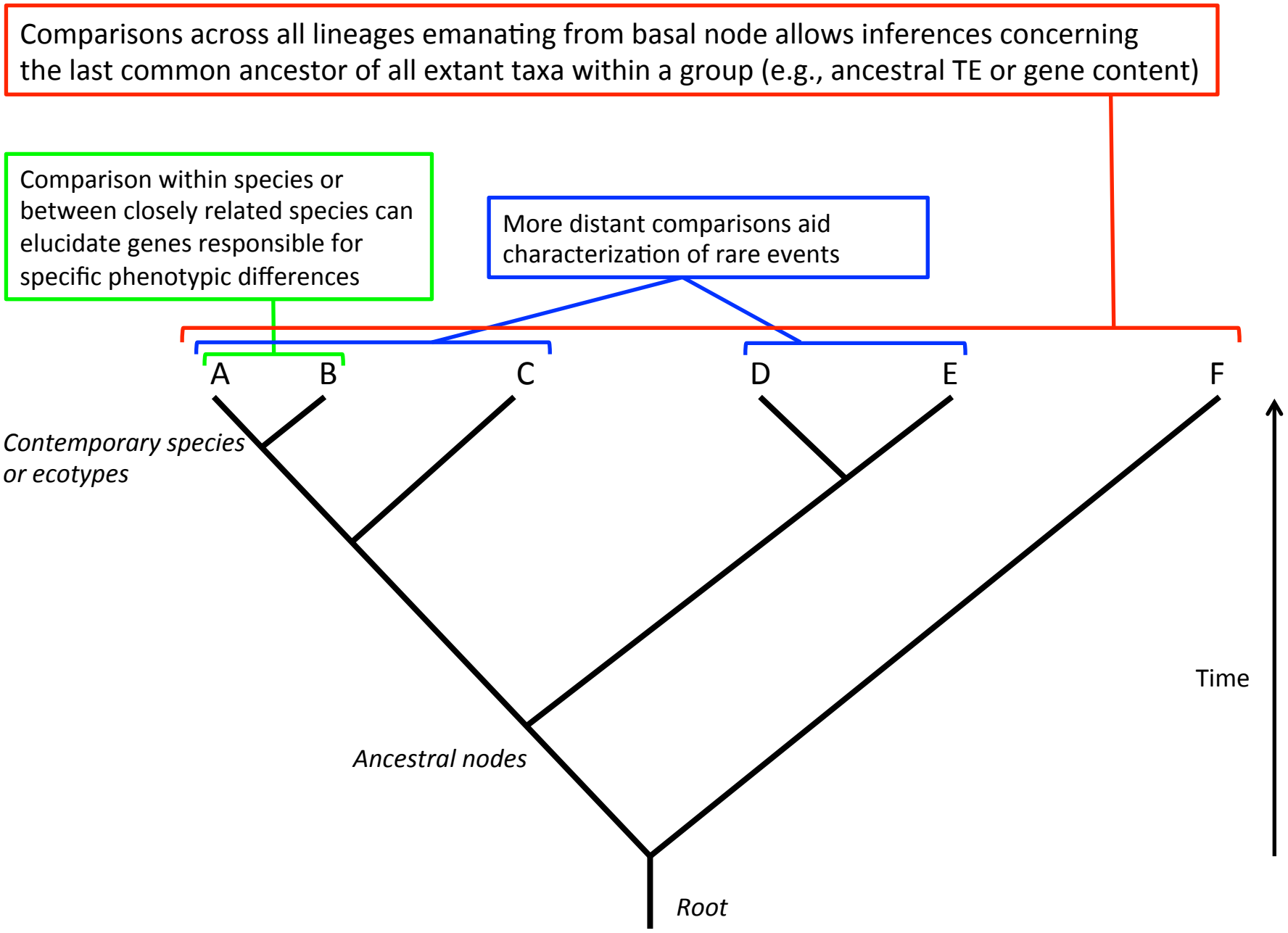
UBC Bioinformatics

Topic 10: Phylogenomics

What is phylogenomics?

1. Using genome-scale data to infer phylogenetic relationships

2. Genome-scale comparisons placed in a phylogenetic context


Comparisons across all lineages emanating from basal node allows inferences concerning the last common ancestor of all extant taxa within a group (e.g., ancestral TE or gene content)
Comparison within species or between closely related species can elucidate genes responsible for specific phenotypic differences
More distant comparisons aid characterization of rare events
A
B
C
D
E
F
Contemporary species or ecotypes
Ancestral nodes
Root
Time

Advances in sequencing technology are fueling phylogenomic studies

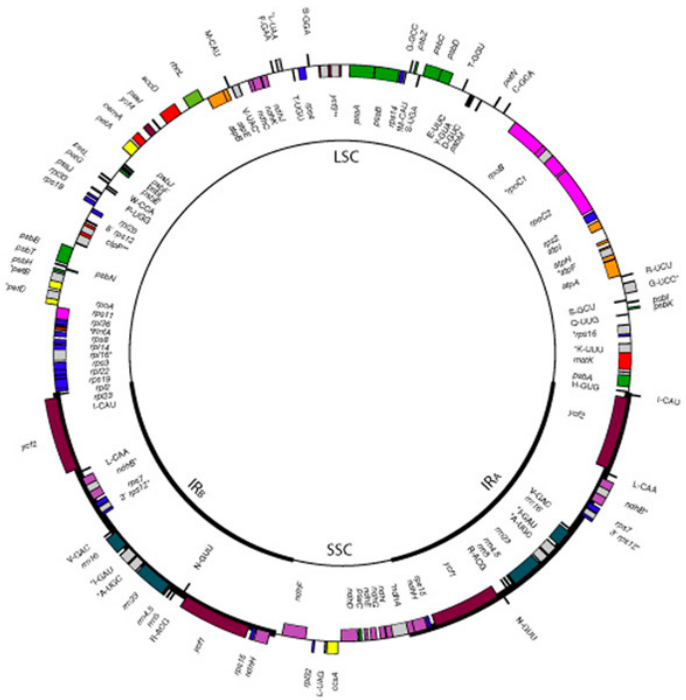
2nd generation sequencing

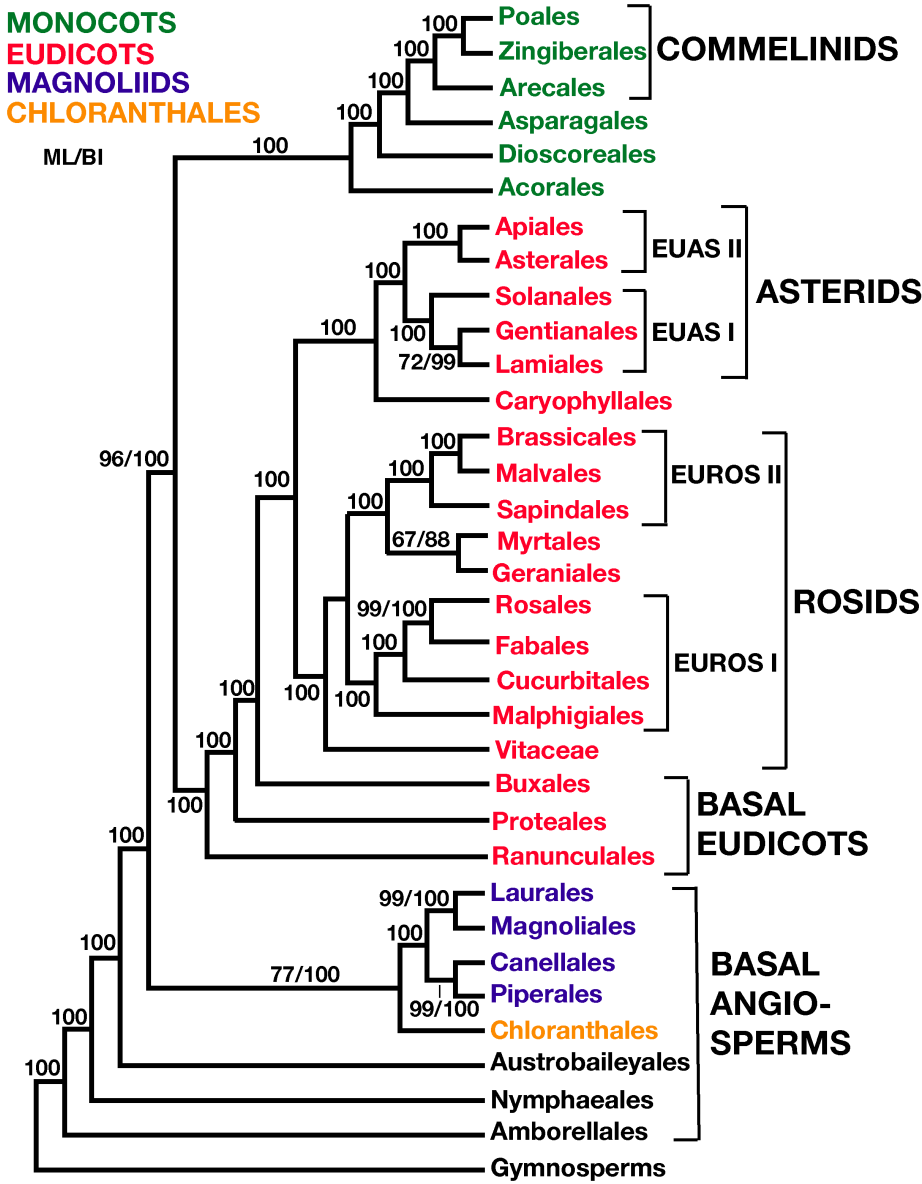
- Whole organelle genomes
- Targeted sequence capture
- Reduced representation sequencing
- Exon arrays
- CGH

3rd generation sequencing

- Whole nuclear genome sequencing
- Improved cytogenetic techniques?

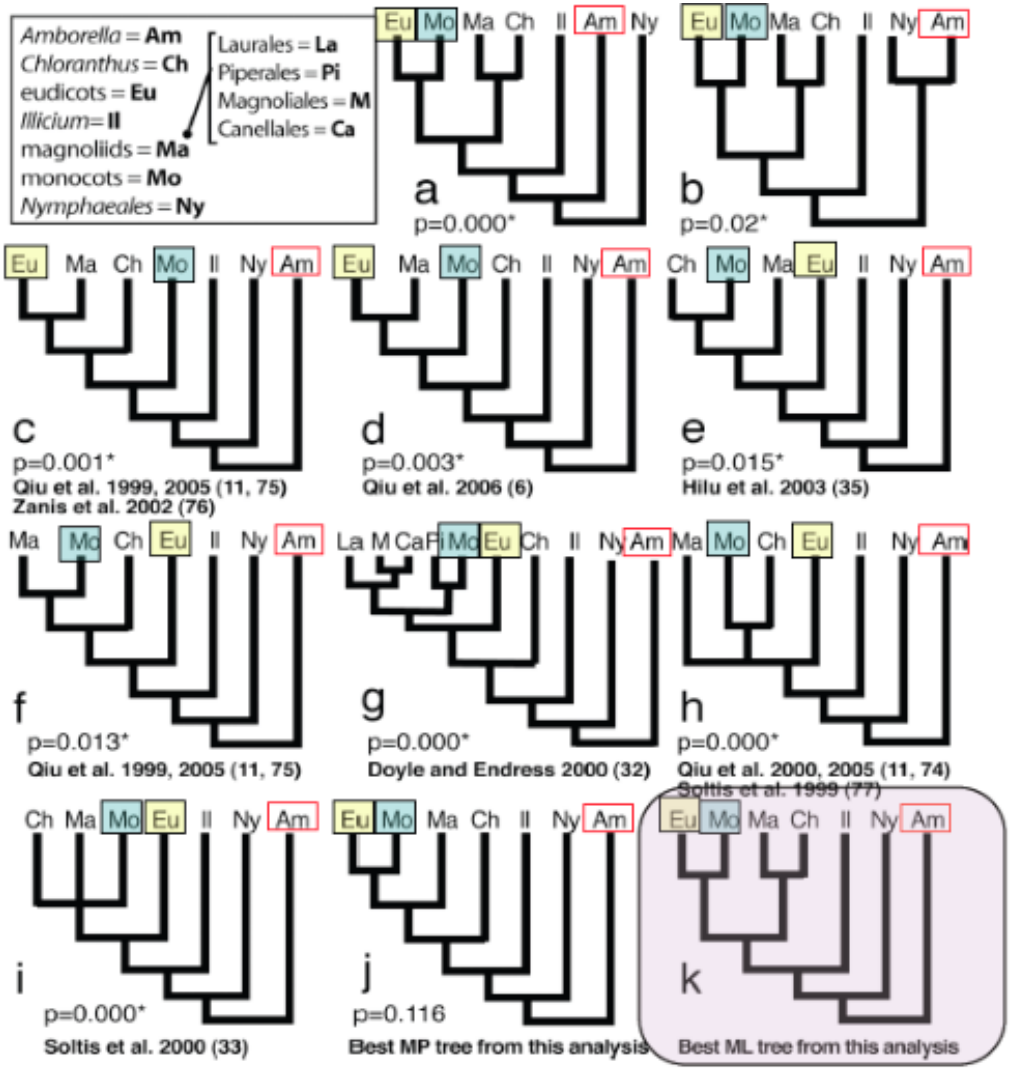
Plastid genome phylogeny allows the resolution of many previously intractable questions





Jansen et al. 2007 *PNAS*

SH, AU tests of alternative topologies

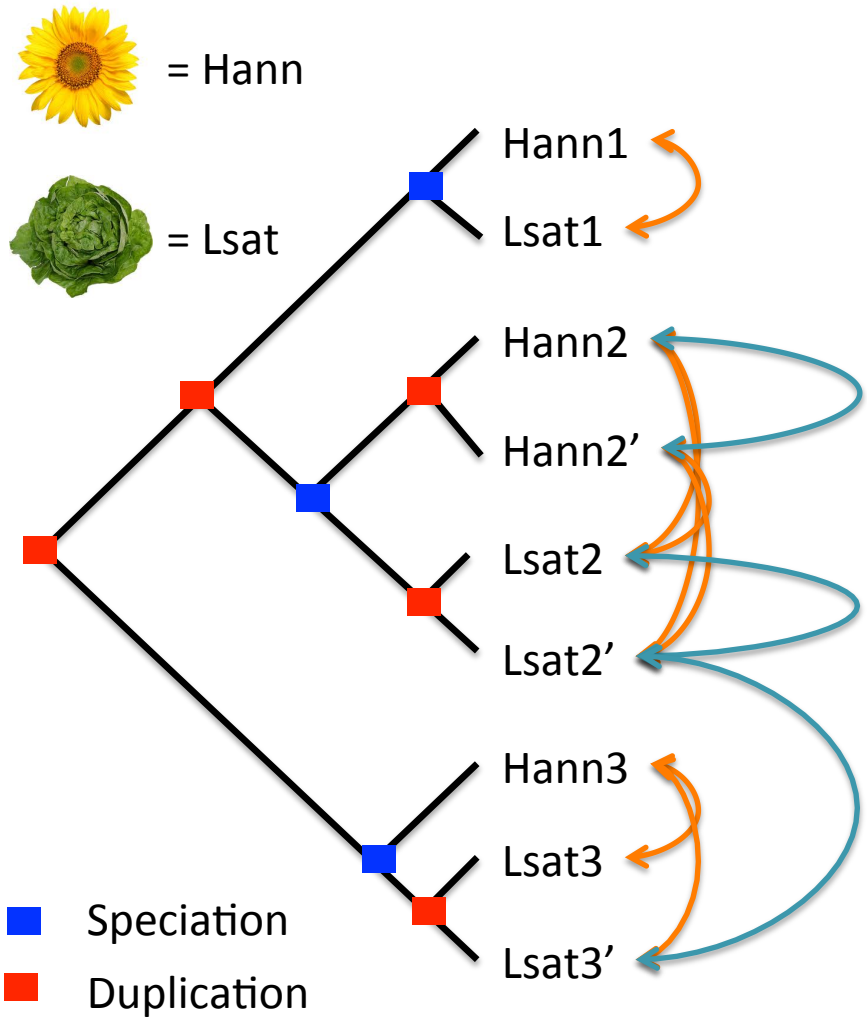


Jansen et al. 2007 *PNAS*

Challenges with nuclear genes

- Hybrid speciation, interspecific gene flow and incomplete lineage sorting
- Gene and genome duplication events (including allopolyploidy)
- Horizontal gene transfer (including endosymbiosis)

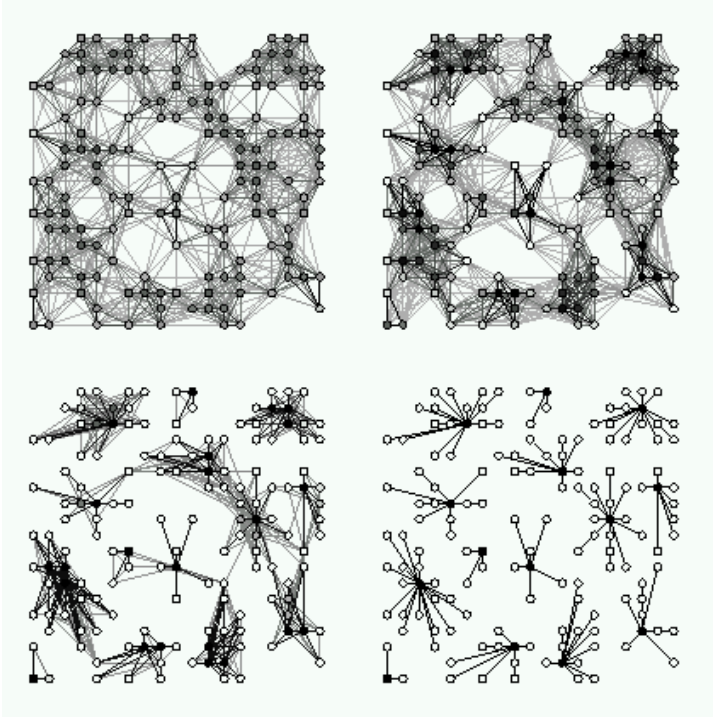
Ortholog identification

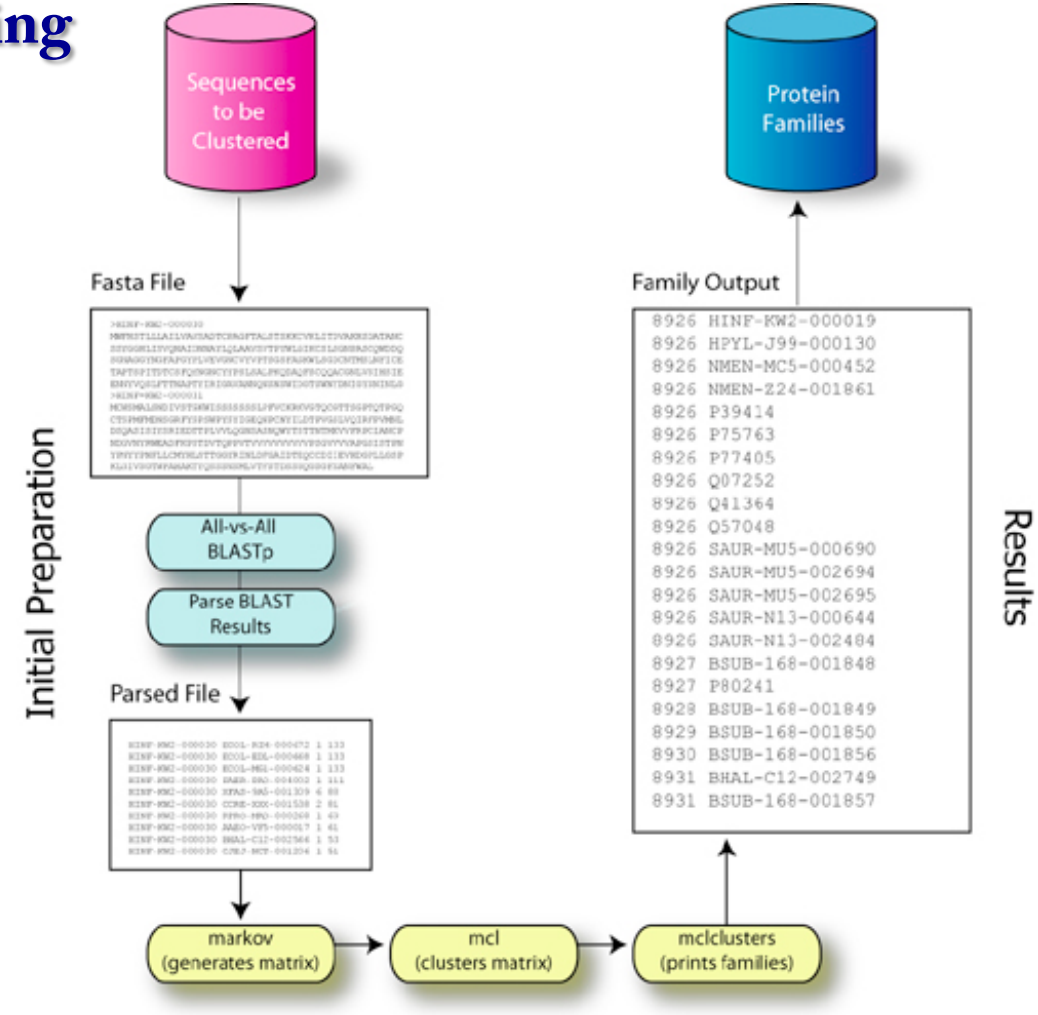


Orthologs are separated by speciation events (1:1 orthologs)

Paralogs are separated by duplication events (In-paralogs)

Markov Clustering (MCL) used for de novo grouping of genes into gene families





Enright 2002

Automatic transfer of manual annotations to orthologs

e.g. Human

Ensembl compara

Macaque

Chimpanzee

Guinea Pig

Rat

Mouse

Cow

Dog

Chicken

...and more

Arabidopsis

Rice

Ensembl compara

Poplar

Maize

Brachypodium

Grape

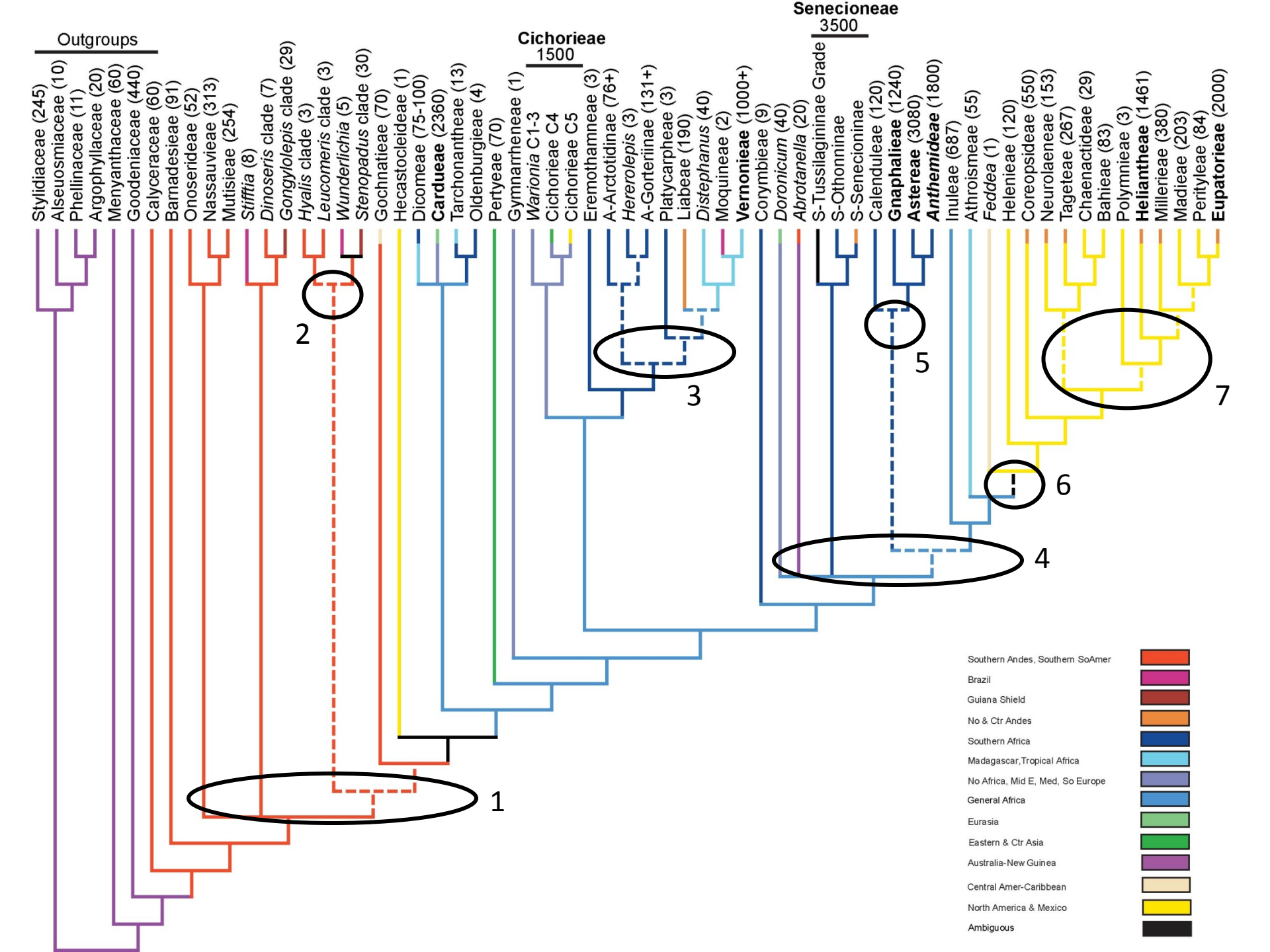
…and more

Ways to capture information about genes

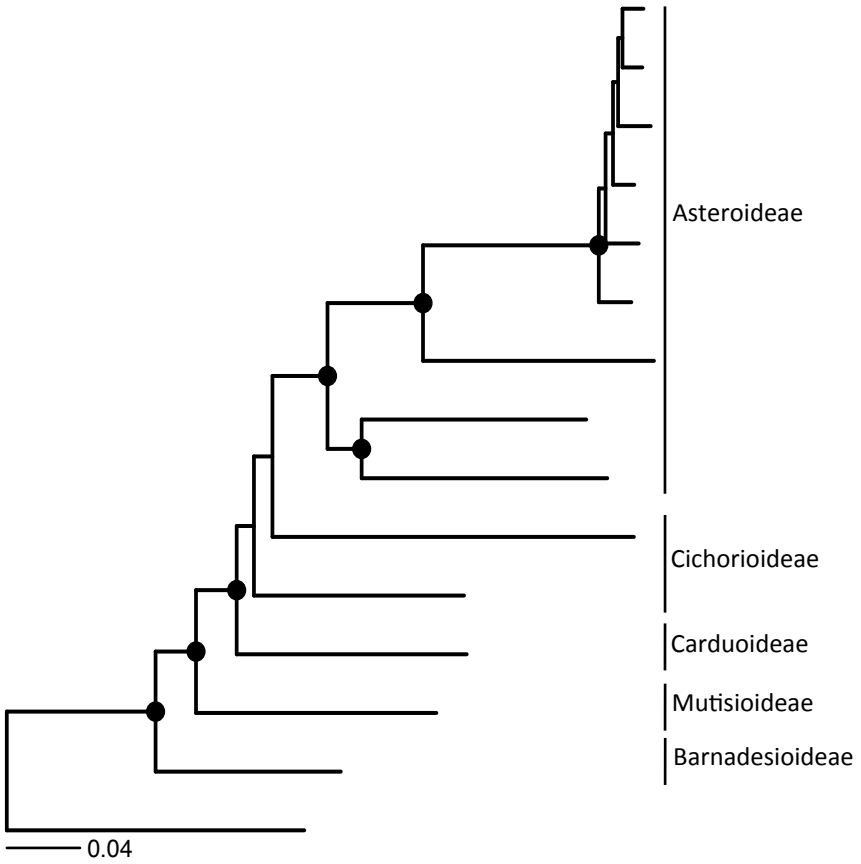
1. Targeted sequence capture
2. Remove repeats enzymatically
3. Sequence the whole genome
4. Sequence the transcriptome

Ways to capture information about genes

1. **Targeted sequence capture**
2. Remove repeats enzymatically
3. Sequence the whole genome
4. Sequence the transcriptome


Outgroups
Cichorieae
1500
Senecioneae
3500
Stylidiaceae (245)
Alseuosmiaceae (10)
Phellinaceae (11)
Argophyllaceae (20)
Menyanthaceae (60)
Goodeniaceae (440)
Calyceraceae (60)
Barnadesieae (91)
Onoserideae (52)
Nassauvieae (313)
Mutisieae (254)
Stifftia (8)
Dinoseris clade (7)
Gongylolepis clade (29)
Hyalis clade (3)
Leucomeris clade (3)
Wunderlichia (5)
Stenopadus clade (30)
Gochnatieae (70)
Hecastocleideae (1)
Dicomeae (75-100)
Cardueae (2360)
Tarchonantheae (13)
Oldenburgieae (4)
Pertyeae (70)
Gymnarrheneae (1)
Warionia C1-3
Cichorieae C4
Cichorieae C5
Eremothamneae (3)
A-Arctotidinae (76+)
Hererolepis (3)
A-Gorteriinae (131+)
Platycarpheae (3)
Liabeae (190)
Distephanus (40)
Moquineae (2)
Vernonieae (1000+)
Corymbieae (9)
Doronicum (40)
Abrotanella (20)
S-Tussilagininae Grade
S-Othonninae
S-Senecioninae
Calenduleae (120)
Gnaphalieae (1240)
Astereae (3080)
Anthemideae (1800)
Inuleae (687)
Athroismeae (55)
Feddea (1)
Helenieae (120)
Coreopsideae (550)
Neurolaeneae (153)
Tageteae (267)
Chaenactideae (29)
Bahieae (83)
Polymnieae (3)
Heliantheae (1461)
Millerieae (380)
Madieae (203)
Perityleae (84)
Eupatorieae (2000)
1
2
3
4
5
6
7
Southern Andes, Southern SoAmer
Brazil
Guiana Shield
No & Ctr Andes
Southern Africa
Madagascar,Tropical Africa
No Africa, Mid E, Med, So Europe
General Africa
Eurasia
Eastern & Ctr Asia
Australia-New Guinea
Central Amer-Caribbean
North America & Mexico
Ambiguous

ML phylogeny from COS genes



Subfamily	Species
Asteroideae	*Helianthus argophyllus*
	Helianthus annuus
	Helianthus niveus ssp. tephrodes
	Helianthus verticillatus
	Helianthus porteri
	Phoebanthus tenufolius
	Conoclinium coelestinum
	Pseduognaphalium helleri
	Senecio vulgaris
Cichorioideae	*Taraxacum kok-saghyz*
	Centrapalus pauciflorus
Carduoideae	*Carthamus tinctorius*
Mutisioideae	*Gerbera hybrida*
Barnadesioideae	*Fulcaldea stuessyi*
	Nasanthus patagonicus

Sequencing

Every major subfamily represented

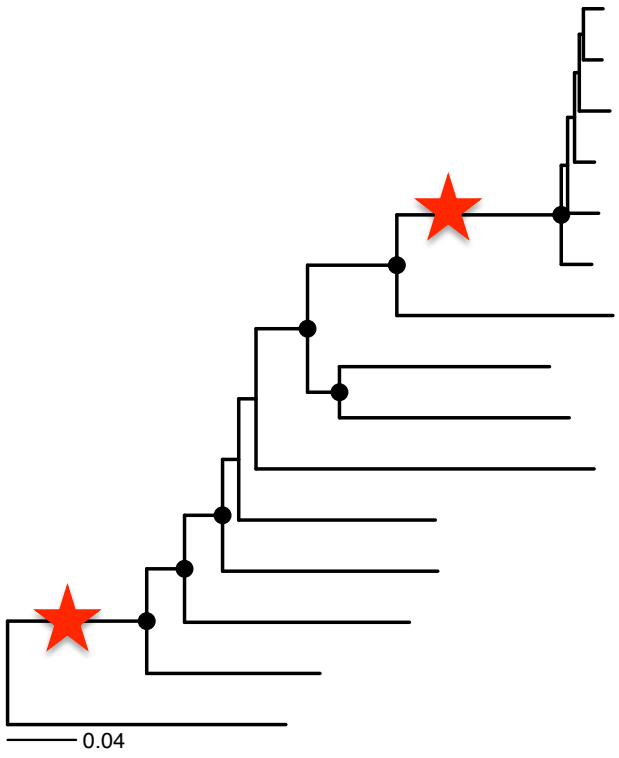
Six species in the tribe Heliantheae

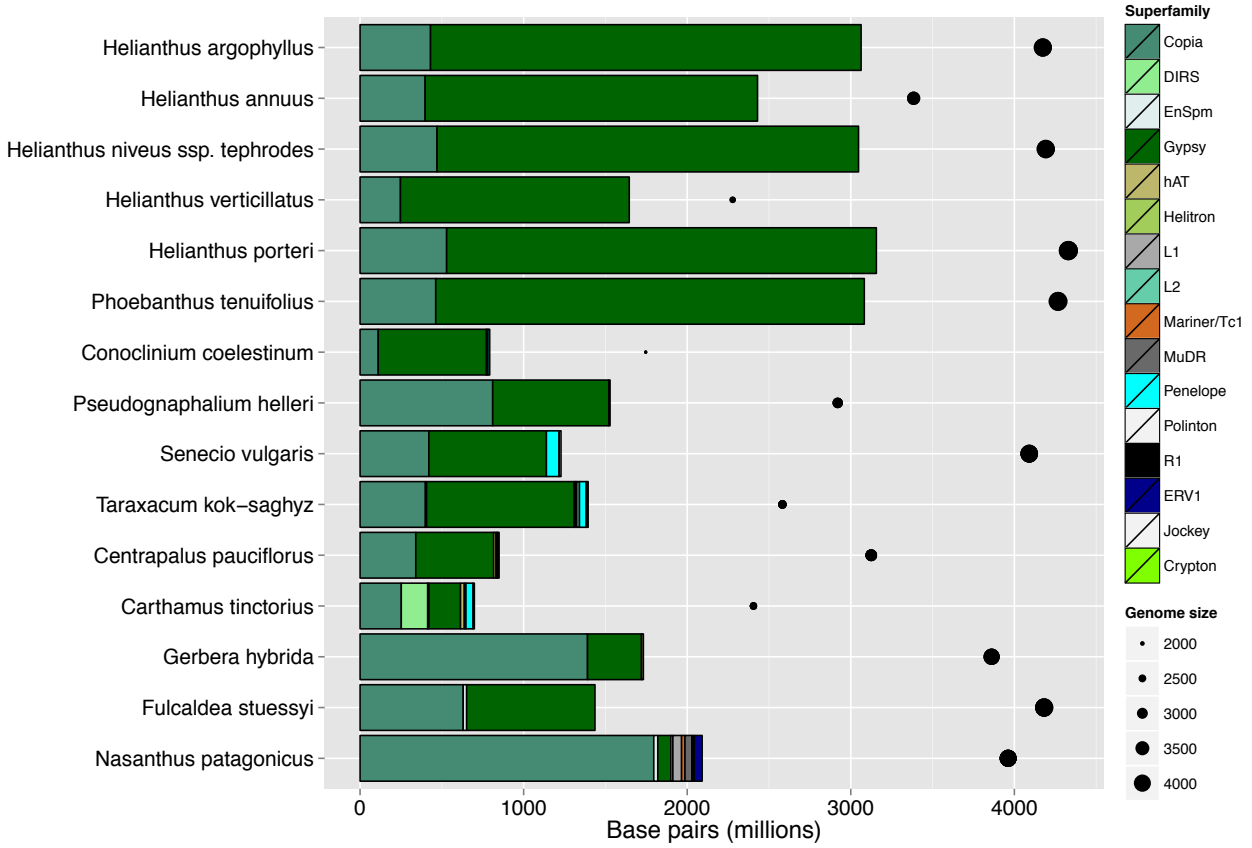
1-5X 100 bp Illumina paired-end WGS

763 nuclear loci targeted by sequence capture
(Mandel et al. 2014)

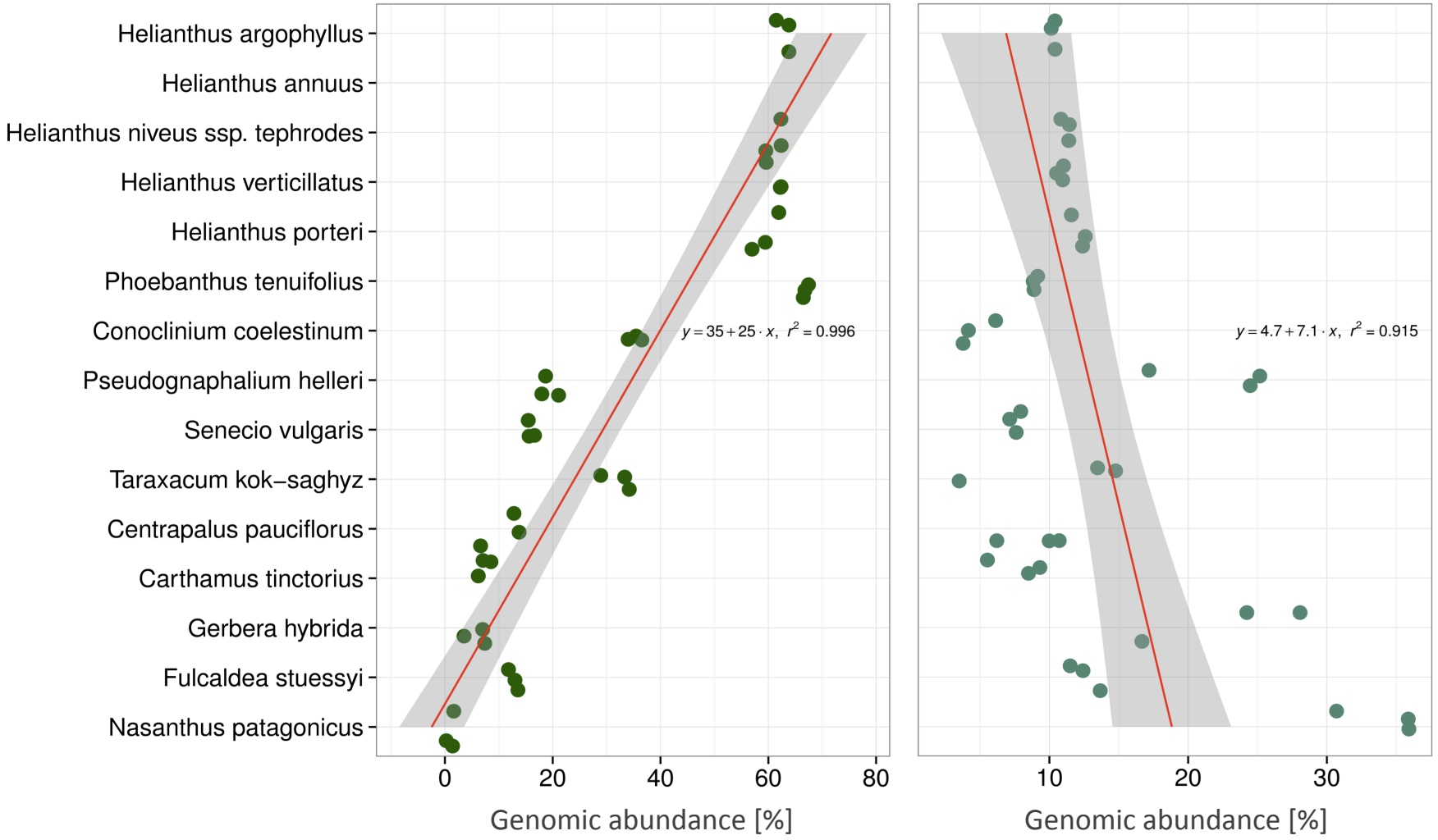
● >75% bootstrap support

Transitions in genome composition coincide with WGD

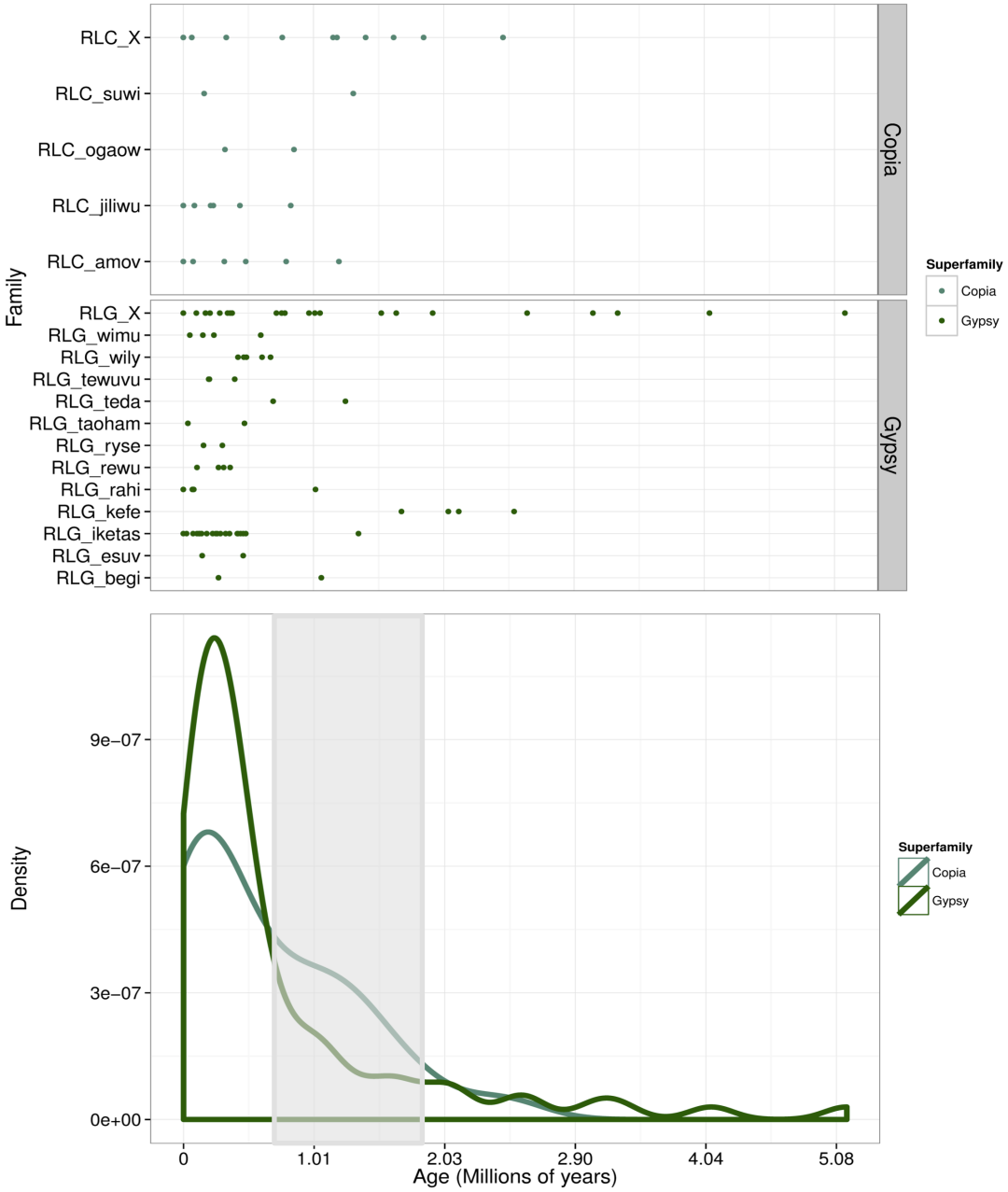




Non-random patterns of change in TE superfamily abundance



What is the time scale over which sunflower LTR elements have been active?



Est. origin of *H. annuus* (Strasburg and Rieseberg 2008)

Ways to capture information about genes

1. Targeted sequence capture
2. **Remove repeats enzymatically**
3. Sequence the whole genome
4. Sequence the transcriptome

Species	Genotype	Module	# reads (millions)	Output (Gbp)
H. annuus[1]	RHA801	76 PE	126.2	16.7
C. tinctorius[2]	AC Sunset	100 PE	166.3	32.8
L. sativa[3]	cv. Salinas	125 PE, 125 SE	347.2	56.7
			639.7	106.2

Genome sizes: [1]3.6 Gb, [2]1.36, [3]2.6

Gene family analysis methods

1) Find ortholog in grape for each gene family based on recip. best hit (~21k)
2) Select only those gene families containing all three species and outgroup (352)

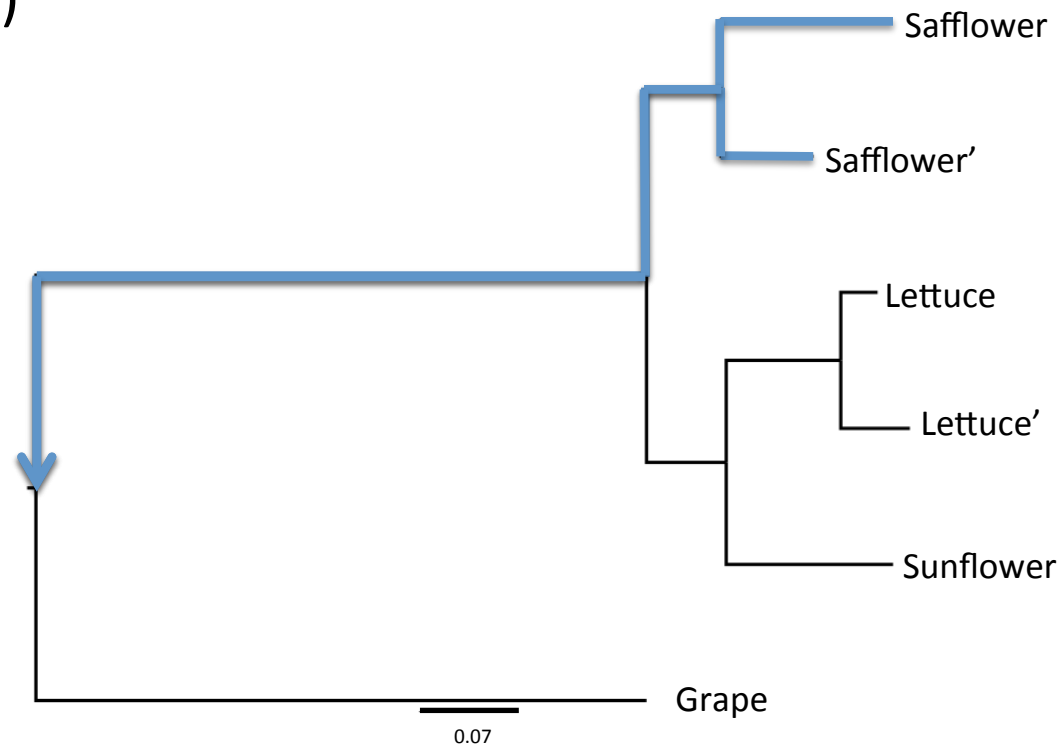
Inference of substitution patterns and gene trees

- Generate codon alignments (from protein sequences and nucleotide alignments)
- Determine most likely model of evolution (JModelTest), infer substitution patterns (codeml)
- “Rapid” ML method used to infer trees (RAxML); compare branch lengths between species

Calculation of evolutionary rates

- Determine the number of tree members for each species (e.g., Safflower = 2)
- Calculate path to root for each member (e.g., Safflower + Safflower')
- Divide by the number of members (n) if n > 1

Gene tree for myrcene synthase
(grape ortholog: GSVIVT01000401001)



Frequency distribution of branch lengths for all trees for each species

Calculation of evolutionary rates

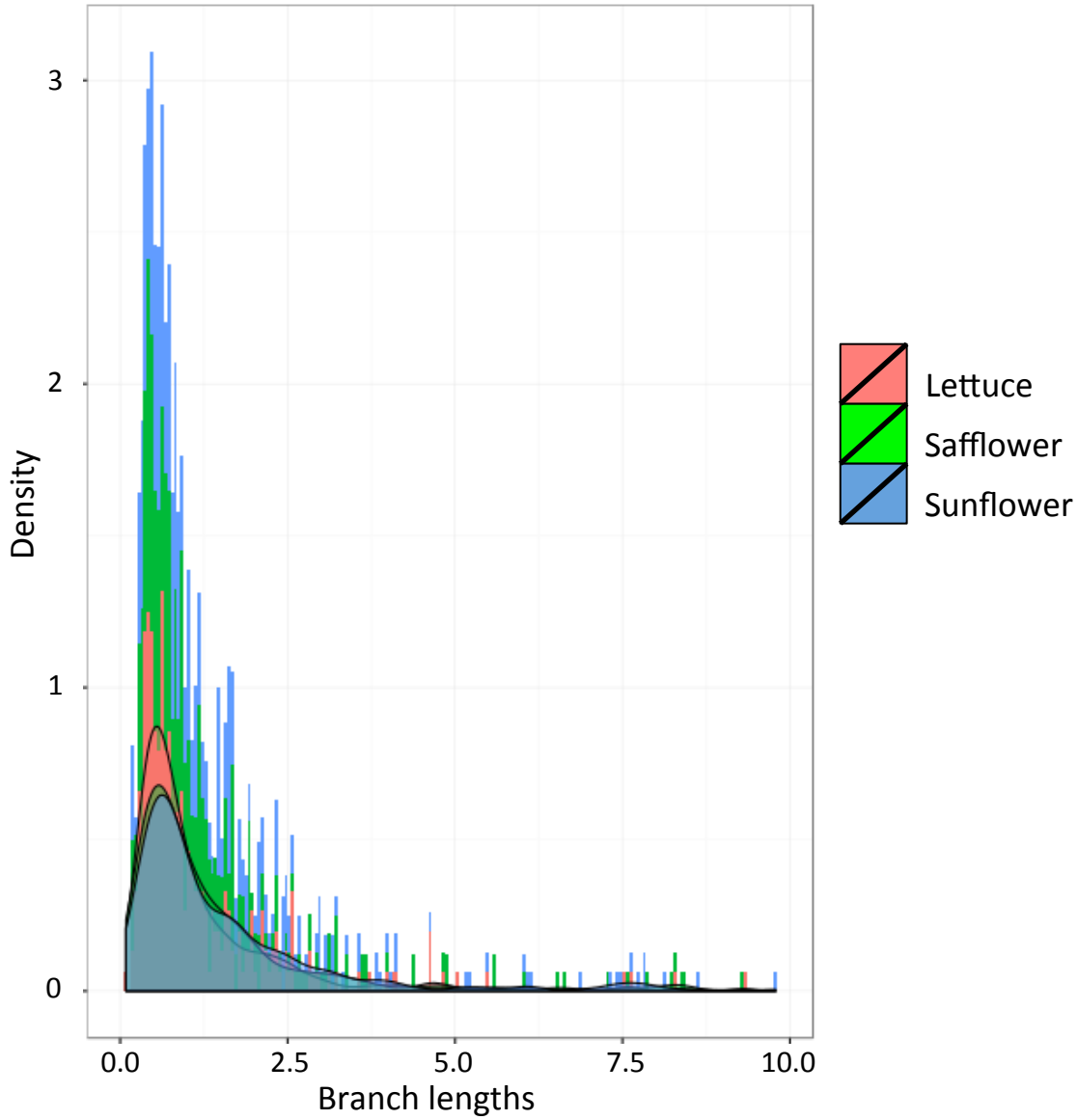
Frequency distribution of branch lengths for 352 gene trees

Species branch lengths per tree:

Short if length < median
Long if length > median

Possible results:

1) All Short branches
2) All Long branches
3) Mixture of Short/Long branches



Calculation of evolutionary rates

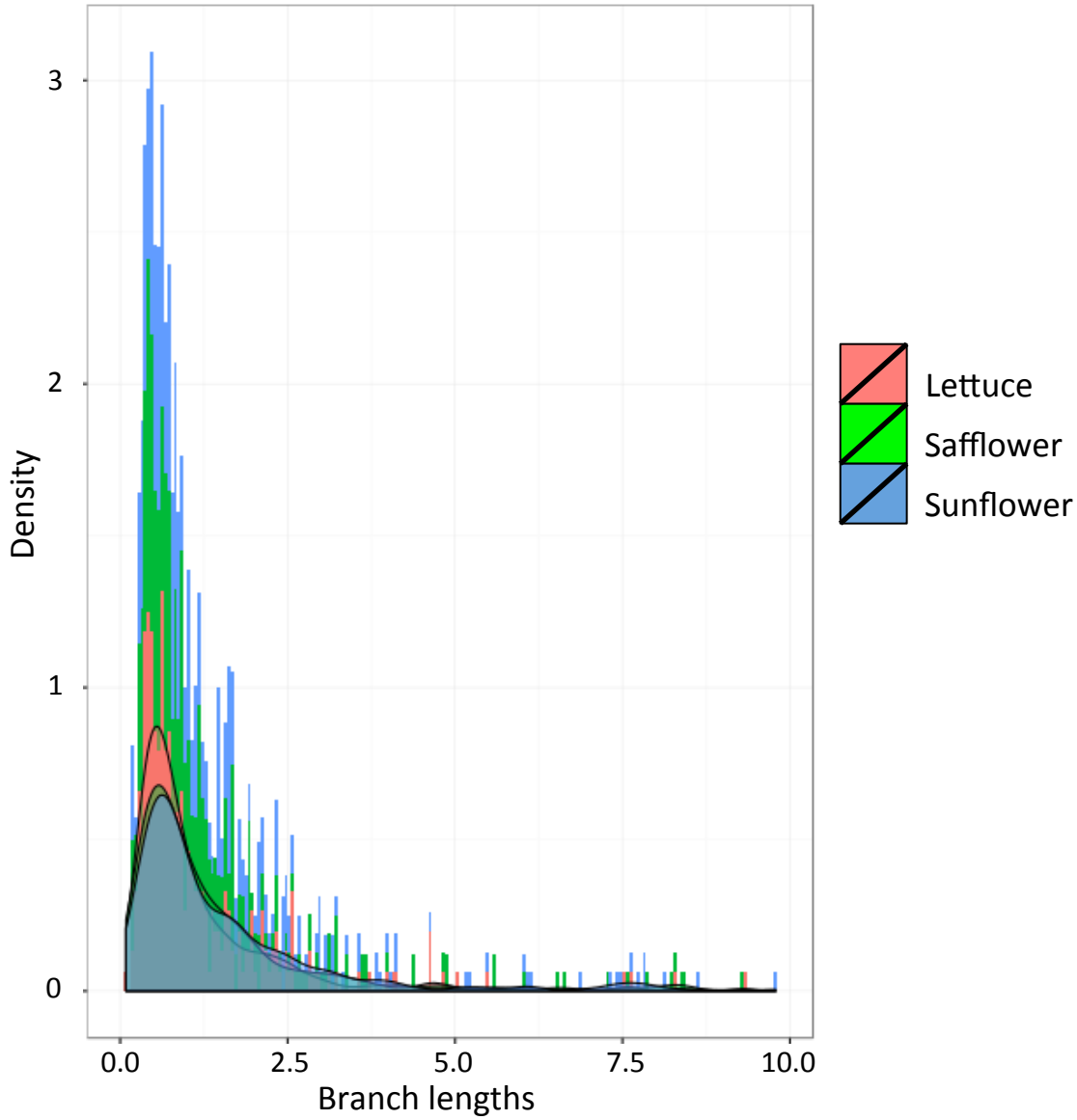
Frequency distribution of branch lengths for 352 gene trees

Species branch lengths per tree:

Short if length < median
Long if length > median

Actual results:

1) 110 trees
2) 111 trees
3) 131 trees



Classifying genes by protein family

- Is there some evolutionary/functional significance to these patterns?

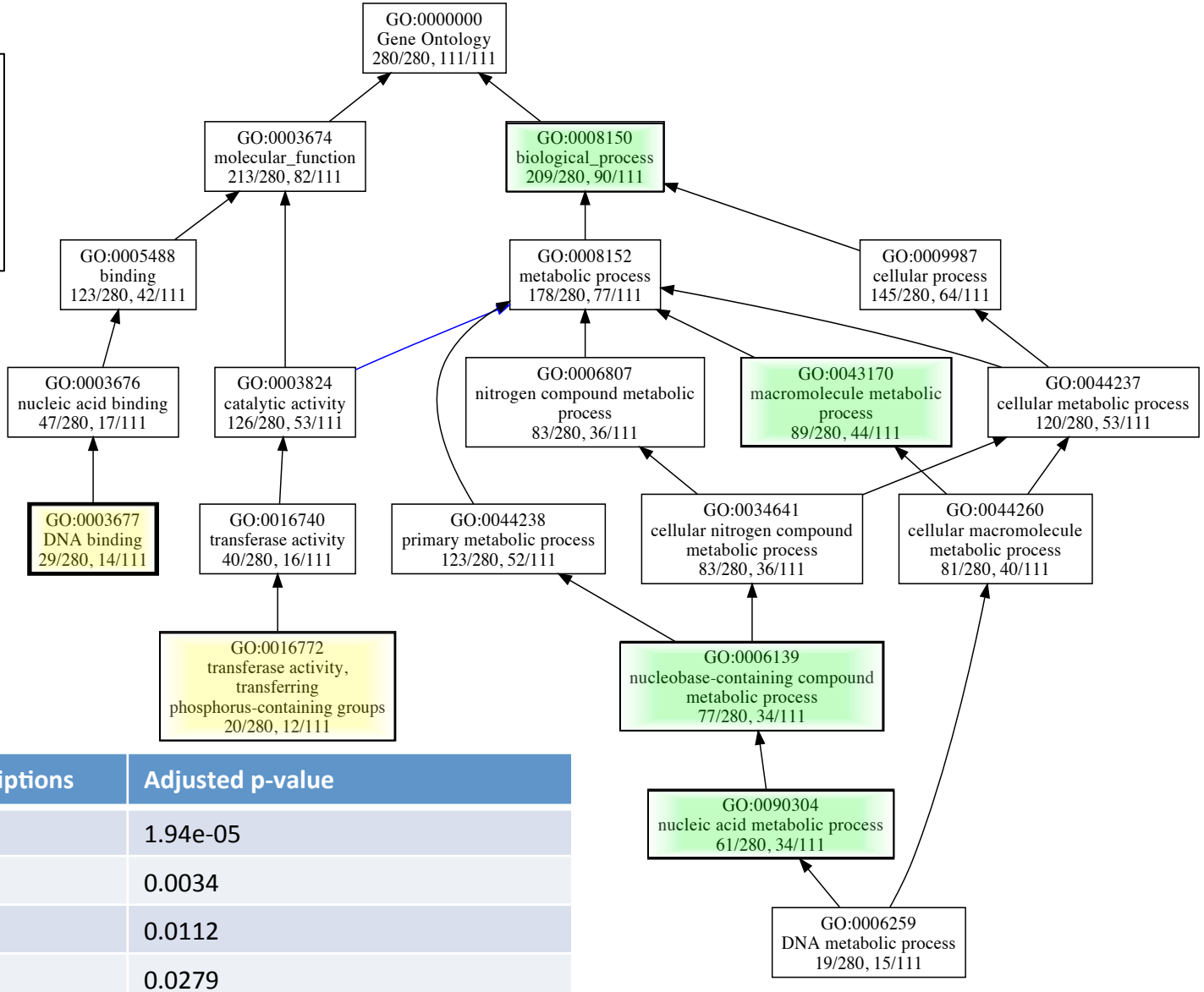
1) Search translated ORFs against Pfam with HMMscan (~12k profile HMMs)

2) Calculate enrichment of terms for each class of gene trees

Composition of trees with long branches

Significant terms from the Molecular Function

Significant terms from the Biological Process

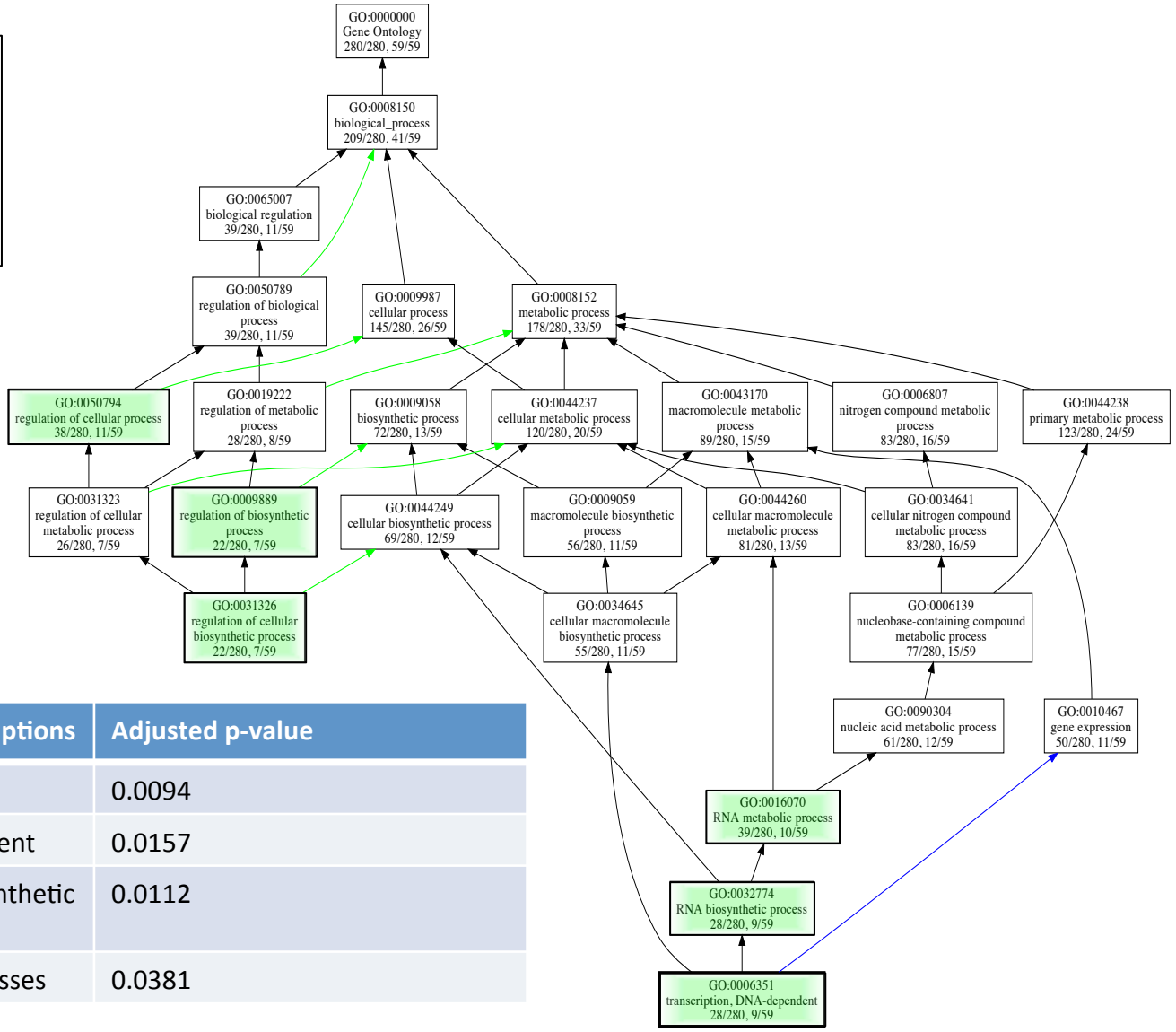


Top-ranked GO term descriptions	Adjusted p-value
Nucleic Acid metabolism	1.94e-05
DNA metabolic processes	0.0034
Transferase activity	0.0112
DNA binding	0.0279

Composition of trees with short branches

Significant terms from the Molecular Function

Significant terms from the Biological Process



Top-ranked GO term descriptions	Adjusted p-value
RNA biosynthetic process	0.0094
Transcription, DNA-dependent	0.0157
Regulation of cellular biosynthetic processes	0.0112
Regulation of cellular processes	0.0381

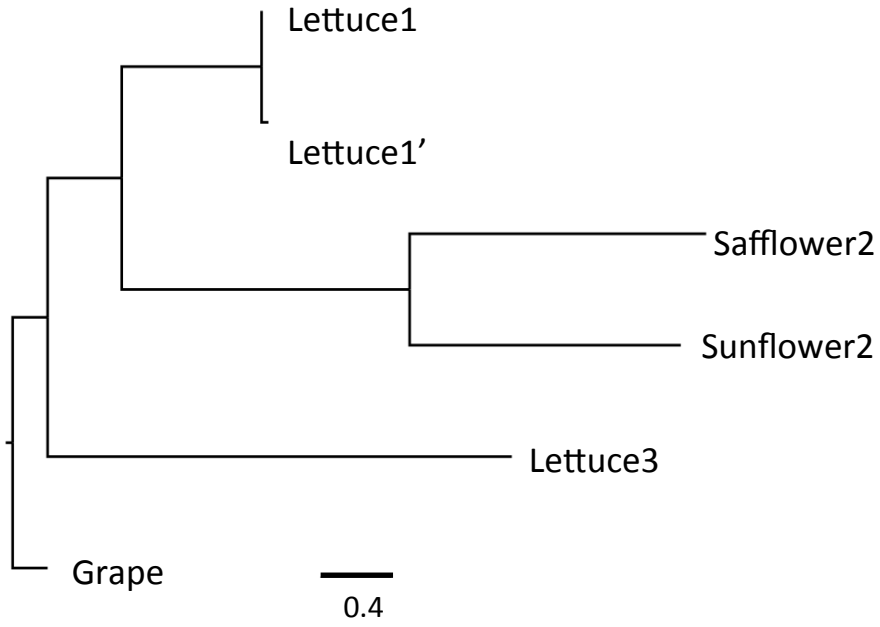
Short branches ~ evolutionary constraint

Are there differences in Ks or Ka between trees with long vs. short branches?

12 trees showed Ka/Ks > 1 (10 have "long" branches)

Short branches = constraint, differences in mutation rate, substitution rate variation

Gene tree for Zinc finger, C3HC4 type (RING finger)
(grape ortholog: GSVIVT01031711001)



Branch len.	Ks (ave)	Ka (ave)	Ka/Ks (ave)	GC (ave)
Long	0.74	0.50	0.43	41.75
Short	0.65	0.15	0.36	41.59

Some findings…

The Asteraceae appear to be evolving a different rates across the genome

~14-fold variation in Ks for Arabidopsis gene pairs (Gaut et al. 2011)

Certain classes of genes show signs of functional constraint, but also lower underlying mutation rate.

…more findings

Genes involved in regulation of cellular process or transcription show signs of constraint.

Those involved in DNA binding, metabolic processes, defense show signs of positive selection.

Species	Domain/Family	Process
Lettuce	Thaumatin	Defense
Lettuce	Zinc Knuckle	DNA binding
Safflower	Myb/SANT-like DNA-binding domain	DNA binding
Sunflower	Agenet	DNA/Protein binding

Phylogenomics summary

1. Using genome-scale data to infer phylogenetic relationships

2. Genome-scale comparisons placed in a phylogenetic context

NGS data is very noisy, often contains contaminants

Tips and tricks

- **Patience** – bioinformatics is mostly informatics

Tips and tricks

- **Patience** – bioinformatics is mostly informatics

- **Where to get help at the command line**
  - Man pages: `man <command>`
  - Accessing the menu: `command; -h; --help`
  - Accessing the documentation: `command -m` or `-man`

- **Where to get help at the online:**
  - Listservs, e.g., bioperl-l
  - biostars.org, seqanswers.com
  - stackoverflow.com, serverfault.com

All of these resources are well indexed by search engines!

Tips and tricks

- **Use cloud computing**
  - Buying a computer is expensive
- **Use all your computer's resources!**
  - parallel processing
- **Don't reinvent the wheel**
  - There is likely an existing method

Tips and tricks

- **Do reinvent the wheel**
  - The only way to know how something works
- **Learn best practices for coding**
  - Use version control and web hosting
  - Write documentation
  - Write tests and use automation / CI
- **Support free Open Source software** ☺

Thank you for being patient!

Any questions?