

Clustering

James B. Pease

University of Michigan

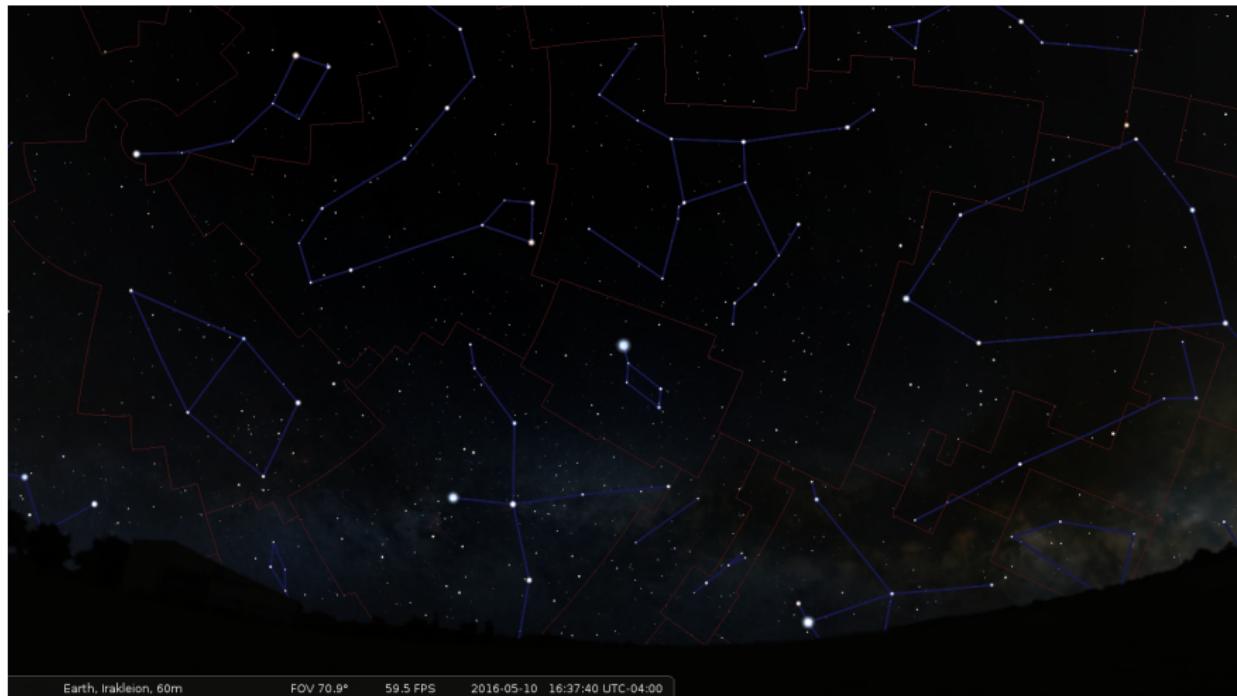


Earth, Irakleion, 60m

FOV 70.9°

59.9 FPS

2016-05-10 16:37:31 UTC-04:00



Earth, Irakleion, 60m

FOV 70.9°

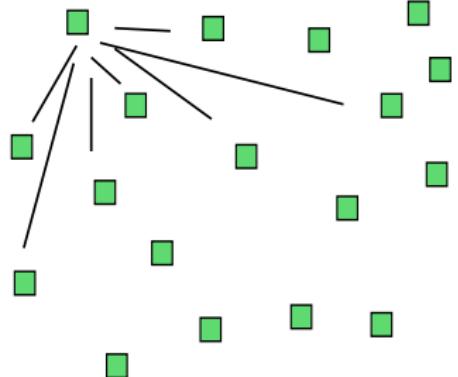
59.5 FPS

2016-05-10 16:37:40 UTC-04:00

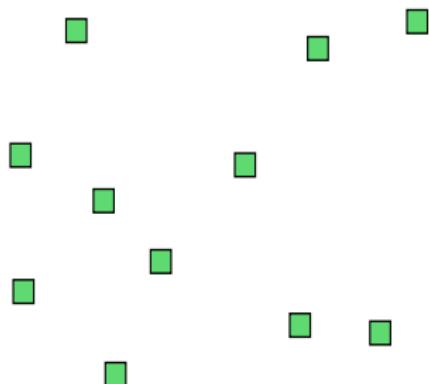
Sets of sequences

- **Goal:** get a set of sequences for a gene region of interest (or set of gene regions)

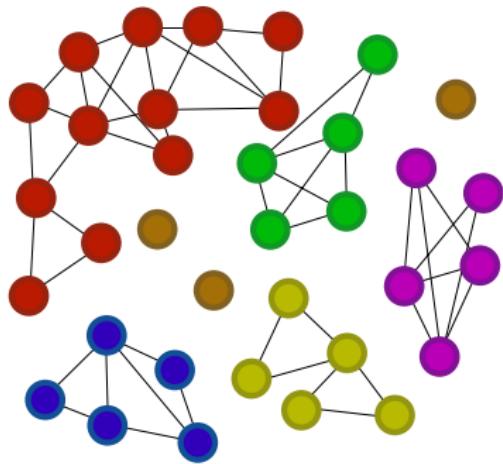
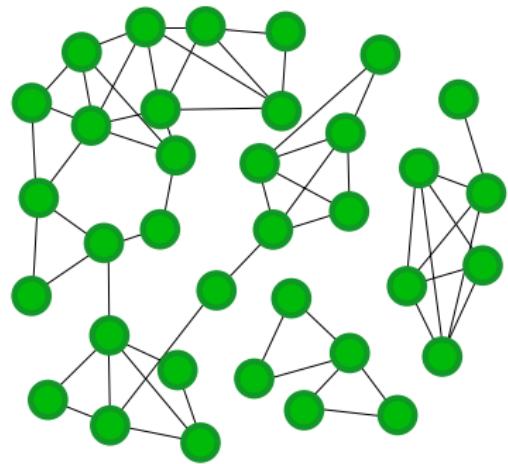
sequences



significant hits



Clustering



Questions

- We already discussed pairwise questions:
 - Are **two** sequences *likely* to be homologous?
 - What sequences in a database are *likely* homologs of a query sequence?
- How do we determine questions of multiple relationships?
 - Are the all sequences in a **set** *likely* to be homologous?
 - Construct a multiple sequence alignment

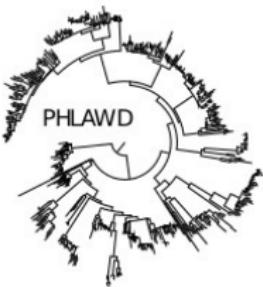
Framework

- Can we “cluster” sets of homologs together?
 - Use (many) pairwise alignment results
 - Identify gene regions that are informative for multiple sequence analysis and phylogenetics
 - Automate and avoid systematic bias
- “Baited” BLAST
- Clustering

- GenBank stores species taxonomy but not gene taxonomy
- Hard to get gene identifications for all of GenBank (e.g., clustering)
 - All genes for species x
 - NOT all species for gene y

PHLAWD: PHyLogenetic dataset Assembly With Databases

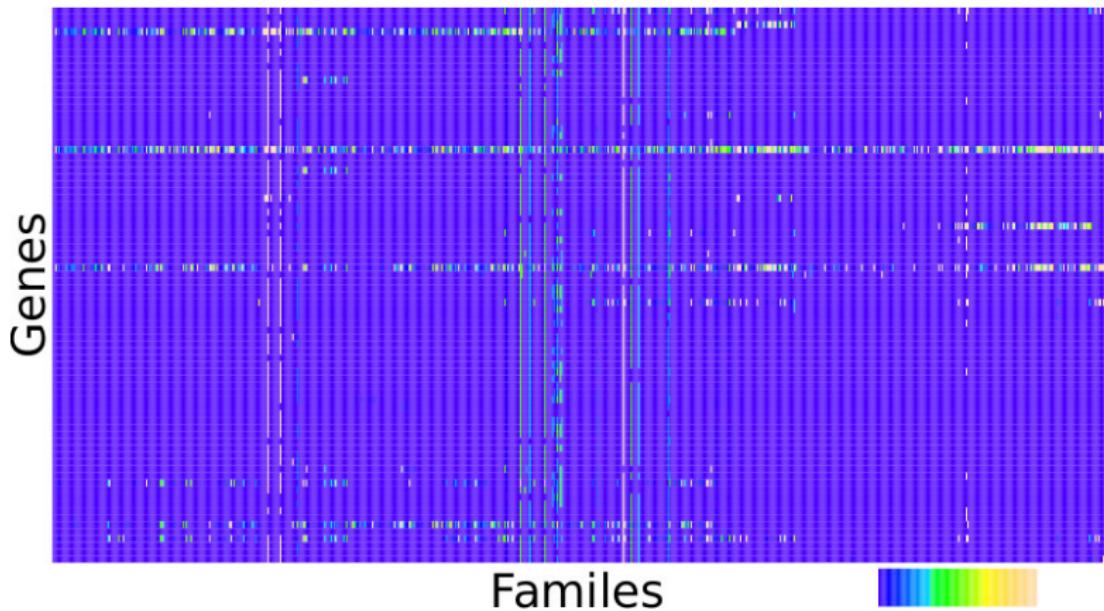
multithreaded c++ program for assembling large phylogenetic datasets

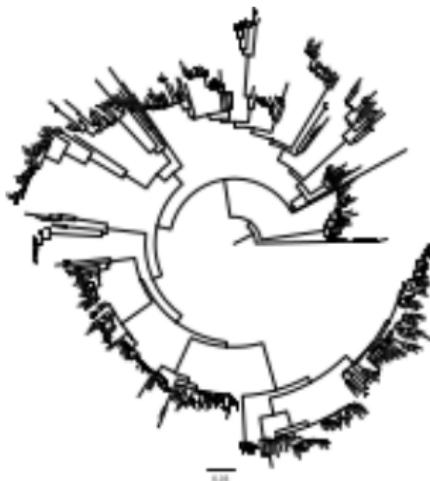
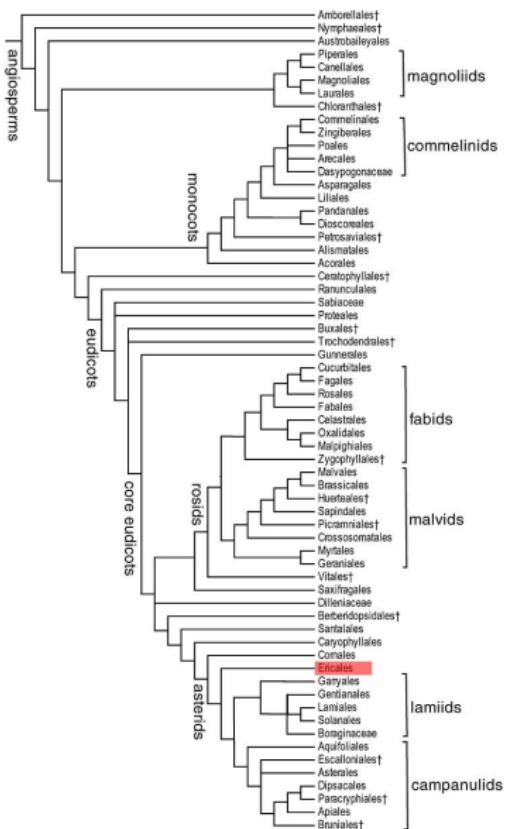


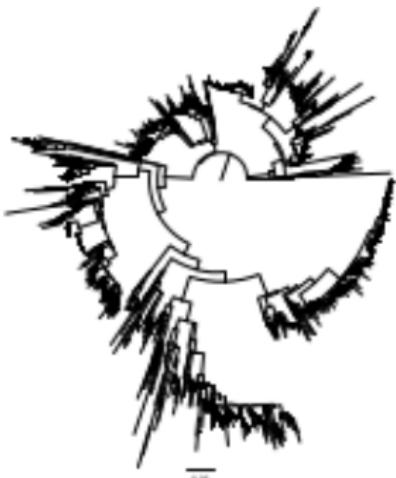
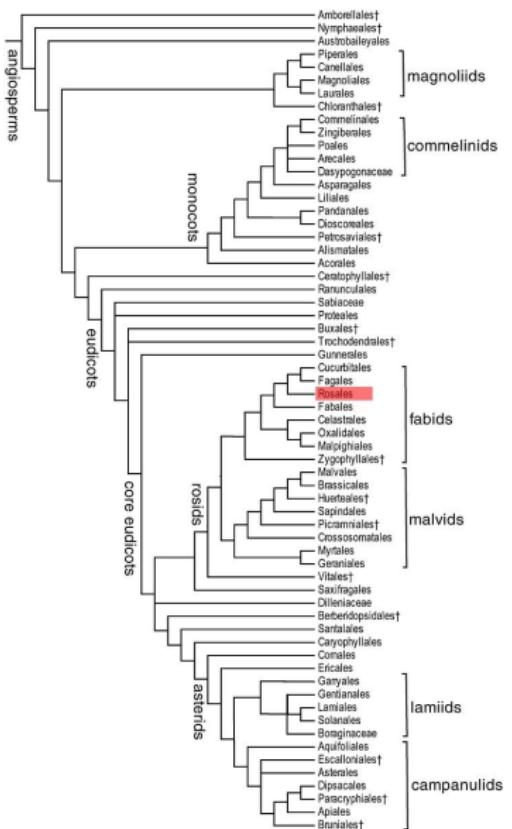
<http://phlawd.net>
Smith et al., 2009, *BMC Evol. Bio.*

- baited Smith-Waterman

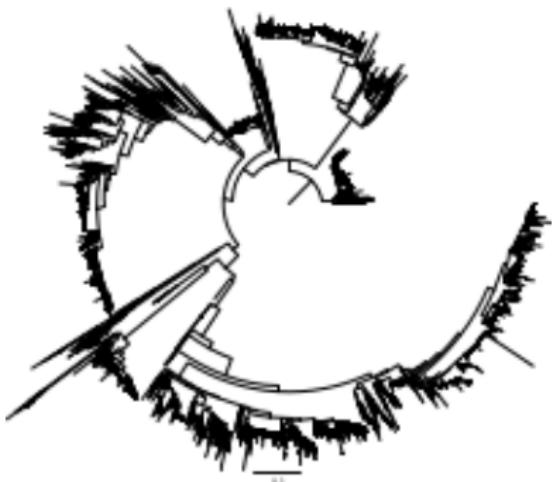
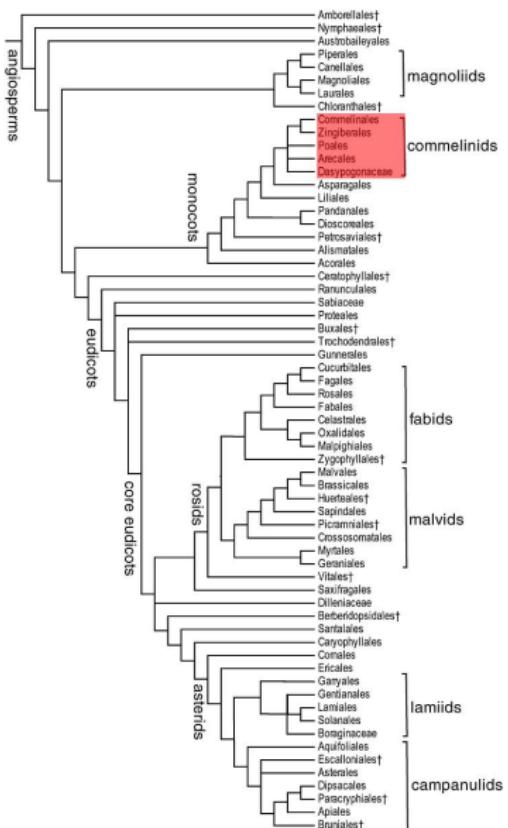
- provide known homologous sequences
- compare a set of unknown sequences to those known sequences
- keep the significant hits





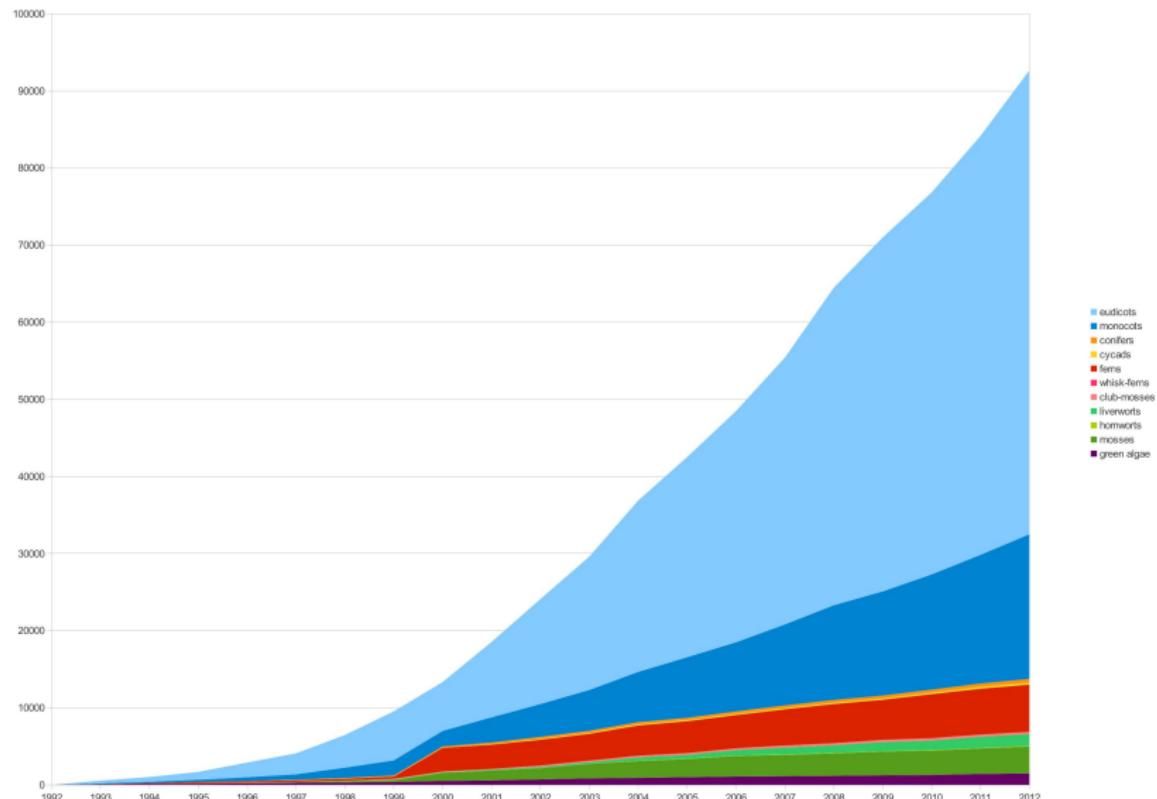


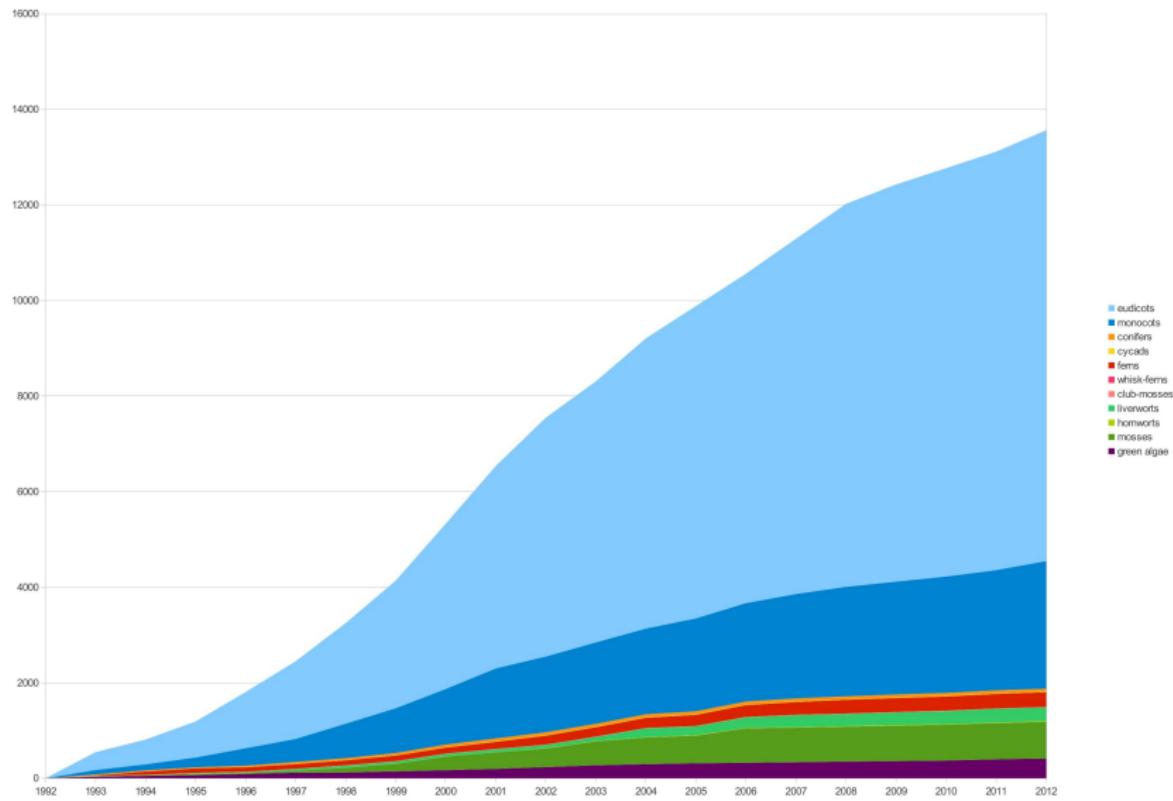
Rosales



Commelinids

Is the data available to construct a tree for all plants?

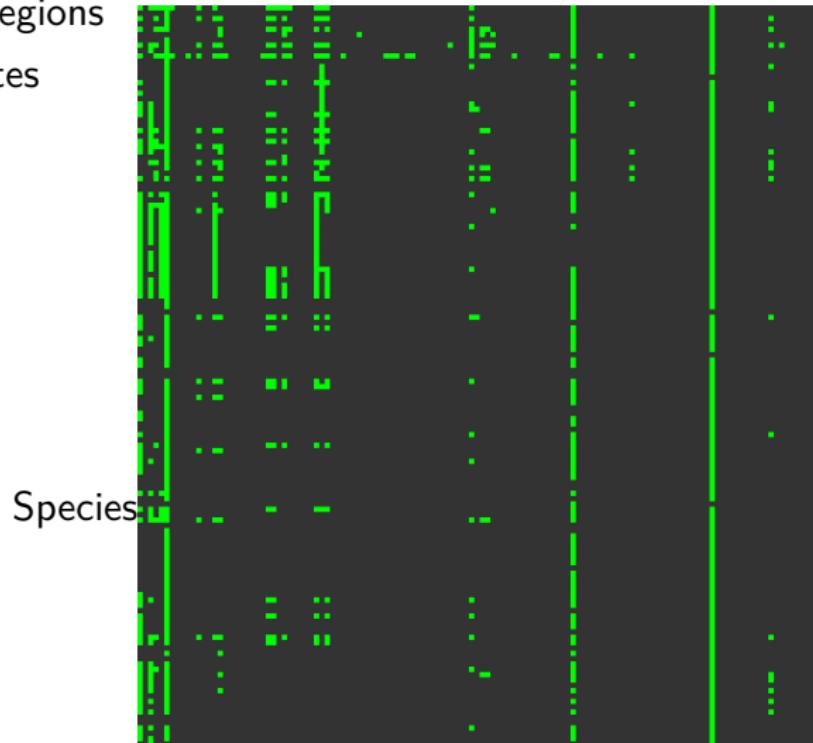


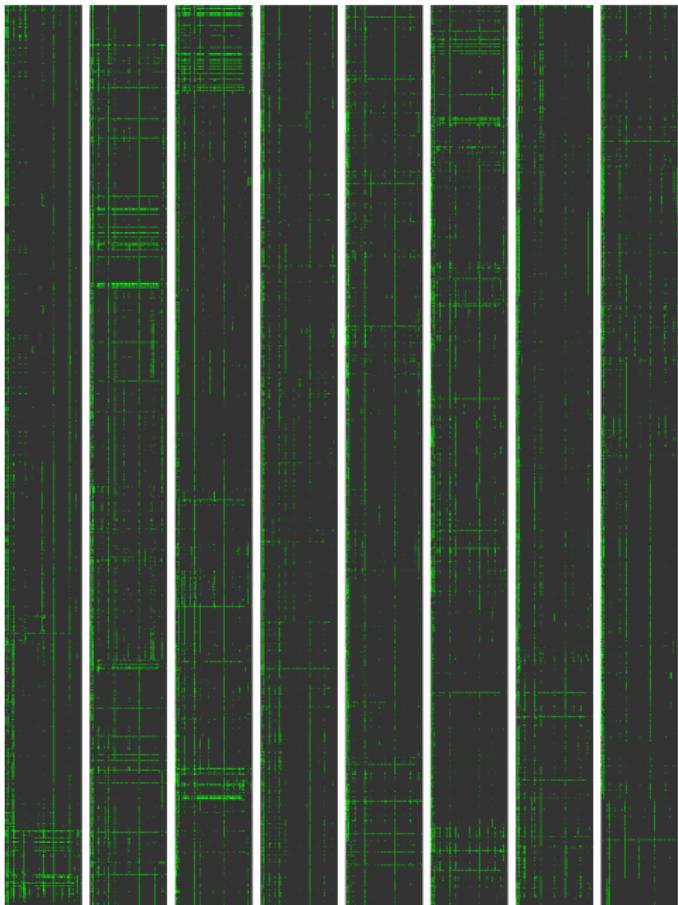


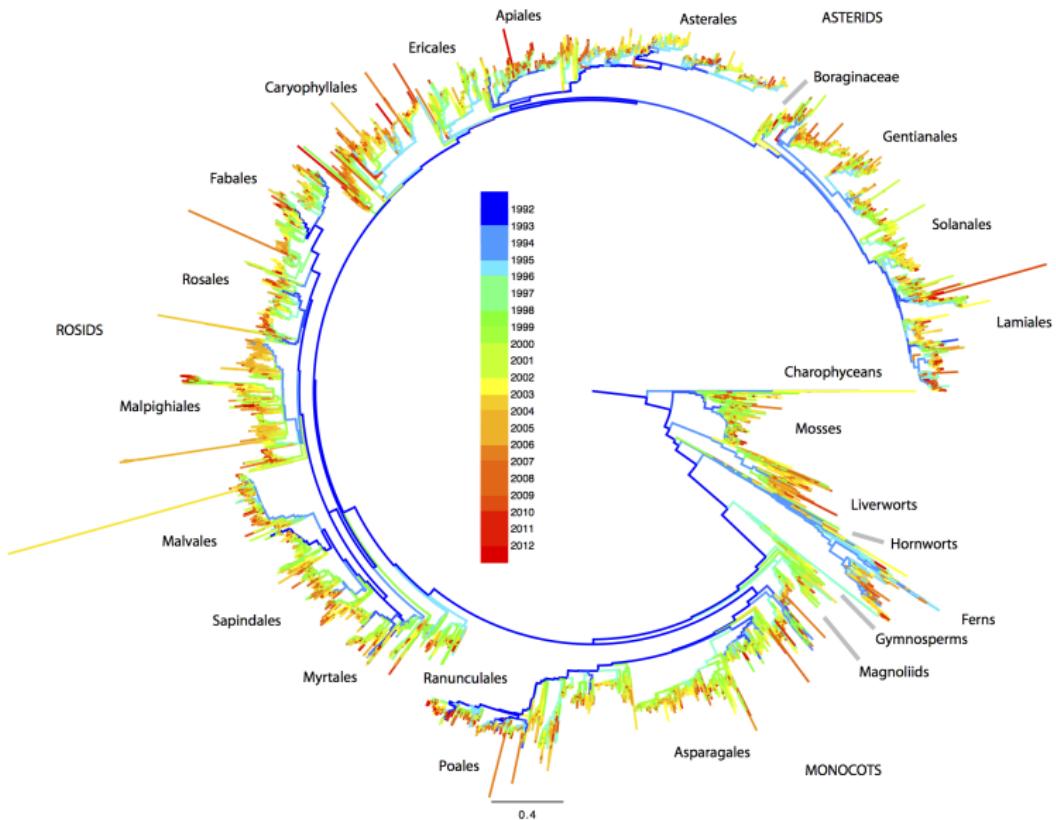
All genera of plants

Genes

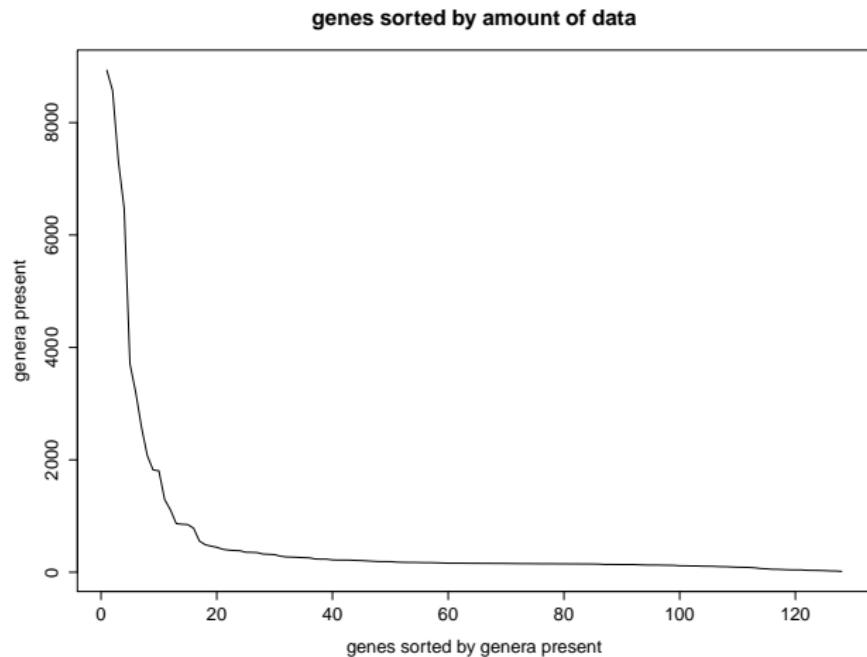
- 13,083 genera
- 128 gene regions
- 148,150 sites

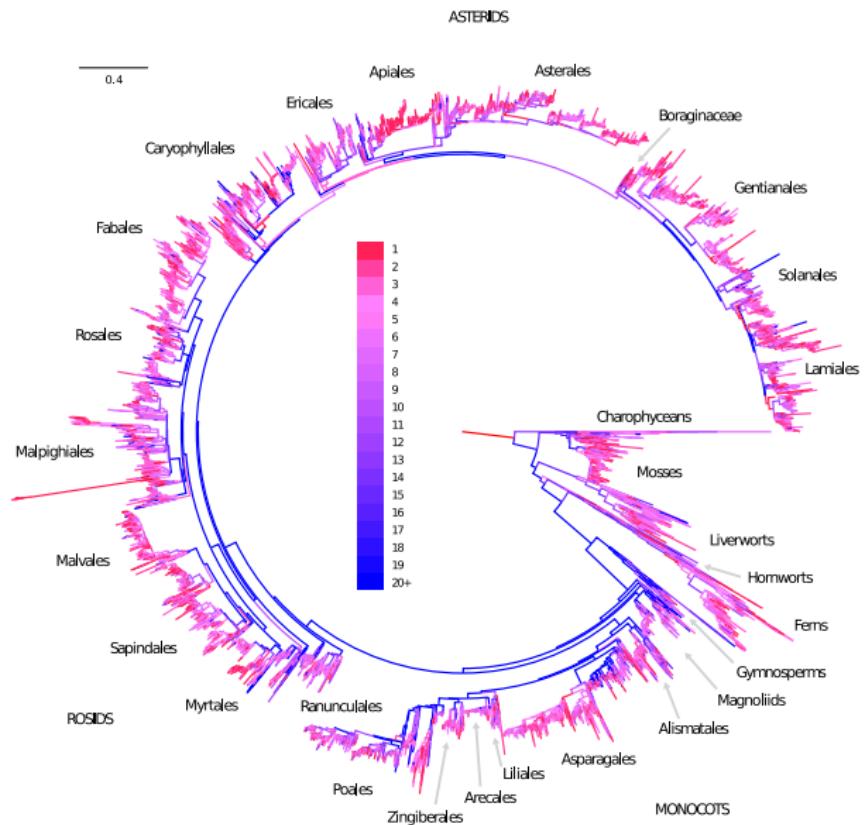






95% missing data





Is the data available to construct a tree for all plants?

- NO
-

- Most of the time the data is more complicated
- We need to add more data (hint: what if we do not have *a priori* knowledge of the gene regions?)

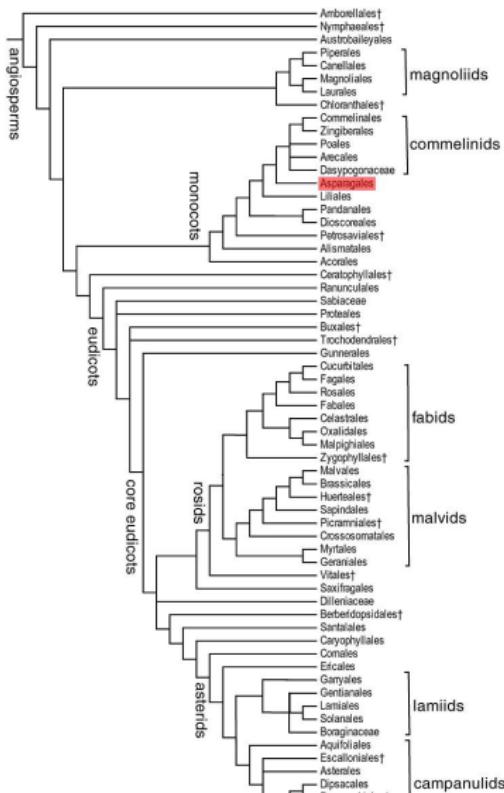
Baited BLAST

- Requires bait (*a priori* identification of not only gene regions but representatives for these regions)
- Relatively simple and works for large datasets
- We will go over some examples in lab

PHLAWD algorithm

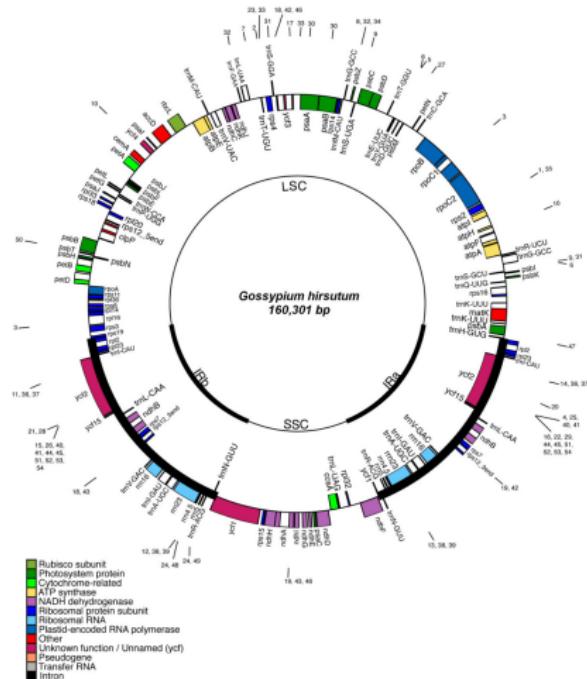
- 1 pick clade
- 2 pick gene region
- 3 select known sequences
- 4 pairwise comparison of possible sequences to known
- 5 test for saturation
- 6 reassemble if separated for saturation
- 7 repeat steps 2-6 for other gene regions
- 8 concatenate all gene regions
- 9 estimate phylogeny

Pick region of the tree of life



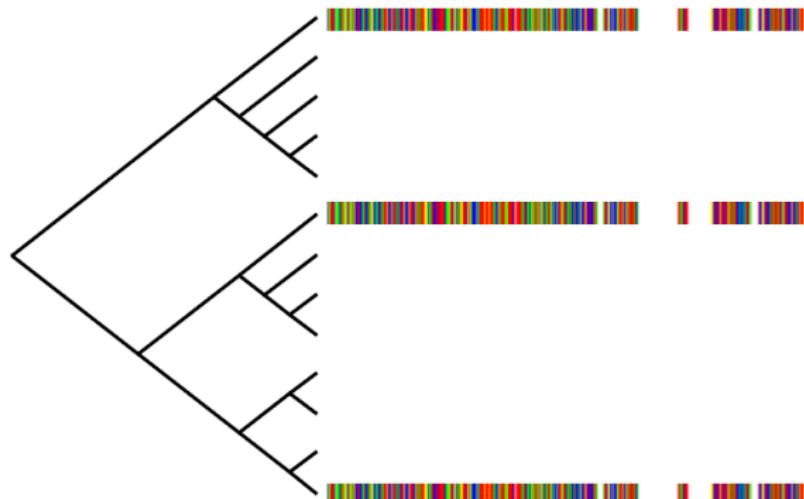
Asparagales

Gene region

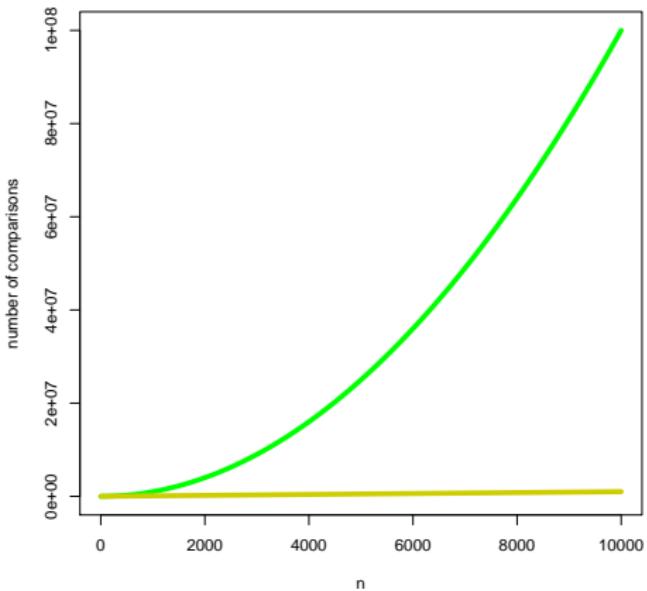
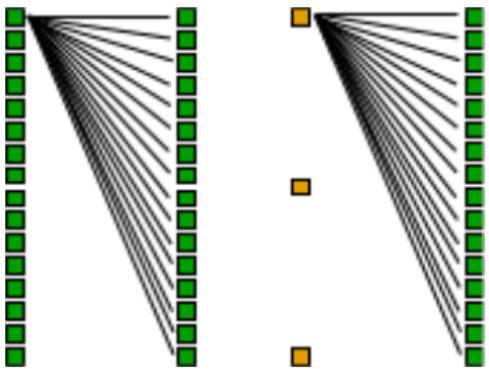


trnL-trnF

Identify set of known representatives for gene region

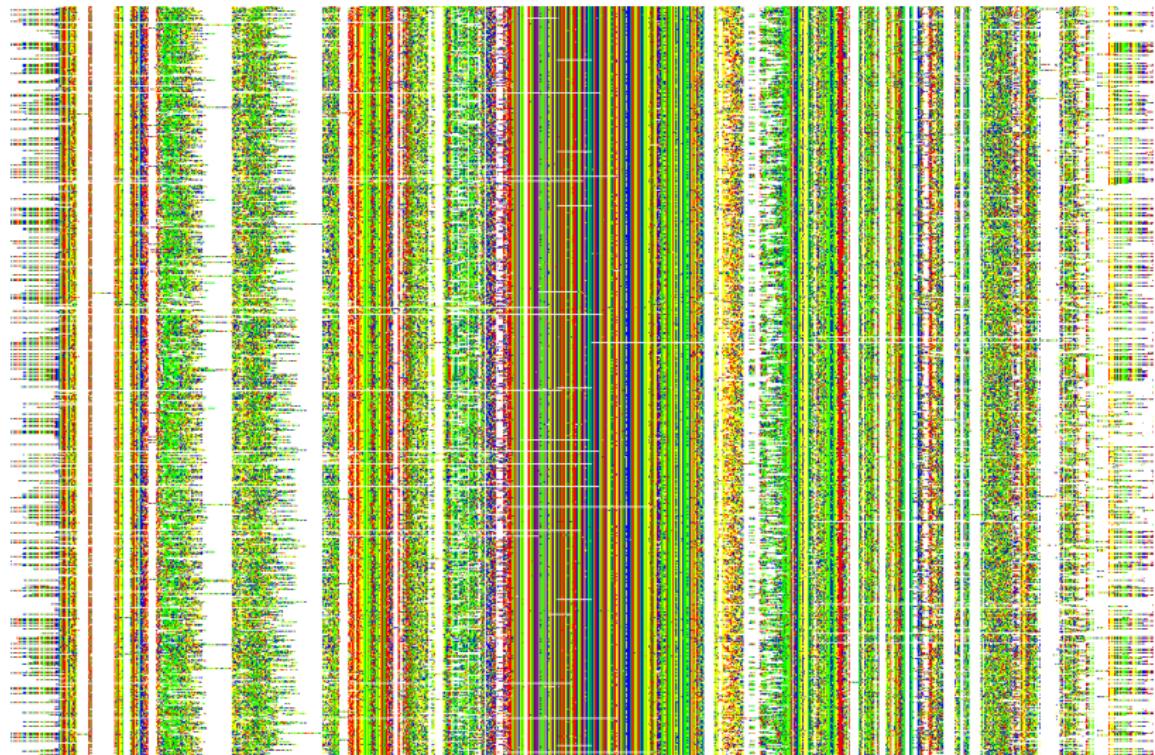


Pairwise comparison

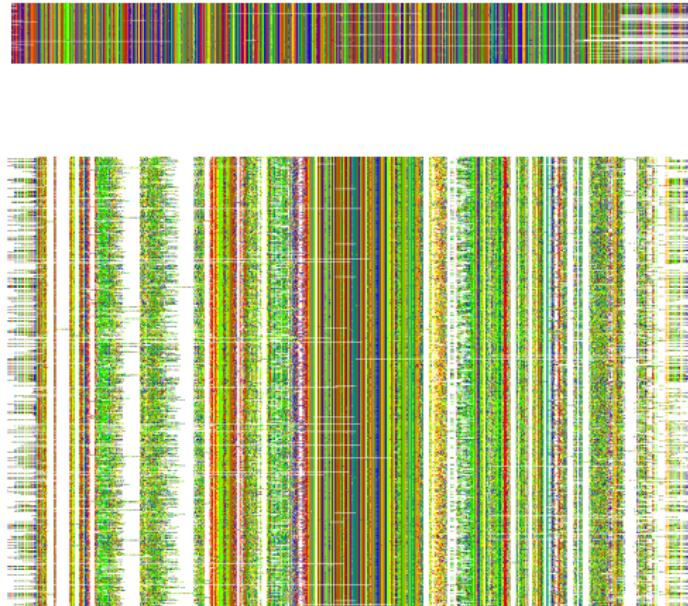


moving from $O(n^2)$ to $O(n)$

Aligning distantly related species

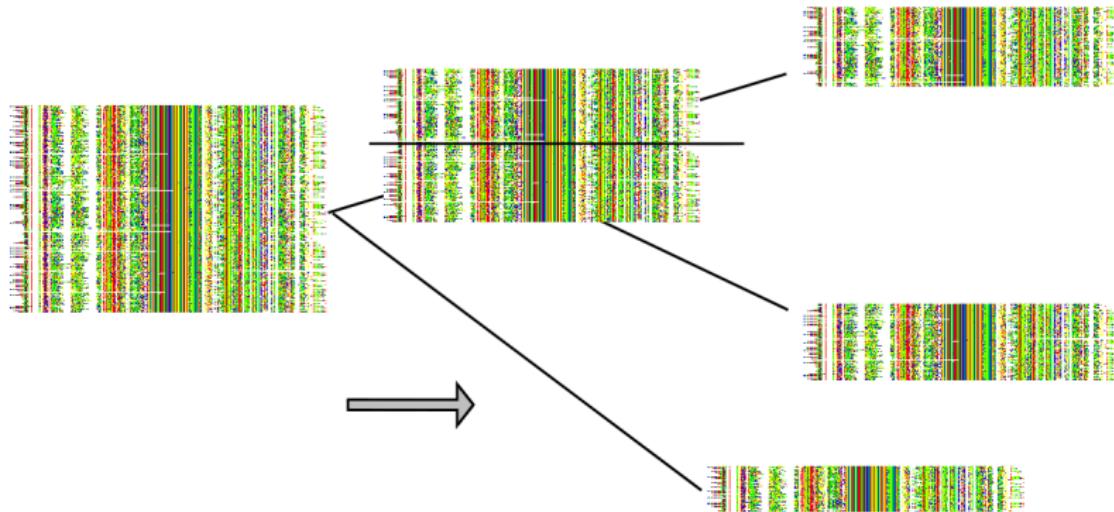


Saturation test

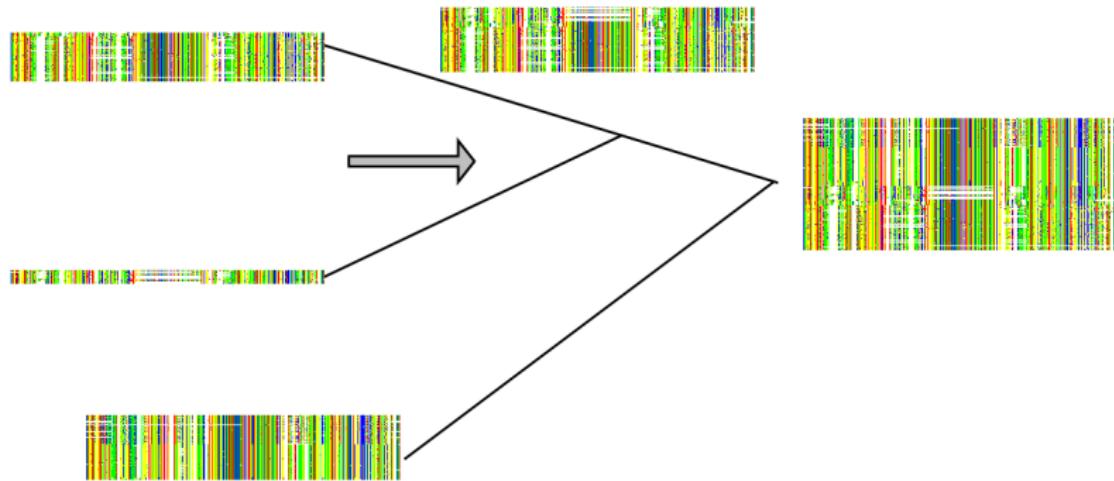


MAD = median absolute deviation ($\text{median}_i (| X_i - \text{median}_j (X_j) |)$)

Saturation test

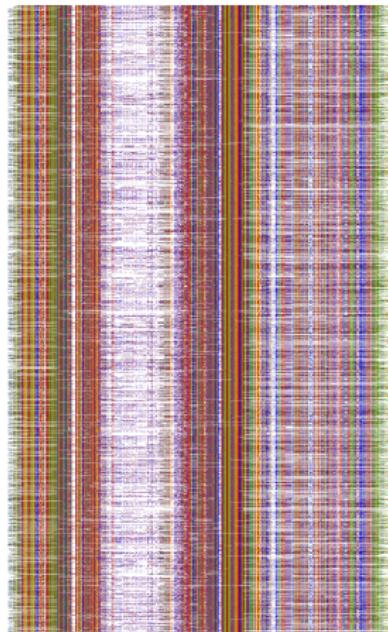


Reassembly

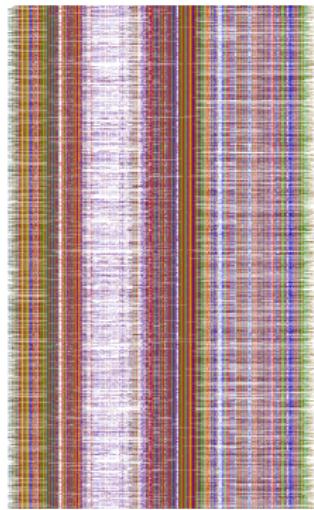


Results

- number of taxa = 2485
- number of sites = 1406
- < 8 minutes
- 26 clades broken down for individual alignments and profile alignment reassembly



Concatenate



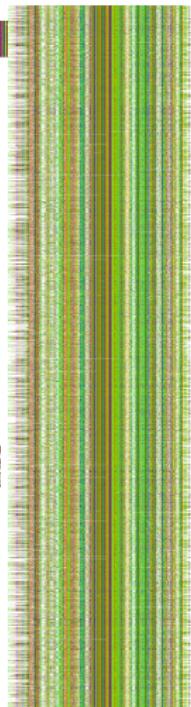
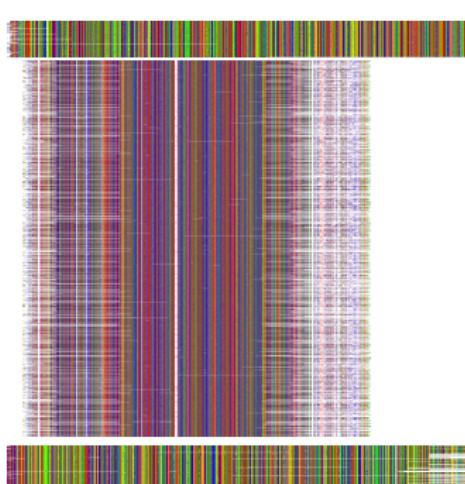
$\text{trnL-trnF} = 2485$

$\text{atpB} = 107$

$\text{matK} = 2182$

$\text{rbcL} = 924$

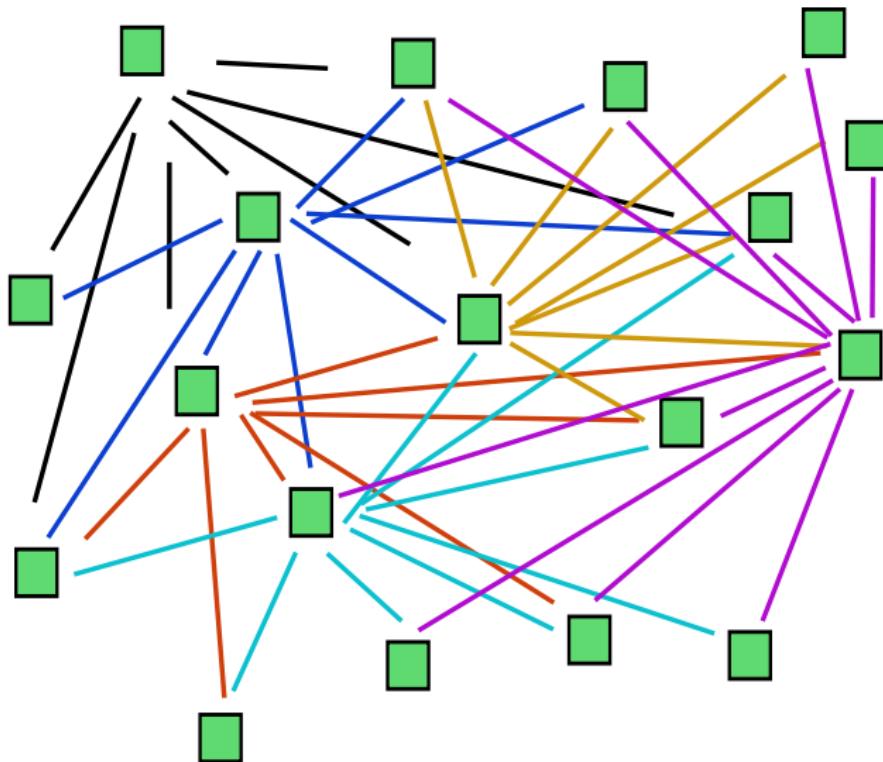
$\text{ITS} = 3725$



Clustering

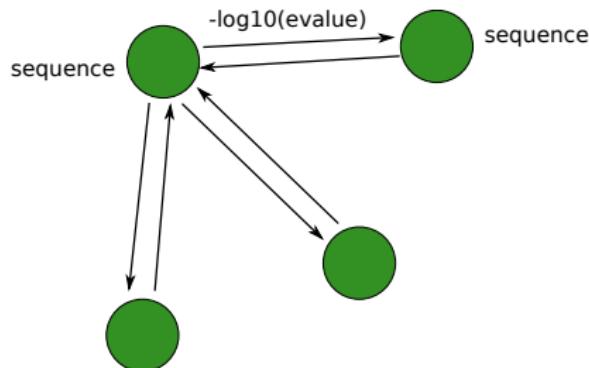
- What if we do not have bait (*no a priori* identification of gene regions)?

All by all comparisons



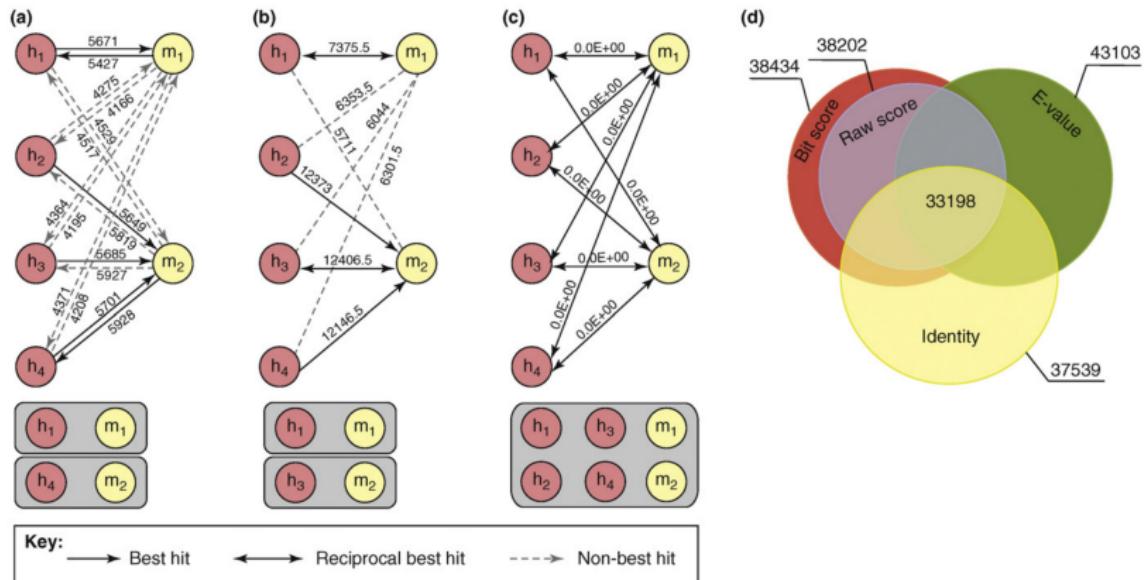
Graph

- Each sequence is connected to the others to which there was a significant hit
- Each vertex is a sequence
- Each edge is a significant hit
- The weight of the edge (as well as the edge itself) is based on *E*-value ($-\log_{10}$ of the *E*-value so that lower *E*-value is a higher / stronger connection)



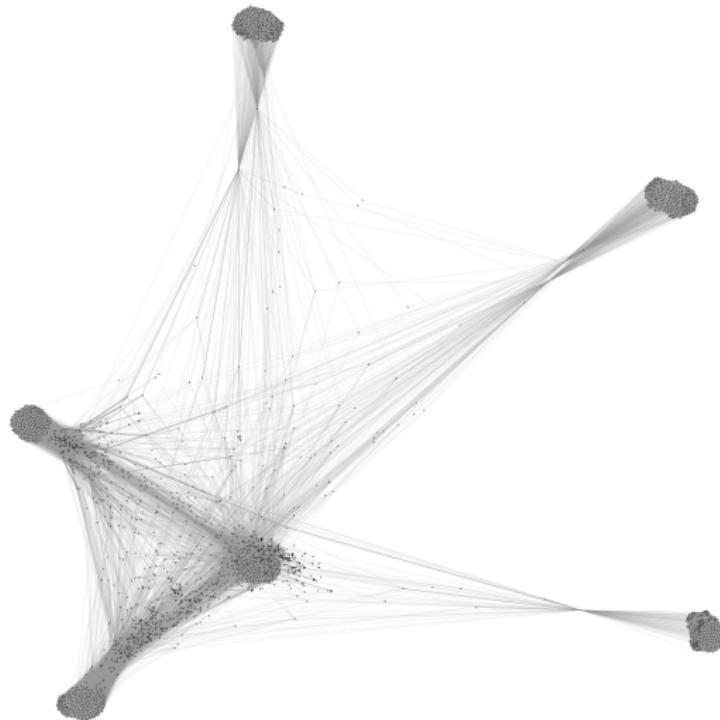
Different measurements

What other measures can you use for the strength of a vertex?

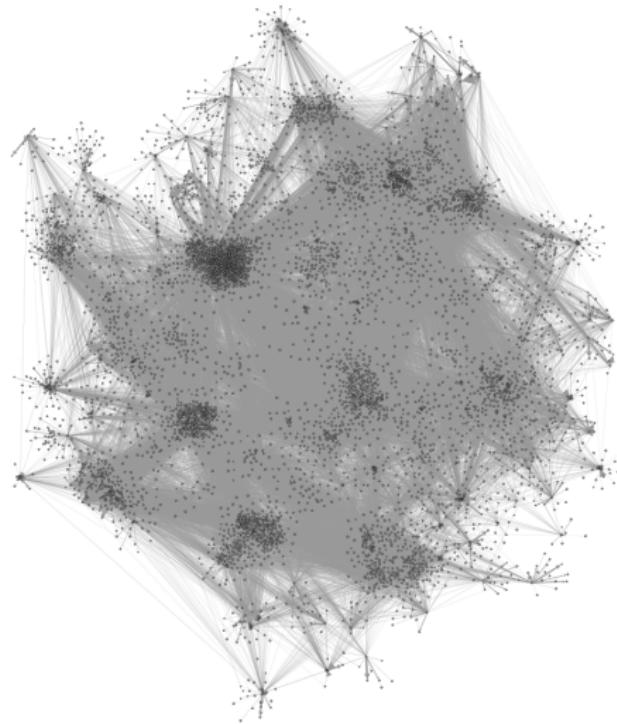


Kuzniar et al., 2008

Visualizing an all-versus-all BLAST



Visualizing an all-versus-all BLAST



Phylota



PhyLoTA Browser (rel. 1.5)

Phylogenetic sequence data sets across 54,671 eukaryotic genera

GenBank rel. 184 clusters, alignments and trees now available

Syst. Biol. 57(3):335–346, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802158688

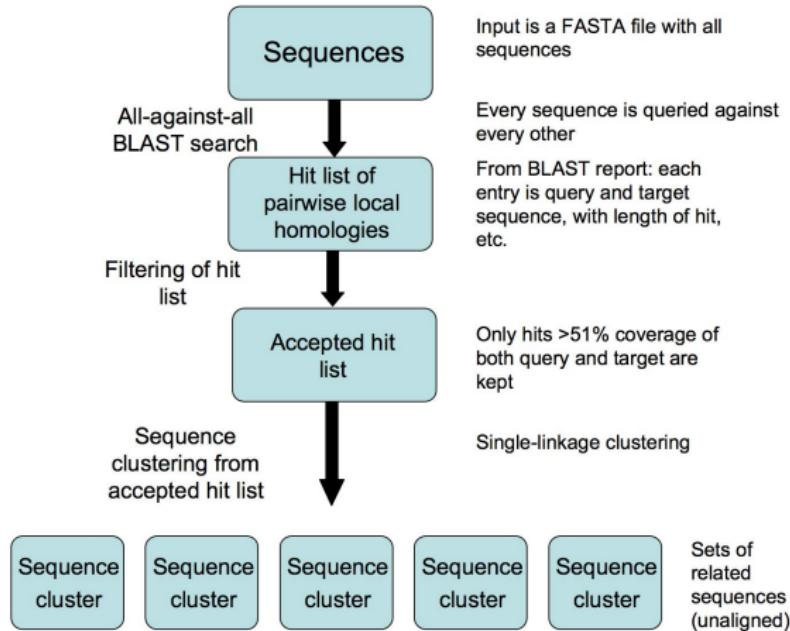
The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research

MICHAEL J. SANDERSON,¹ DARREN BOSS,¹ DUHONG CHEN,² KAREN A. CRANSTON,¹ AND ANDRE WEHE²

¹*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson 85721, USA; E-mail: sanderm@email.arizona.edu (M.J.S.); dboss@email.arizona.edu (D.B.); cranston@email.arizona.edu (K.C.)*

²*Department of Computer Science, Iowa State University, Ames, IA, USA; E-mail: duhong@iastate.edu (D.C.); andre@wehe.us (A.W.)*

Phylota



Phylota



Sequence diversity and cluster set summaries (rel. 184)

Sequence tallies include those from "model" organisms. To *exclude* model organisms, click [here](#).

To show changes between this release and the previous one, click [here](#).

NCBI taxon name ¹		Descendant species ²	Descendant terminals	Descendant genera	Sequences (GIs)	Seq. clusters ³	Phylog. inform. seq. clusters ⁴
Fabaceae	up to Fabales	6205	6459	644	71104	-	-
Caesalpinoideae		789	813	165	4802	278	39
Mimoideae		1017	1048	70	6583	553	38
Papilioideae		4399	4598	409	59719	2238	364

¹Names refer to node and its subtree unless the term "node only" appears.

²Taxa at species rank as determined by NCBI (including this node). Note that some of these have not been formally named yet and are not retrieved in certain NCBI Taxonomy (*Marina* sp. Lavin 5341), but they are associated with sequence(s) and are counted as species here.

³Clusters built by exhaustive all-against-all BLAST searches. For further information of about this see [here](#). A dash (-) means there were too many sequences to cluster or no se

⁴Phylogenetically informative clusters have four or more taxa (not GIs) represented.

Phylota



Cluster set at subtree whose root is Papilioideae (rel. 184)



Click to see [complete](#) taxon coverage matrix or taxon coverage by [genus](#) (very large matrices may load slowly!).

Cluster ID	Parent cluster	TaxIDs	GIs	Genera	L _{min}	L _{max}	MAD ¹	Defline of longest sequence
0	-	51	70	10	446	941	0.887	Pediomelum megalanthum var. epipodium voucher Egan & Egan 146 (BRY) granule bound starch synthase (GBSS1) gene, exons 10 through 13 and partial cds.
1	-	1645	2321	275	229	1409	0.443	Broad bean chloroplast genes for tRNA-Leu(CAA) and (UAA) and tRNA-Phe.
2	-	68	132	14	613	1600	0.881	Coursetia rostrata isolate DNA 2129 tmD-tmY intergenic spacer, partial sequence; tRNA-Tyr (tmY)-tmE intergenic spacer, and tRNA-Glu (tmE) gene, complete sequence; and tmE-tmT intergenic spacer, partial sequence; chloroplast.
4	-	975	1289	249	343	3021	0.573	Medicago sativa tRNA-Lys (tmK) gene, complete sequence; maturase K (matK) gene, complete chloroplast genes for chloroplast products.
8	-	689	910	200	436	1797	0.684	Lathyrus cirrhosus ATP synthase CF1 beta subunit (atpB) and ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) genes, partial cds; chloroplast.
10	-	3779	7286	328	131	1354	0.419	V.faba 5.8S, 18S and 25S ribosomal RNA genes and ITS regions.
15	-	28	28	8	241	531	0.786	Psorothamnus fremontii CNGC4-like gene, partial sequence.
30	-	25	33	19	214	382	0.901	Flemingia macrophylla voucher PS0222MT03 Ycf5 (ycf5) gene, partial cds; chloroplast.
31	-	94	123	39	457	735	0.737	Andira inermis voucher LA8239 RNA polymerase beta subunit (rpoC1) gene, partial cds; chloroplast.
32	-	38	38	26	673	1863	0.868	Arachis batizocoi 18S ribosomal RNA gene, partial sequence.
33	-	7	7	7	662	814	0.929	L.langustifolius 26S rRNA (partial).
37	-	92	96	16	446	766	0.760	Pultenaea trinervis voucher de Kok 801 NADH dehydrogenase subunit F (ndhF) gene, partial cds; chloroplast.

Phylota



Papilioideae (rel. 184): Cluster 0

Number of sequences (GIs)	70
Number of distinct taxon IDs (TIs)	51
Shortest sequence	446 nt -- <i>Orbexilum simplex</i> voucher Thomas & Pias 58668 (UT) granule bound starch synthase (GBSSI) gene, exons 10 and partial cds.
Longest sequence	941 nt -- <i>Pediomelum megalanthum</i> var. <i>epipsillum</i> voucher Egan & Egan 146 (BRY) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
Maximum alignment density ¹	0.887
Cluster status	Cluster has 3 child clusters
View tree	

¹The fraction of non-missing nucleotides in an ideal alignment if no gaps had to be introduced. Low values indicate that cluster sequences are heterogeneous in length, with possibly different numbers of homologies (for example, if one sequence is a complete mitochondrial genome, and others are single mt genes).

Download cluster...

- Unaligned cluster in Fasta format

Download alignments...

- Muscle alignment in Fasta format ([Color view](#))

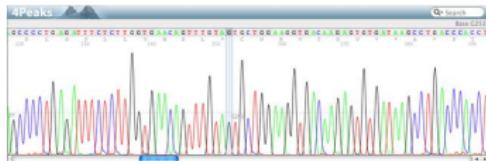
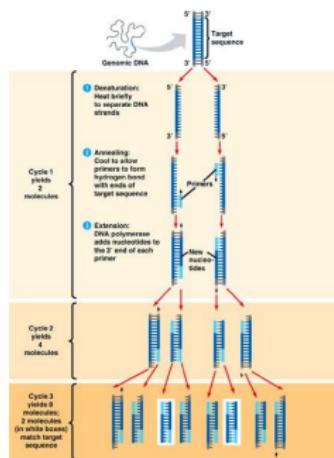
Phylota

Download a midpoint rooted RAxML optimal tree with branch lengths (Muscle alignment)...

- Use gi# as taxon label

NCBI taxon name	GI	TI	Length	Defline
<i>Abrus precatorius</i>	156615502	3816	670	<i>Abrus precatorius</i> voucher Thorne et al 6791 (BRY) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Bituminaria bituminosa</i>	156615494	53836	790	<i>Bituminaria bituminosa</i> voucher Hobbs 1 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Cullen australasicum</i>	156615496	109219	882	<i>Cullen australasicum</i> voucher Grimes 3188 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Cullen cinereum</i>	156615374	100168	447	<i>Cullen cinereum</i> voucher Henry 264 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 12 and partial cds.
<i>Cullen discolor</i>	156615498	458332	898	<i>Cullen discolor</i> voucher Grimes 3213 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Cullen tenax</i>	156615500	458333	898	<i>Cullen tenax</i> voucher Grimes 3159 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Desmodium floridanum</i>	156615480	458388	454	<i>Desmodium floridanum</i> granule bound starch synthase (GBSSI) gene, exons 10 through 12 and partial cds.
<i>Hoita macrostachya</i>	156615402	458335	713	<i>Hoita macrostachya</i> voucher Egan & Egan 271 (BRY) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Hoita orbicularis</i>	156615414	458336	851	<i>Hoita orbicularis</i> granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Hoita orbicularis</i>	156615416	458336	851	<i>Hoita orbicularis</i> voucher Egan & Egan 279 (BRY) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Orbexilum lupinellum</i>	156615398	458338	922	<i>Orbexilum lupinellum</i> voucher Egan & Egan 257 (BRY) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Orbexilum lupinellum</i>	156615482	458338	922	<i>Orbexilum lupinellum</i> granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
<i>Orbexilum melanocarpum</i>	156615392	458339	910	<i>Orbexilum melanocarpum</i> voucher Grimes 2287 (TEX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.
				<i>Orbexilum melanocarpum</i> voucher Felicianno et al 153 (TFX) granule bound starch synthase (GBSSI) gene, exons 10 through 13 and partial cds.

Massive parallel sequencing



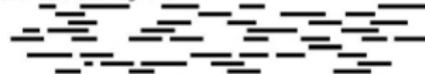
a) Multiple copies of genome



b) Sheared random fragments



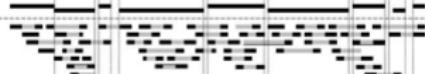
c) Size fractionated fragments



d) Reads



e) Contigs



f) Scaffolds(Super contigs)



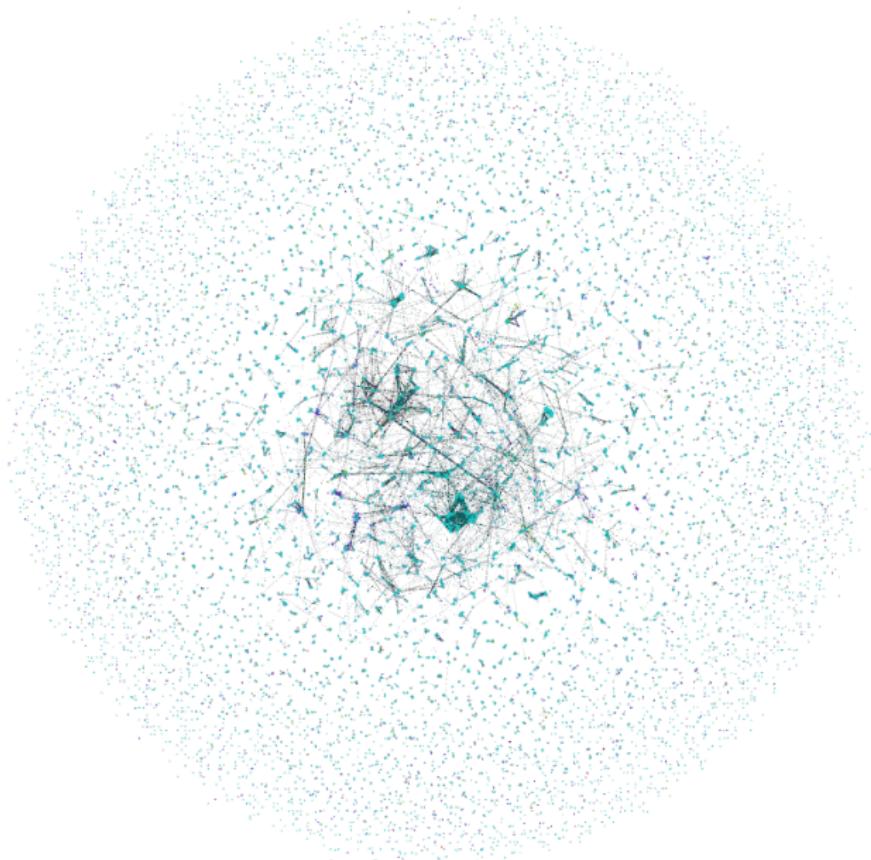
Number of genes

Typical phylogenetic analyses

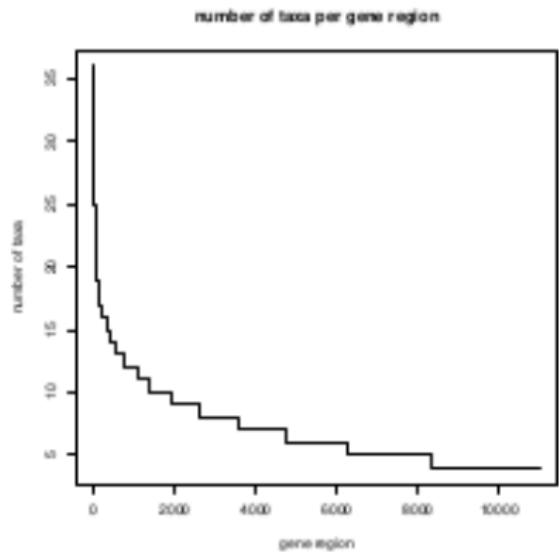
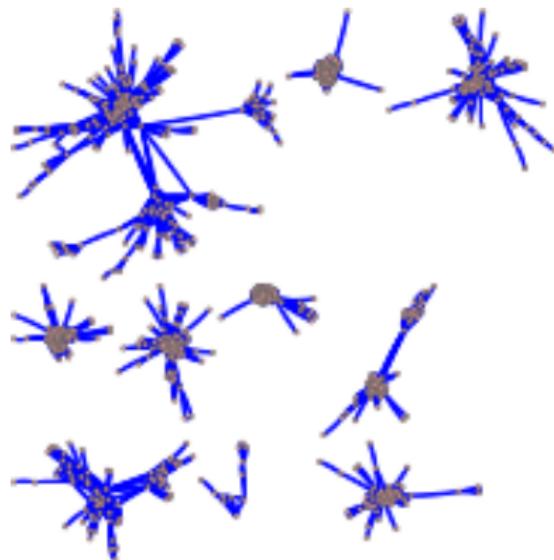
- 1 to 10 genes
- 17 genes. Plants (Soltis et al. 2011)
- 19 genes. Birds (Hackett et al. 2007)

Transcriptomic and genomic phylogenetic analyses

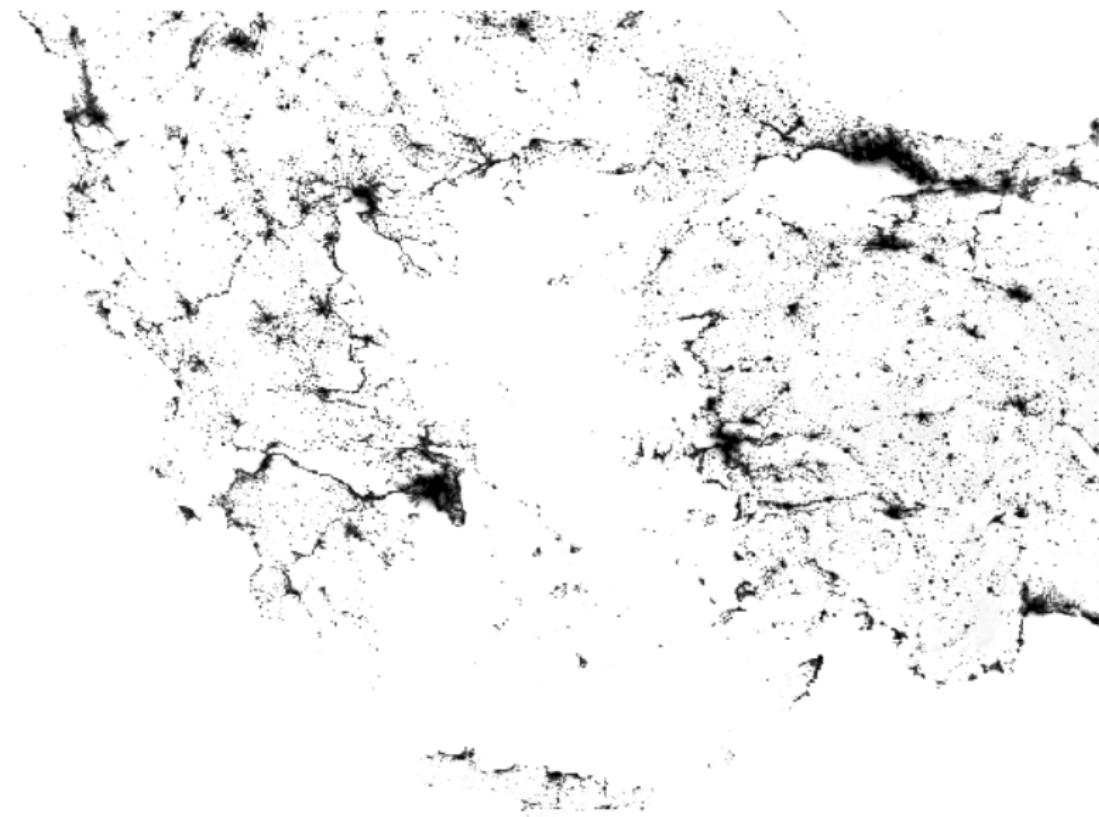
- 140 genes. Metazoa (Dunn et al. 2008)
- 1185 genes. Molluscs (Smith et al. 2011)
- 2970 genes. Seed plants (Lee et al. 2011)
- 8251 genes. Birds (Jarvis et al. 2014)
- 20,374 genes. Equids. (Jónsson et al. 2014)



Clustering



What is this?



What cities are connected by land?

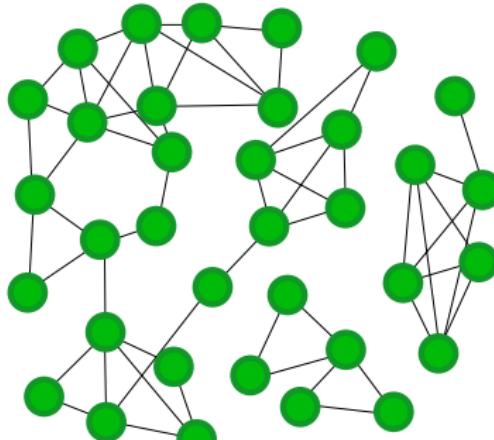


Methods

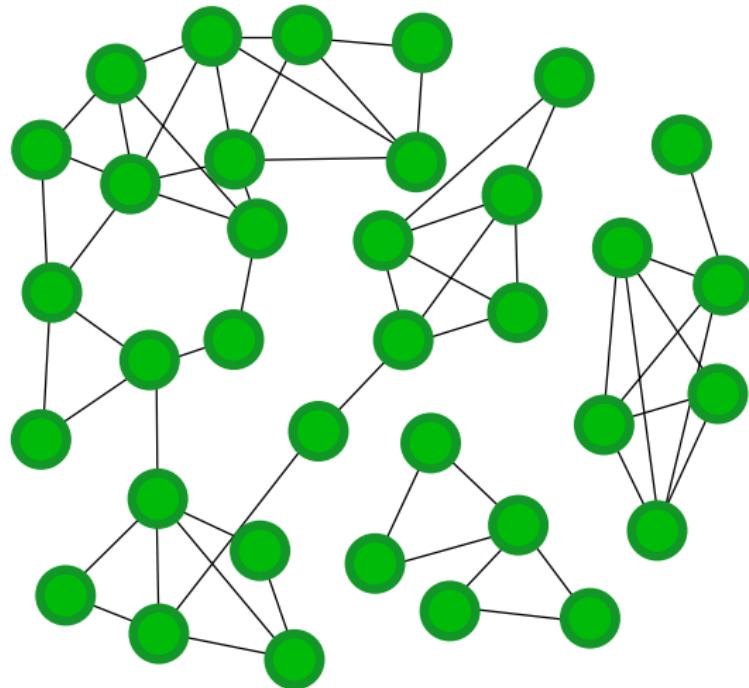
- Graph based methods
 - COG: Clusters of Orthologous Groups
 - Paranoid methods (inParanoid and MultiParanoid):
 - MCL methods (Markov CLustering)
- Tree based methods
 - COCO-CL: given homologs, calculates paralogs and orthologs

Simple concept

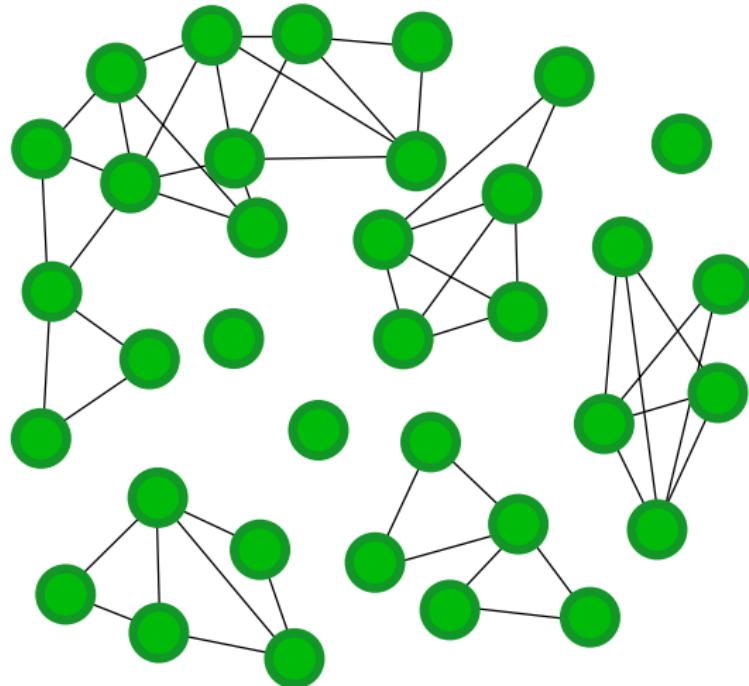
- Start at a node, then randomly travel to connected nodes.
- We will (on average) be more likely to stay within a cluster than to travel between clusters
- As we iterate, edges that are traveled frequently might increase in weight, while edges traveled rarely might be removed



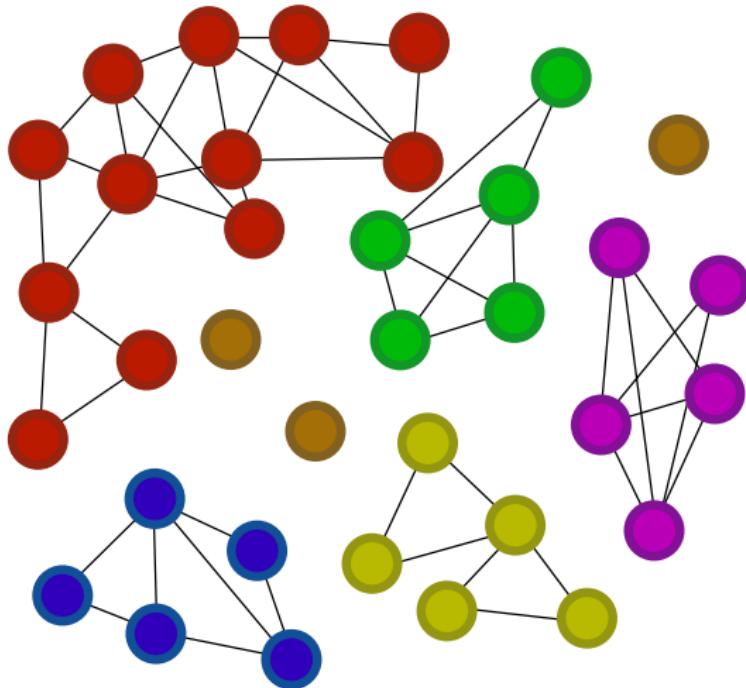
Simple visual example of clustering



Simple visual example of clustering

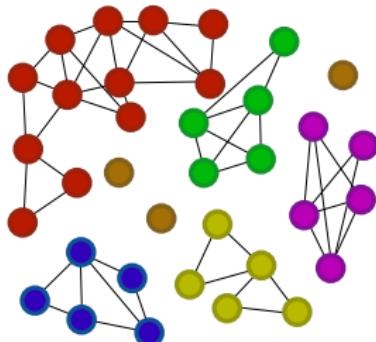
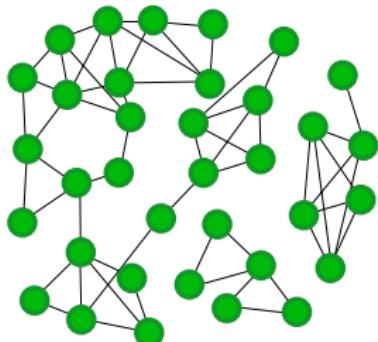


Simple visual example of clustering

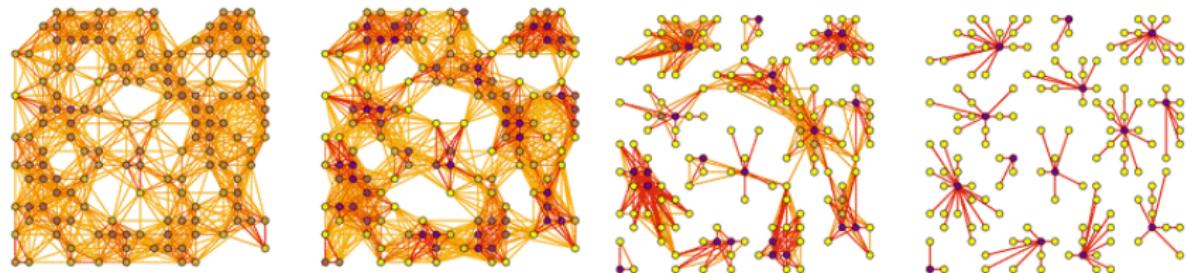


Markov clustering algorithm (MCL)

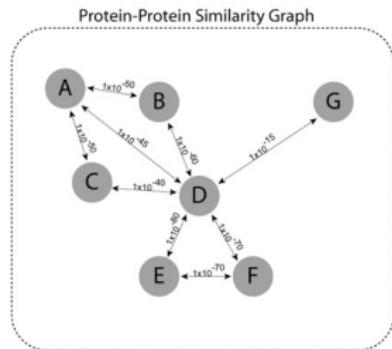
- Take a graph with vertices and edges
- Assume that there are clusters
- Start at a node and randomly travel between connected nodes
- We are more likely to stay within a cluster than to move between clusters
- These random walks (Markov chains) will help us discover clusters
- Stijn van Dongen, Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, May 2000.
(<http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm>)



Markov chain



Usage

A

Generate weighted transition matrix using BLAST E-Values as weights (-logE)

B

Weighted Transition Matrix

	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	80	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

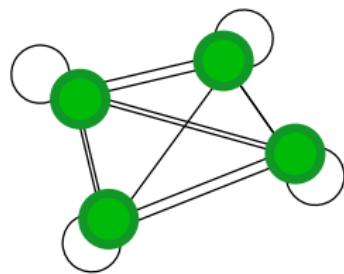
Transform weights into column-wise transition probabilities

Markov Matrix

	A	B	C	D	E	F	G
A	0.42	0.24	0.20	0.11	0.00	0.00	0.00
B	0.20	0.48	0.24	0.15	0.00	0.00	0.00
C	0.20	0.00	0.40	0.10	0.00	0.00	0.00
D	0.18	0.28	0.16	0.24	0.32	0.29	0.13
E	0.00	0.00	0.00	0.19	0.40	0.29	0.00
F	0.00	0.00	0.00	0.17	0.28	0.42	0.00
G	0.00	0.00	0.00	0.04	0.00	0.00	0.87

Markov chain

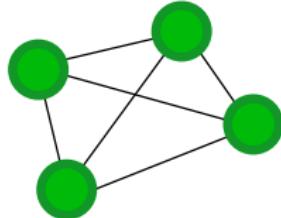
- We can describe the connections between nodes by a matrix



	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

Markov chain

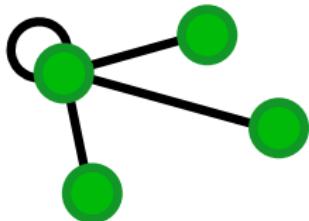
- Start at time step t_0
- Randomly travel between connected nodes
- Examine time steps $t_0 \rightarrow t_1 \rightarrow t_2$



	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

Markov chain

- Take one cell and do time step $t_0 \rightarrow t_1 \rightarrow t_2$
- We need to take into account
 - traveling on the A→A circuit = $(A \rightarrow A) * (A \rightarrow A) = 0.6 * 0.6$
 - **and** A→B then B→A = $(A \rightarrow B) * (B \rightarrow A) = 0.2 * 0.1$
 - **and** A→C then C→A = $(A \rightarrow C) * (C \rightarrow A) = 0.1 * 0.1$
 - **and** A→D then D→A = $(A \rightarrow D) * (D \rightarrow A) = 0.1 * 0.1$



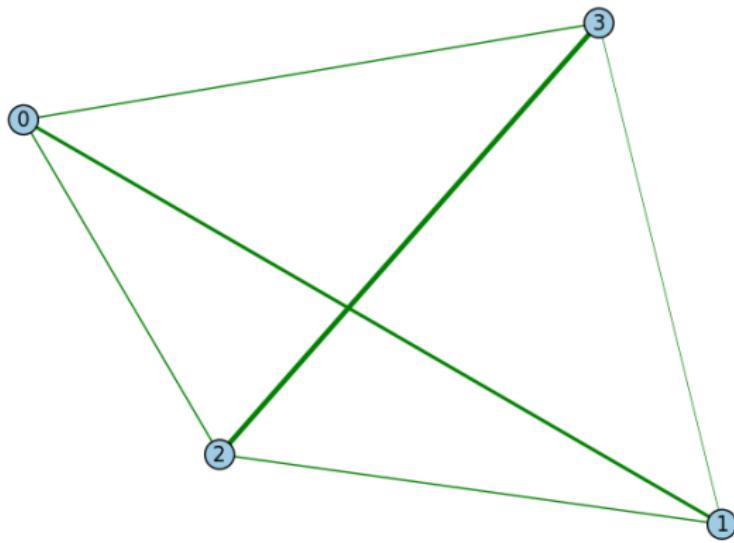
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0.6	0.1	0.1	0.1
<i>B</i>	0.2	0.8	0	0.05
<i>C</i>	0.1	0.1	0.6	0.15
<i>D</i>	0.1	0	0.3	0.7

$$0.6 \times 0.6 + 0.1 \times 0.2 + 0.1 \times 0.1 + 0.1 \times 0.1 = 0.4$$

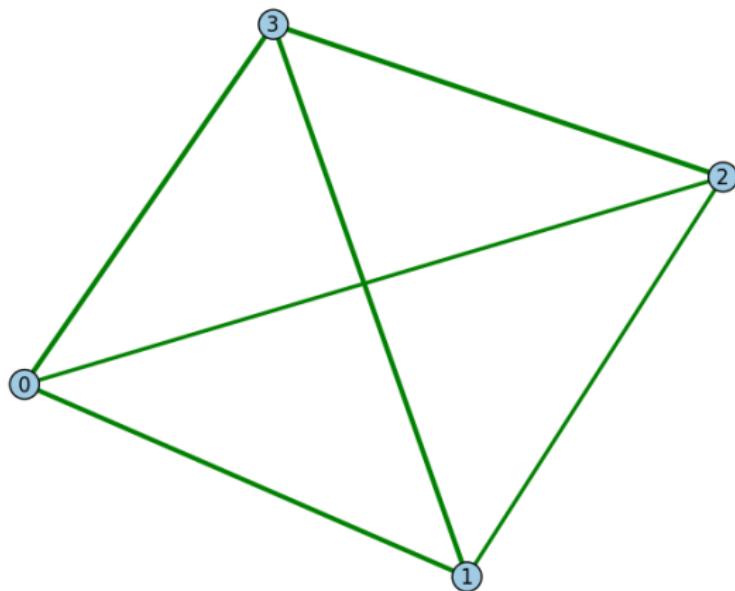
Markov chain

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0.6	0.1	0.1	0.1	<i>A</i>	0.4	0.15	0.15	0.15
<i>B</i>	0.2	0.8	0	0.05	<i>B</i>	0.285	0.66	0.035	0.095
<i>C</i>	0.1	0.1	0.6	0.15	<i>C</i>	0.155	0.15	0.415	0.21
<i>D</i>	0.1	0	0.3	0.7	<i>D</i>	0.16	0.04	0.4	0.545

Markov chain



Markov chain



Markov chain

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0.6	0.1	0.1	0.1	<i>A</i>	0.2	0.2	0.2	0.2
<i>B</i>	0.2	0.8	0	0.05	<i>B</i>	0.274	0.274	0.274	0.274
<i>C</i>	0.1	0.1	0.6	0.15	<i>C</i>	0.229	0.229	0.229	0.229
<i>D</i>	0.1	0	0.3	0.7	<i>D</i>	0.296	0.296	0.296	0.296

MCL Algorithm

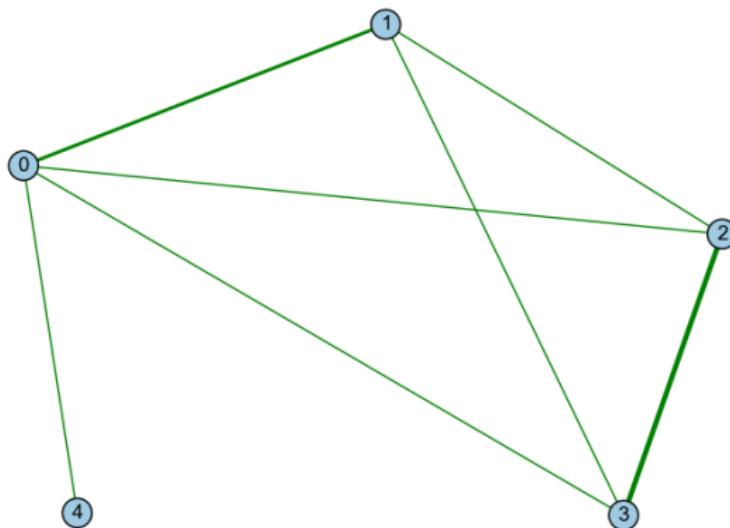
- Seems like the connections get more diffuse
- MCL does more than just the Markov walks

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0.6	0.1	0.1	0.1
<i>B</i>	0.2	0.8	0	0.05
<i>C</i>	0.1	0.1	0.6	0.15
<i>D</i>	0.1	0	0.3	0.7

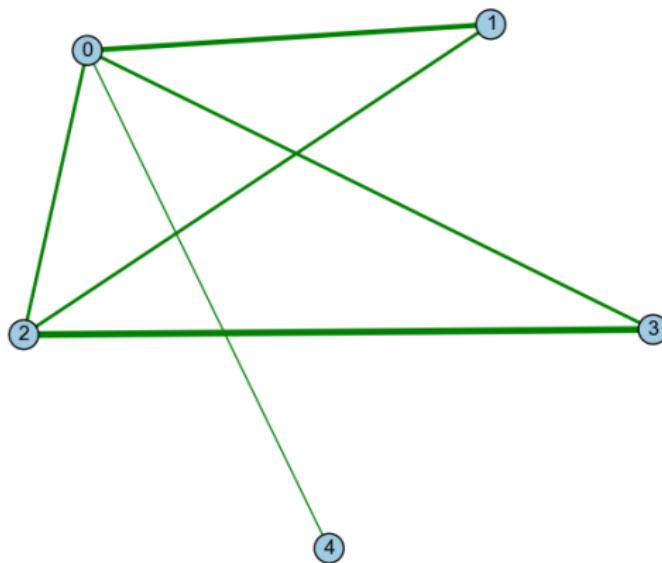
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0.2	0.2	0.2	0.2
<i>B</i>	0.274	0.274	0.274	0.274
<i>C</i>	0.229	0.229	0.229	0.229
<i>D</i>	0.296	0.296	0.296	0.296

- The clusters are more pronounced in the first than the last matrix
- MCL slightly changes the procedure to emphasize these divisions

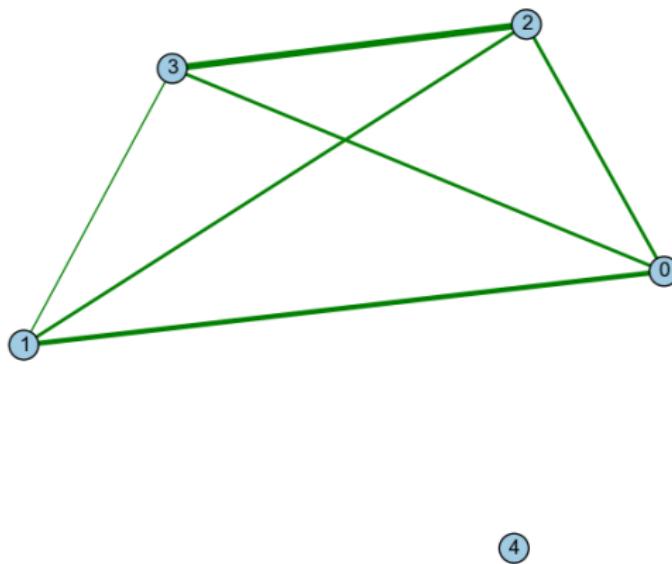
Markov chain



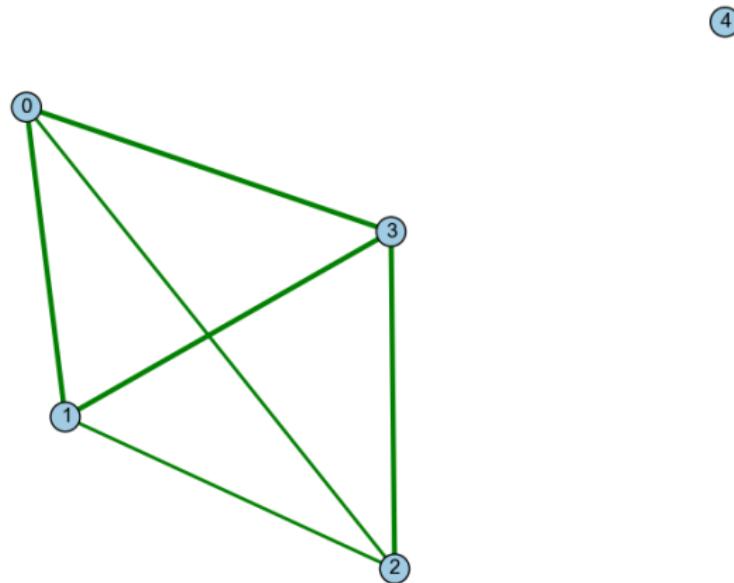
Markov chain



Markov chain



Markov chain



MCL Algorithm

- Major ways that MCL does this is in each cycle
 - adjusts the columns by raising them to a power (inflating them)
 - strengthens the strong neighbors
 - weakens the weak neighbors
- It “inflates” the clustering effect by raising each column to a non-negative power and then normalizing back so that they sum to one
- The Markovian multiplication is then called “expansion”
 - that matrix multiplication step we already saw
- Keep doing this until the matrix values do not change anymore

Algorithm

- 1 Input the all by all sequence comparisons (graph) and expansion parameter ϵ , and inflation parameter r
- 2 Construct the matrix (e.g., add self loops, normalize the matrix)
- 3 Until convergence or maximum number of iterations
 - 1 expand the matrix to the ϵ^{th} power (standard Markov chain)
 - 2 inflate the each column and normalize by raising the column to the power of r
- 4 Report the clusters

Inflation parameter

- r , the inflation parameter strengthens strong connections (current) and weakens already weak currents
- the larger the inflation parameter the greater the effect
- here is a view of the difference

	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

	A	B	C	D
A	0.4	0.15	0.15	0.15
B	0.285	0.66	0.035	0.095
C	0.155	0.15	0.415	0.21
D	0.16	0.04	0.4	0.545

	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

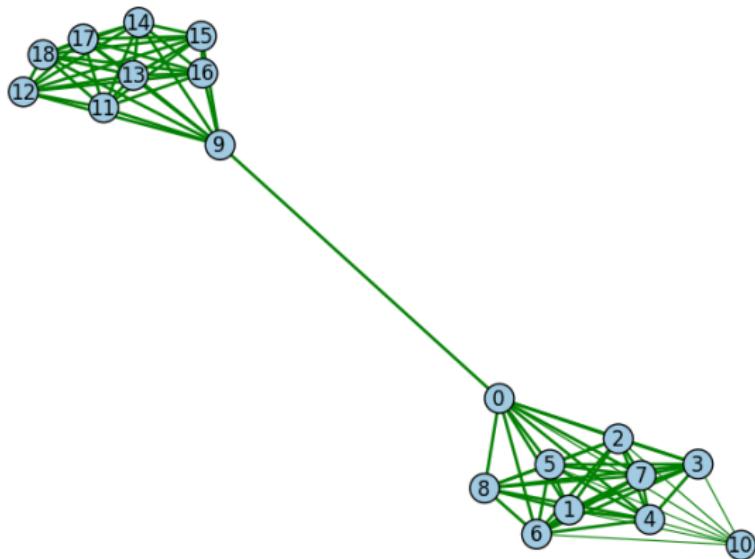
	A	B	C	D
A	0.85	0.015	0.02	0.019
B	0.09	0.96	0.	0.004
C	0.02	0.015	0.78	0.04
D	0.02	0.	0.19	0.93

Typical usage

- r inflation parameter will emphasize the strong connections
 - the higher the value, the more the clusters are split up
- e expansion parameter will dissipate the matrix (smooth things out)
 - this is the standard Markov walk
 - the higher the value, the less the clusters are emphasized
- it is typical that inflation parameters are 1.1~3

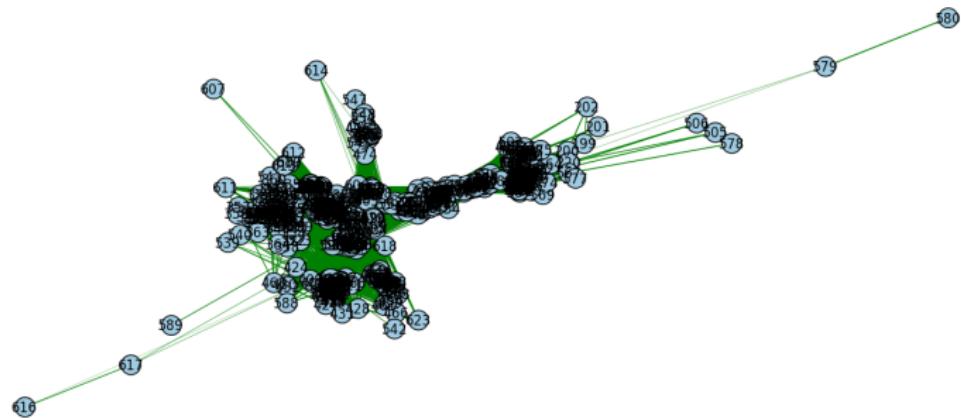
Trivial example

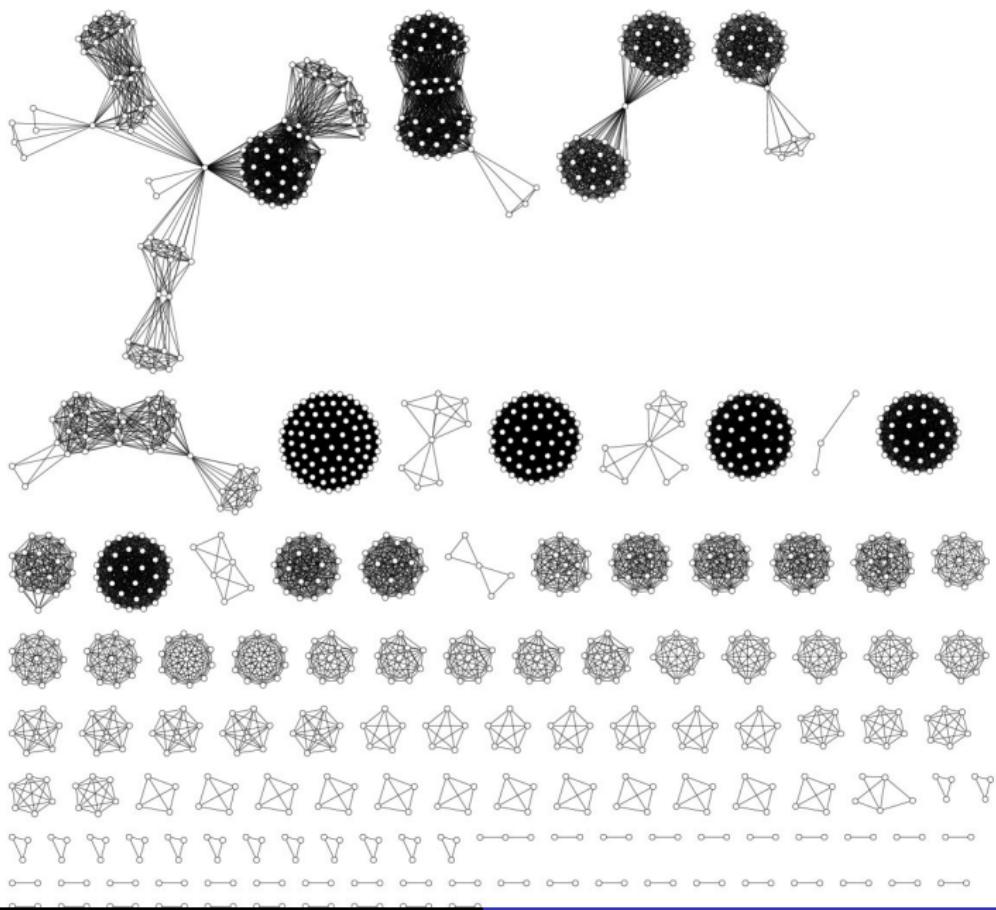
- Two gene regions with one spurious connection

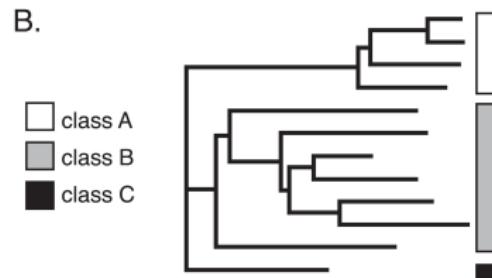
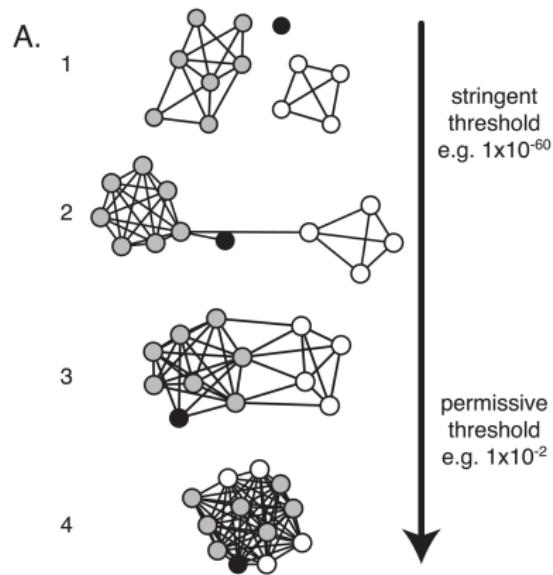


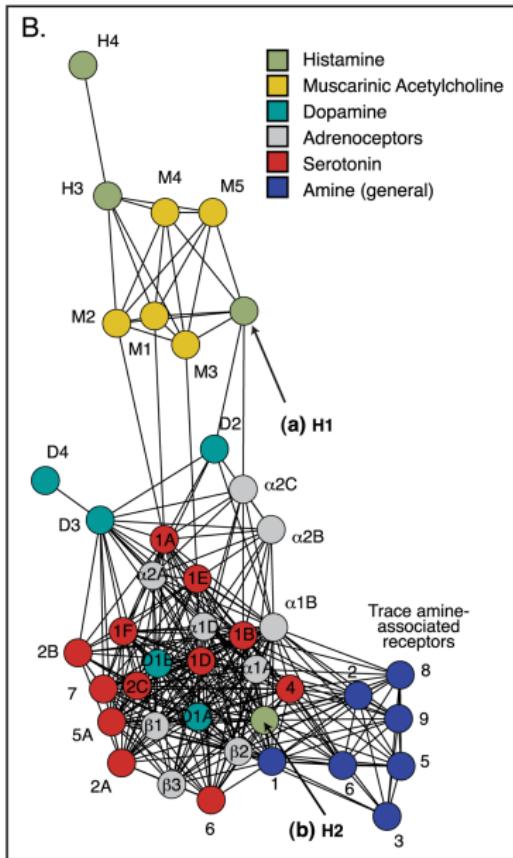
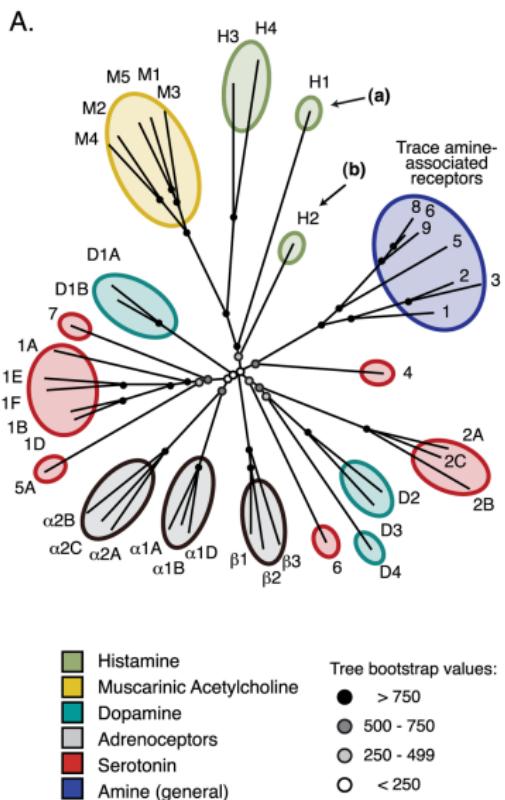
Complex example

- One gene regions many weak connections
- Results in 16 clusters







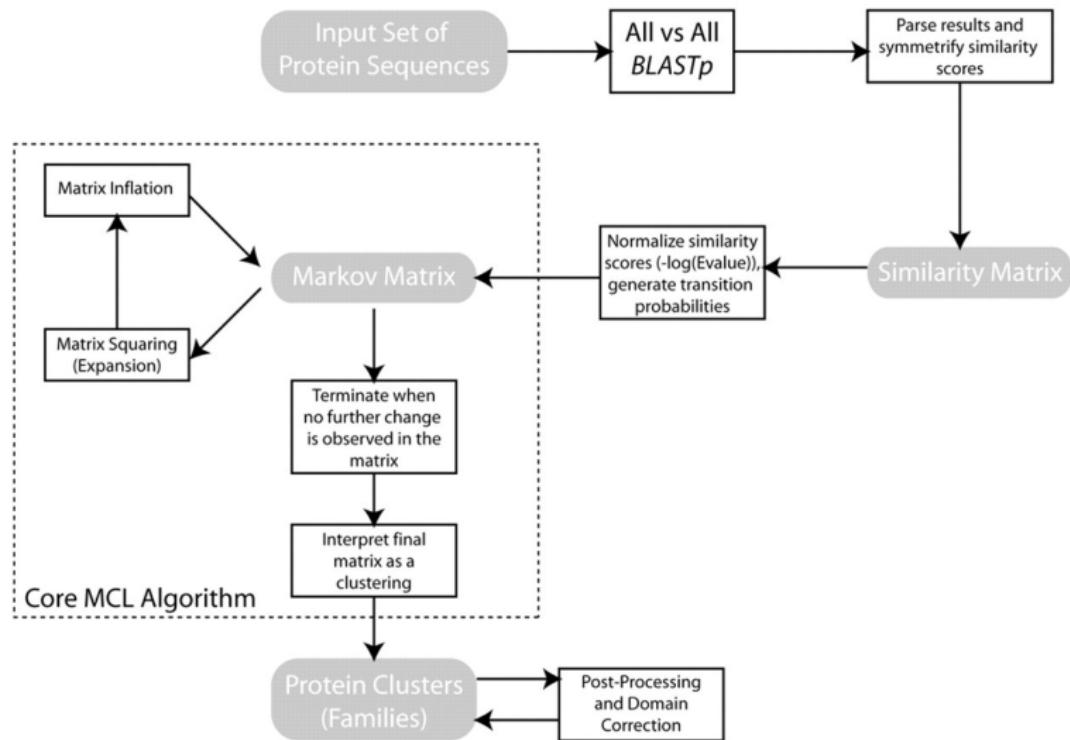


Atkinson et al., 2009

James B. Pease

Clustering

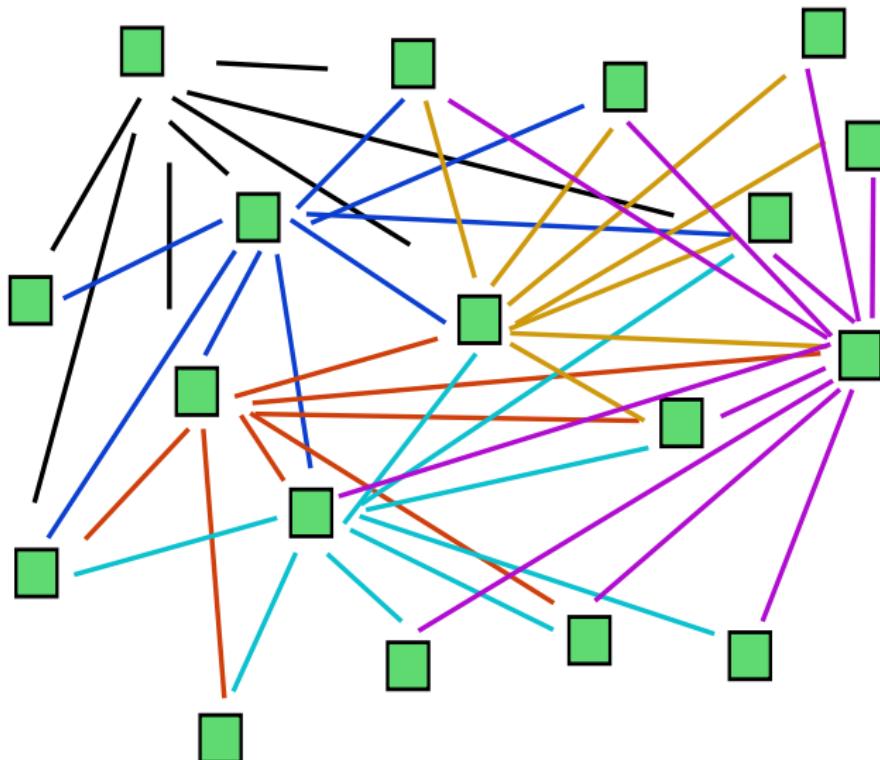
Usage



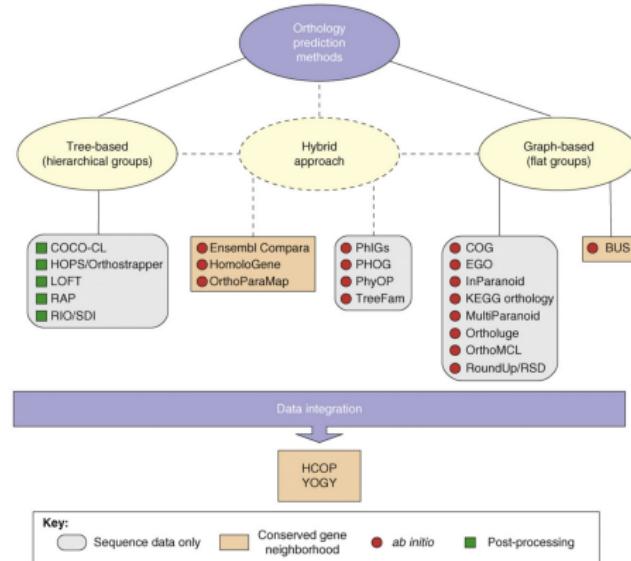
MCL

- MCL can be successful, in part due to simplicity and speed
- One of the most commonly used methods
- There are situations where MCL does not perform as well

- All-versus-all comparisons can take a very long time
- They also scale poorly



Other options



Kuzniar et al., 2008

- There is a lot of room for development
 - Develop your own!

Caryophyllales phylogeny



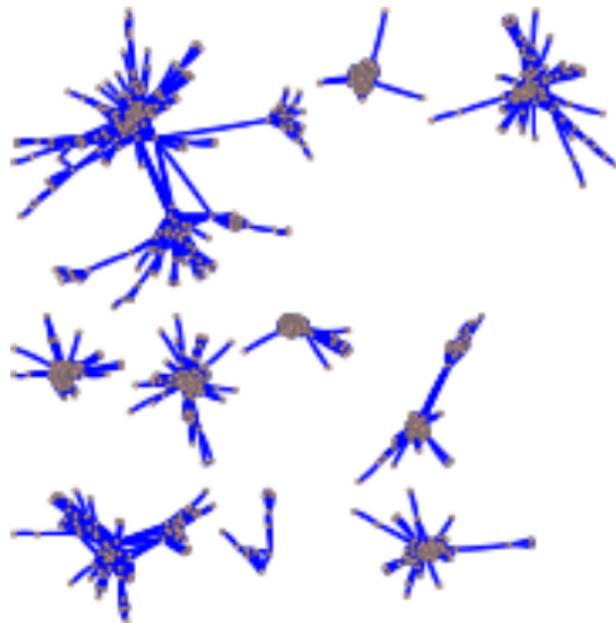
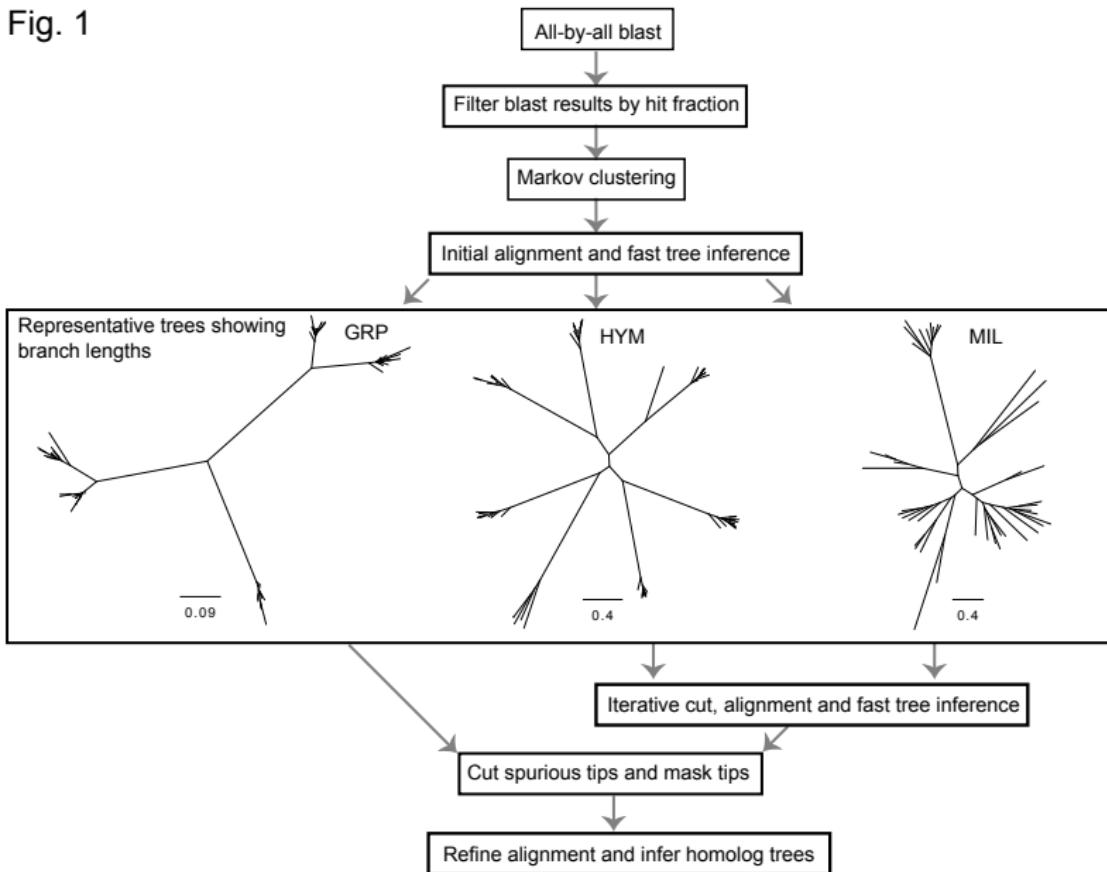
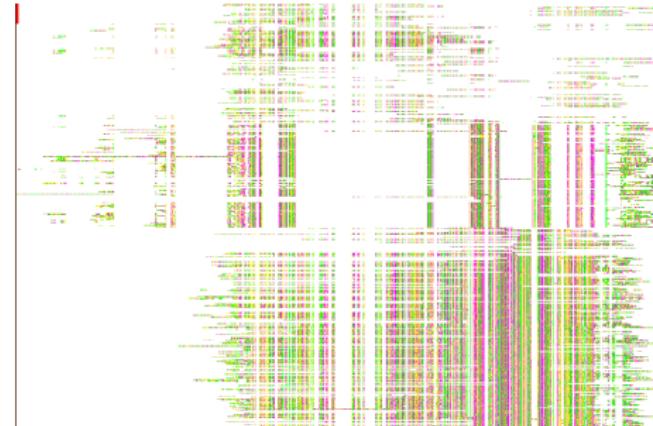
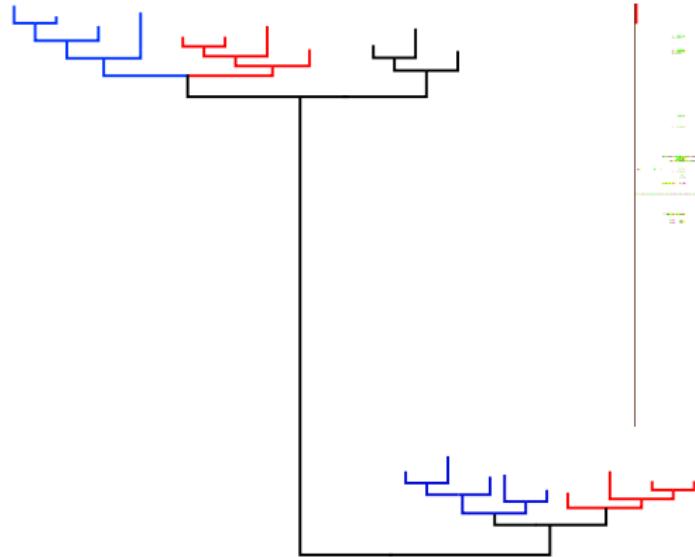


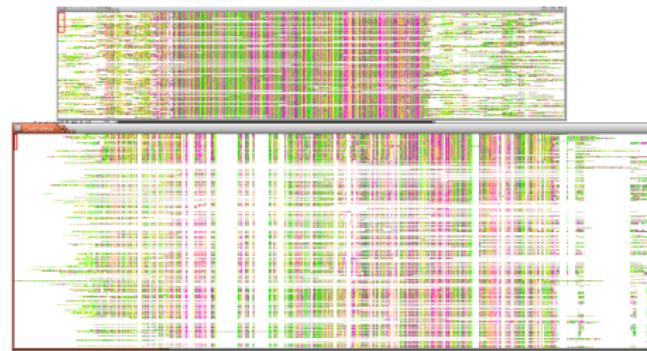
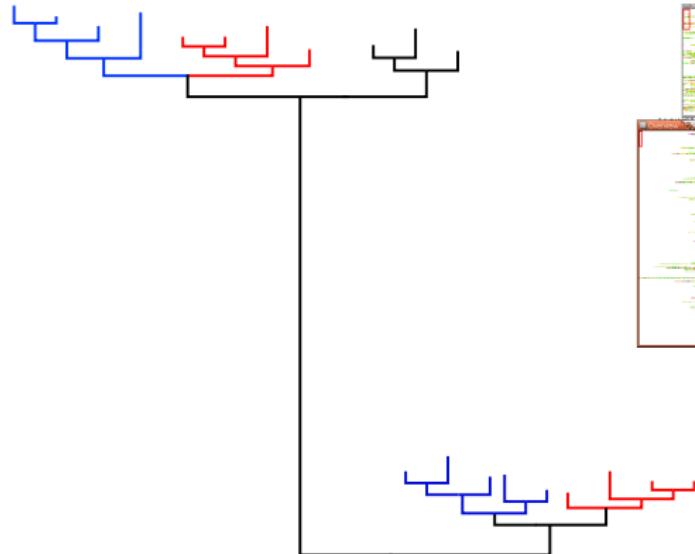
Fig. 1



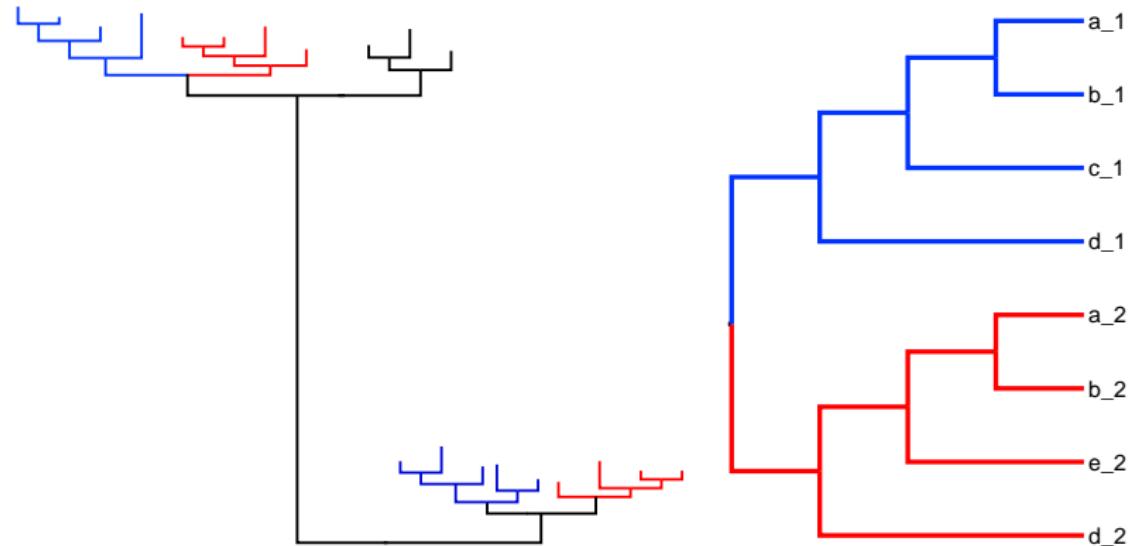
Homolog assignment



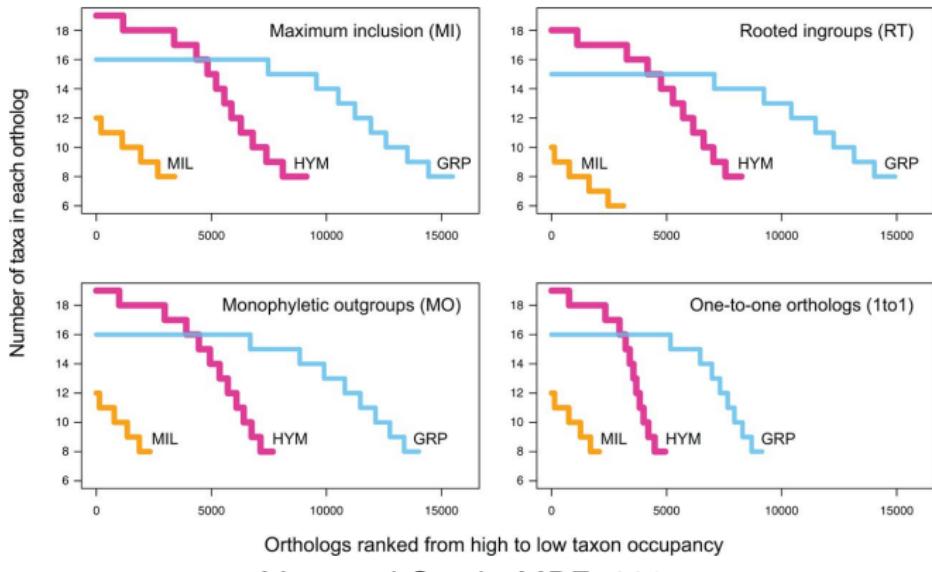
Homolog assignment



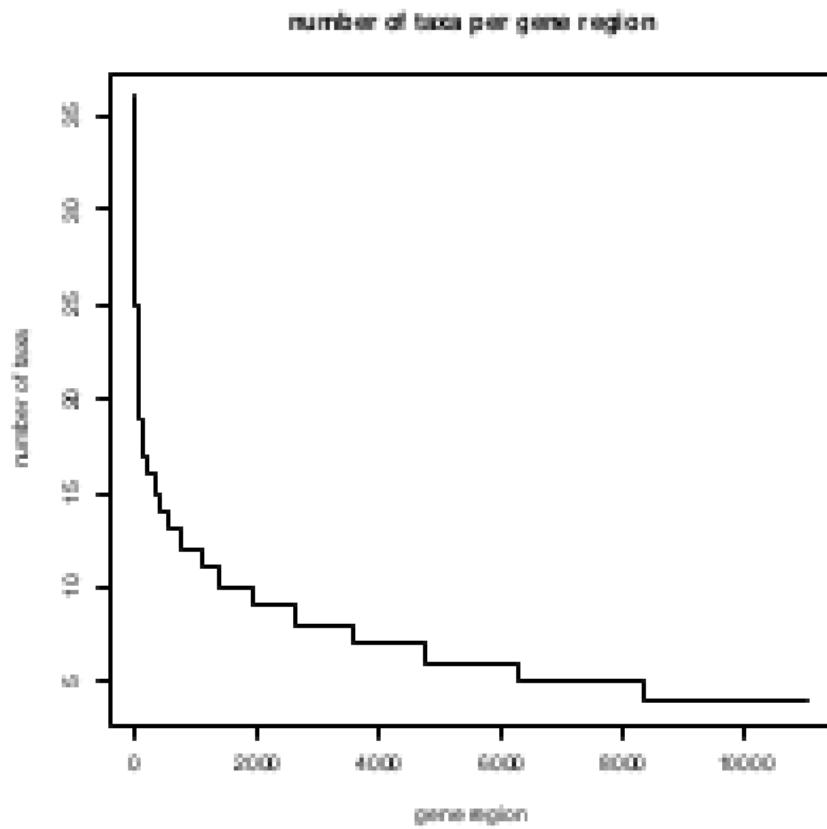
Orthology assignment



Ingroups: A, B, C Outgroup: O Inferred orthologs: ○	Maximum inclusion (MI)	Rooted ingroups (RT)	Monophyletic outgroups (MO)	One-to-one orthologs (1to1)
Outgroup present but non-monophyletic duplicated taxa present				
Outgroup present and monophyletic duplicated taxa present				
Outgroup absent duplicated taxa present				
Outgroup present or absent duplicated taxa absent				

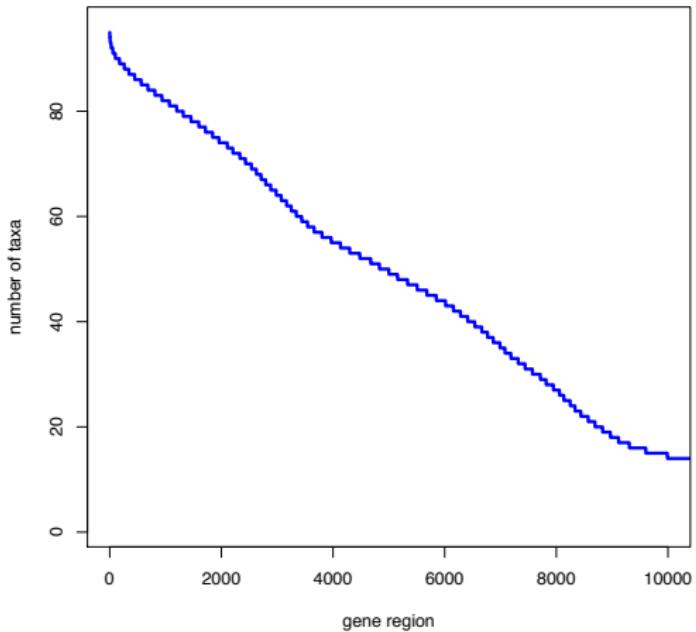


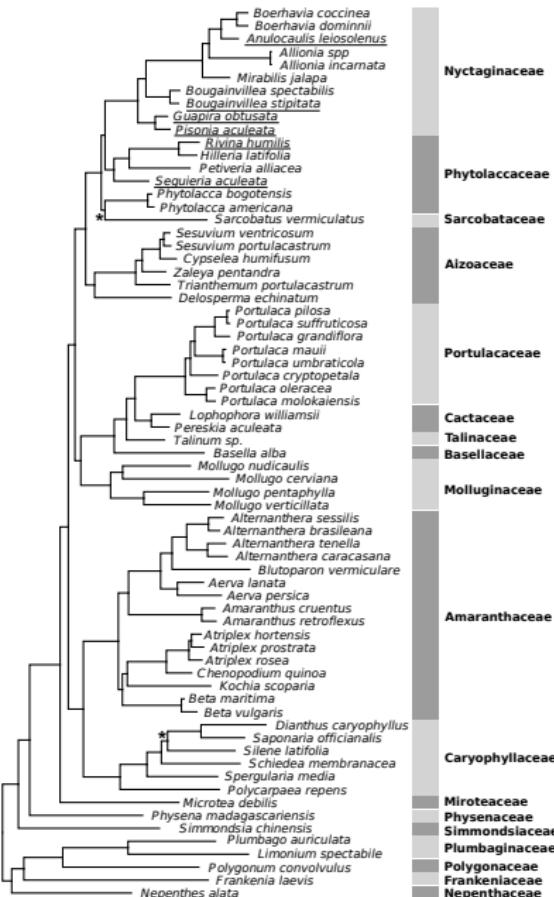
Yang and Smith, MBE, 2014



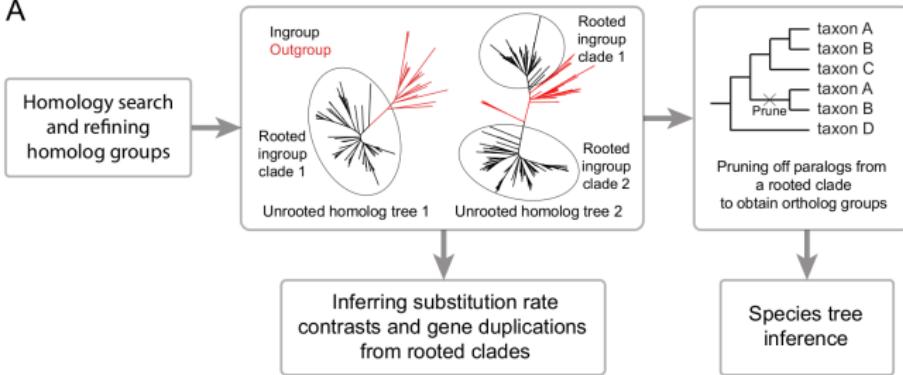
# genes	occupancy
10457	50%
5507	70%
1068	90%

(compare to with 300 genes at 50%)

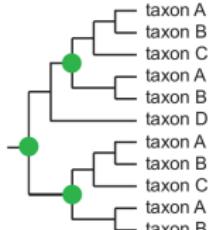




- 96 species (68 ingroup), Illumina (1kp and our own)
- 1000-10000 gene regions (nucleotide and amino acid)
- Stephen Smith, Ya Yang, Michael Moore, and Sam Brockington

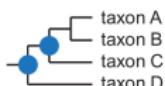
A**B**

● = Locations of gene duplication detected from a homolog tree



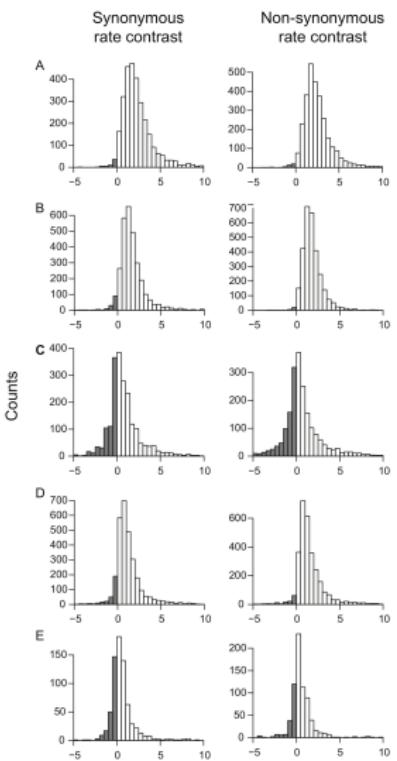
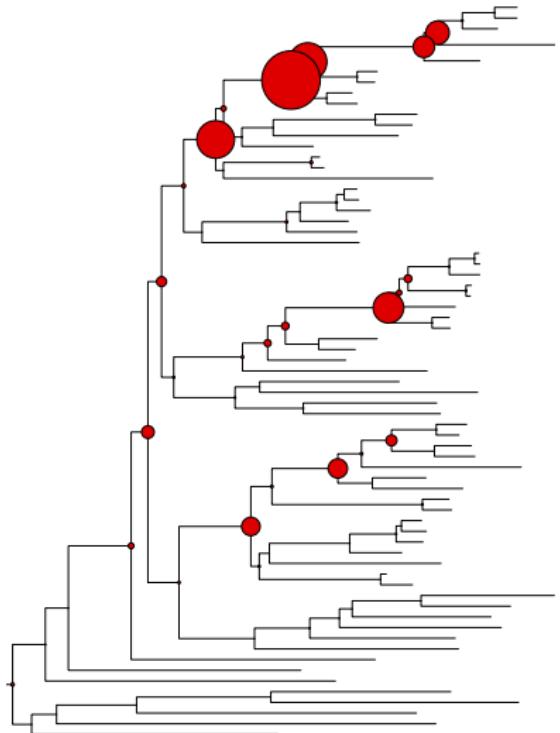
Rooted clades extracted from homologs

● = Gene duplication events from a homolog tree mapped to the species tree. When nested duplications are detected, only one event per node is counted

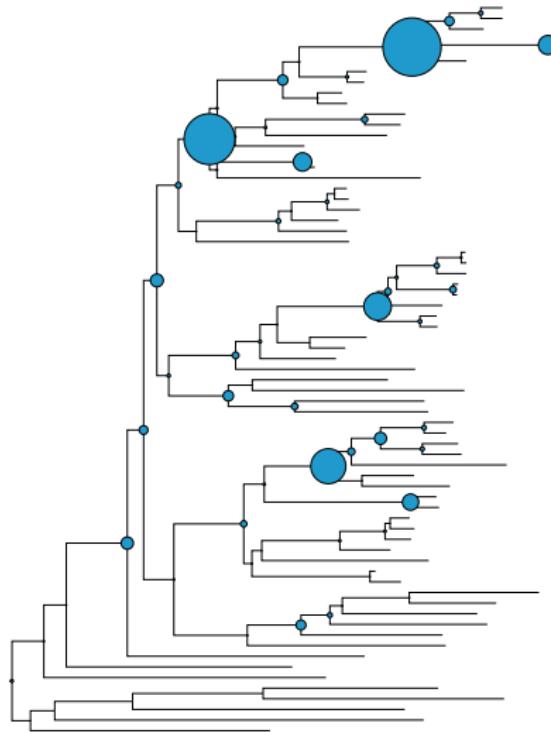


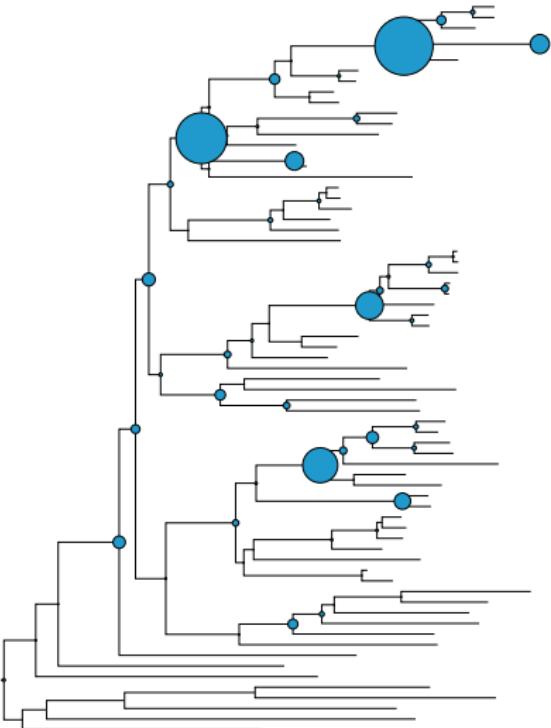
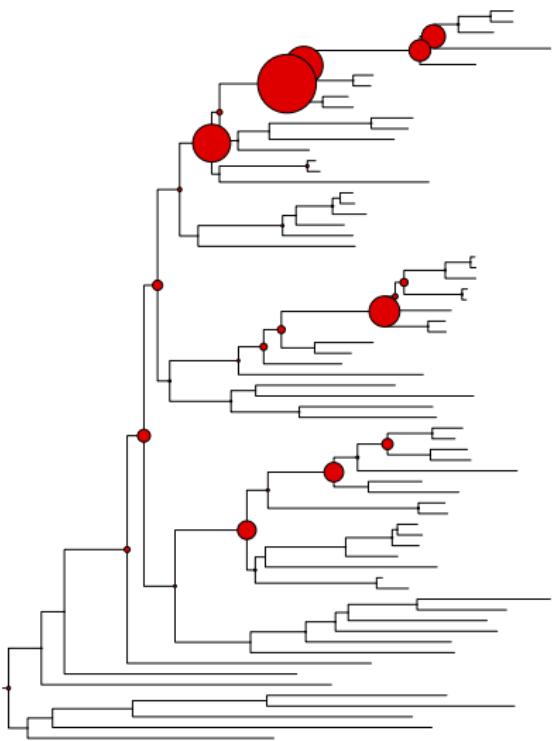
Species tree

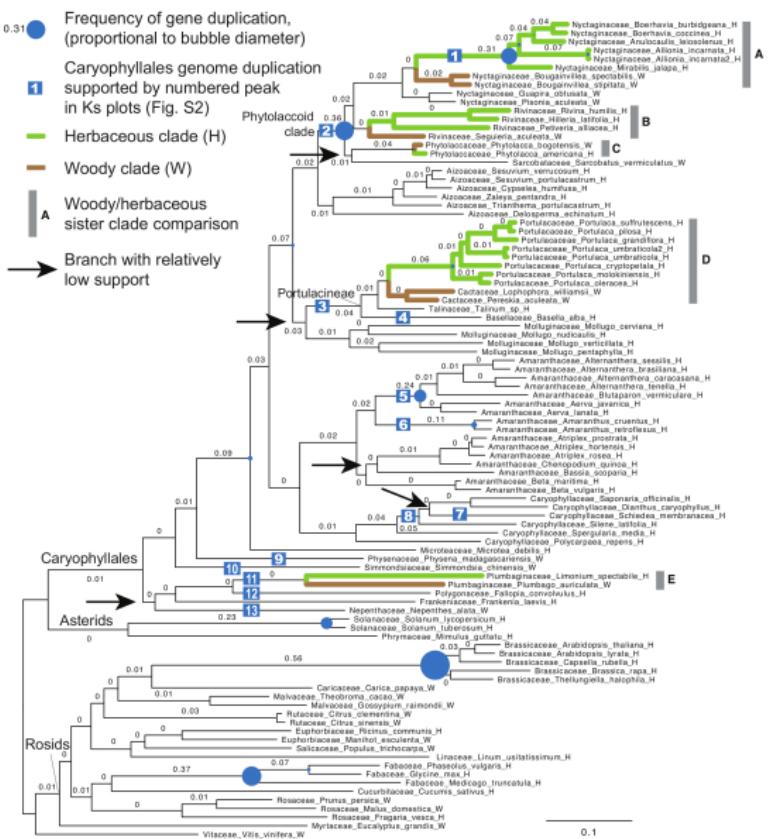
Rate shifts

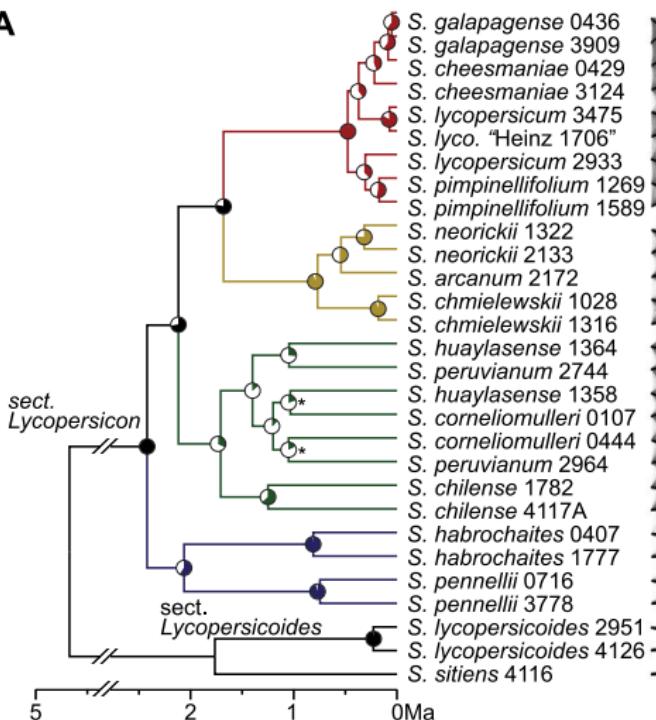
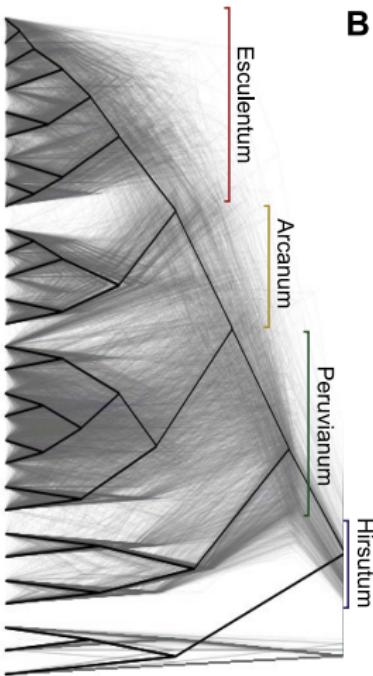


Gene duplications







A**B**

How to approach clustering

- When gene regions are well defined and understood:
 - Can use methods like baited BLAST
 - Other methods (hidden Markov model based methods)
- When gene regions are not well understood (naïve):
 - Can use MCL (most common solution)
 - Other methods not discussed here
 - Develop your own method!

Other topics to look into

- Hierarchical clustering methods
 - taxonomies
 - other hierarchical clustering methods
- Methods that require multiple sequence alignments
 - using hidden Markov models
- Methods that require phylogenetic trees ←←
- Species tree reconciliation