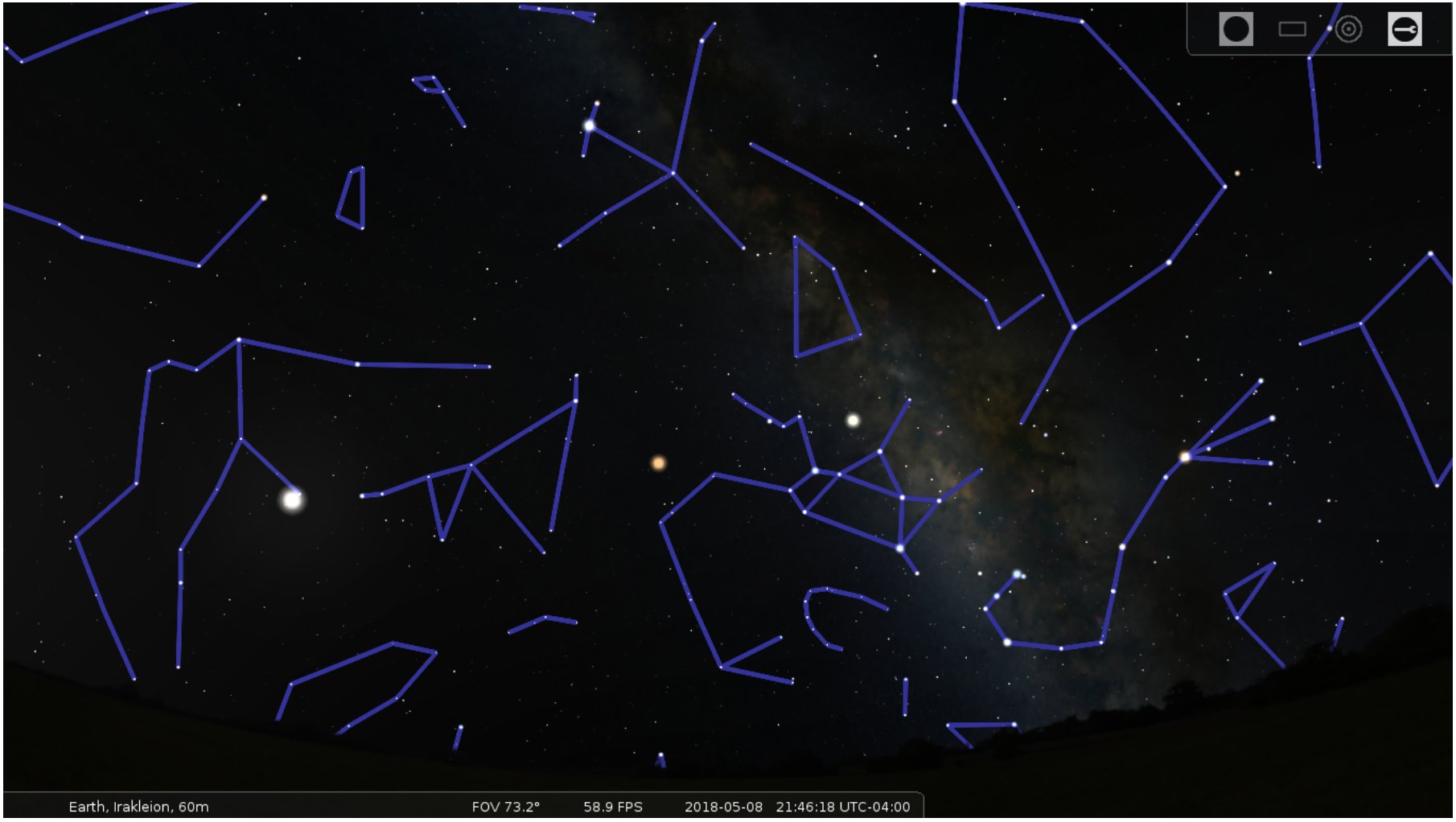
Clustering

James B. Pease

Wake Forest University

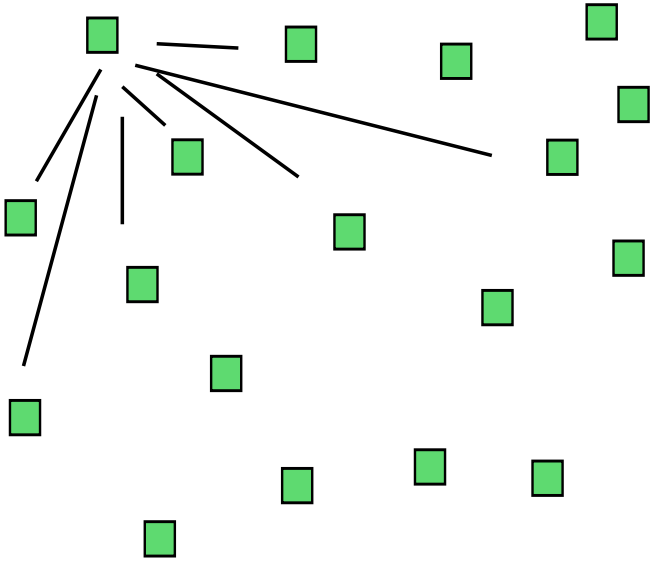

Earth, Irakleion, 60m
FOV 73.2°
59.1 FPS
2018-05-08 21:46:22 UTC-04:00


Earth, Irakleion, 60m
FOV 73.2°
58.9 FPS
2018-05-08 21:46:18 UTC-04:00

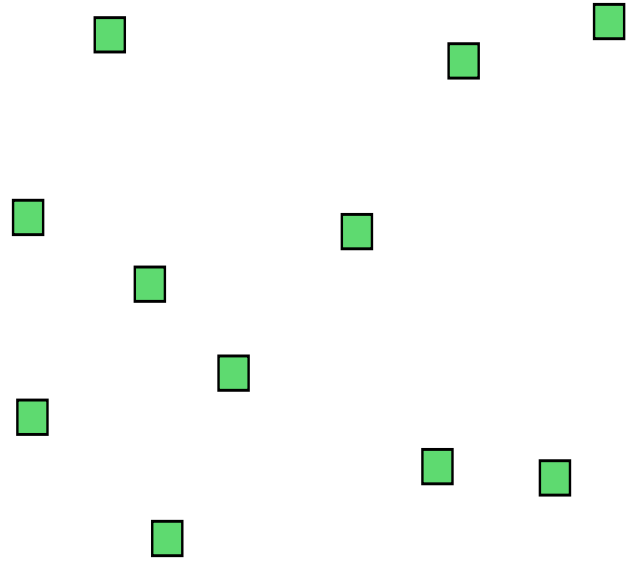
Sets of sequences

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

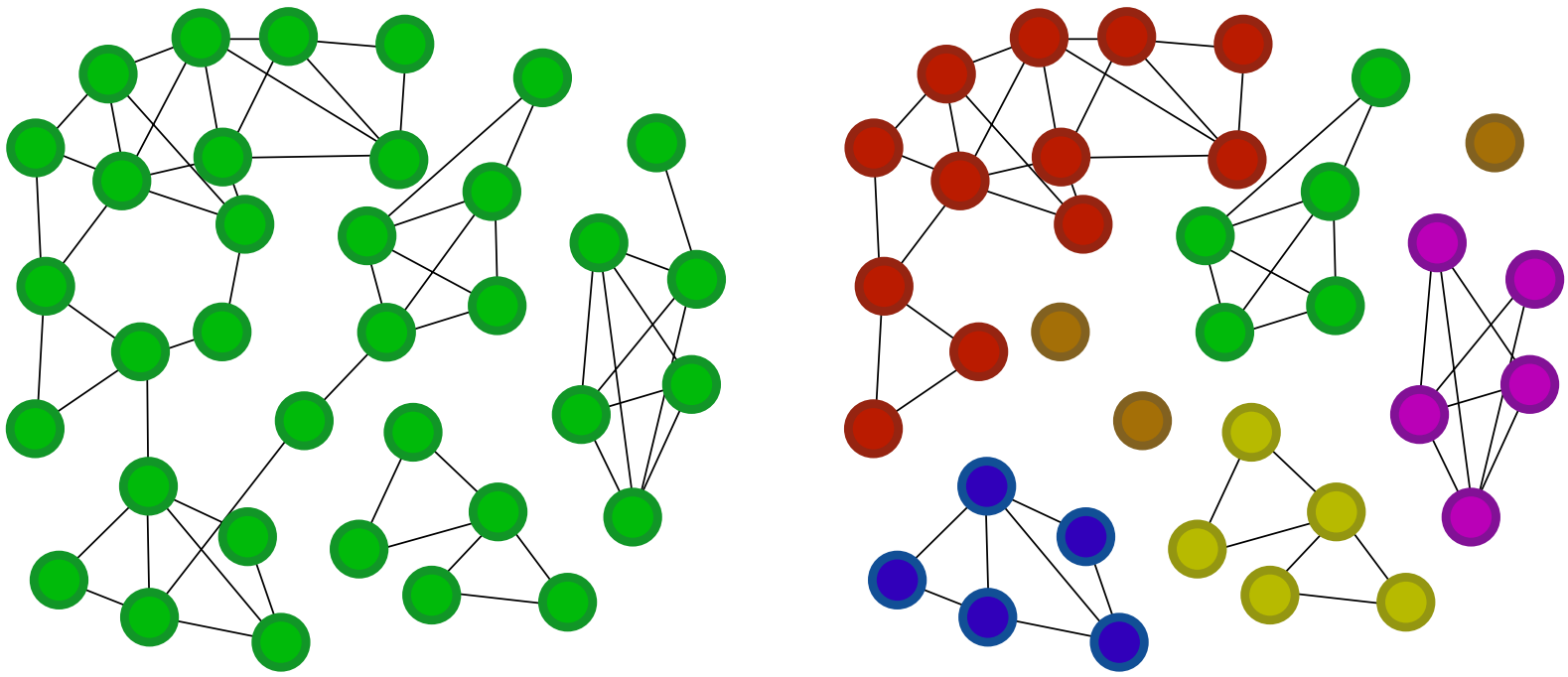
sequences



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

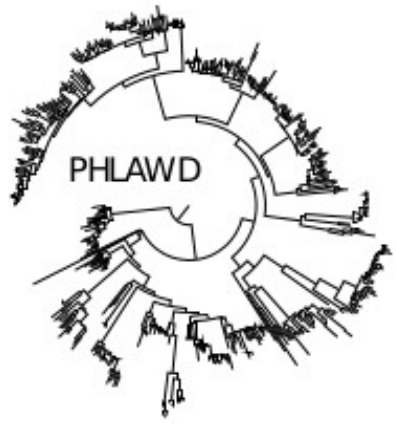
Hard to get gene identifications for all of GenBank (e.g., clustering)

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`

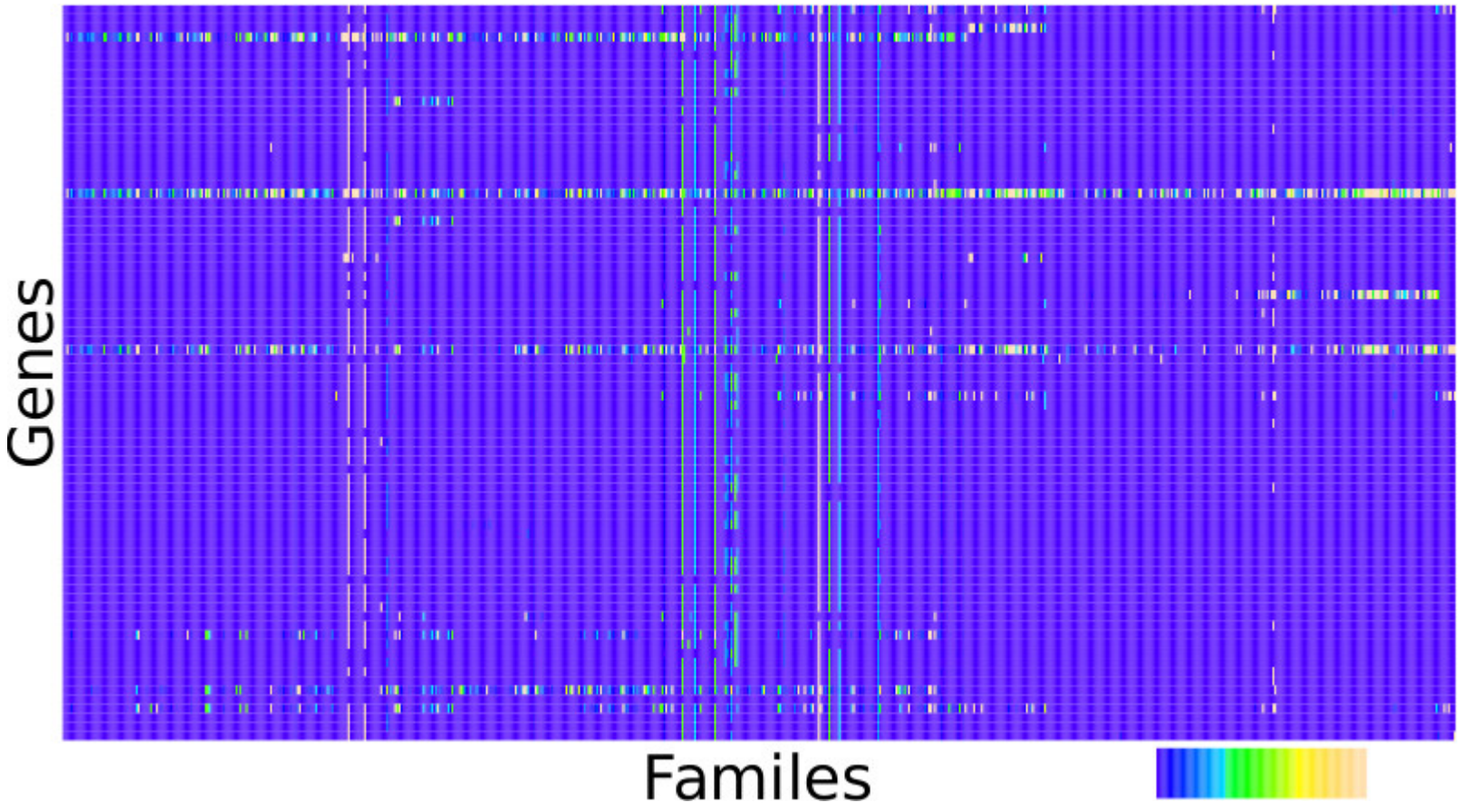
Smith et al., 2009, *BMC Evol. Bio.*

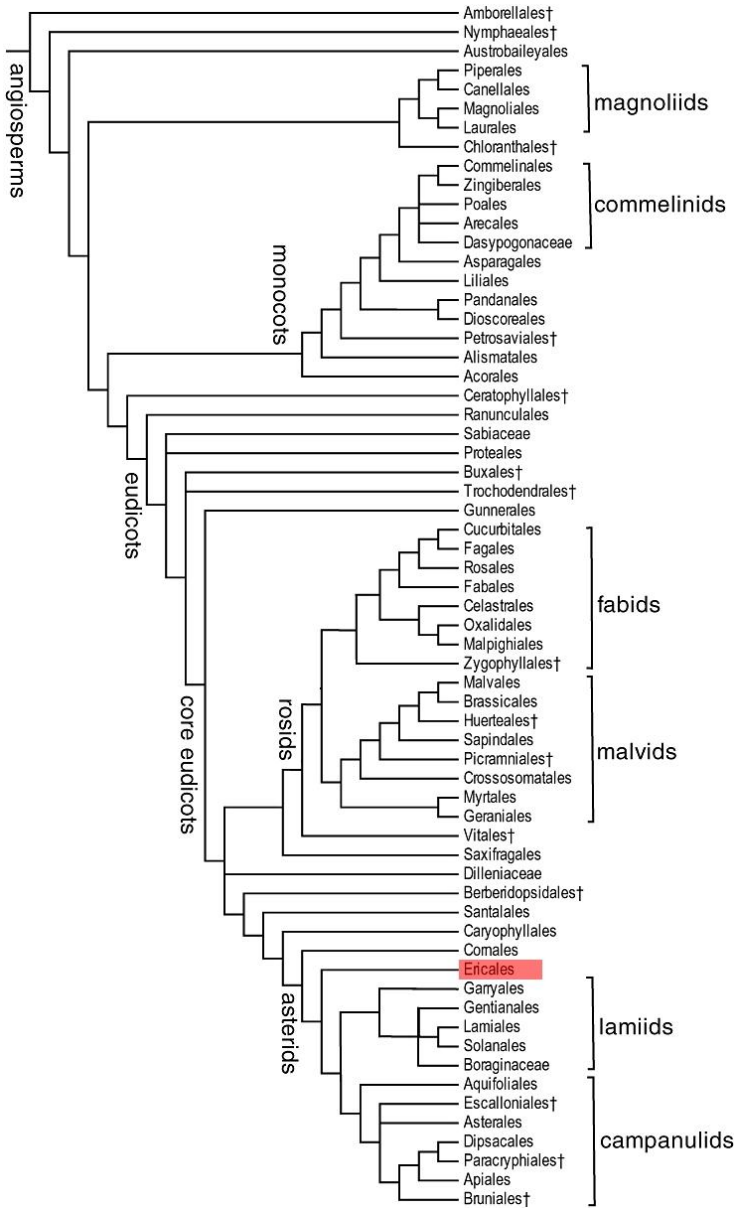
baited Smith-Waterman

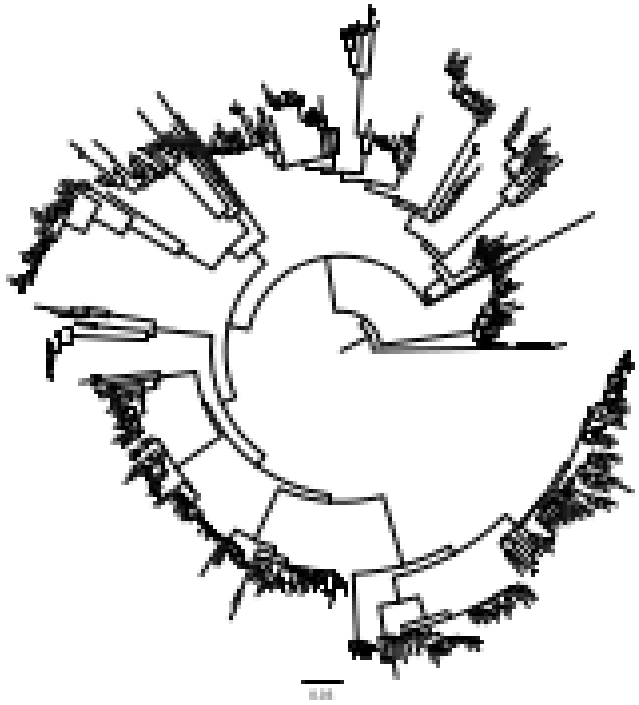
provide known homologous sequences

compare a set of unknown sequences to those known sequences

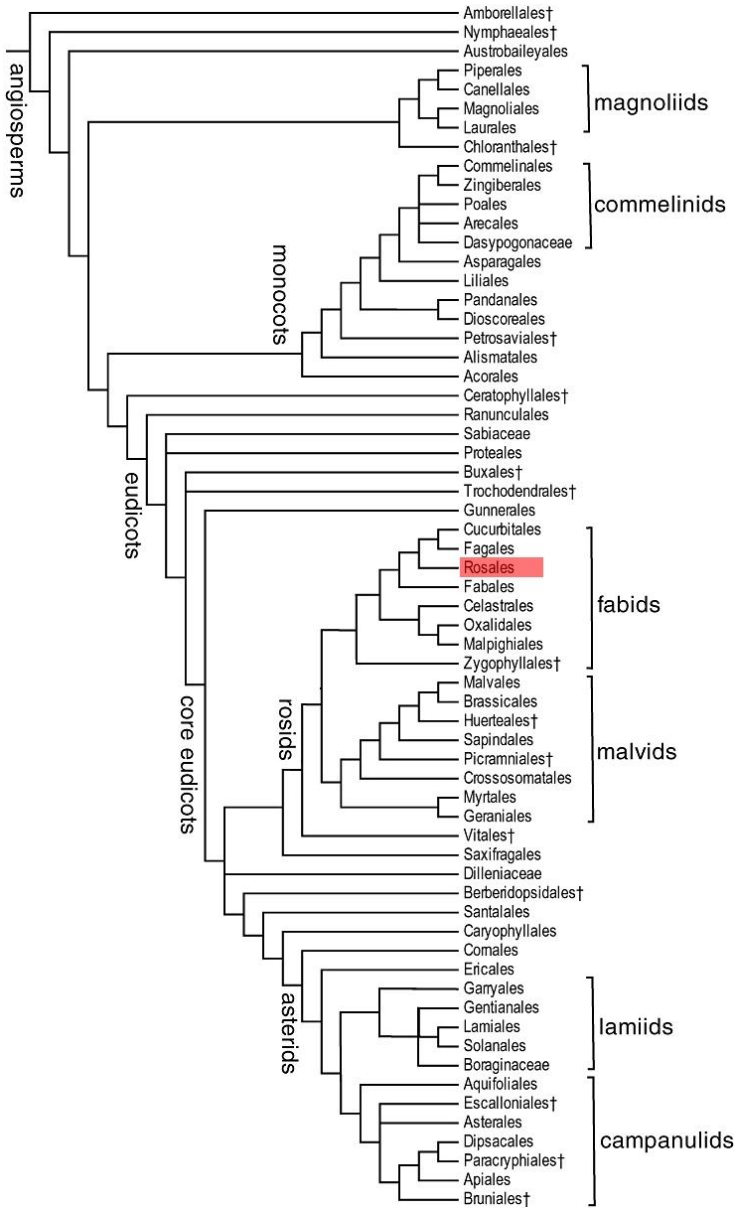
keep the significant hits

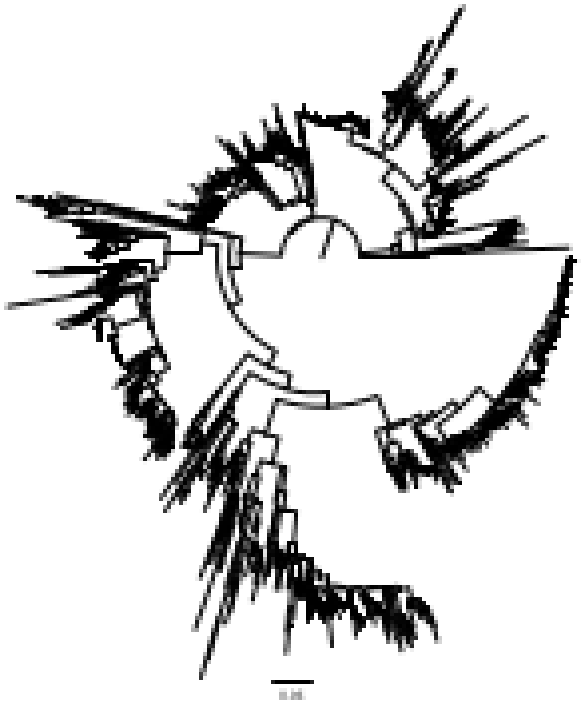

Genes
Familes


Amborellales†
Nymphaeales†
Austrobaileyales
Piperales
Canellales
Magnoliales
Laurales
Chloranthales†
Commelinales
Zingiberales
Poales
Arecales
Dasypogonaceae
Asparagales
Liliales
Pandanales
Dioscoreales
Petrosaviales†
Alismatales
Acorales
Ceratophyllales†
Ranunculales
Sabiaceae
Proteales
Buxales†
Trochodendrales†
Gunnerales
Cucurbitales
Fagales
Rosales
Fabales
Celastrales
Oxalidales
Malpighiales
Zygophyllales†
Malvales
Brassicales
Huertealesl†
Sapindales
Picramniales†
Crossosomatales
Myrtales
Geraniales
Vitales†
Saxifragales
Dilleniaceae
Berberidopsidales†
Santalales
Caryophyllales
Cornales
Ericales
Garryales
Gentianales
Lamiales
Solanales
Boraginaceae
Aquifoliales
Escalloniales†
Asterales
Dipsacales
Paracryphiales†
Apiales
Bruniales†
magnoliids
commelinids
fabids
malvids
lamiids
campanulids
angiosperms
monocots
eudicots
core eudicots
rosids
asterids

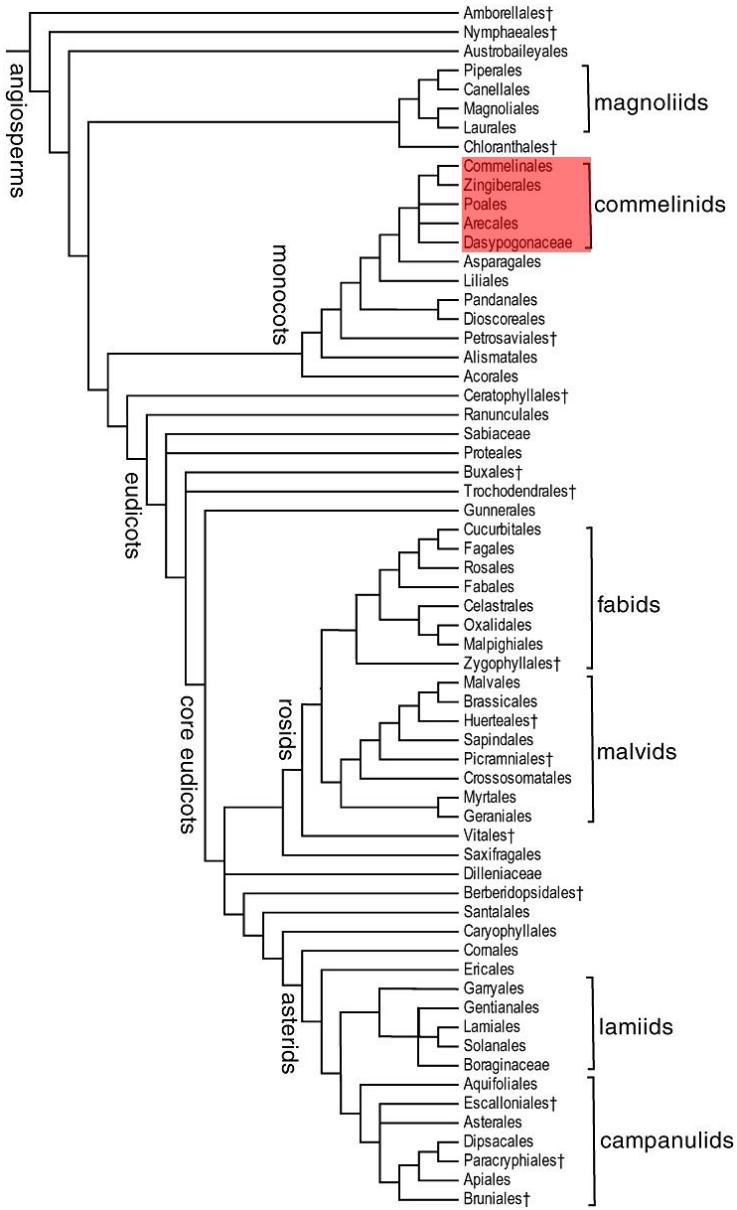


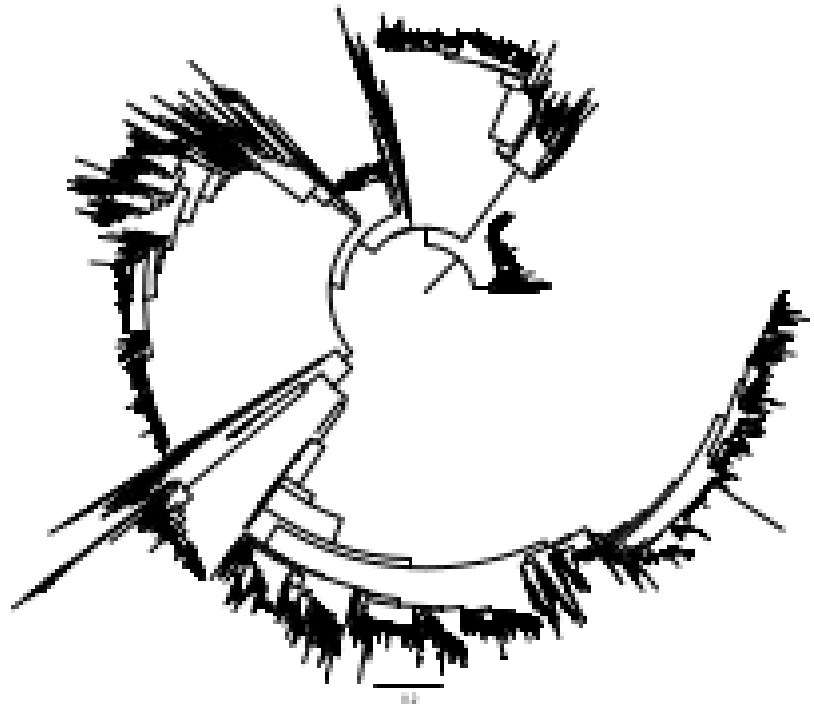
Ericales


Amborellales†
Nymphaeales†
Austrobaileyales
Piperales
Canellales
Magnoliales
Laurales
Chloranthales†
Commelinales
Zingiberales
Poales
Arecales
Dasypogonaceae
Asparagales
Liliales
Pandanales
Dioscoreales
Petrosaviales†
Alismatales
Acorales
Ceratophyllales†
Ranunculales
Sabiaceae
Proteales
Buxales†
Trochodendrales†
Gunnerales
Cucurbitales
Fagales
Rosales
Fabales
Celastrales
Oxalidales
Malpighiales
Zygophyllales†
Malvales
Brassicales
Huertealesٹ
Sapindales
Picramniales†
Crossosomatales
Myrtales
Geraniales
Vitales†
Saxifragales
Dilleniaceae
Berberidopsidales†
Santalales
Caryophyllales
Cornales
Ericales
Garryales
Gentianales
Lamiales
Solanales
Boraginaceae
Aquifoliales
Escalloniales†
Asterales
Dipsacales
Paracryphiales†
Apiales
Bruniales†
magnoliids
commelinids
fabids
malvids
lamiids
campanulids
angiosperms
monocots
eudicots
core eudicots
rosids
asterids



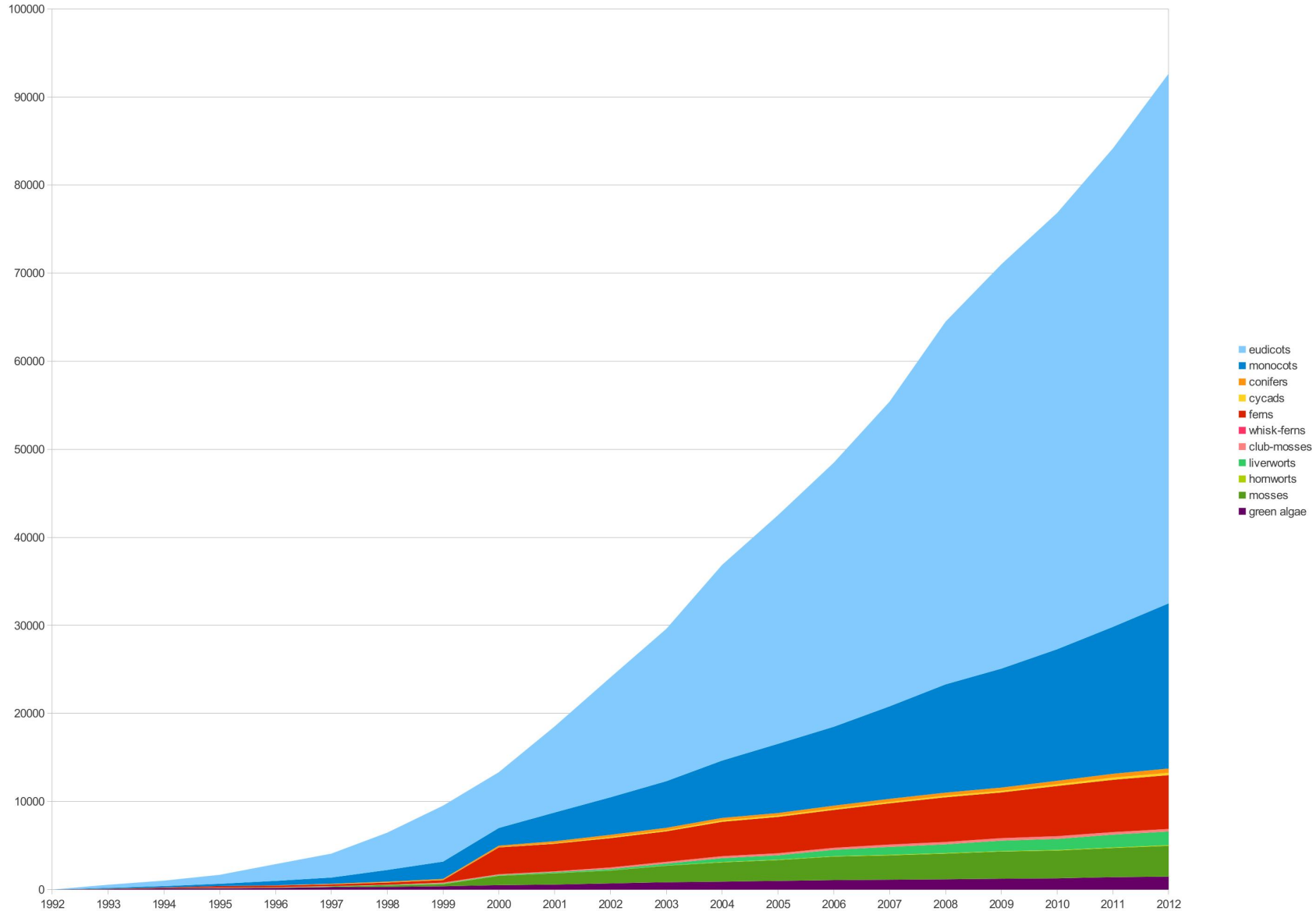
Rosales

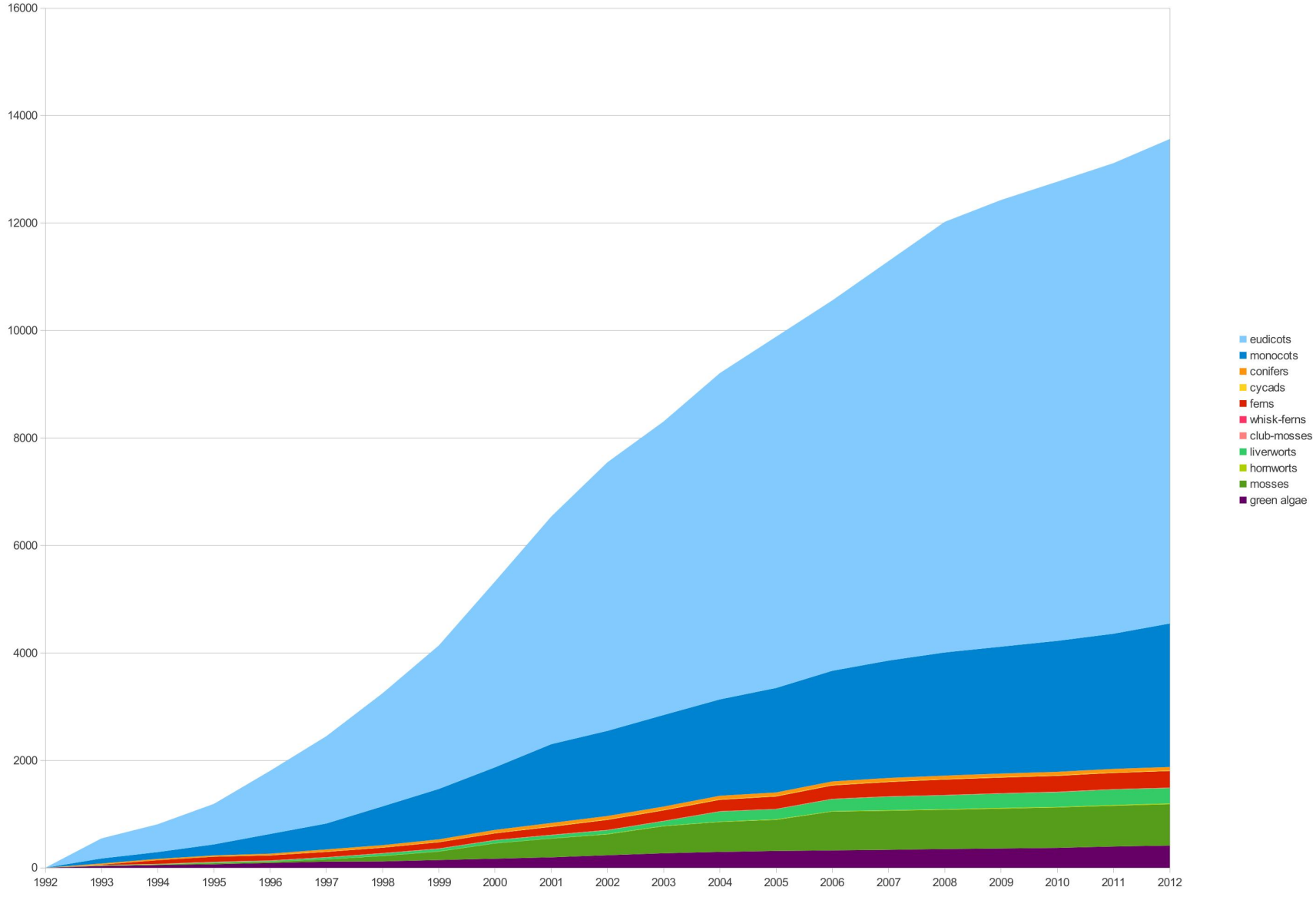




Commelinids

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


100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
eudicots
monocots
conifers
cycads
ferns
whisk-ferns
club-mosses
liverworts
hornworts
mosses
green algae

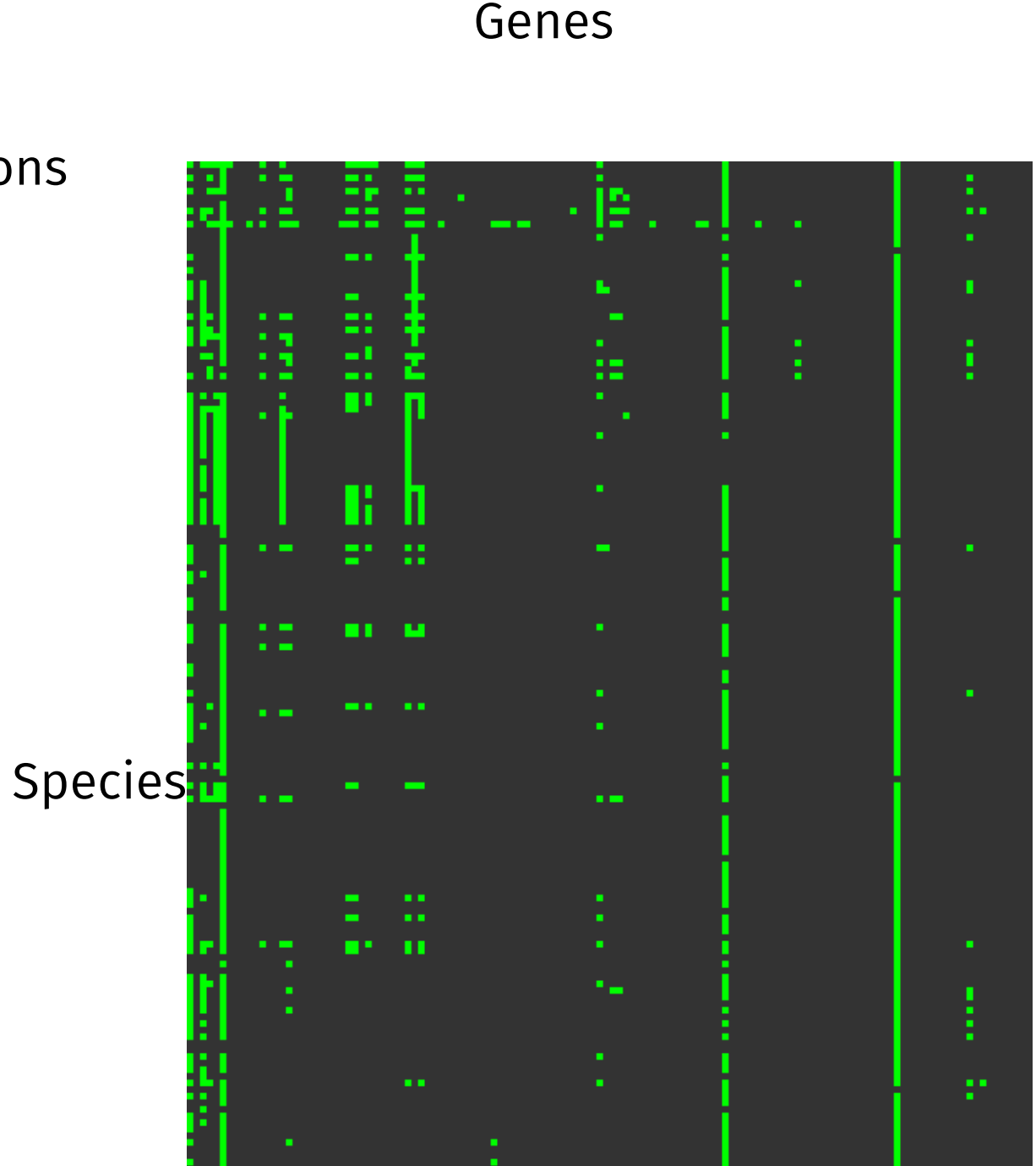

16000
14000
12000
10000
8000
6000
4000
2000
0
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
eudicots
monocots
conifers
cycads
ferns
whisk-ferns
club-mosses
liverworts
hornworts
mosses
green algae

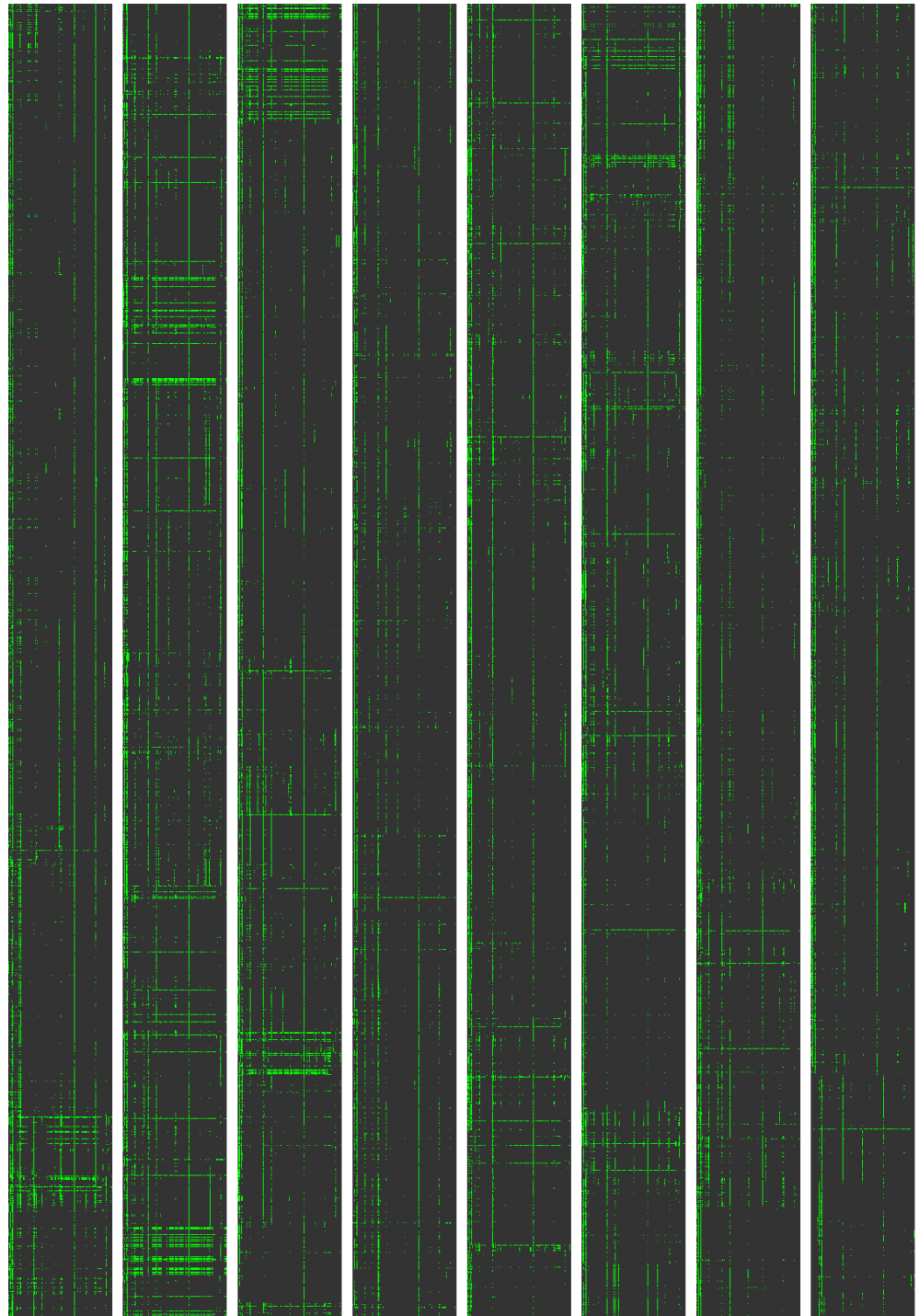
All genera of plants

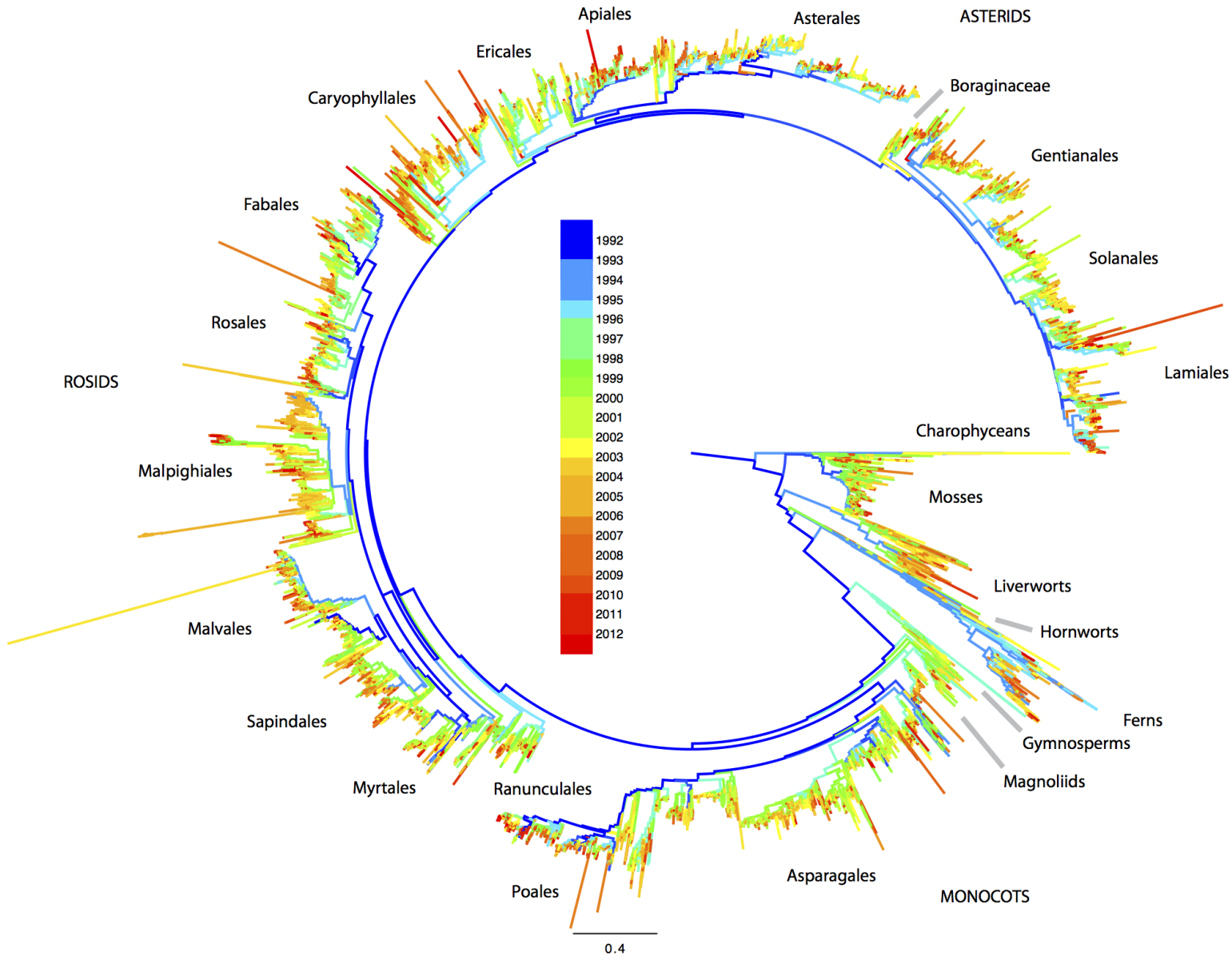
13,083 genera

128 gene regions

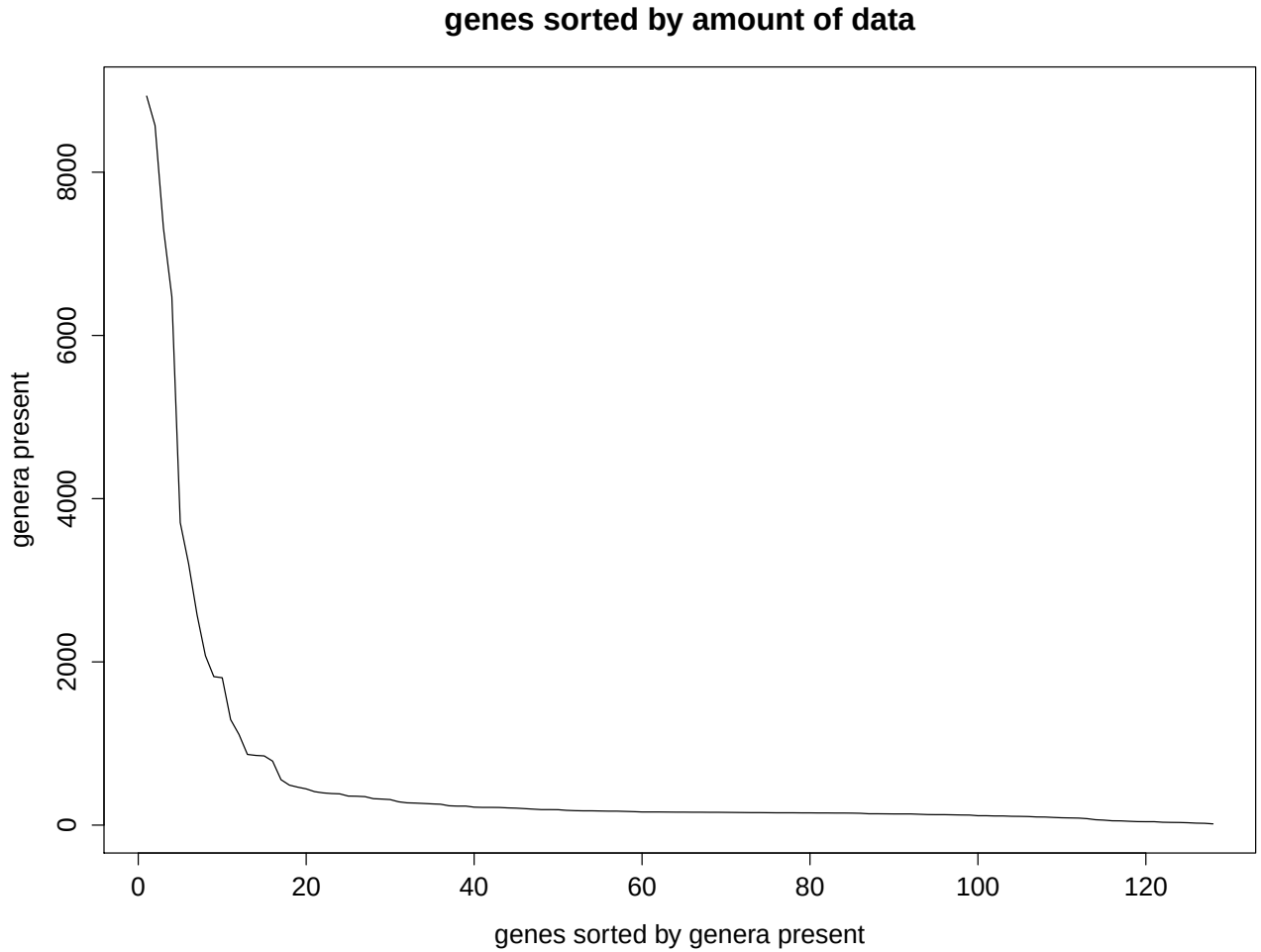
148,150 sites

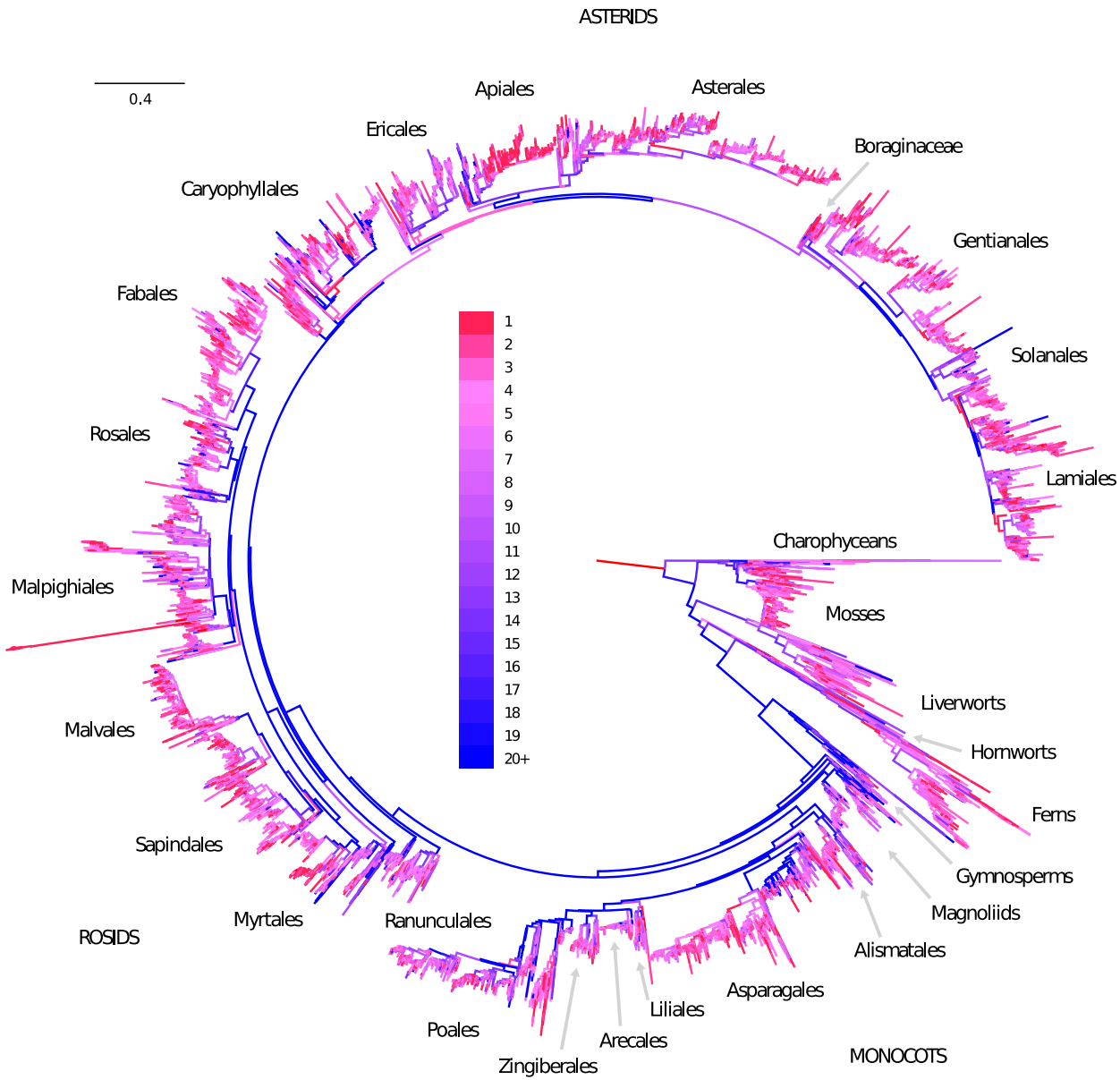





Apiales
Asterales
ASTERIDS
Ericales
Boraginaceae
Caryophyllales
Gentianales
Fabales
Solanales
Rosales
Lamiales
ROSIDS
Charophyceans
Malpighiales
Mosses
Liverworts
Malvales
Hornworts
Ferns
Sapindales
Gymnosperms
Magnoliids
Myrtales
Ranunculales
Asparagales
Poales
MONOCOTS
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
0.4

95% missing data




ASTERIDS
0.4
Apiales
Asterales
Ericales
Boraginaceae
Caryophyllales
Gentianales
Fabales
Solanales
Rosales
Lamiales
Charophyceans
Malpighiales
Mosses
Liverworts
Malvales
Hornworts
Ferns
Sapindales
Gymnosperms
Magnoliids
Myrtales
Ranunculales
ROSIDS
Alismatales
Asparagales
Liliales
Poales
Arecales
Zingiberales
MONOCOTS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20+

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?)

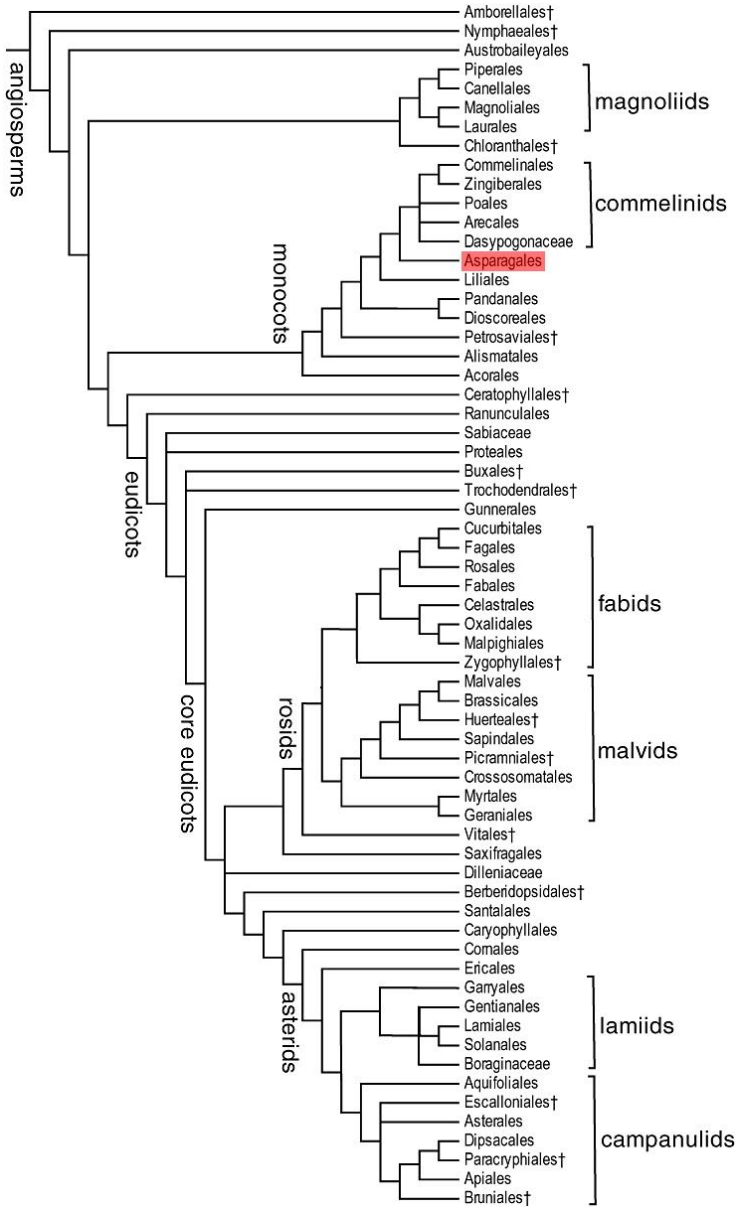
Baited BLAST

- 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

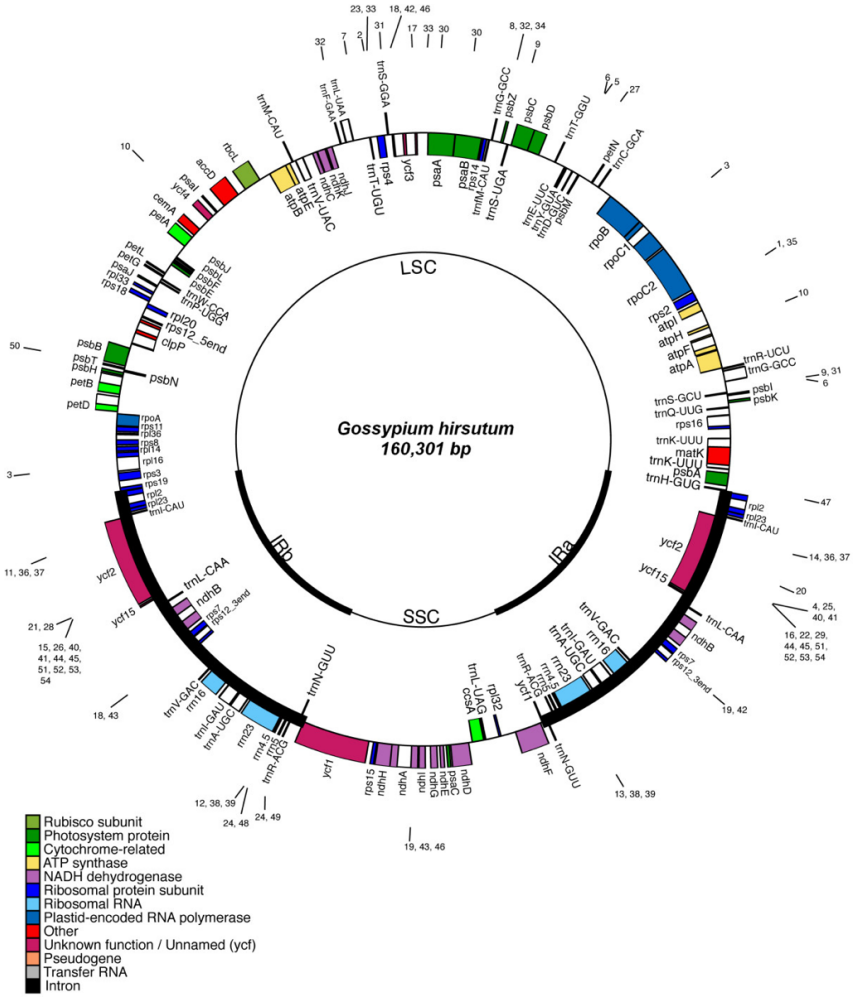
1. pick clade
2. pick gene region
3. select known sequences
4. pairwise comparison of possible sequences to known
5. test for saturation
6. reassemble if separated for saturation
7. repeat steps 2-6 for other gene regions
8. concatenate all gene regions
9. estimate phylogeny

Pick region of the tree of life



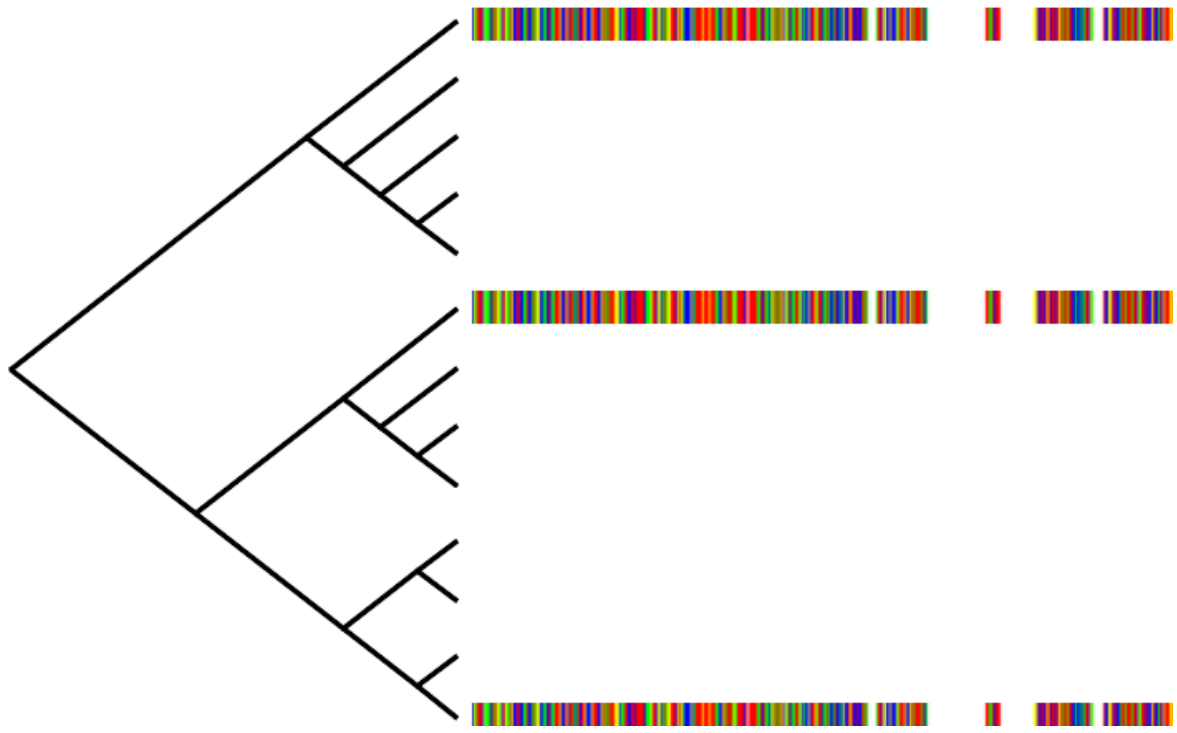
Asparagales

Gene region

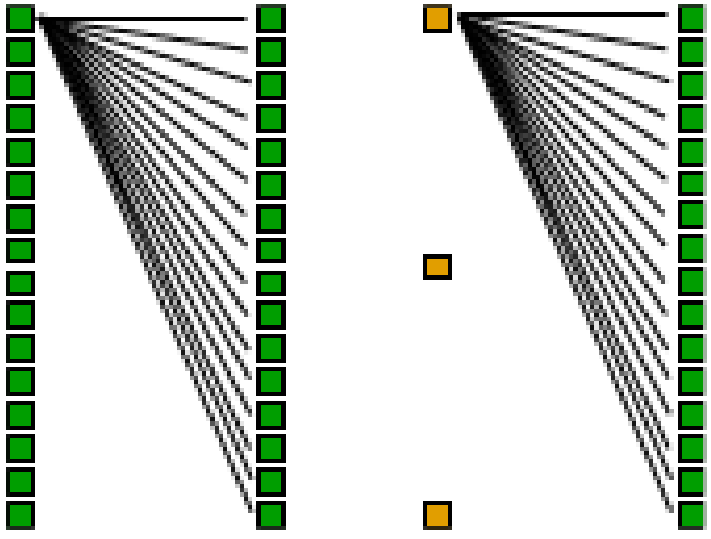


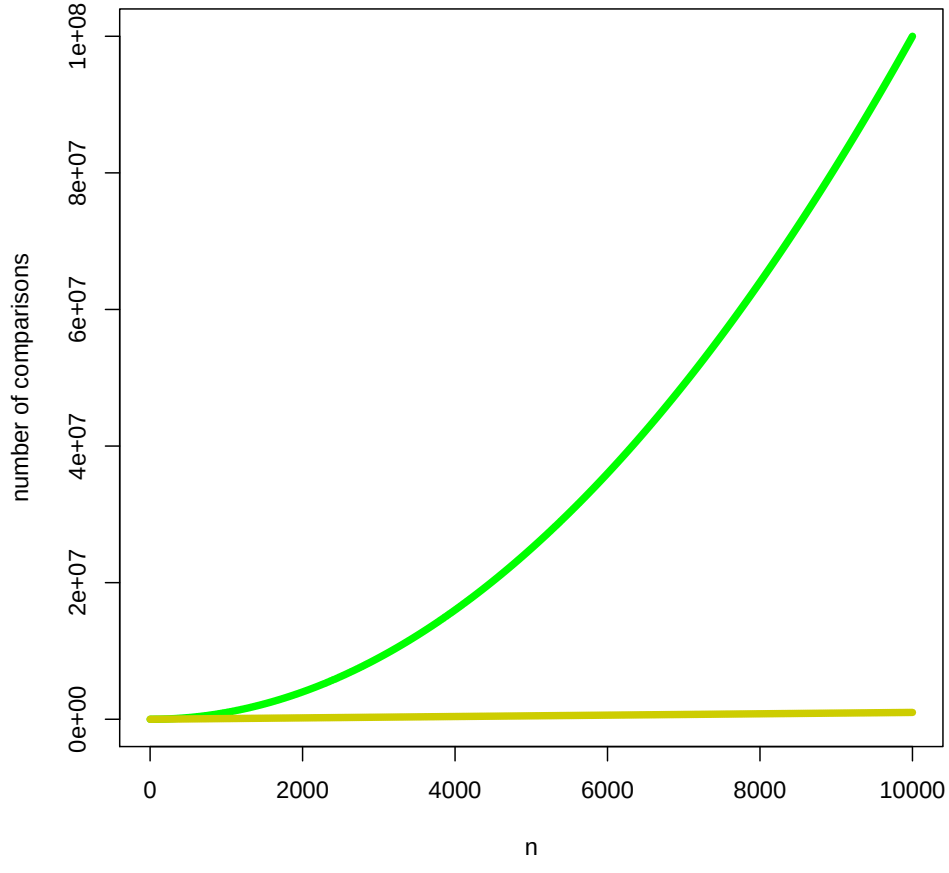
trnL-trnF

Identify set of known representatives for gene region



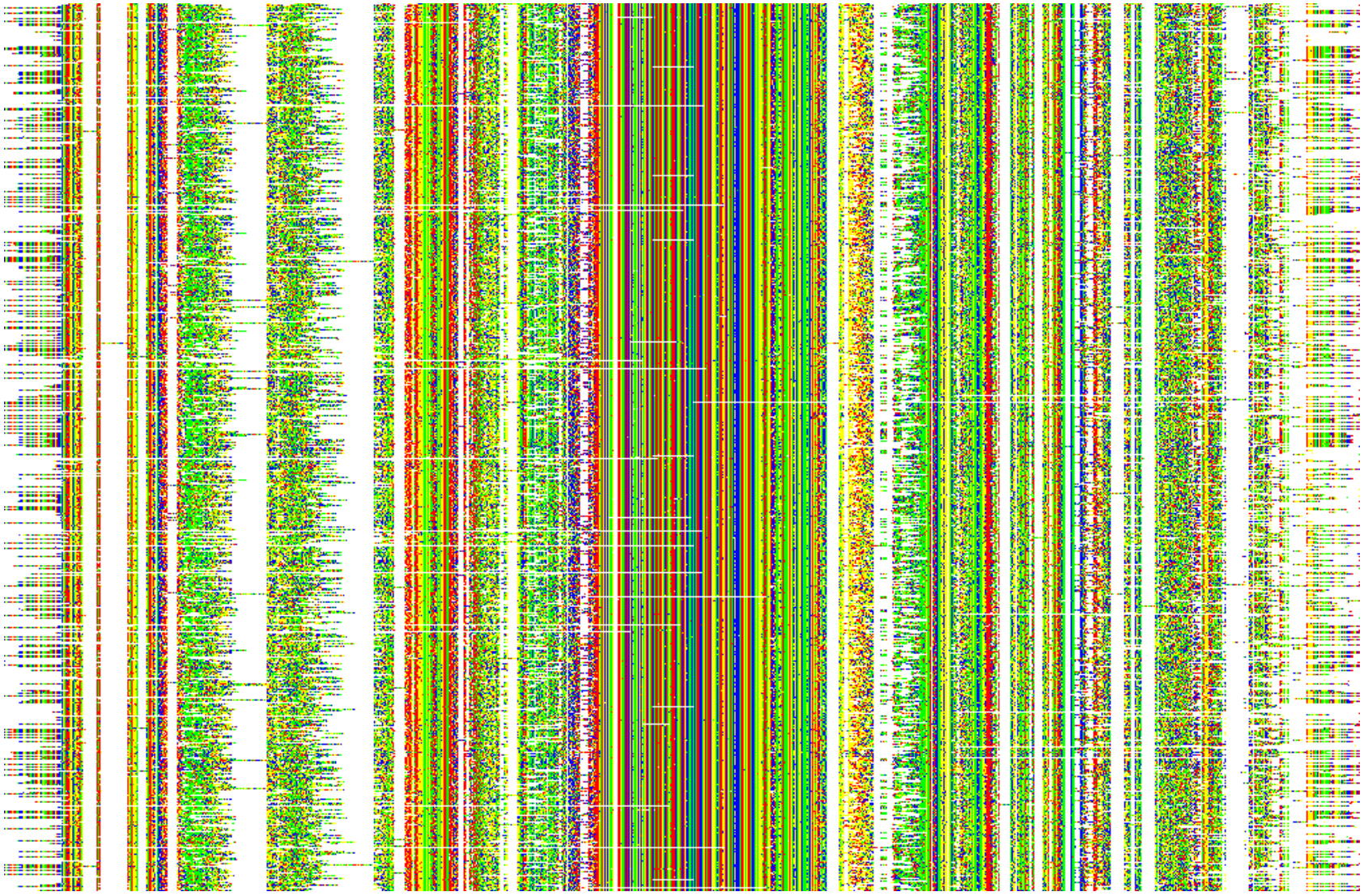
Pairwise comparison



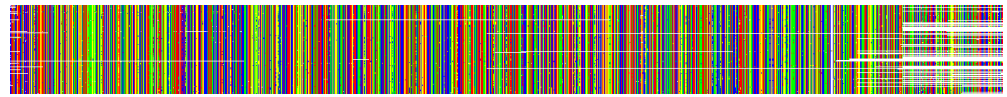


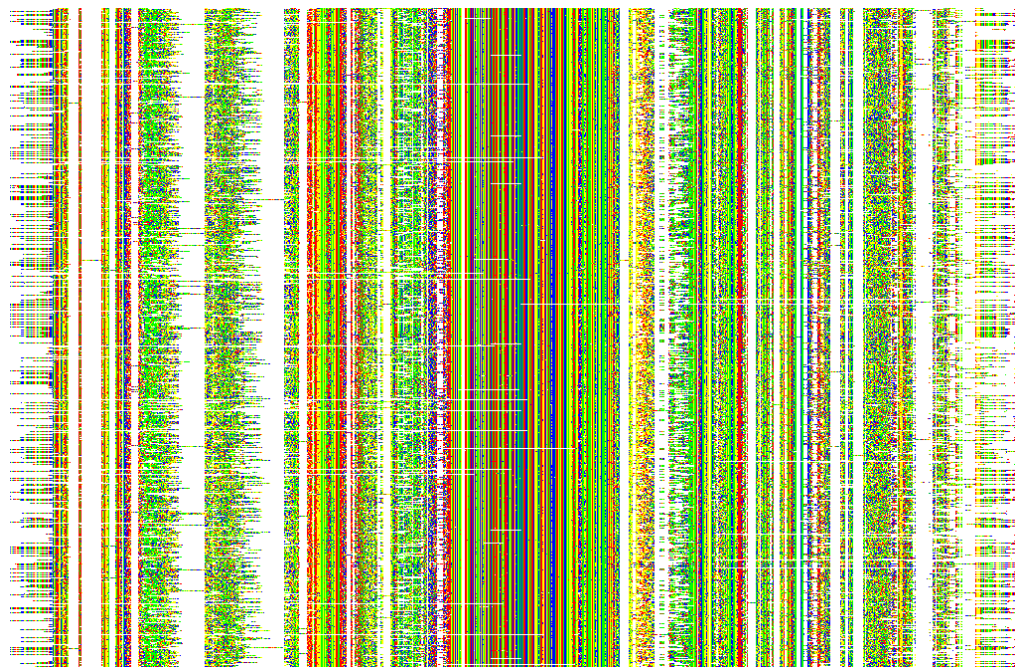
moving from $O(n^2)$ to $O(n)$

Aligning distantly related species



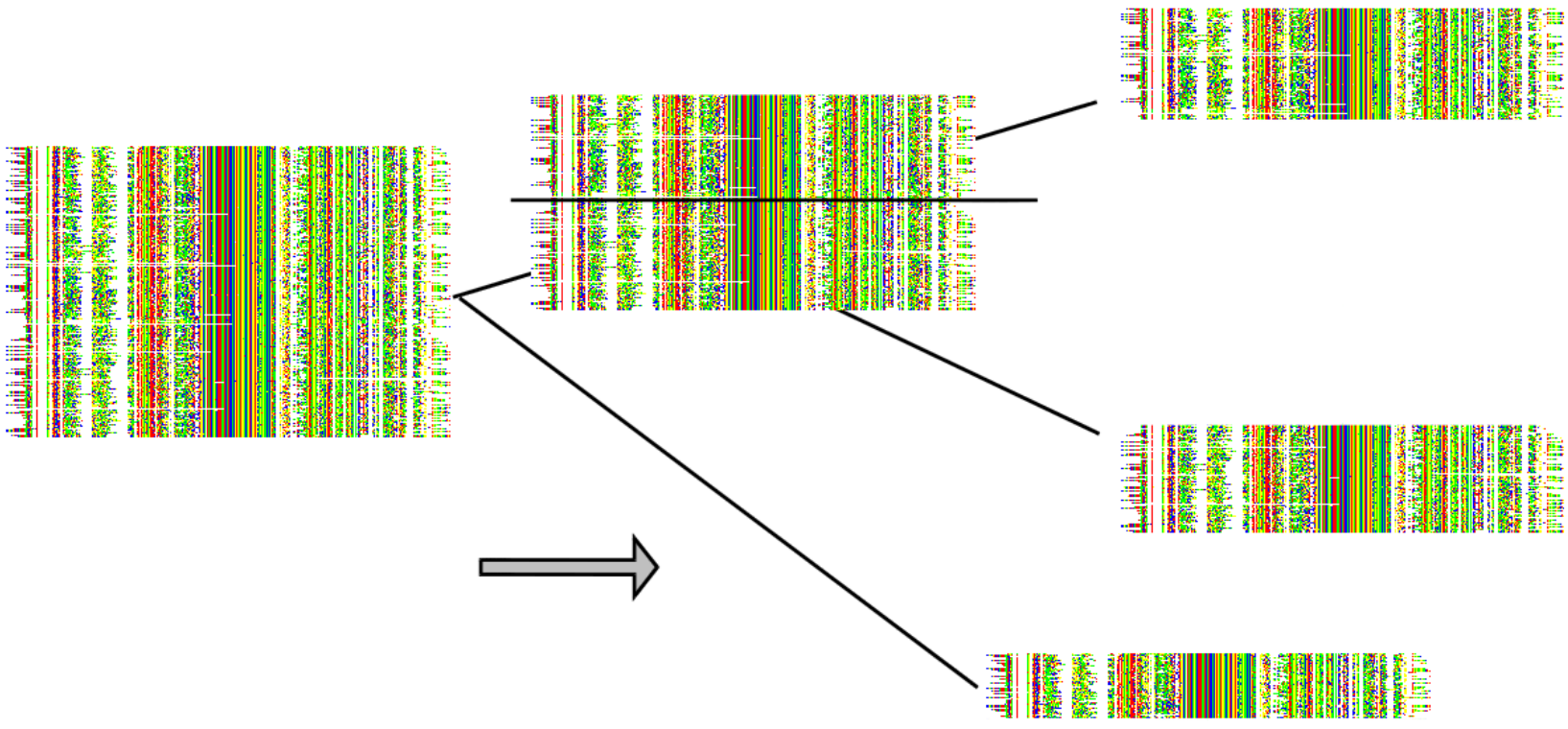
Saturation test



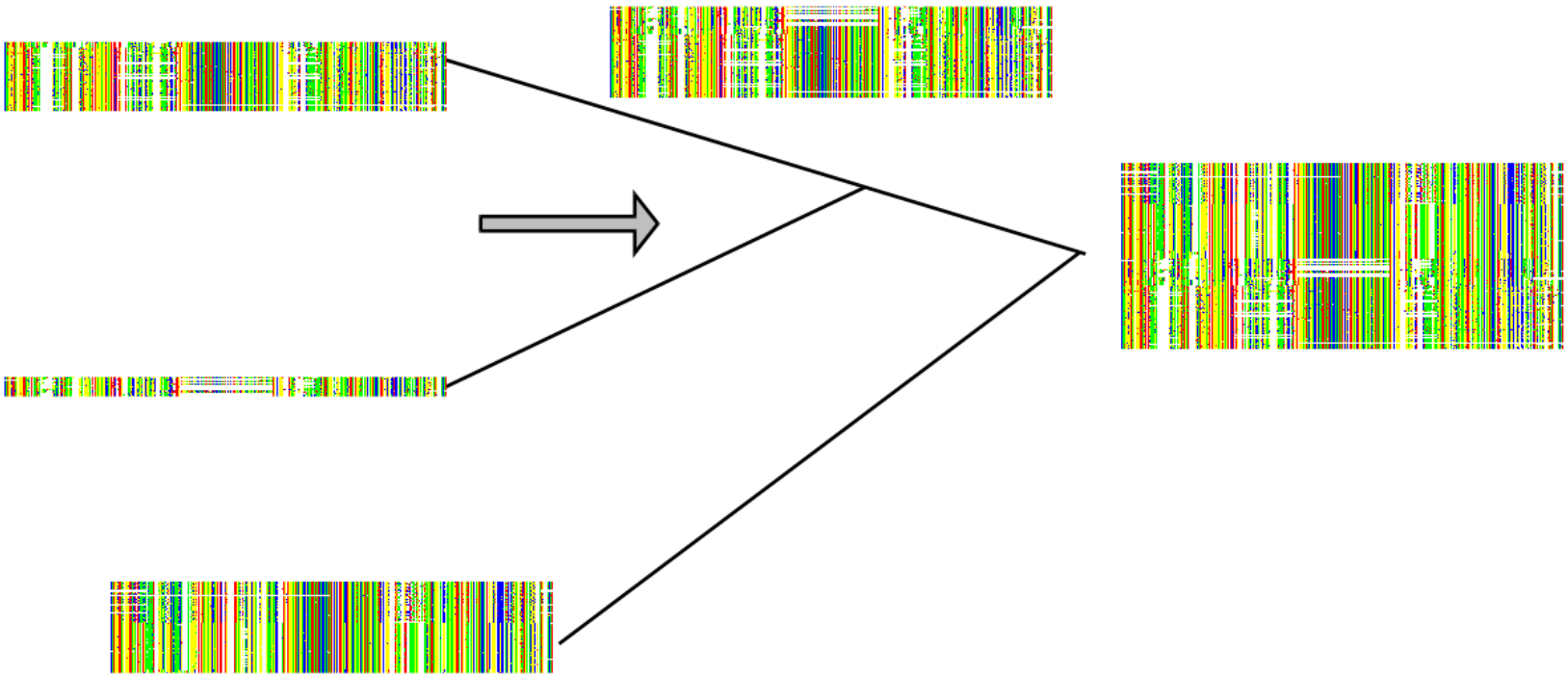


MAD = median absolute deviation ($median_i \left(| X_i - median_j(X_j) | \right)$)

Saturation test



Reassembly



Results

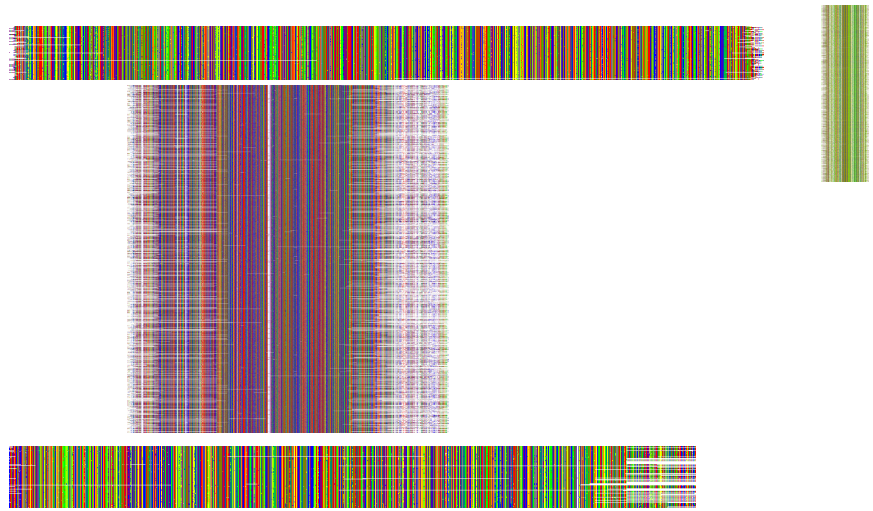
number of taxa = 2485

number of sites = 1406

< 8 minutes

26 clades broken down for individual alignments and profile alignment reassembly

Concatenate



trnL-trnF = 2485

atpB = 107

matK = 2182

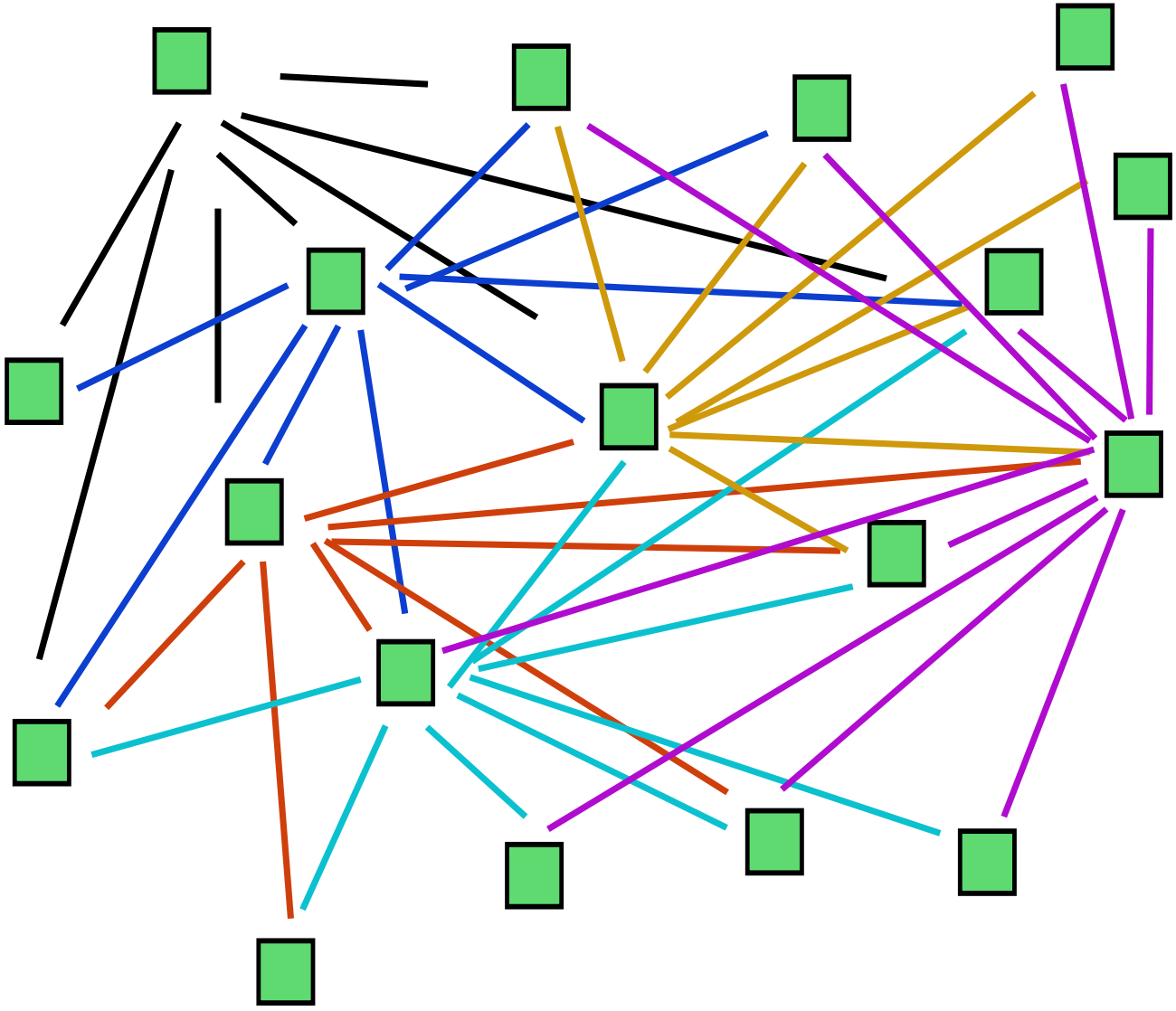
rbcL = 924

ITS = 3725

Clustering

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

All by all comparisons



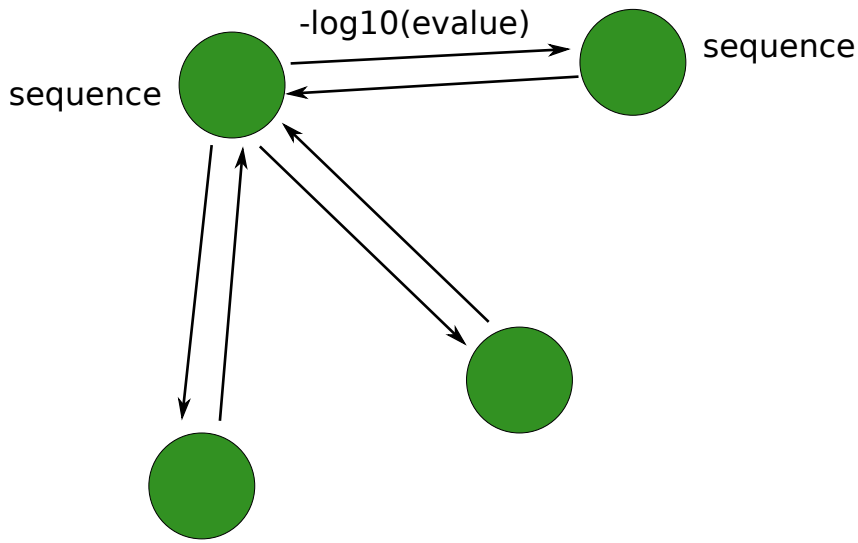
Graph

Each sequence is connected to the others to which there was a significant hit

Each vertex is a sequence

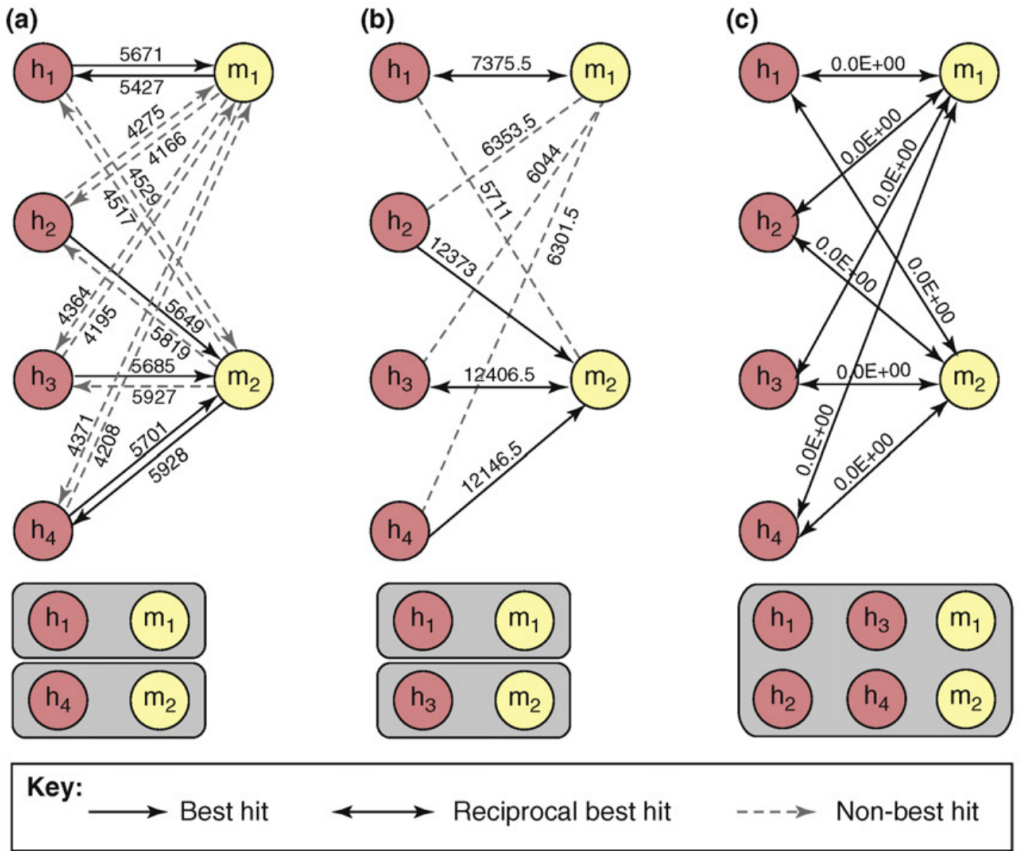
Each edge is a significant hit

The weight of the edge (as well as the edge itself) is based on *E*-value (-log10 of the *E*-value so that lower *E*-value is a higher / stronger connection)



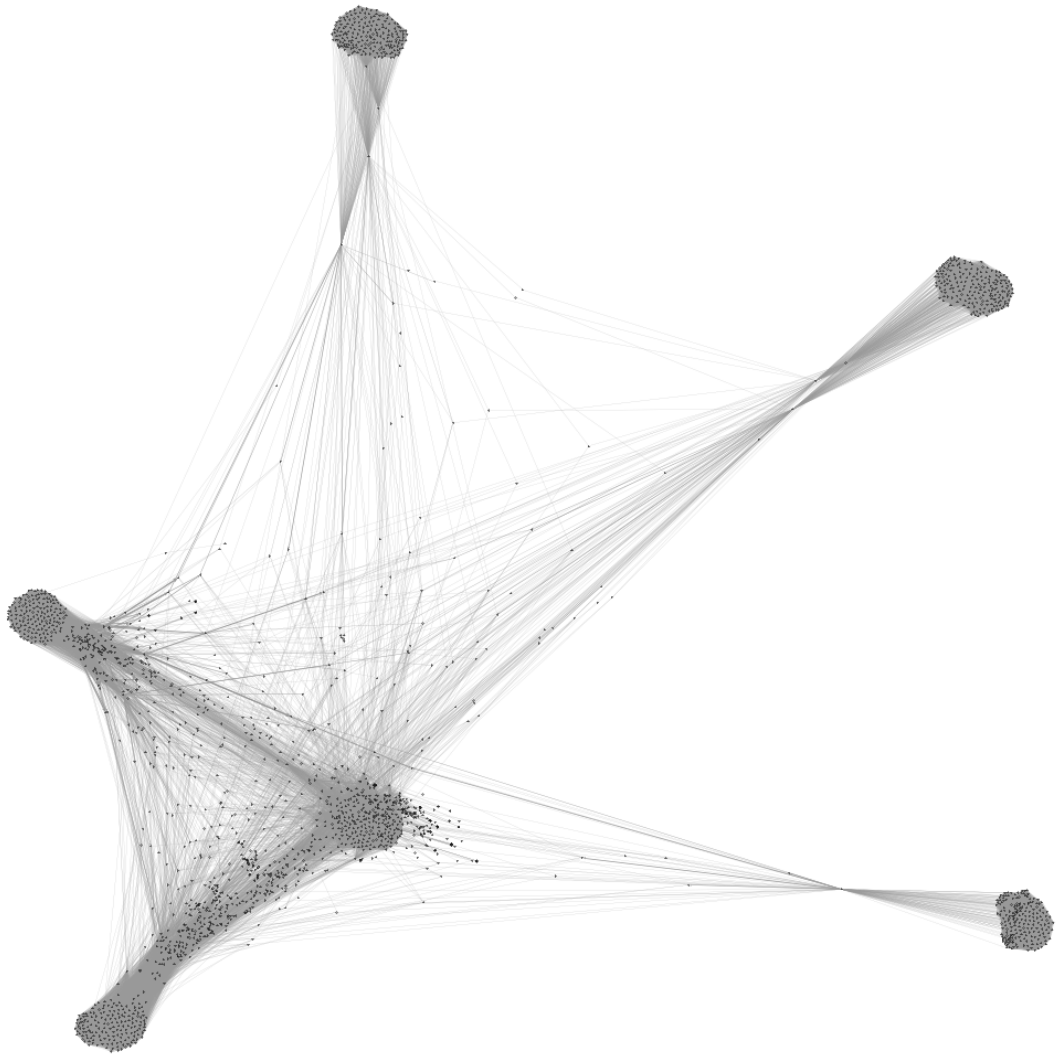
Different measurements

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

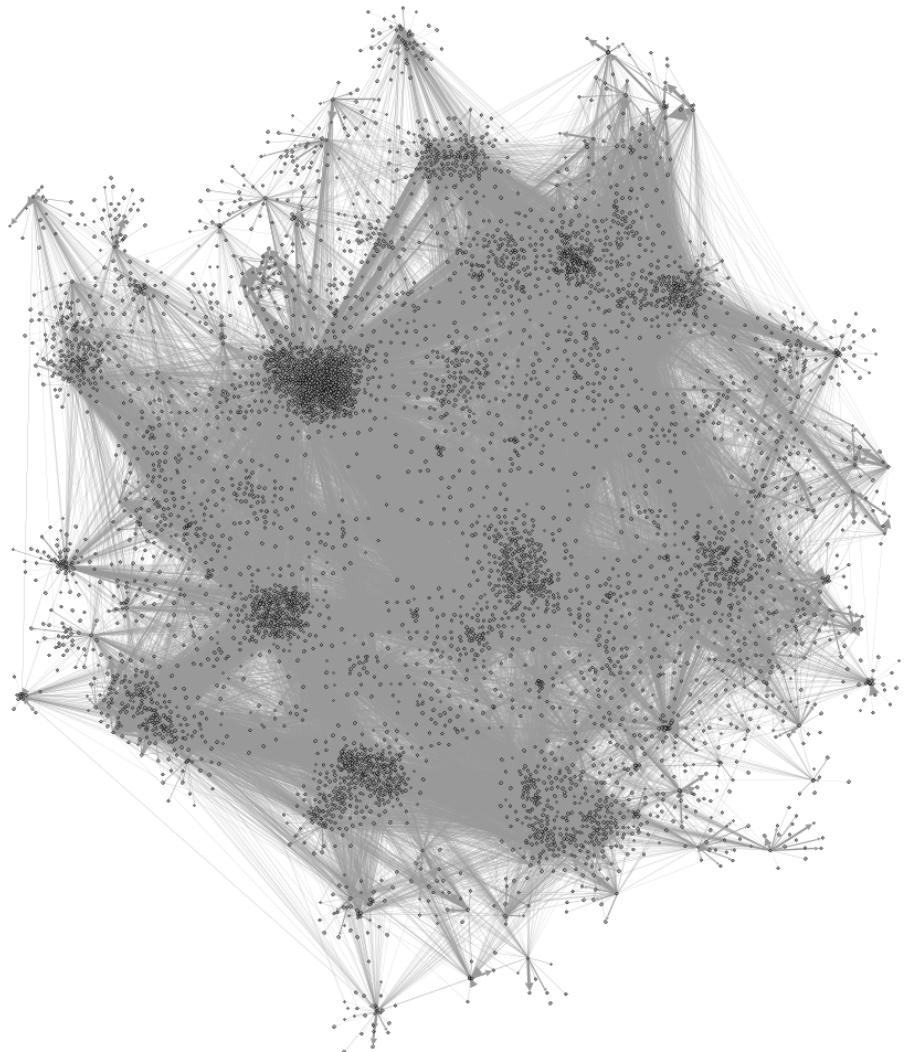


Kuzniar et al., 2008

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

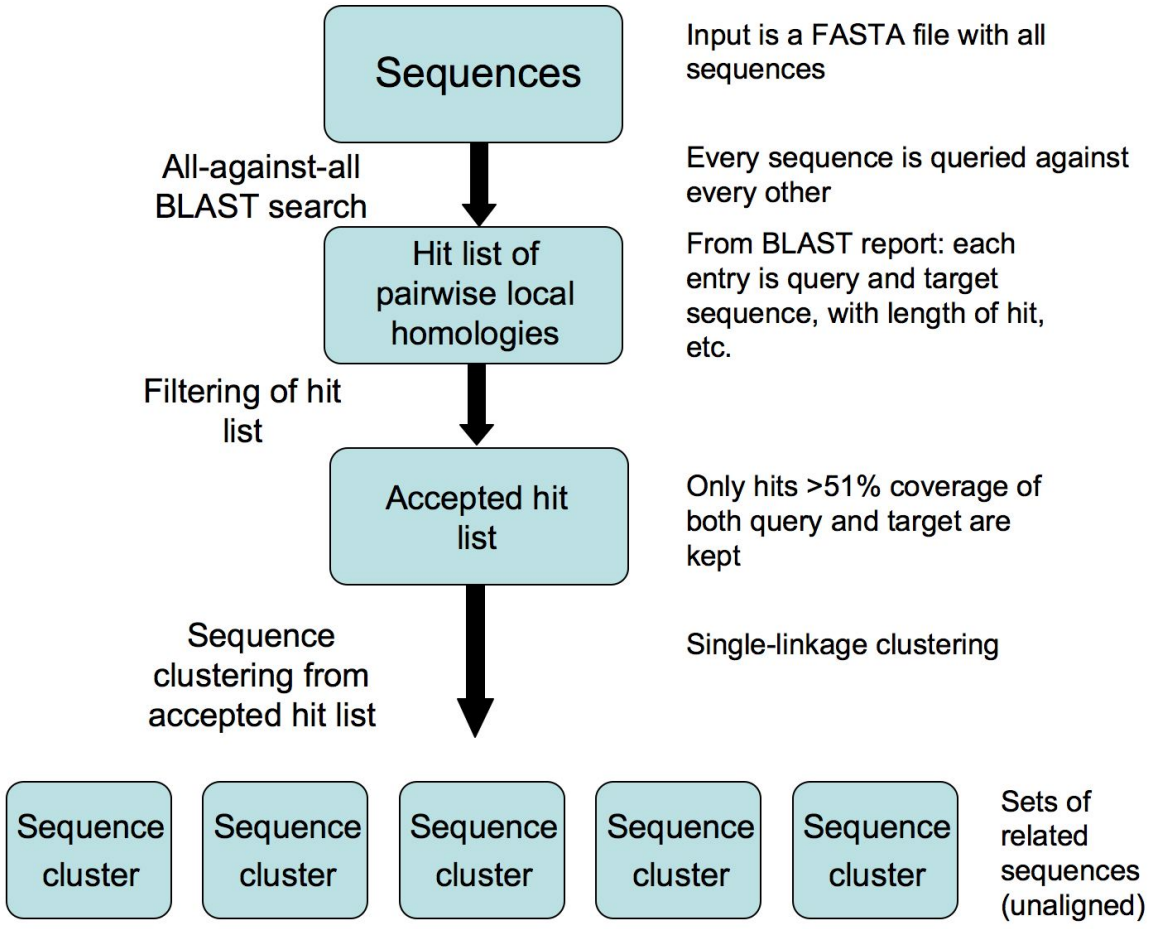
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802158688

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 WEHE[2]

[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


Sequences
Input is a FASTA file with all sequences
All-against-all BLAST search
Every sequence is queried against every other
Hit list of pairwise local homologies
From BLAST report: each entry is query and target sequence, with length of hit, etc.
Filtering of hit list
Accepted hit list
Only hits >51% coverage of both query and target are kept
Sequence clustering from accepted hit list
Single-linkage clustering
Sequence cluster
Sequence cluster
Sequence cluster
Sequence cluster
Sequence cluster
Sets of related sequences (unaligned)

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	-	-
Caesalpinioideae		789	813	165	4802	278	39
Mimosoideae		1017	1048	70	6583	553	38
Papilionoideae		4399	4598	409	59719	2238	364

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

[2]Taxa 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.

[3]Clusters 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

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

Phylota

Cluster set at subtree whose root is Papilionoideae (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. epipsilum voucher Egan & Egan 146 (BRY) granule bound starch synthase (GBSSI) 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 trnD-trnY intergenic spacer, partial sequence; tRNA-Tyr (trnY trnY-trnE intergenic spacer, and tRNA-Glu (trnE) gene, complete sequence; and trnE-trnT interge spacer, partial sequence; chloroplast.
4	-	975	1289	249	343	3021	0.573	Medicago sativa tRNA-Lys (trnK) 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; chloropl
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.angustifolius 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

Papilionoideae (rel. 184): Cluster 0

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

[1]The 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 po homologies (for example, if one sequence is a complete mitochondrial genome, and others are single mt genes).

Download cluster...

- Unaligned cluster in Fasta format
  Use defline as taxon label Submit

Download alignments...

- Muscle alignment in Fasta format (Color view)
  Use gi# as taxon label Submit

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

- Use gi# as taxon label Submit

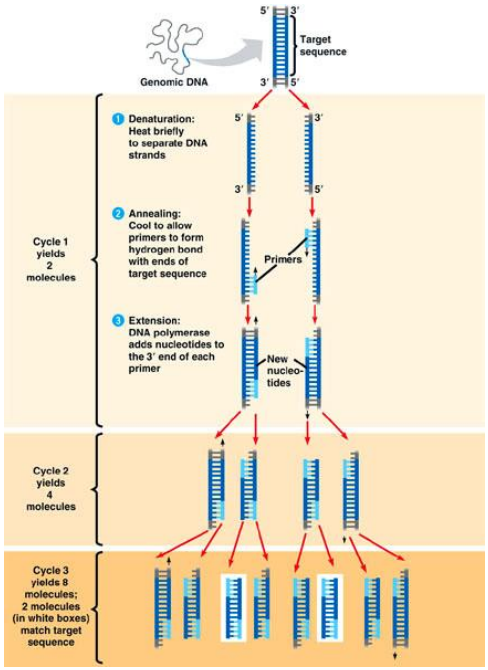
Phylota

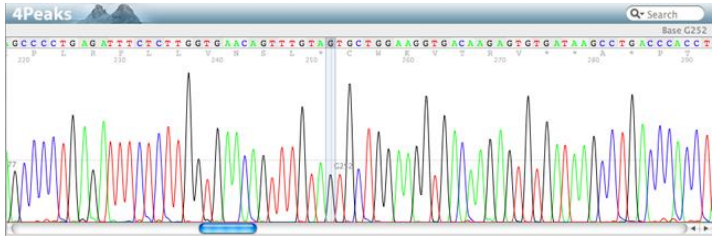
Download a midpoint rooted RAxML optimal tree with branch lengths (Muscle alignment)...

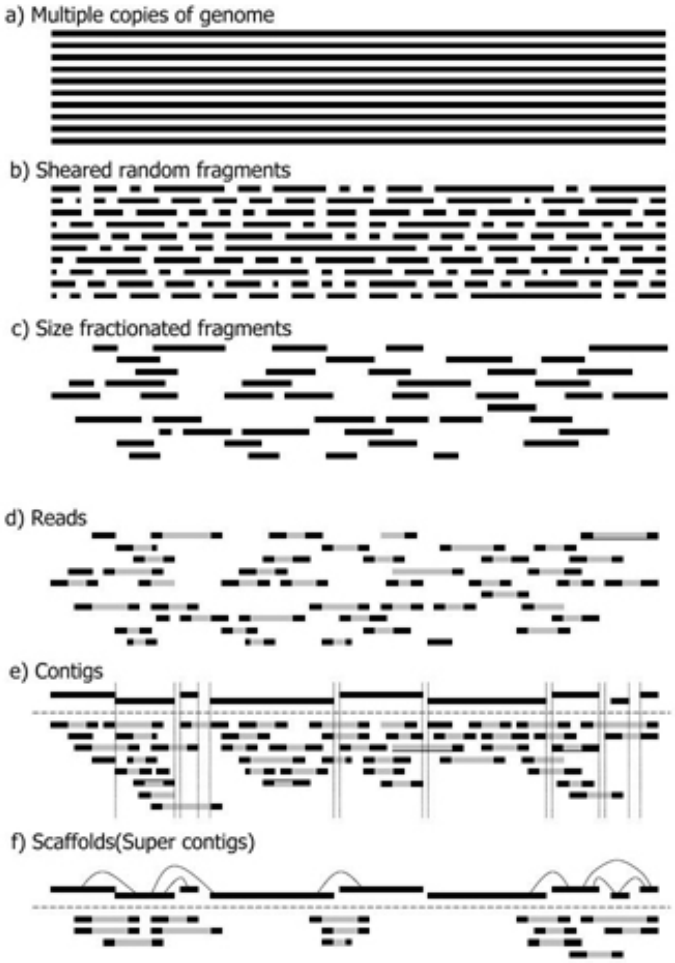
- Use gi# as taxon label Submit

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

Massive parallel sequencing







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

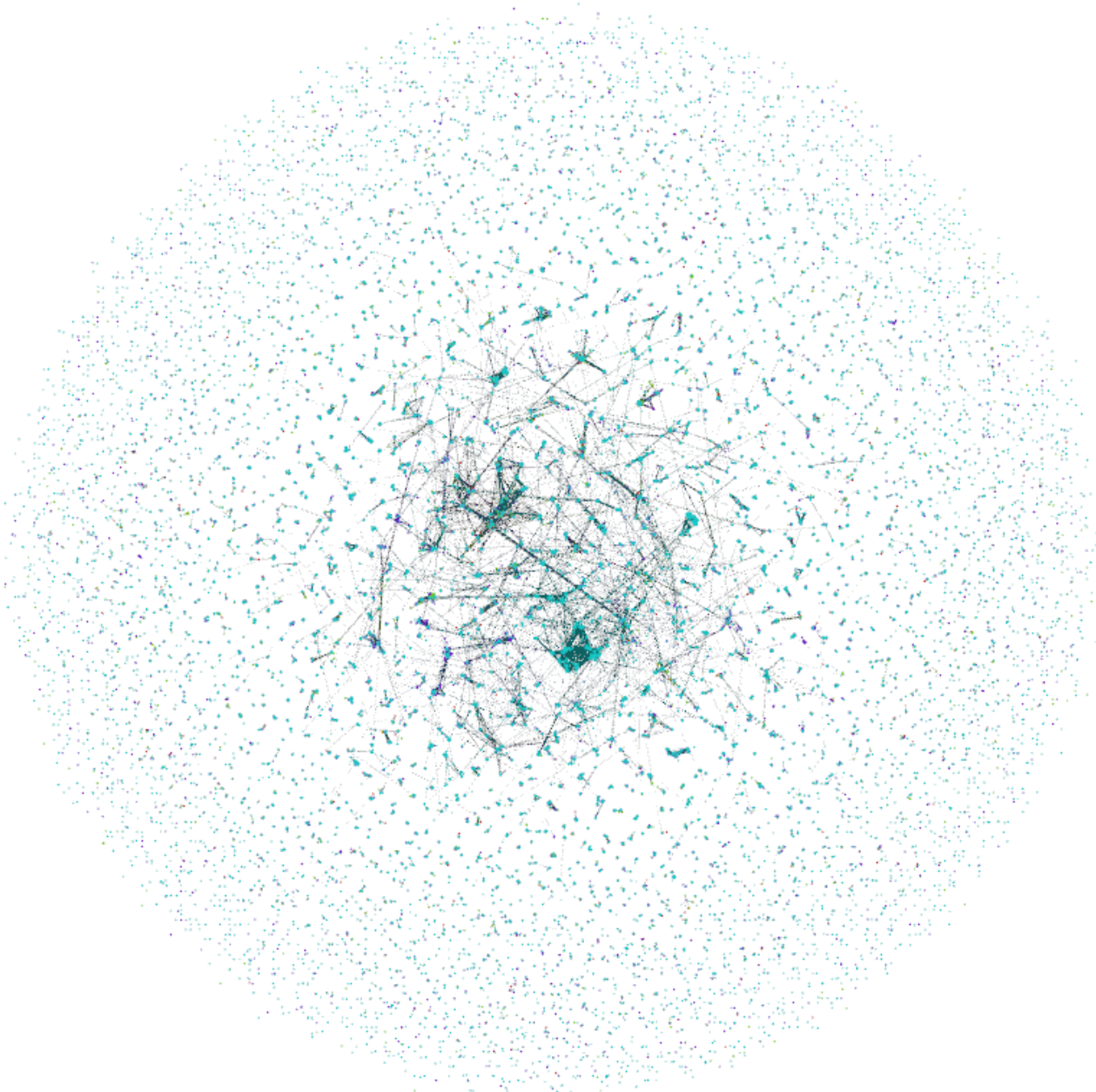
140 genes. Metazoa (Dunn et al. 2008)

1185 genes. Molluscs (Smith et al. 2011)

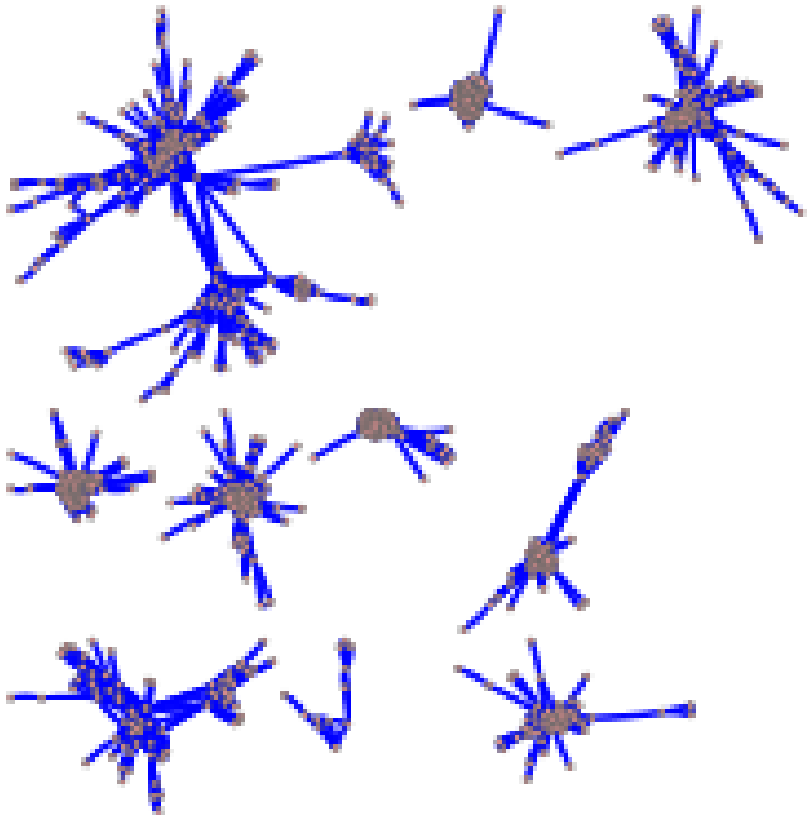
2970 genes. Seed plants (Lee et al. 2011)

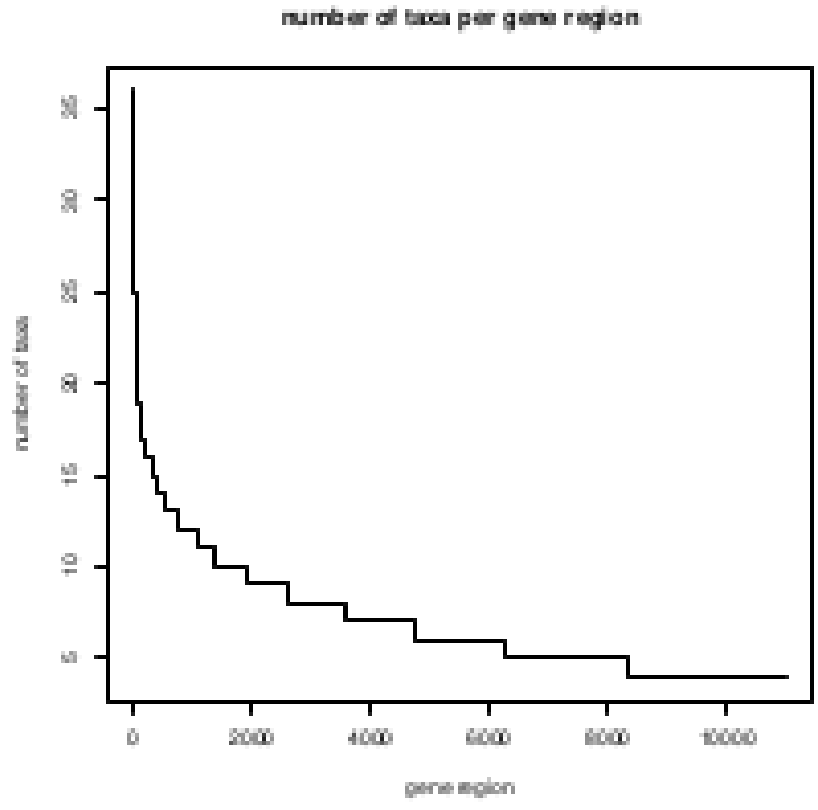
8251 genes. Birds (Jarvis et al. 2014)

20,374 genes. Equids. (Jónsson et al. 2014)

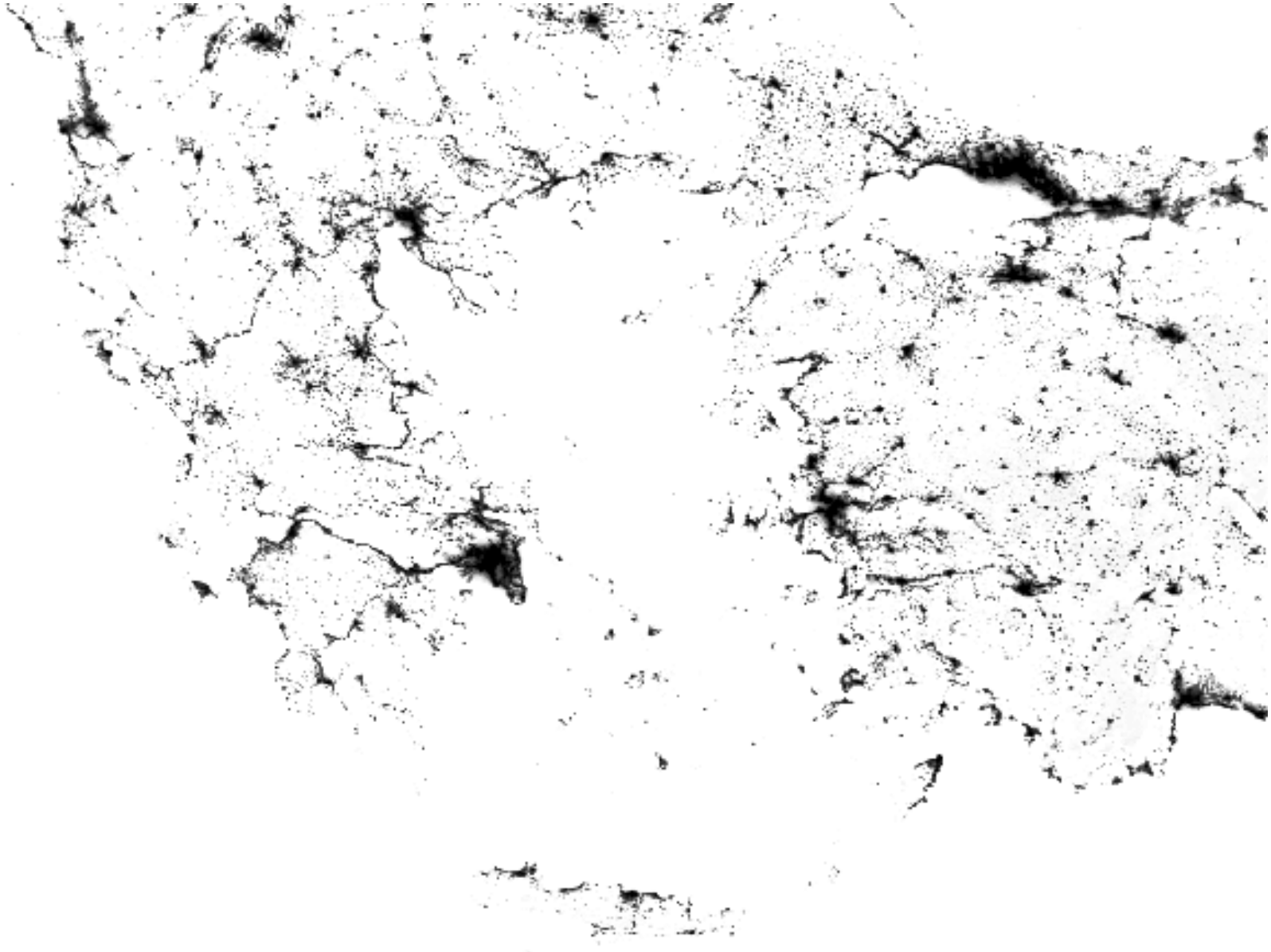


Clustering




number of taxa per gene region
number of taxa
5
10
15
20
25
30
35
0
2000
4000
6000
8000
10000
gene region

What is this?



What cities are connected by land?

Methods

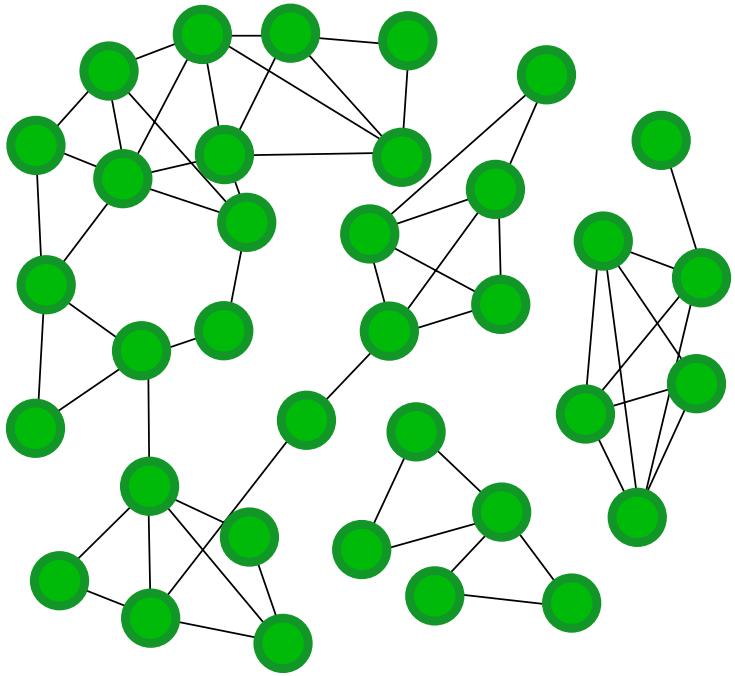
- Graph based methods
  - COG: Clusters of Orthologous Groups
    Paranoid methods (inParanoid and MultiParanoid):
    MCL methods (Markov CLustering)
- Tree based methods
  - COCO-CL: given homologs, calculates paralogs and orthologs

Simple concept

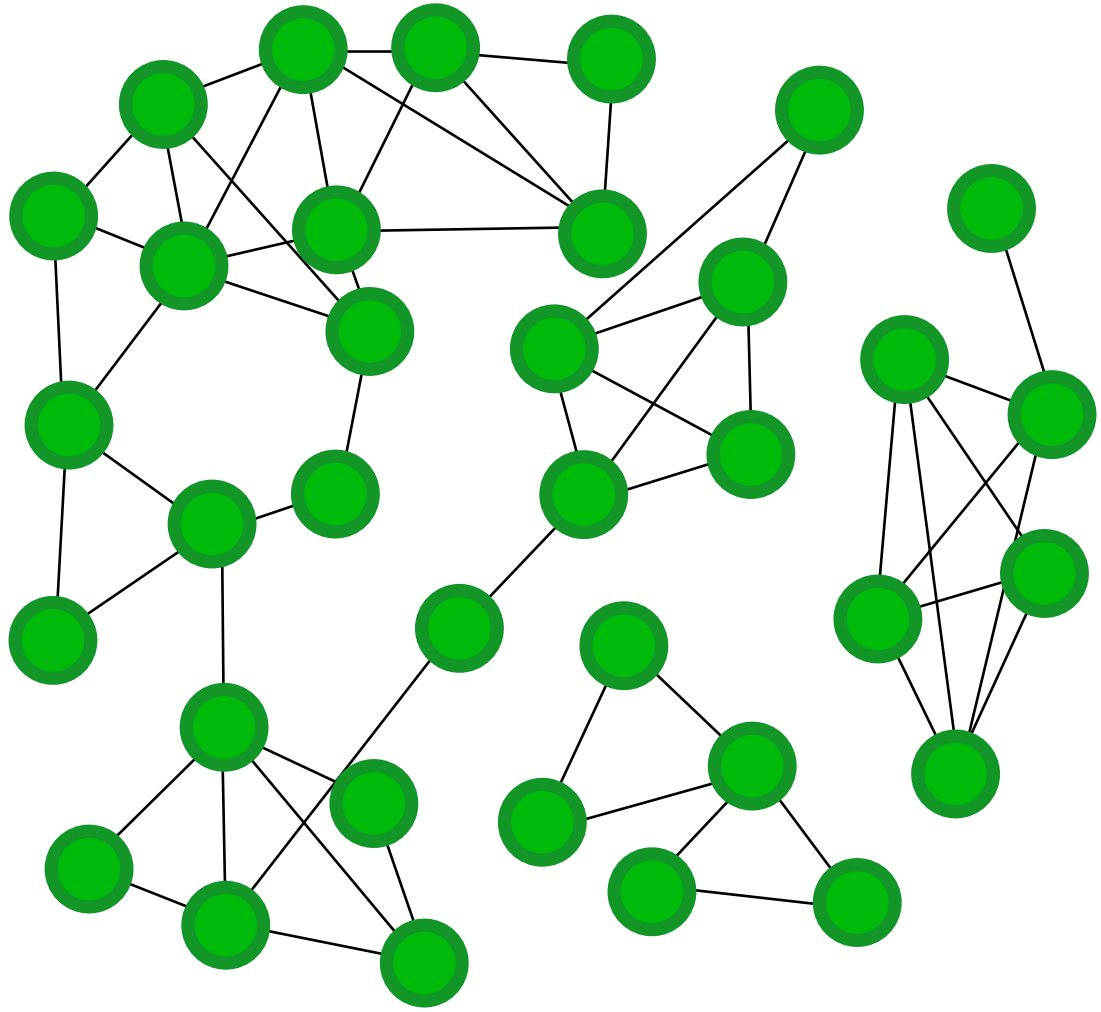
Start at a node, then randomly travel to connected nodes.

We will (on average) be more likely to stay within a cluster than to travel between clusters

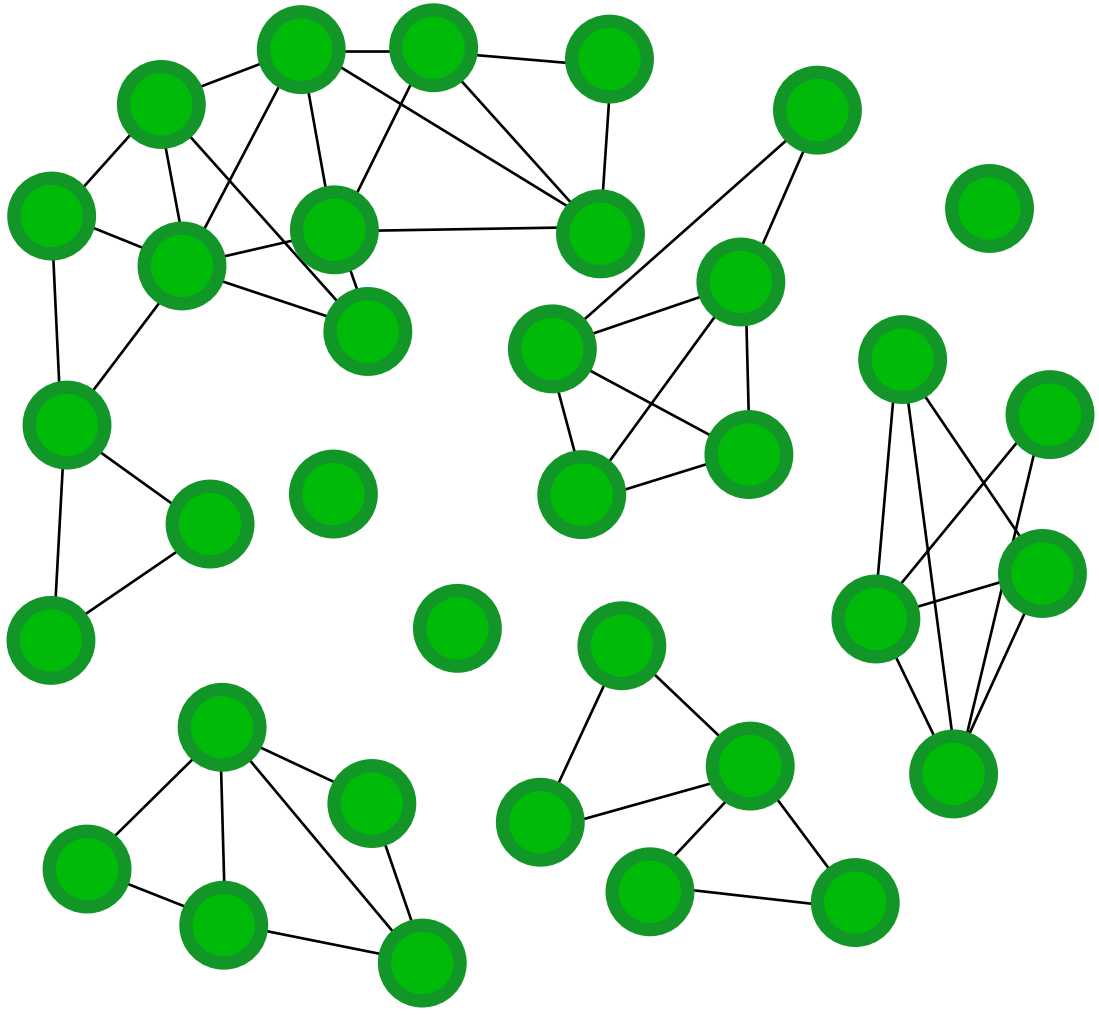
As we iterate, edges that are traveled frequently might increase in weight, while edges traveled rarely might be removed



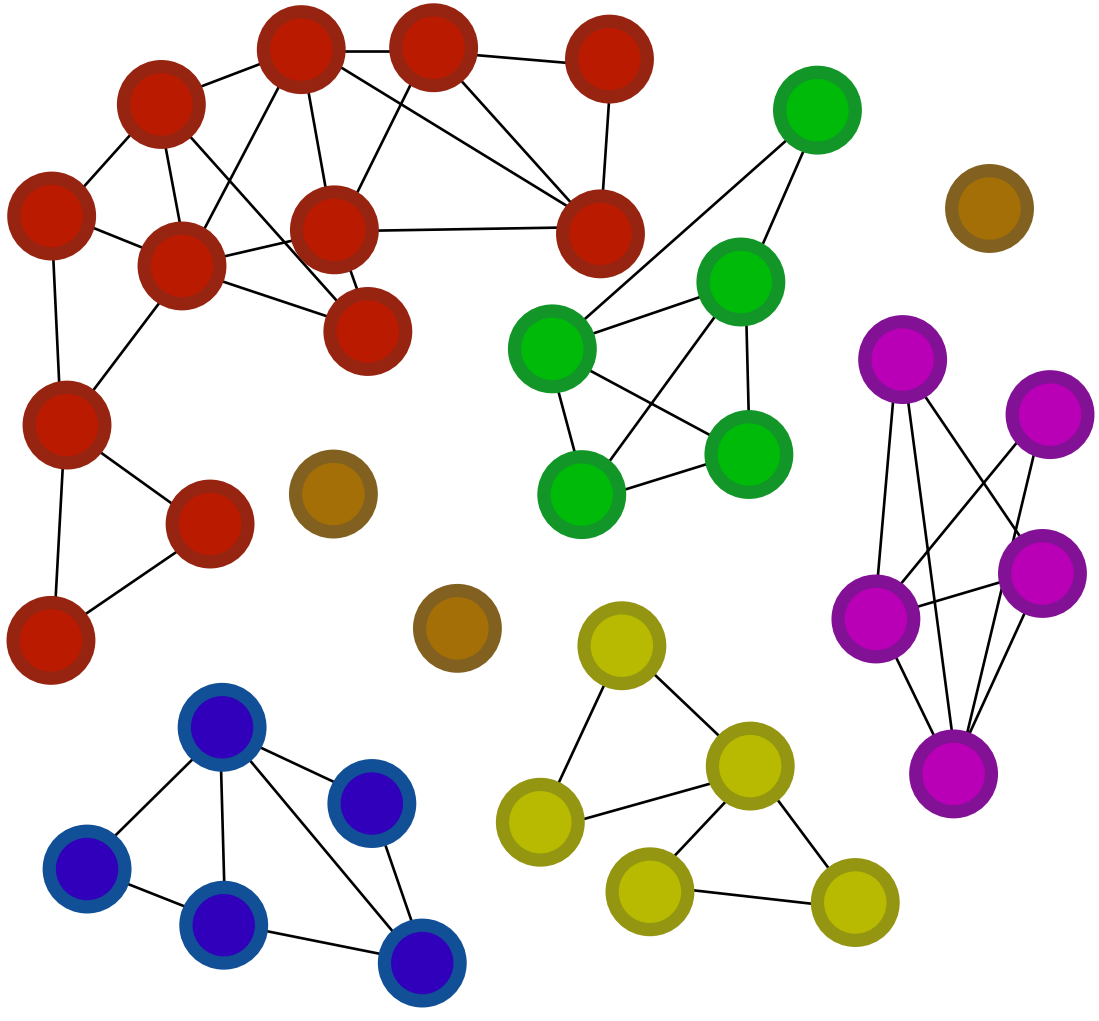
Simple visual example of clustering



Simple visual example of clustering



Simple visual example of clustering



Markov clustering algorithm (MCL)

Take a graph with vertices and edges

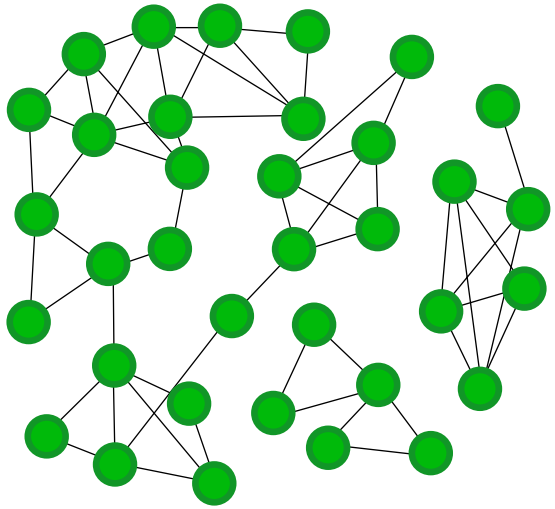
Assume that there are clusters

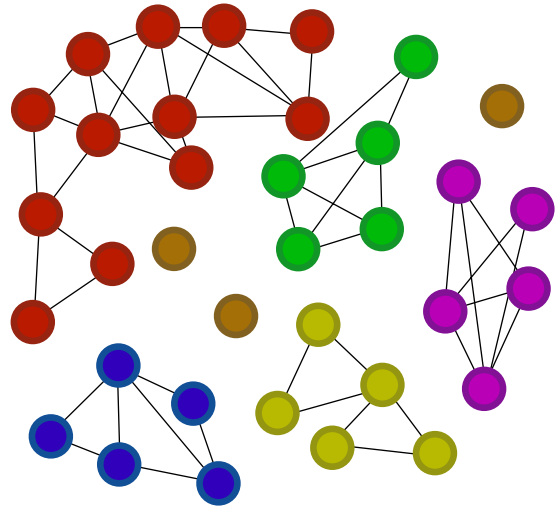
Start at a node and randomly travel between connected nodes

We are more likely to stay within a cluster than to move between clusters

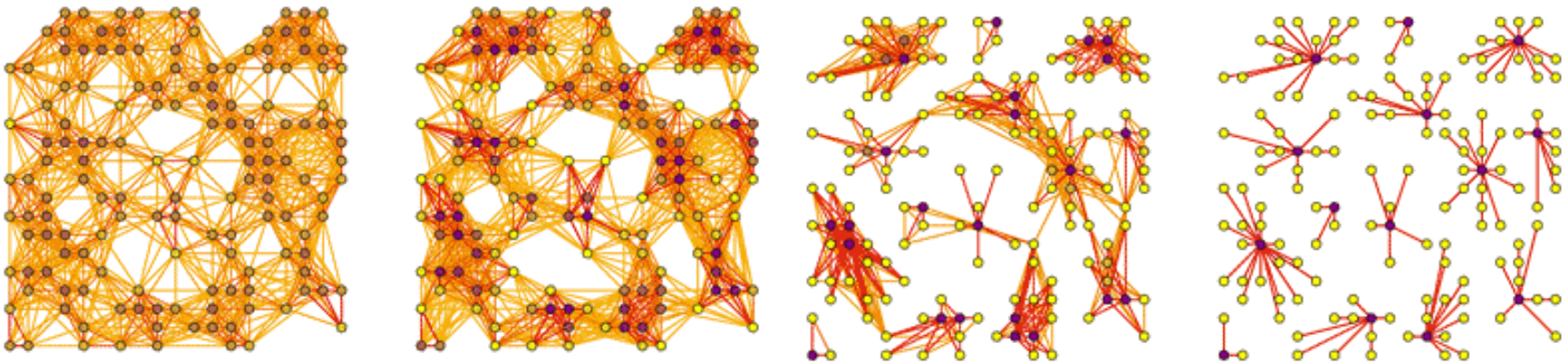
These random walks (Markov chains) will help us discover clusters

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)

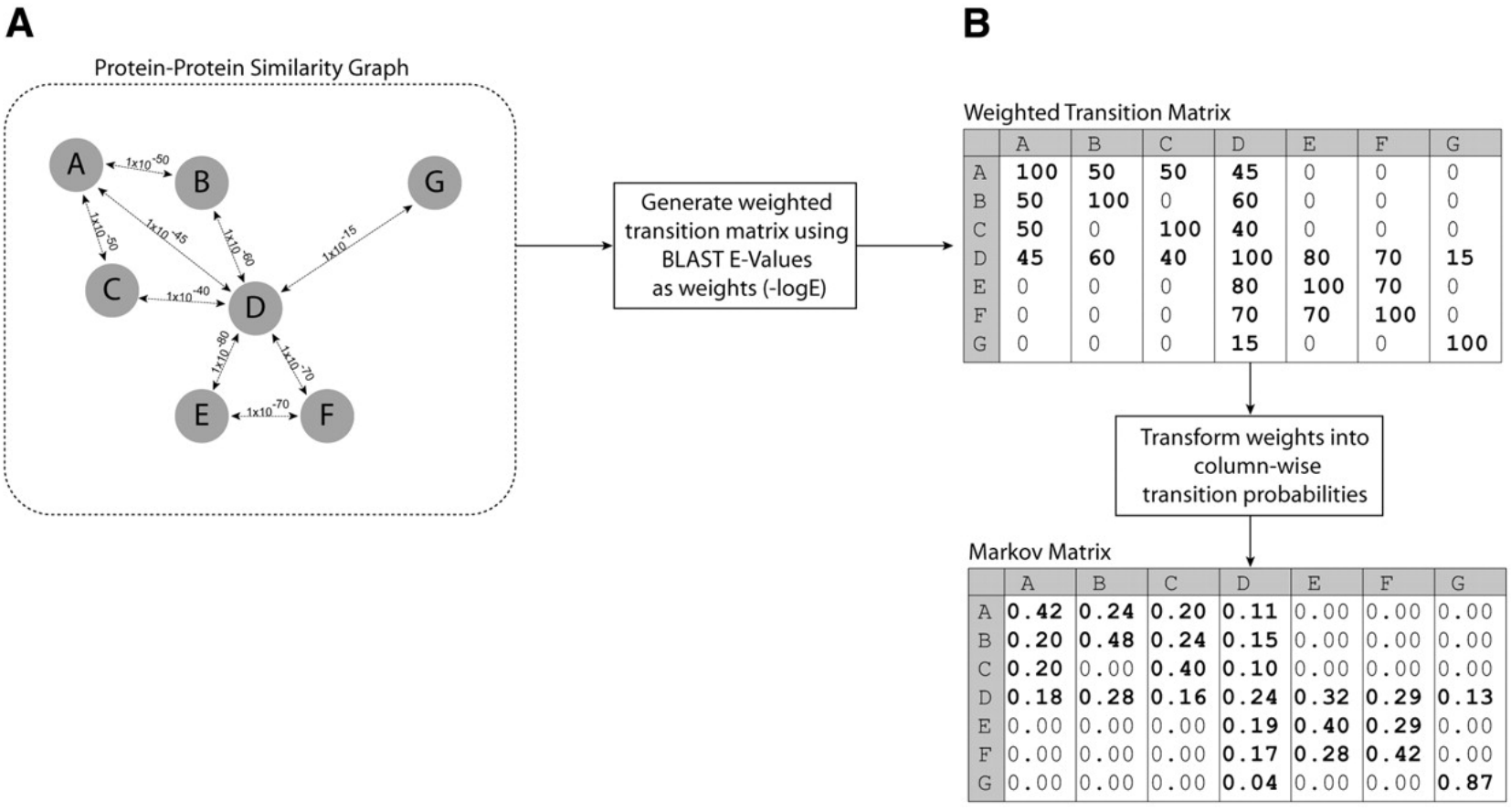




Markov chain



Usage

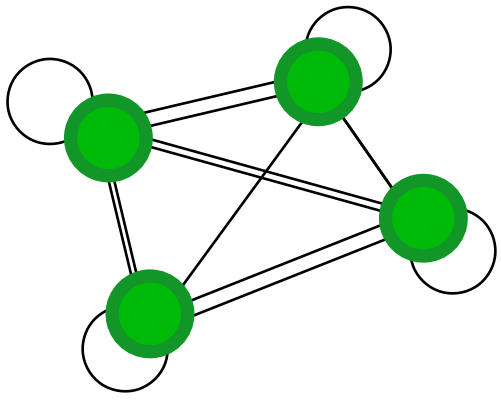


	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	80	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

	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



copy.pdf

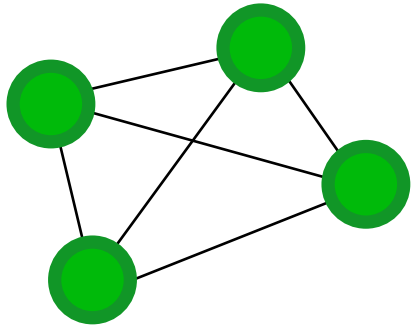
	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

Markov chain

Start at time step t_0

Randomly travel between connected nodes

Examine time steps $t_0 \rightarrow t_1 \rightarrow t_2$



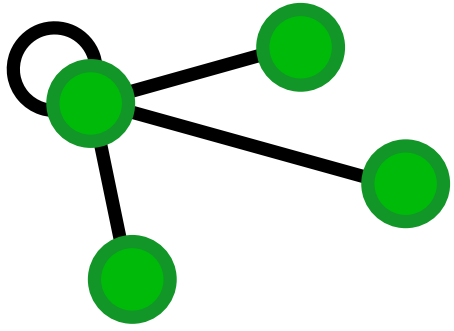
	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

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→A) * (A→A) = 0.6 * 0.6
and A→B then B→A = (A→B) * (B→A) = 0.2 * 0.1
and A→C then C→A = (A→C) * (C→A) = 0.1 * 0.1
and A→D then D→A = (A→D) * (D→A) = 0.1 * 0.1



	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

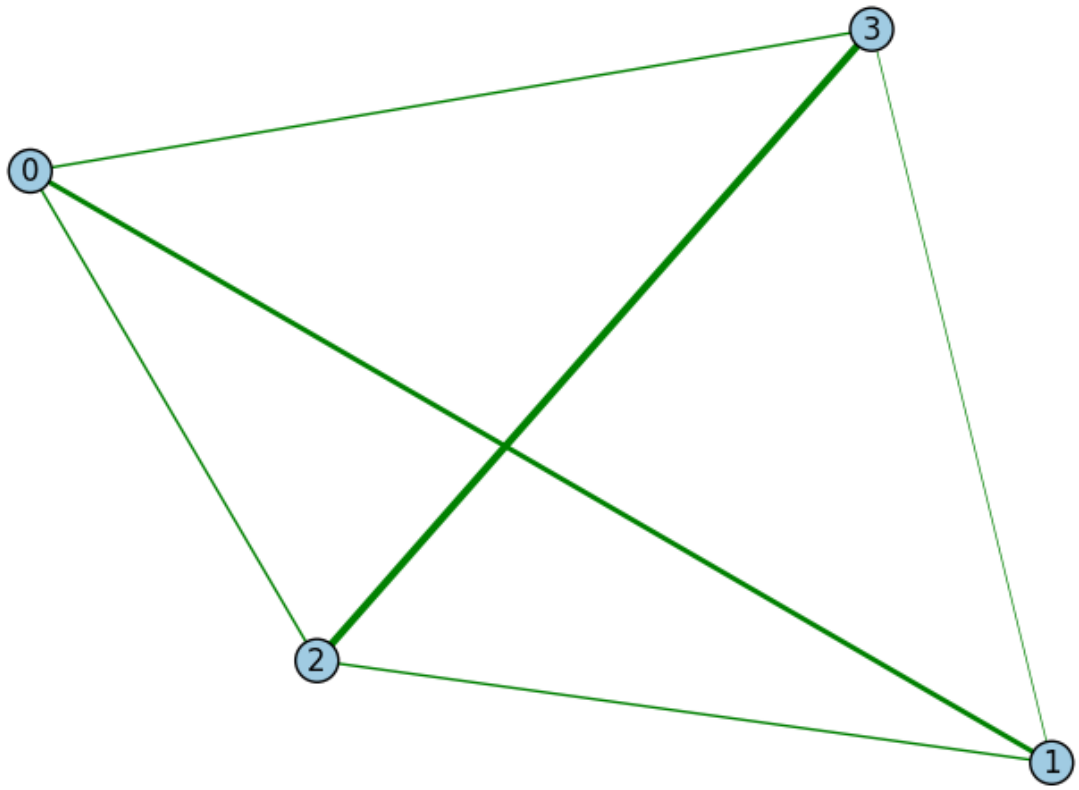
$$0.6 \times 0.6 + 0.1 \times 0.2 + 0.1 \times 0.1 + 0.1 \times 0.1 = 0.4$$

Markov chain

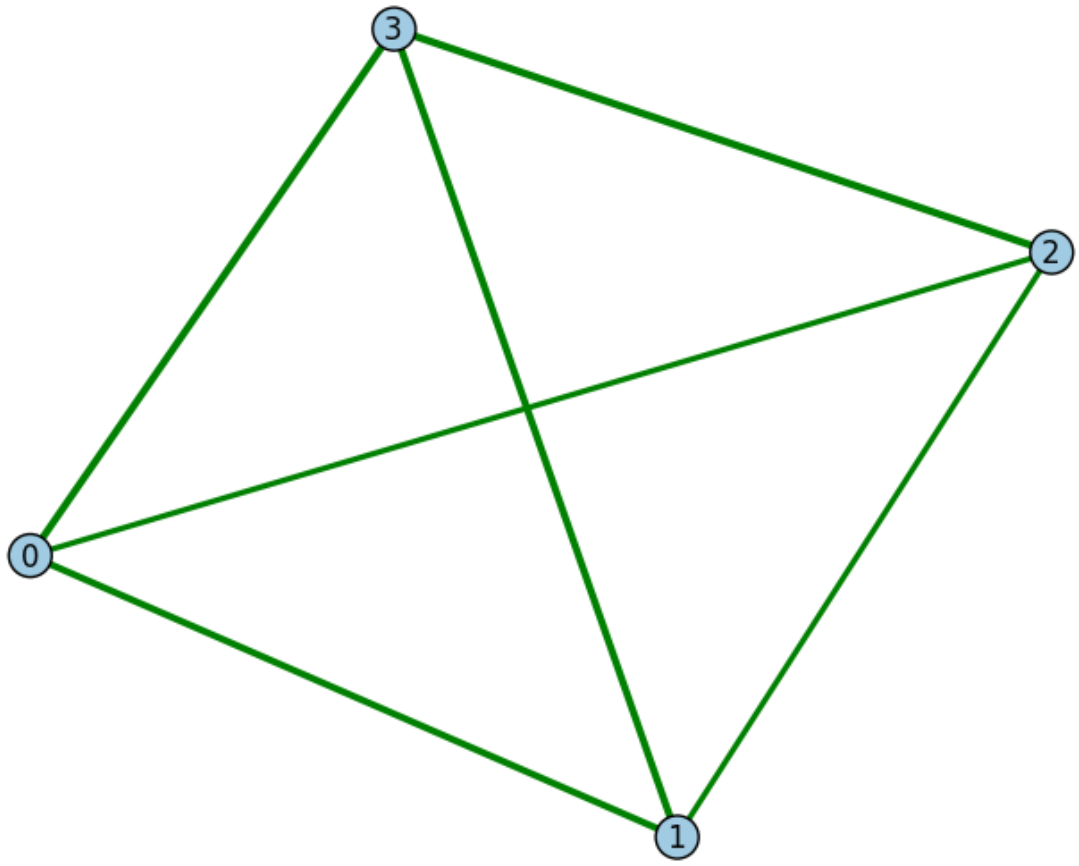
	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

	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

Markov chain



Markov chain



Markov chain

	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	0.2	0.2	0.2	0.2
B	0.274	0.274	0.274	0.274
C	0.229	0.229	0.229	0.229
D	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

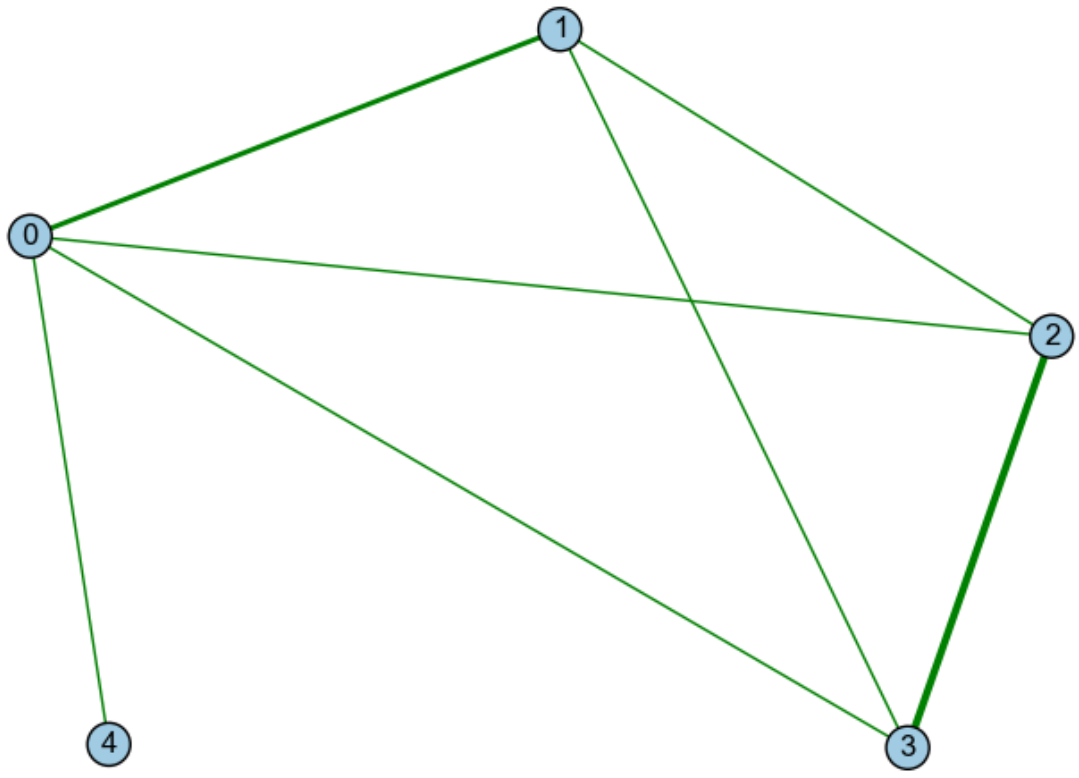
	A	B	C	D
A	0.6	0.1	0.1	0.1
B	0.2	0.8	0	0.05
C	0.1	0.1	0.6	0.15
D	0.1	0	0.3	0.7

	A	B	C	D
A	0.2	0.2	0.2	0.2
B	0.274	0.274	0.274	0.274
C	0.229	0.229	0.229	0.229
D	0.296	0.296	0.296	0.296

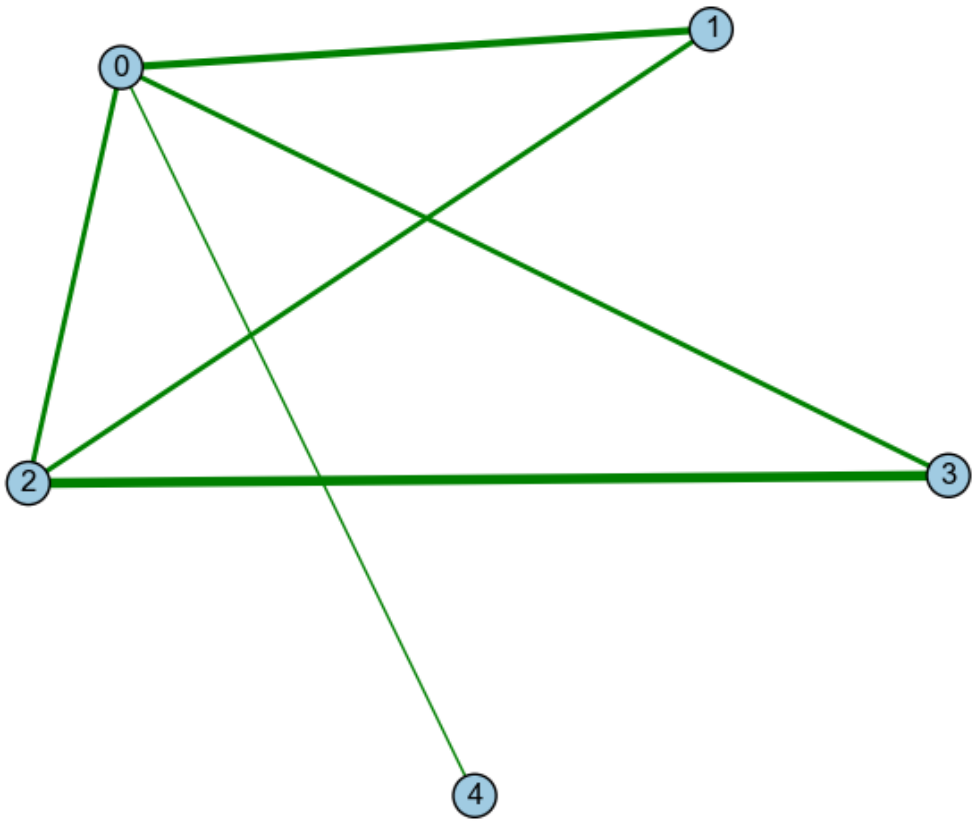
The clusters are more pronounced in the first than the last matrix

MCL slightly changes the procedure to emphasize these divisions

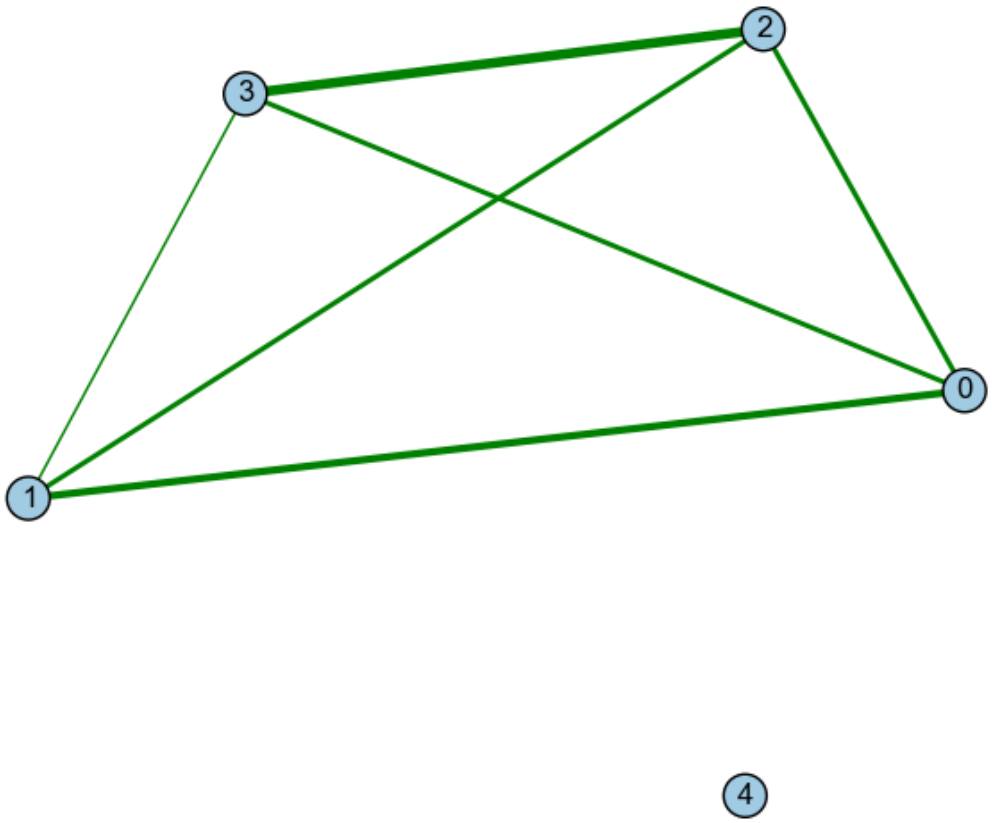
Markov chain



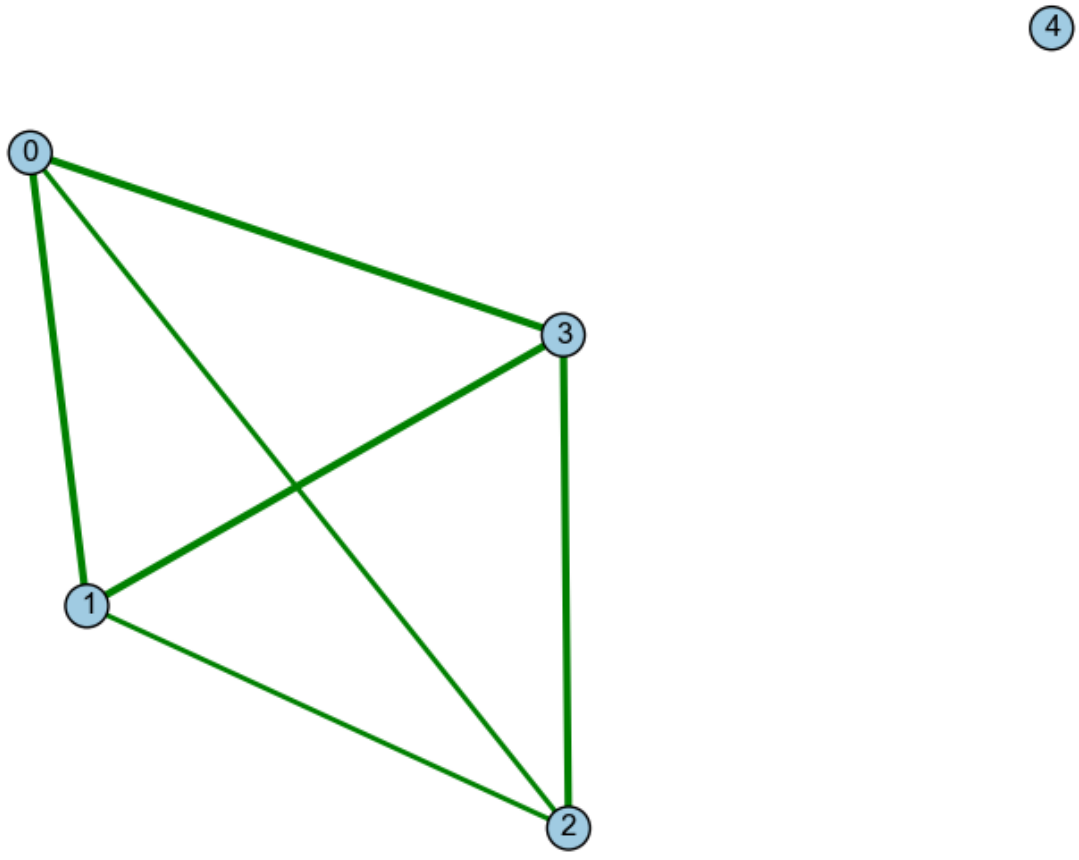
Markov chain



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 e, and inflation parameter r
2. Construct the matrix (e.g., add self loops, normalize the matrix)
3. Until convergence or maximum number of iterations
   - 3.1 expand the matrix to the e^{th} power (standard Markov chain)
   - 3.2 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

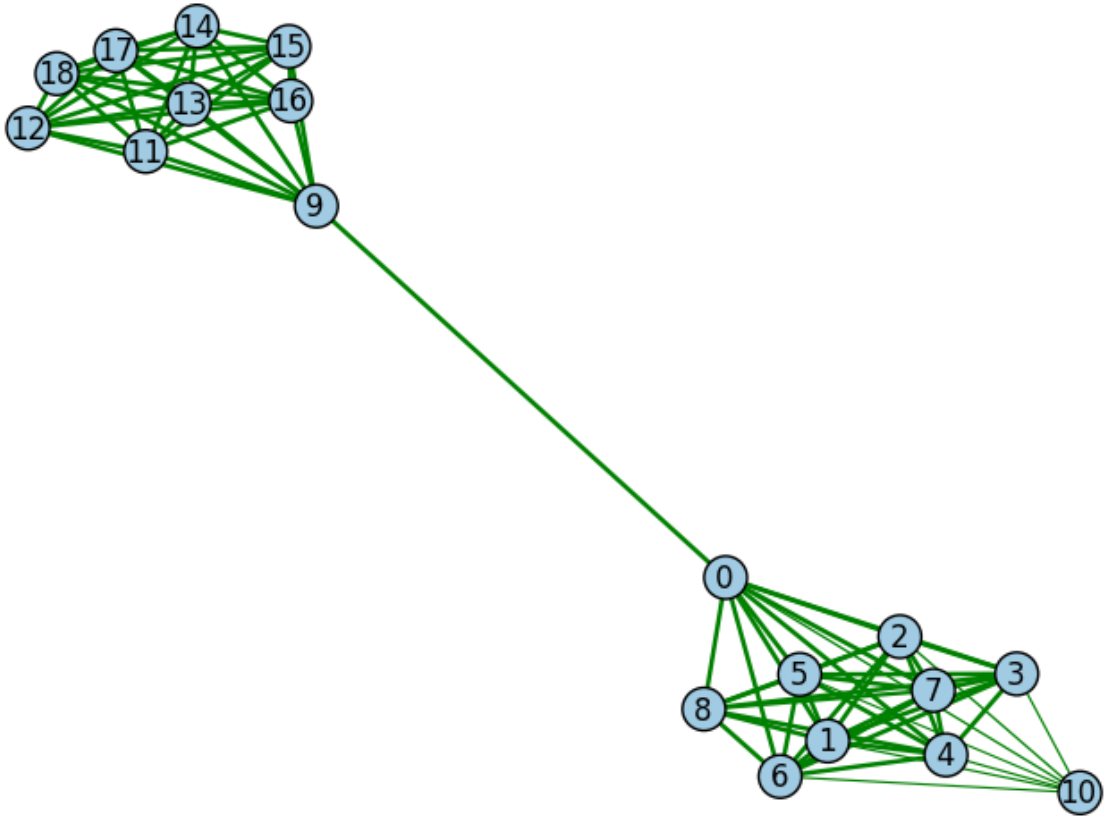
e expansion parameter will dissipate the matrix (smooth things out)

- 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

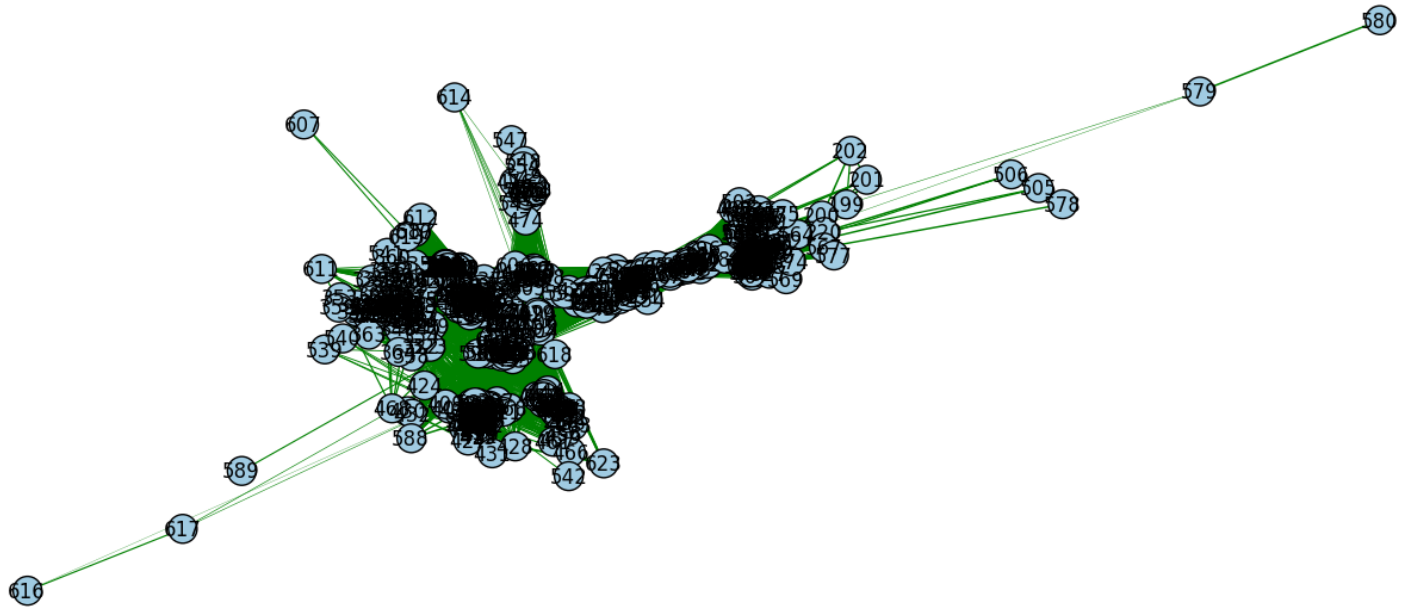
Two gene regions with one spurious connection

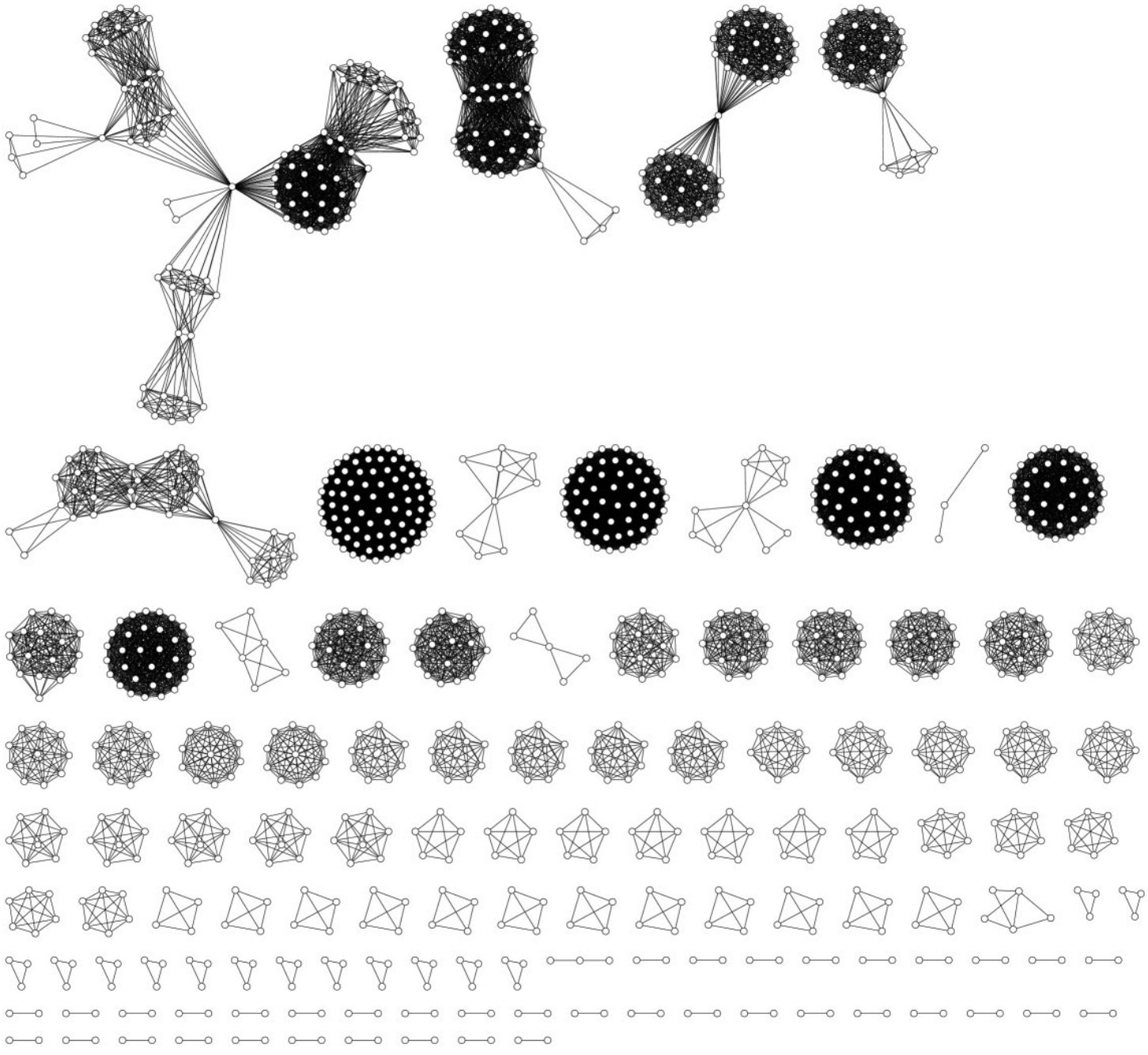


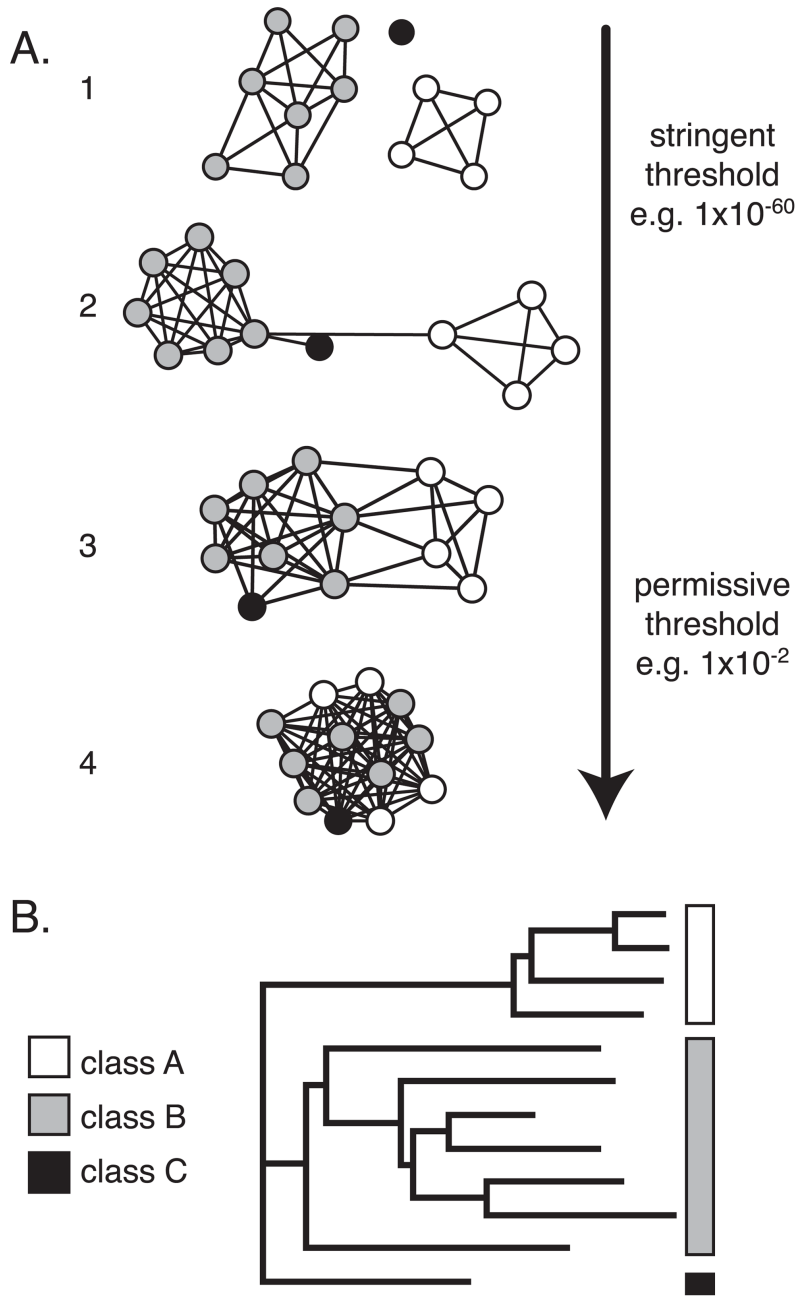
Complex example

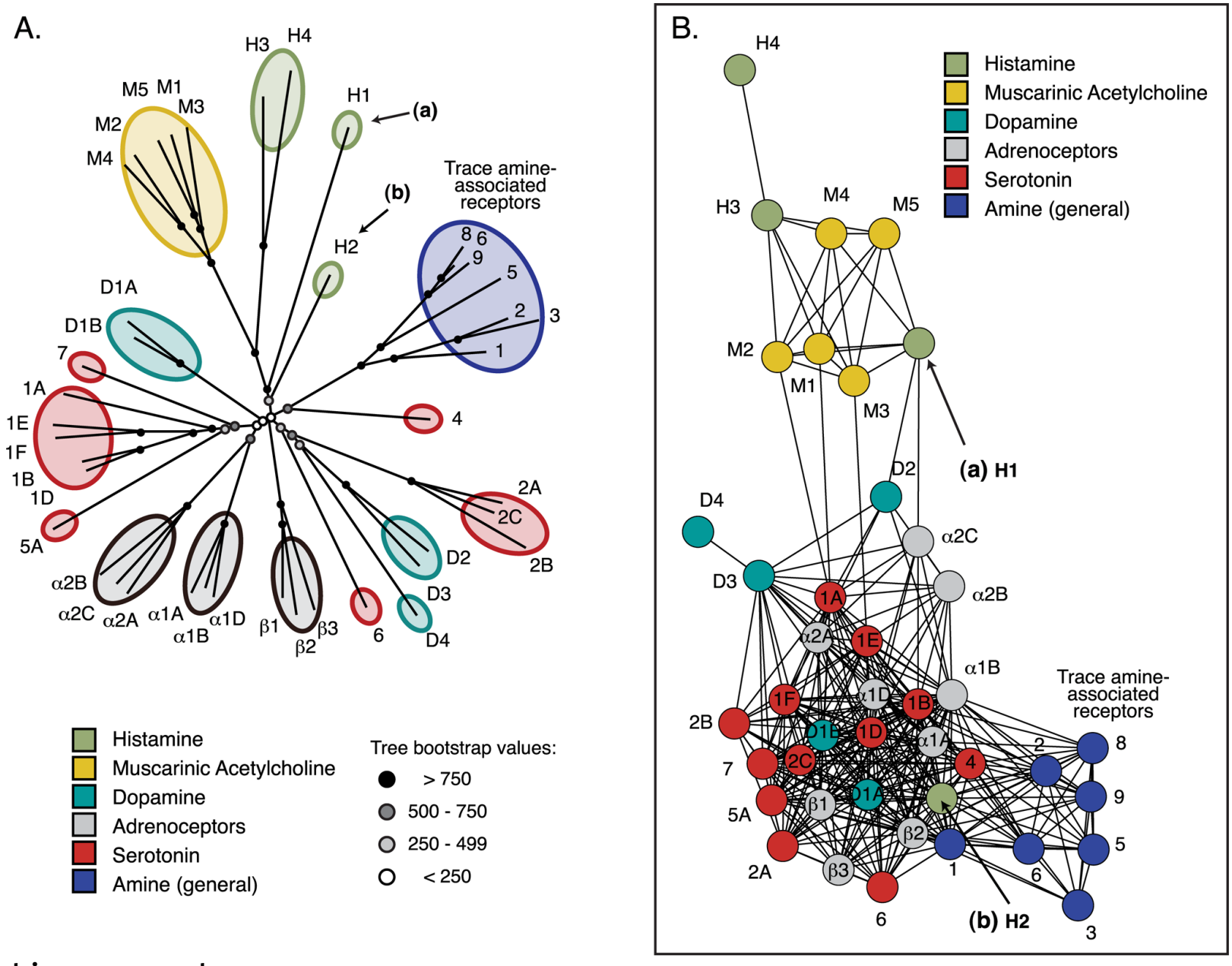
One gene regions many weak connections

Results in 16 clusters



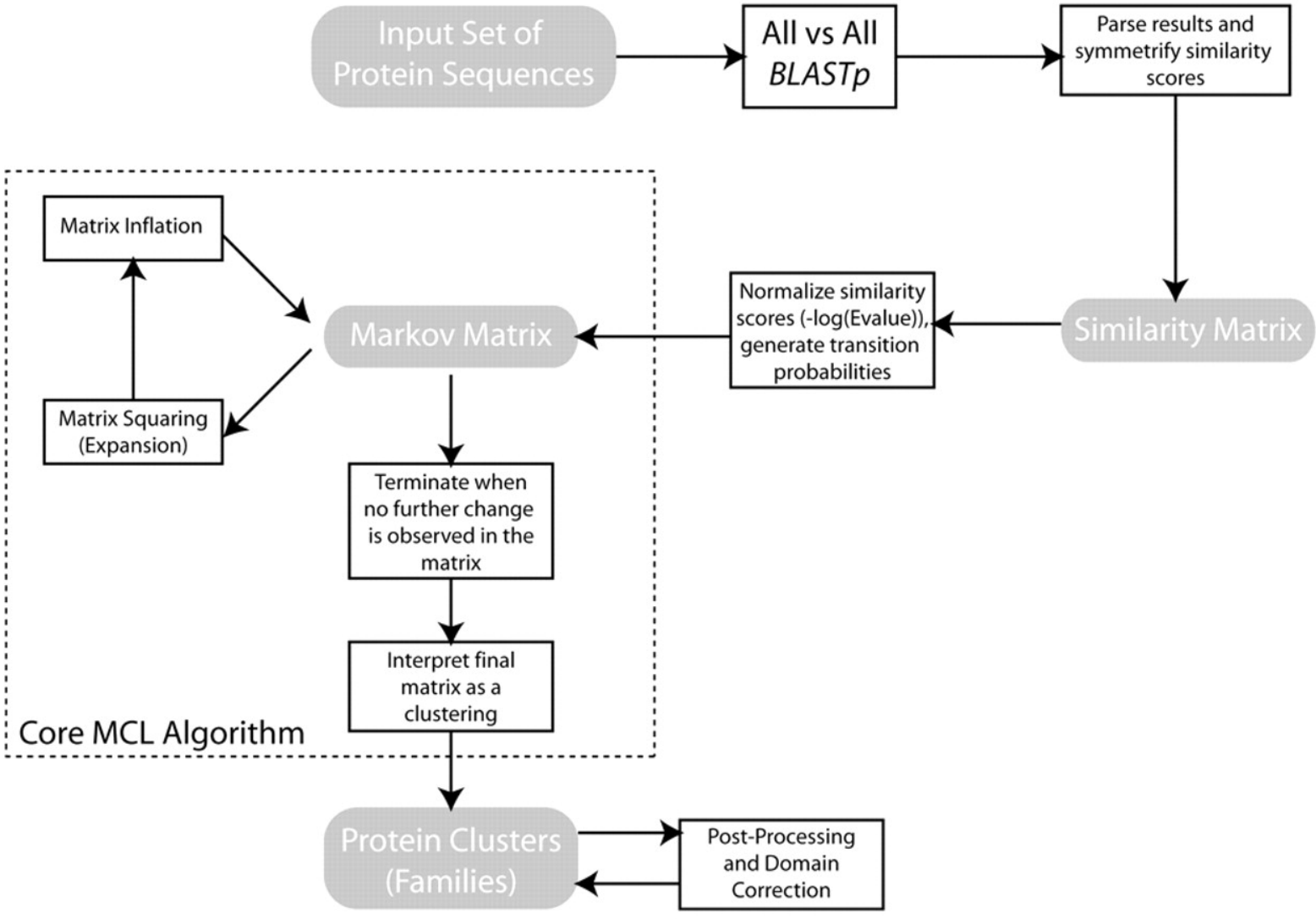



A.
1
2
3
4
stringent
threshold
e.g. 1x10-60
permissive
threshold
e.g. 1x10-2
B.
class A
class B
class C



Atkinson et al., 2009

Usage


Input Set of Protein Sequences
All vs All BLASTp
Parse results and symmetrify similarity scores
Similarity Matrix
Normalize similarity scores (-log(Evalue)), generate transition probabilities
Markov Matrix
Matrix Inflation
Matrix Squaring (Expansion)
Terminate when no further change is observed in the matrix
Interpret final matrix as a clustering
Core MCL Algorithm
Protein Clusters (Families)
Post-Processing and Domain Correction

MCL

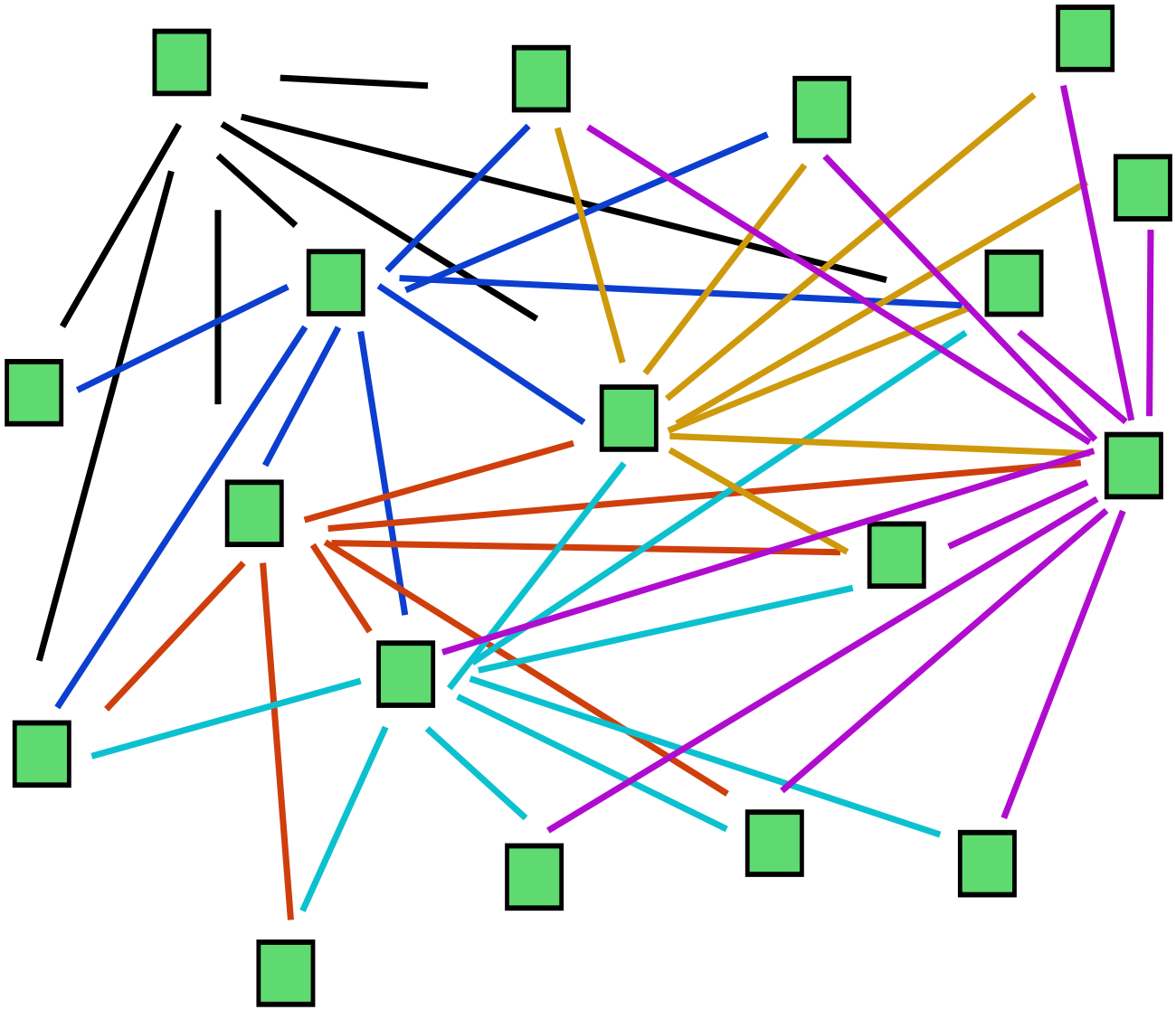
MCL can be successful, in part due to simplicity and speed

One of the most commonly used methods

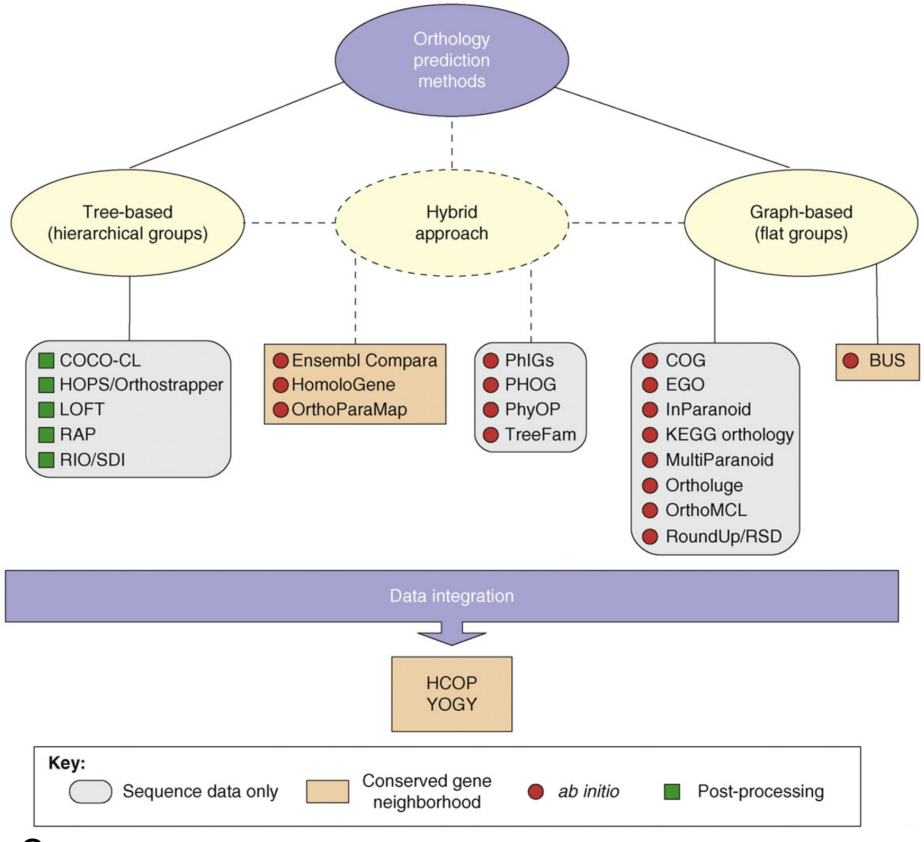
There are situations where MCL does not perform as well

All-*versus*-all comparisons can take a very long time

They also scale poorly



Other options

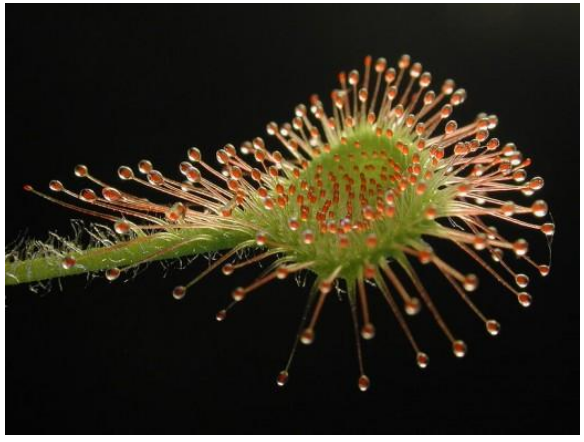


Kuzniar et al., 2008

There is a lot of room for development

Develop your own!

Caryophyllales phylogeny



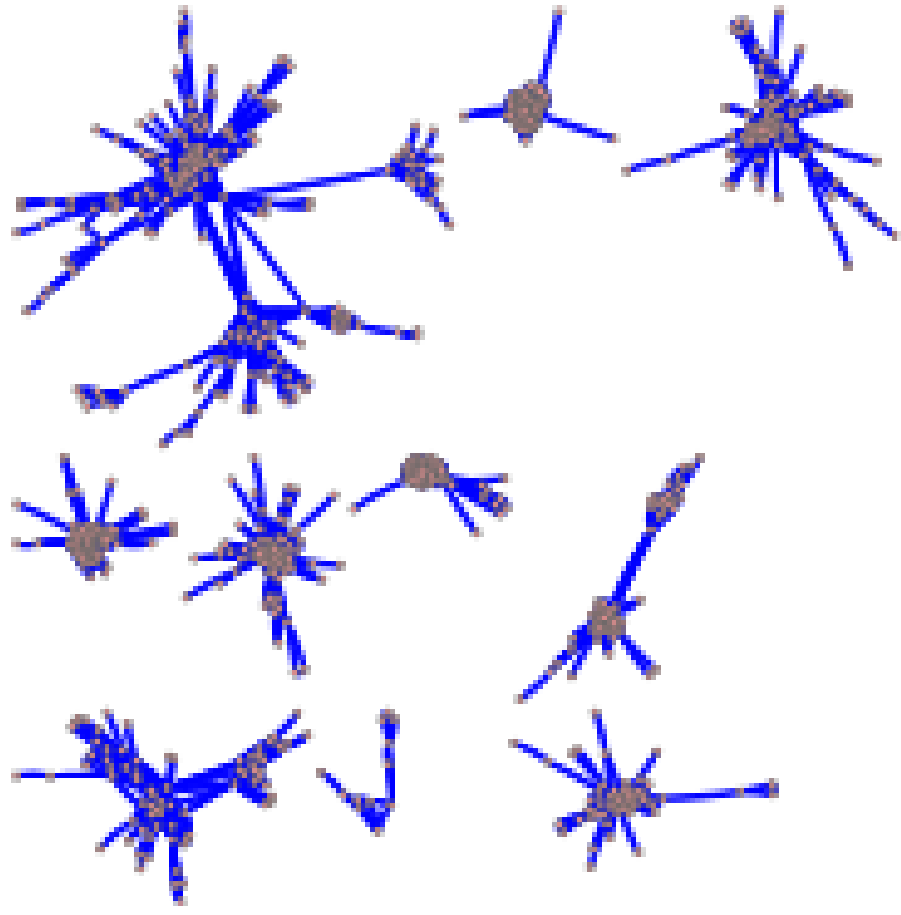
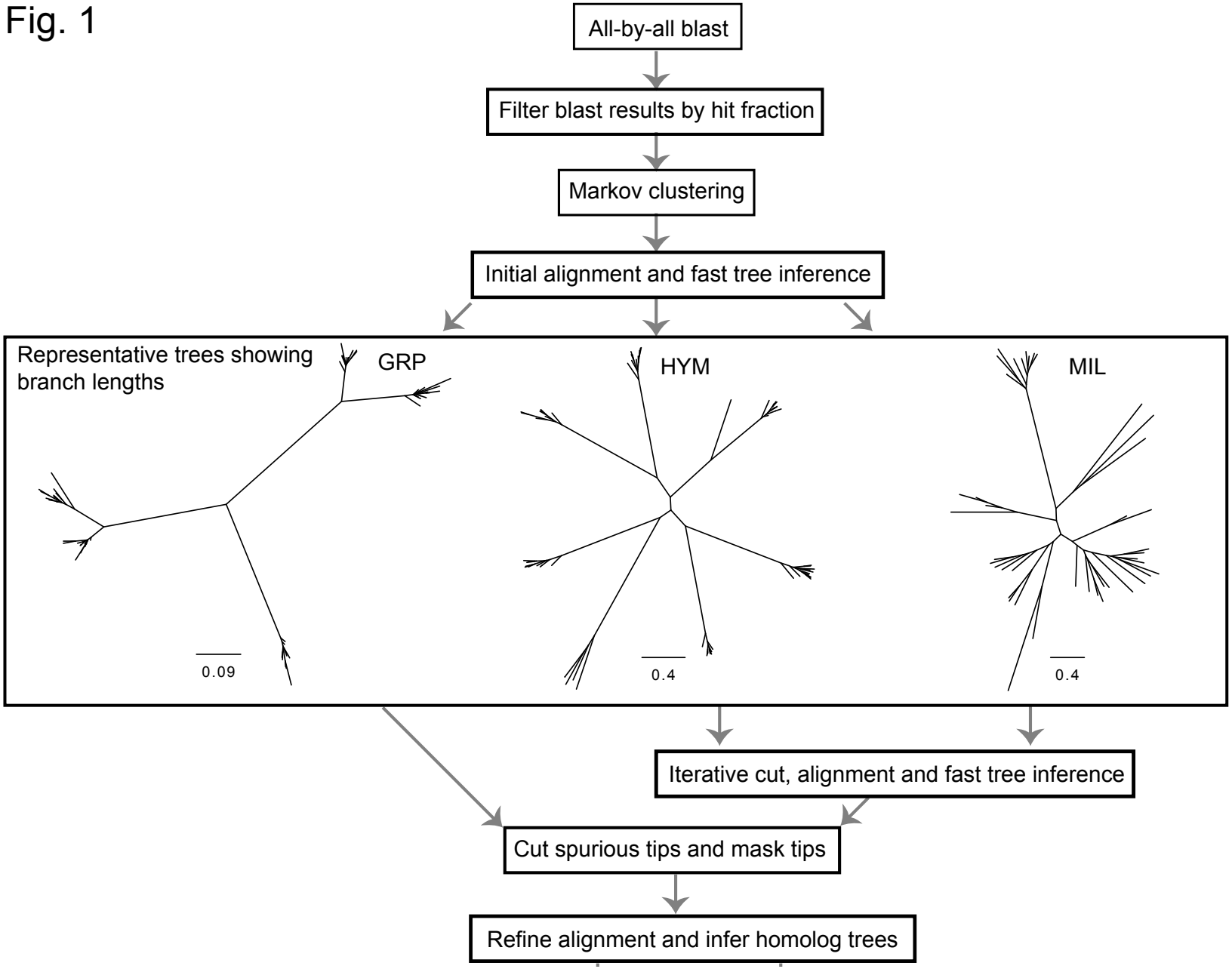
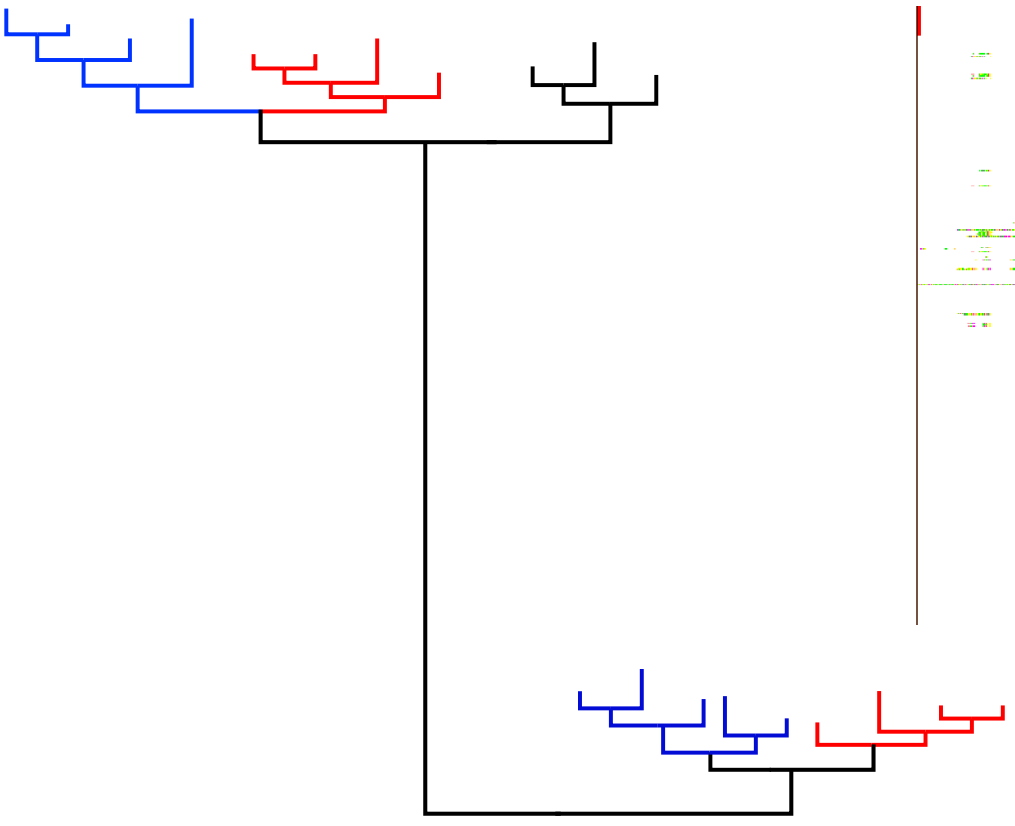


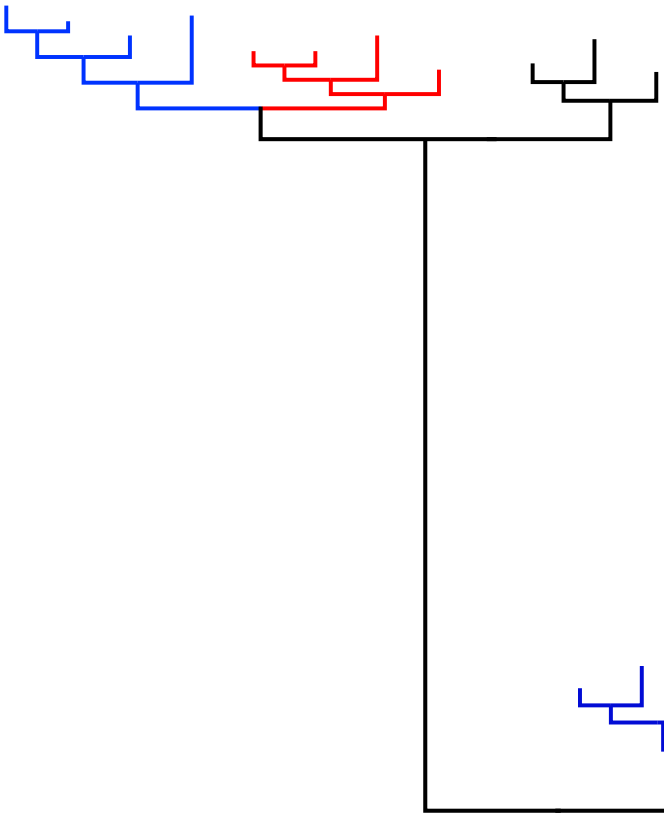
Fig. 1

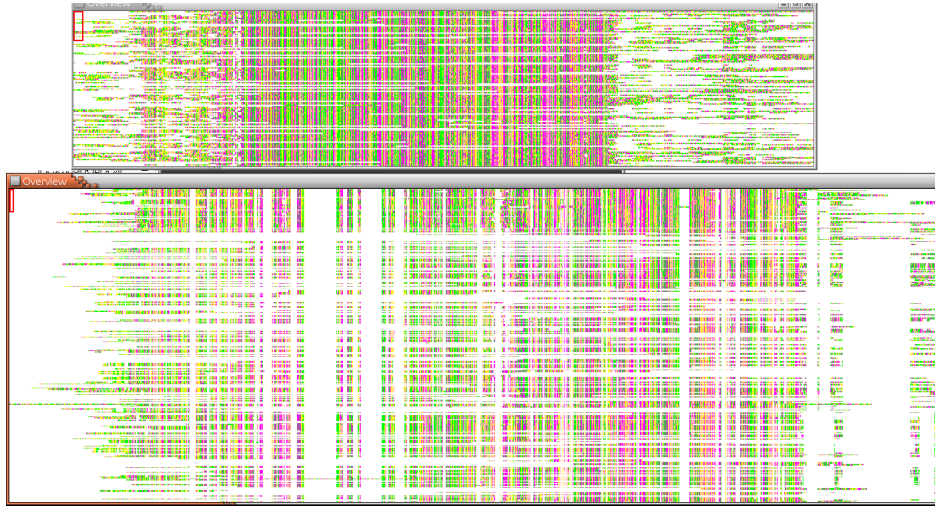

All-by-all blast
Filter blast results by hit fraction
Markov clustering
Initial alignment and fast tree inference
Representative trees showing branch lengths
GRP
HYM
MIL
0.09
0.4
0.4
Iterative cut, alignment and fast tree inference
Cut spurious tips and mask tips
Refine alignment and infer homolog trees

Homolog assignment

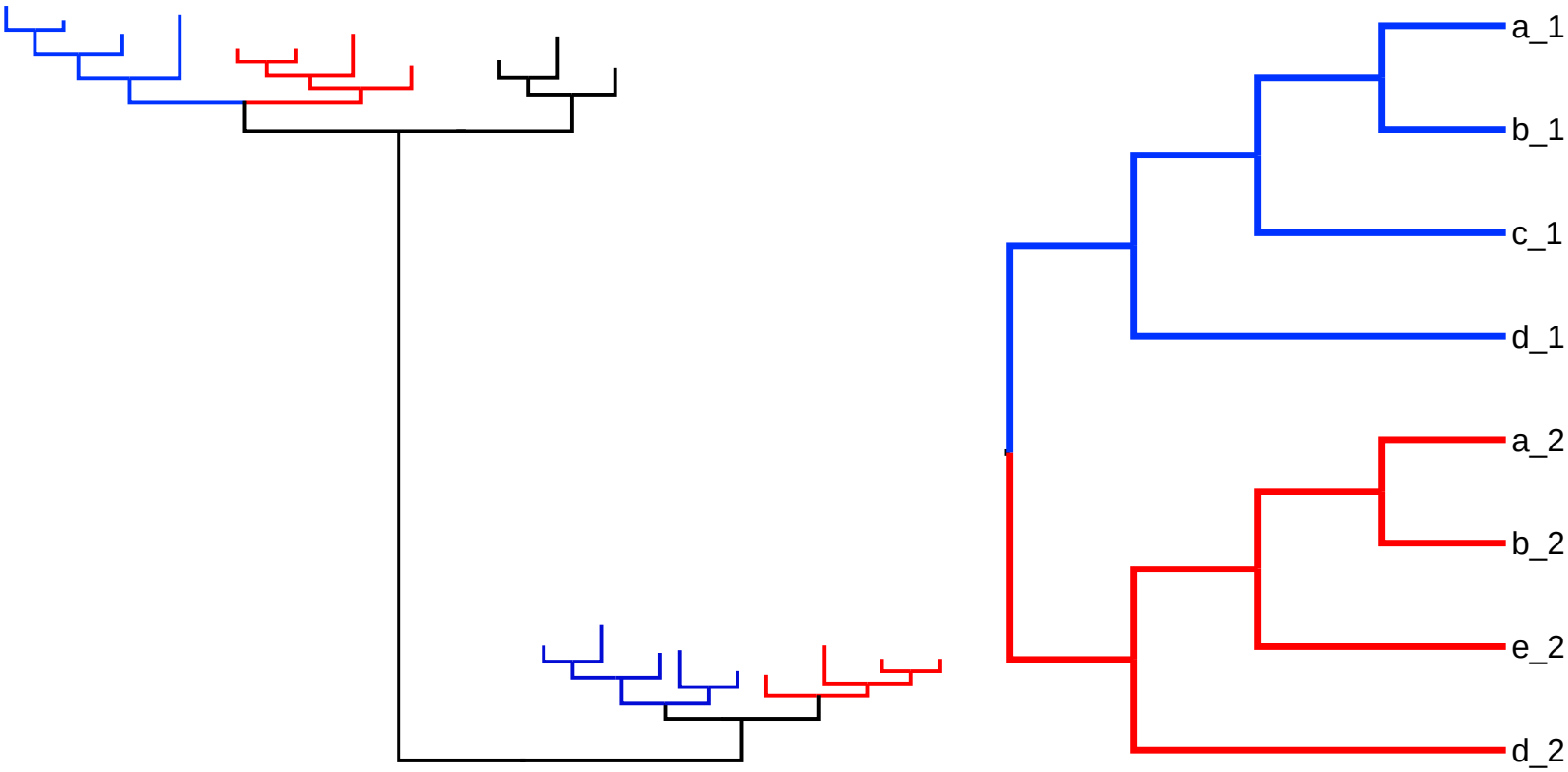


Homolog assignment

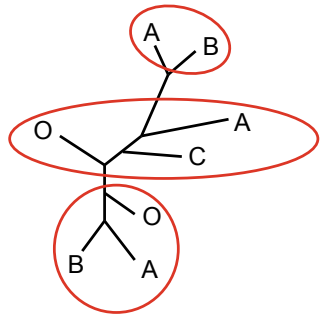
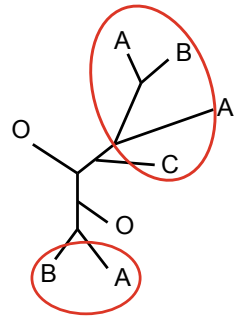
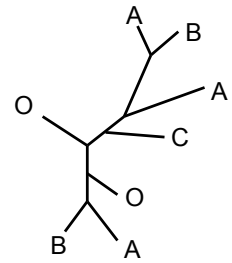
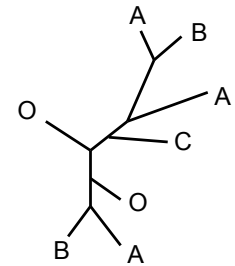
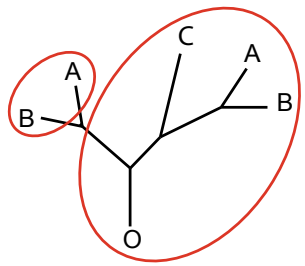
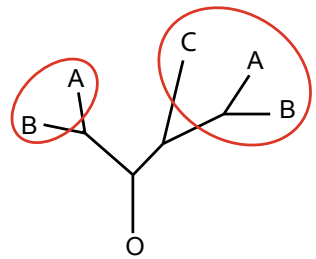
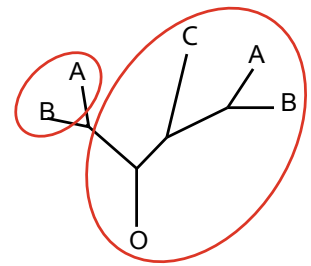
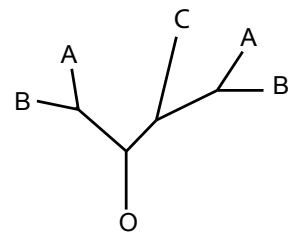
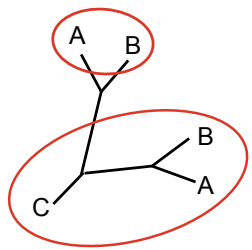
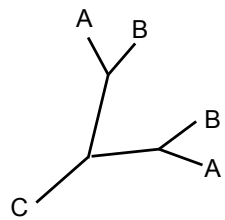
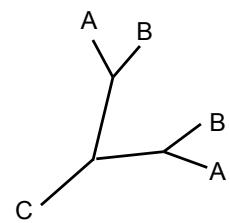
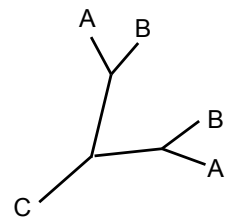
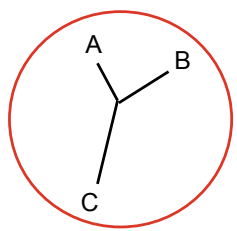
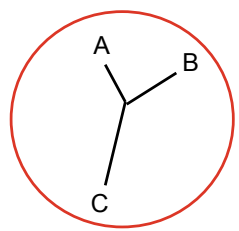
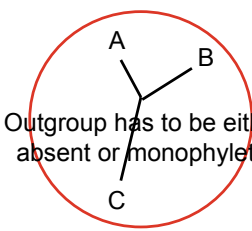
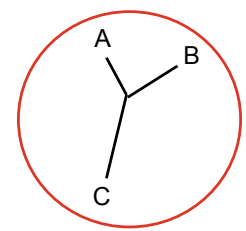


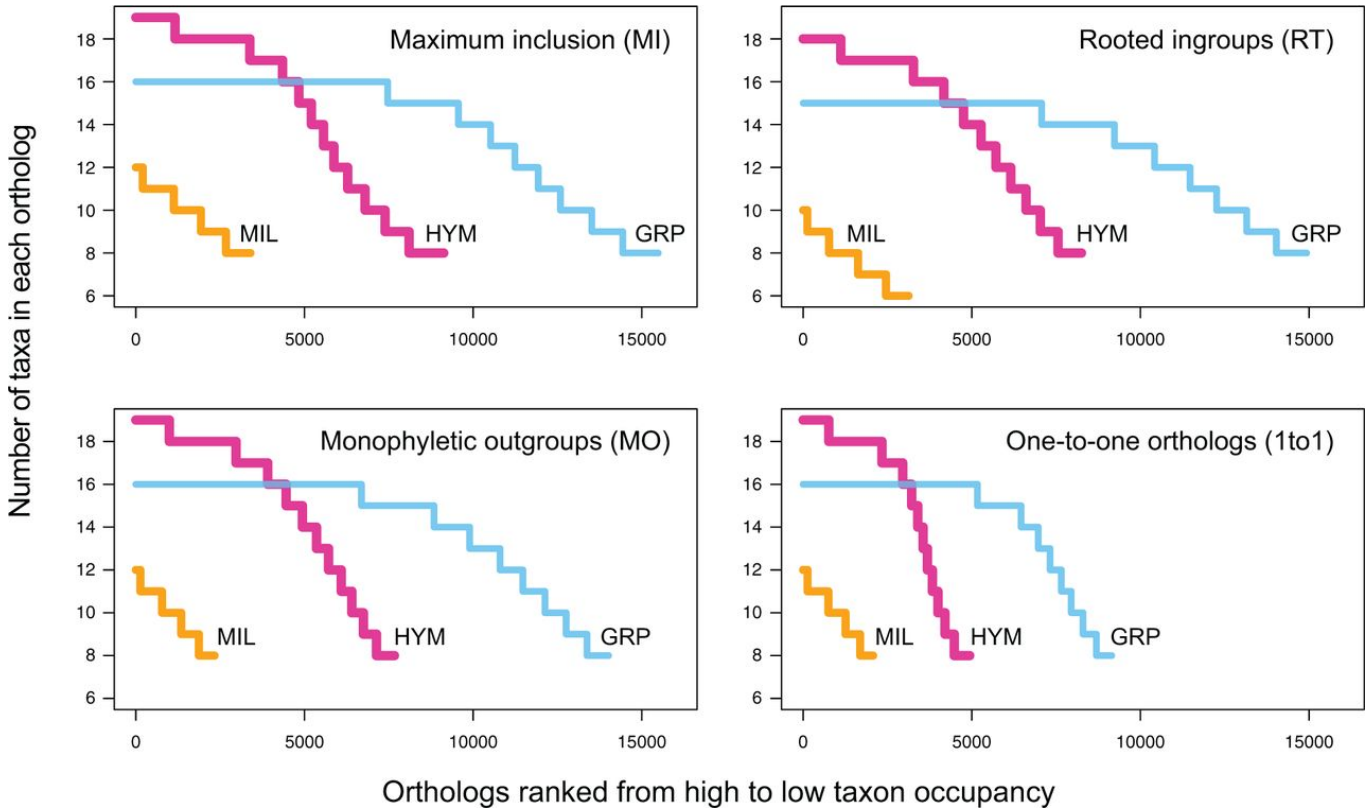


Orthology assignment

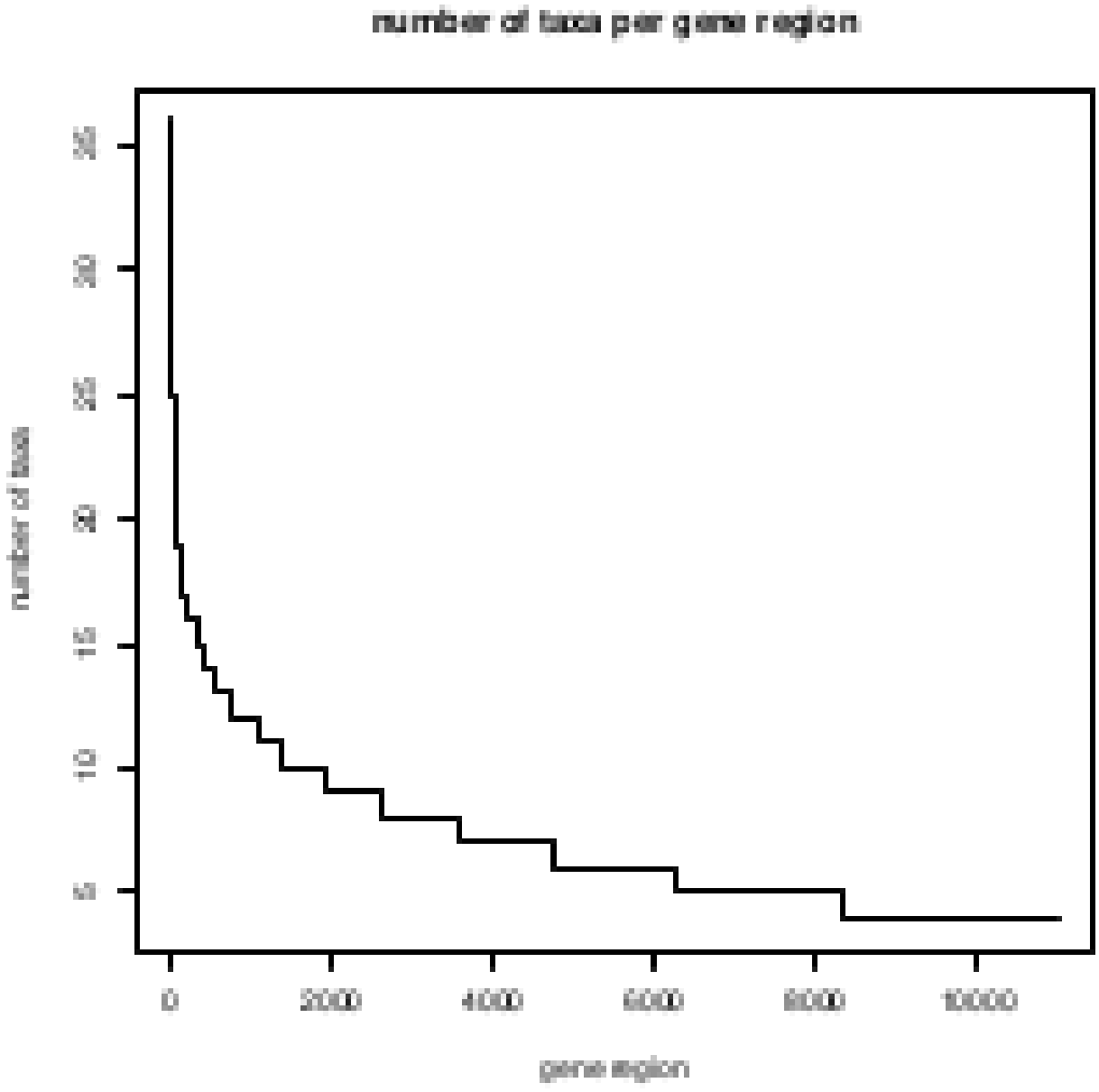


Ingroups: A, B, C
Outgroup: O
Infered orthologs:

	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			 Outgroup has to be either absent or monophyletic	

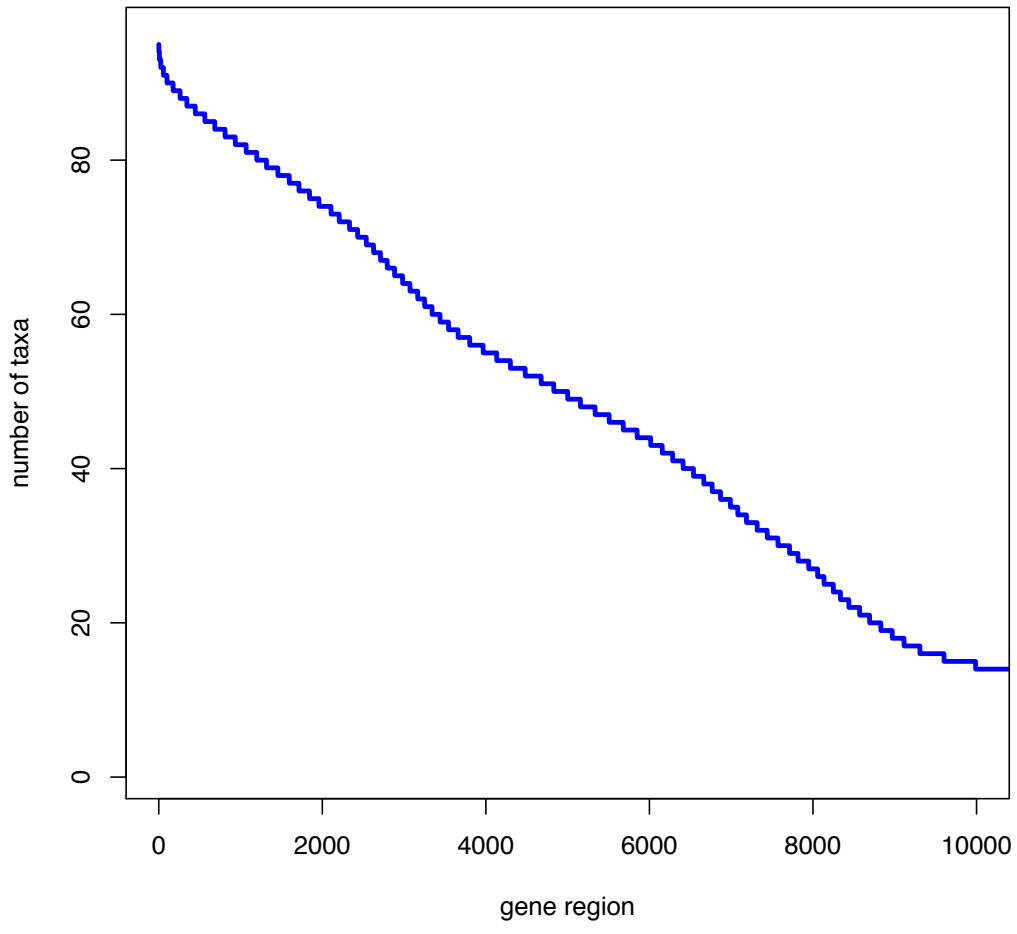


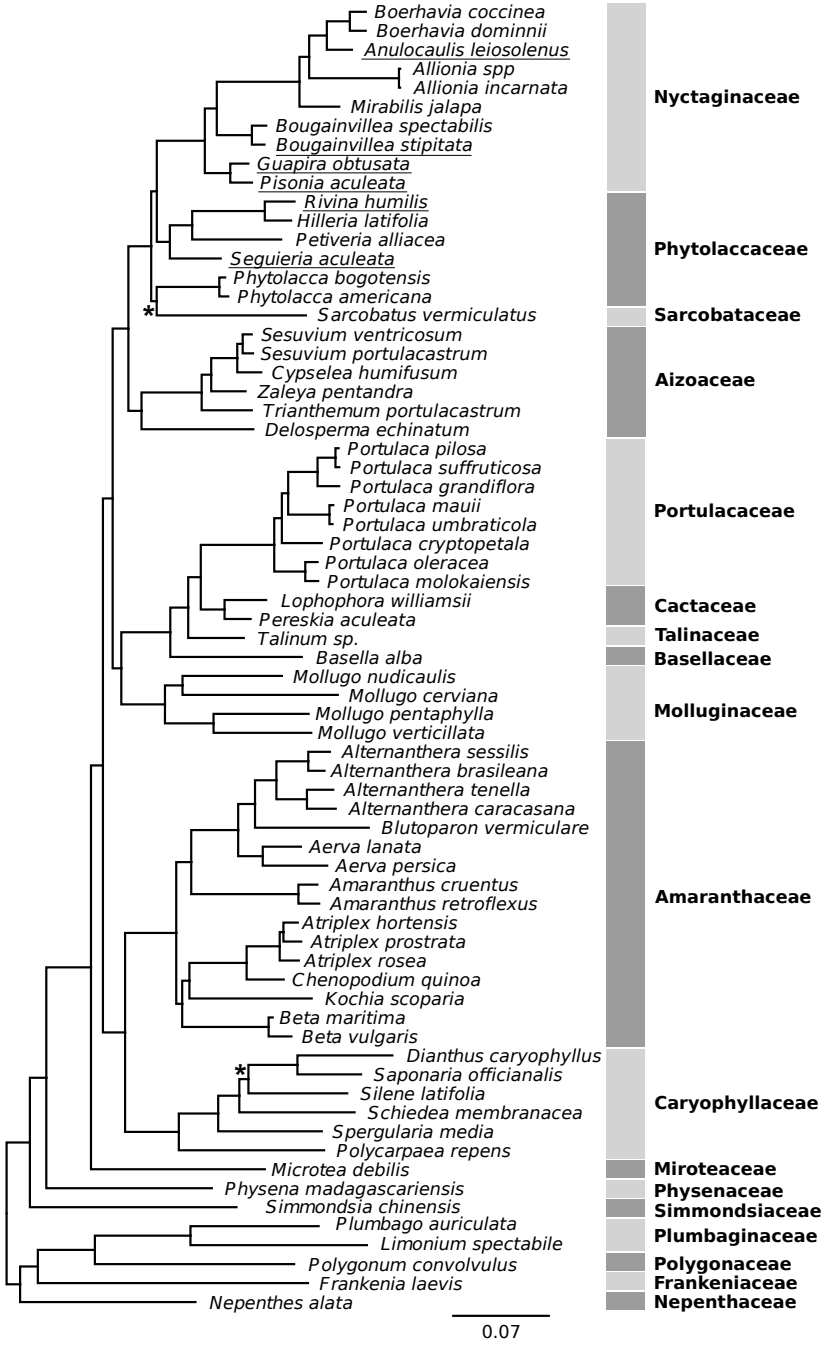
Yang and Smith, MBE, 2014


number of bins
gene region
0
2000
4000
6000
8000
10000

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

(compare to with 300 genes at 50%)

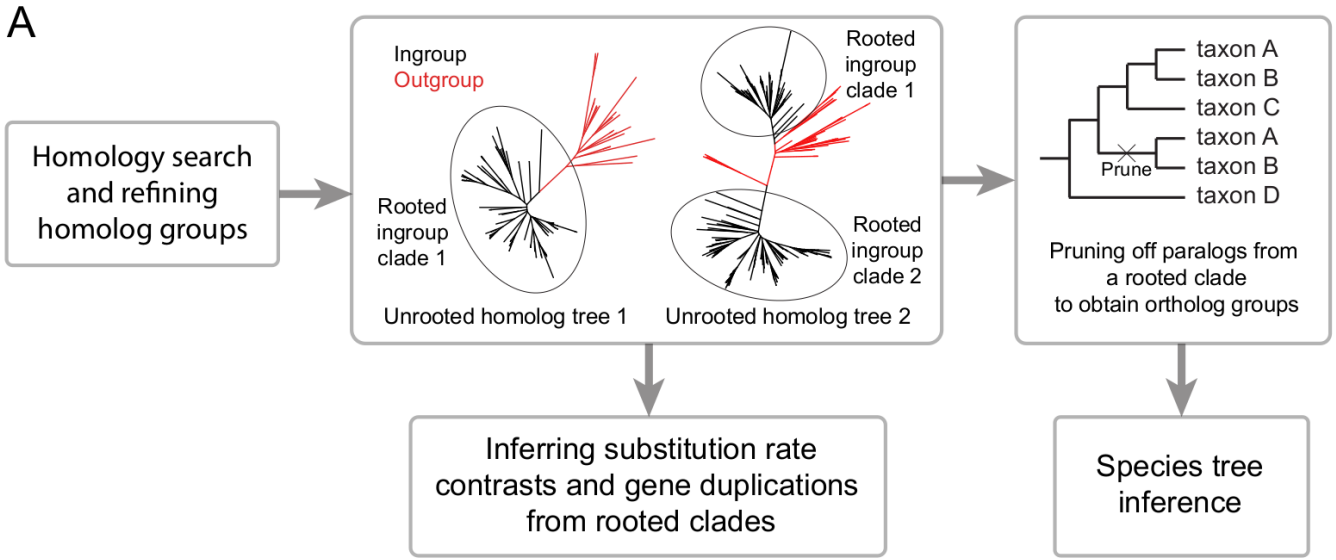


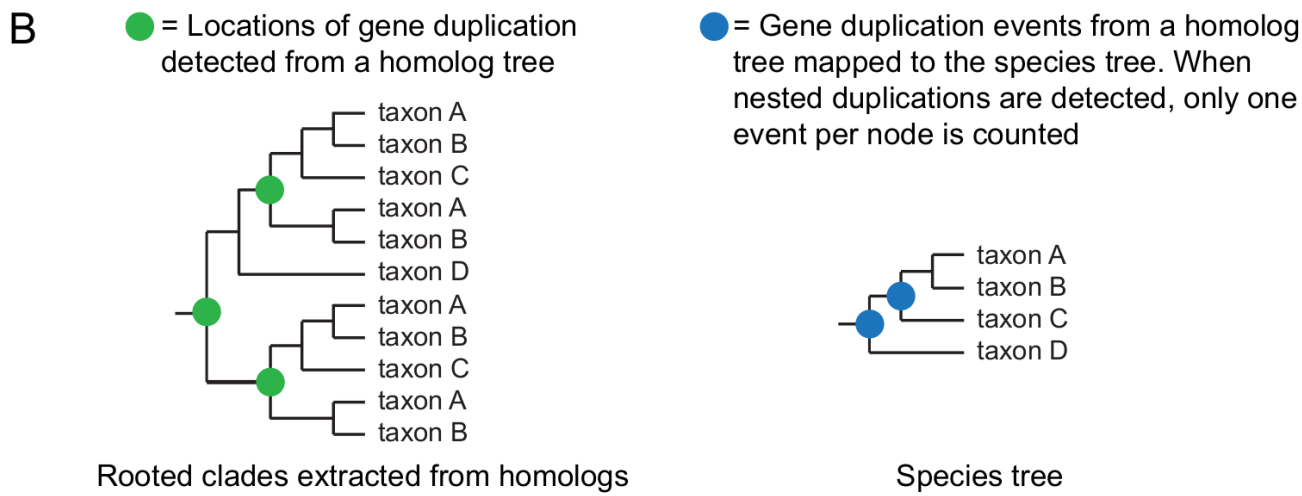


96 species (68 ingroup), Illumina (1kp and our own)

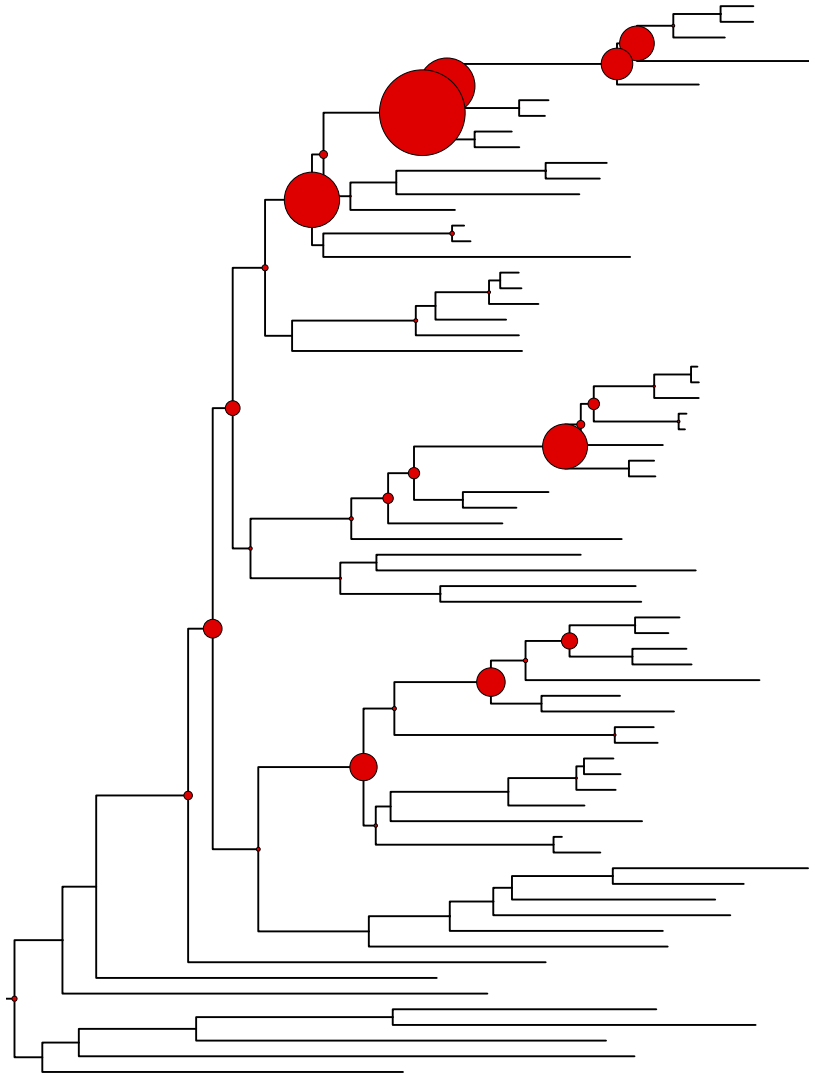
1000-10000 gene regions (nucleotide and amino acid)

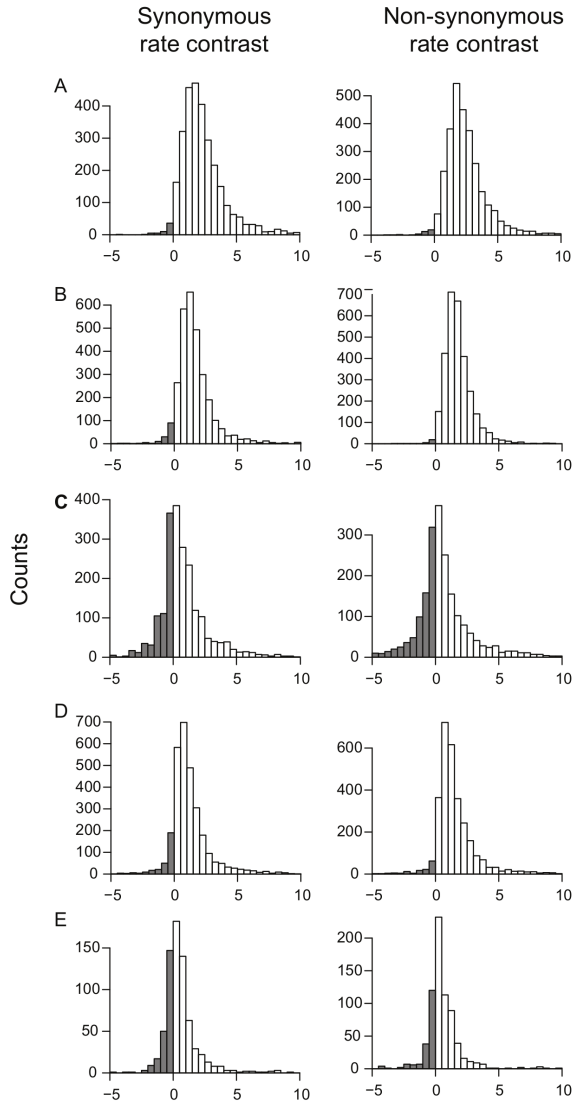
Stephen Smith, Ya Yang, Michael Moore, and Sam Brockington


A
Homology search and refining homolog groups
Ingroup
Outgroup
Rooted ingroup clade 1
Rooted ingroup clade 1
Rooted ingroup clade 2
Unrooted homolog tree 1
Unrooted homolog tree 2
taxon A
taxon B
taxon C
taxon A
taxon B
taxon D
Prune
Pruning off paralogs from a rooted clade to obtain ortholog groups
Inferring substitution rate contrasts and gene duplications from rooted clades
Species tree inference

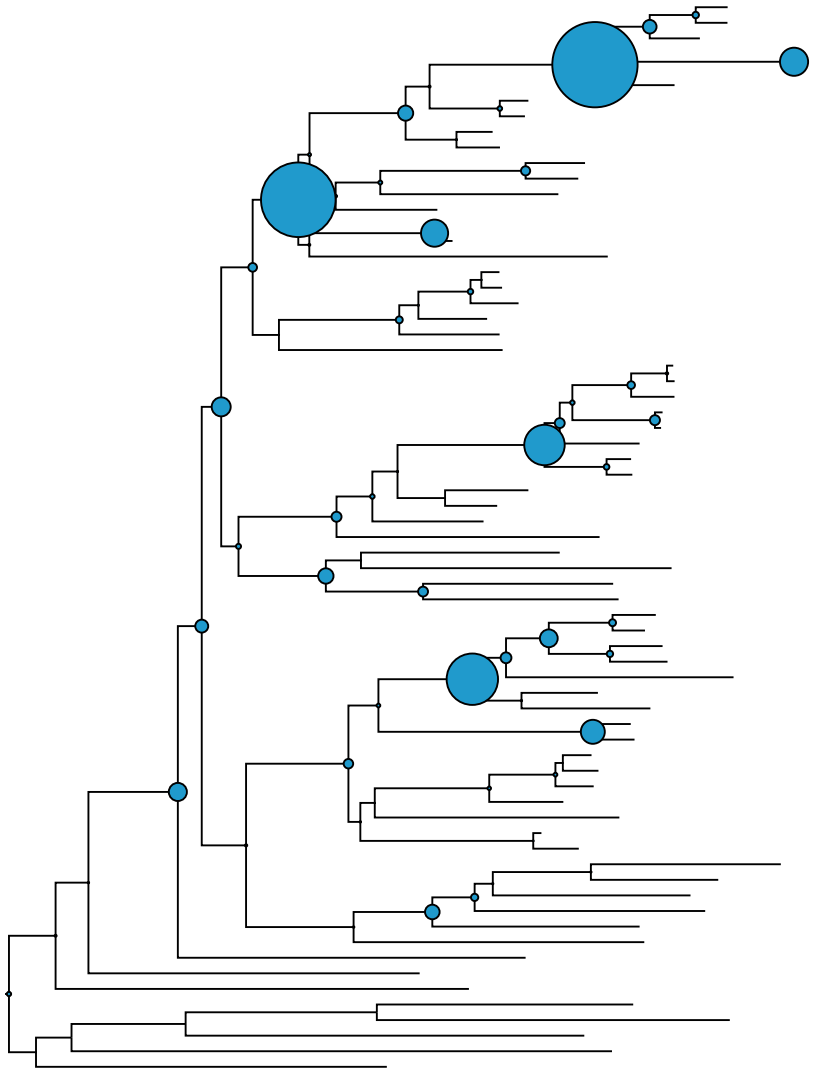

B
= Locations of gene duplication detected from a homolog tree
= Gene duplication events from a homolog tree mapped to the species tree. When nested duplications are detected, only one event per node is counted
taxon A
taxon B
taxon C
taxon A
taxon B
taxon D
taxon A
taxon B
taxon C
taxon A
taxon B
Rooted clades extracted from homologs
taxon A
taxon B
taxon C
taxon D
Species tree

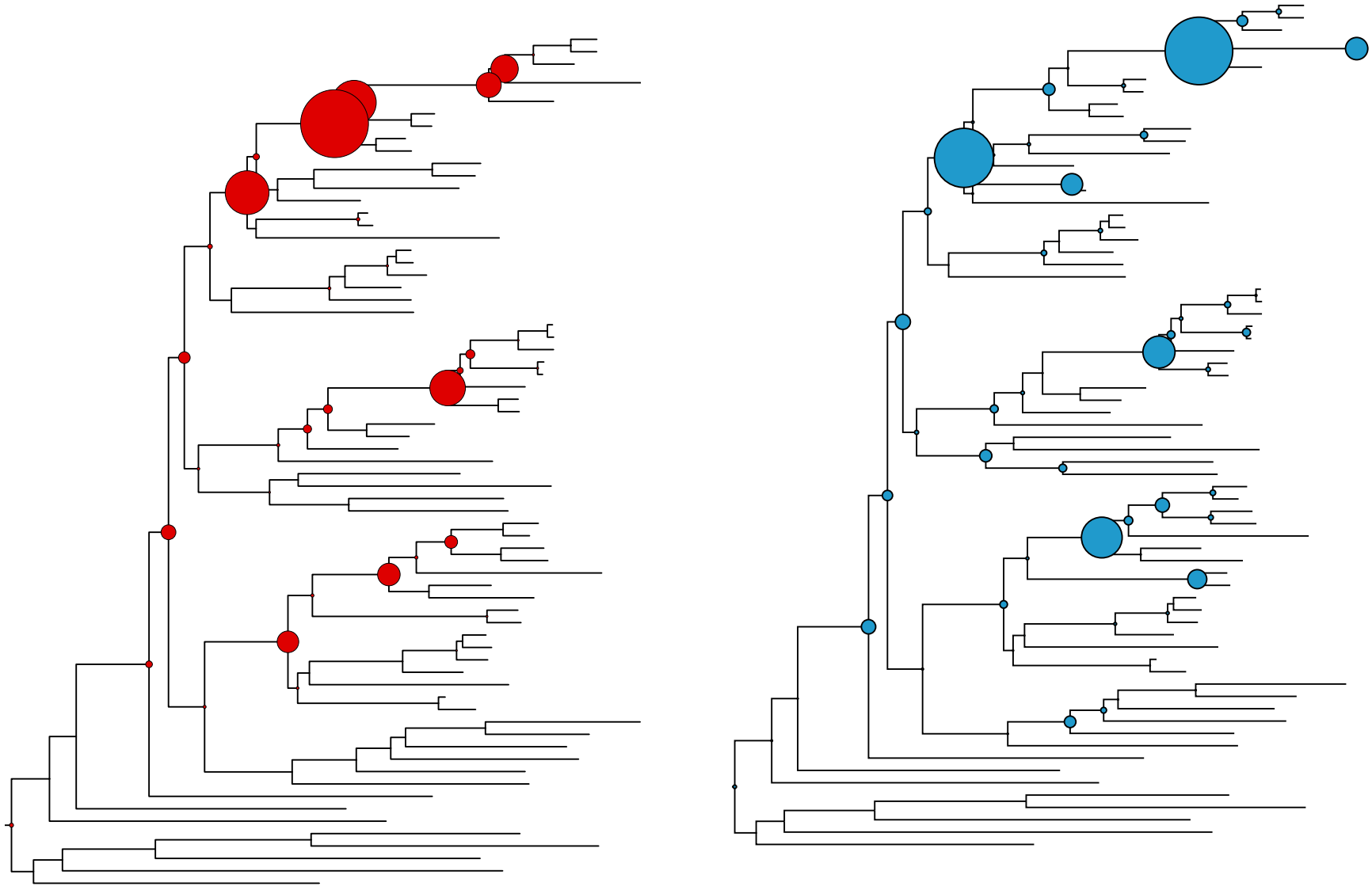
Rate shifts

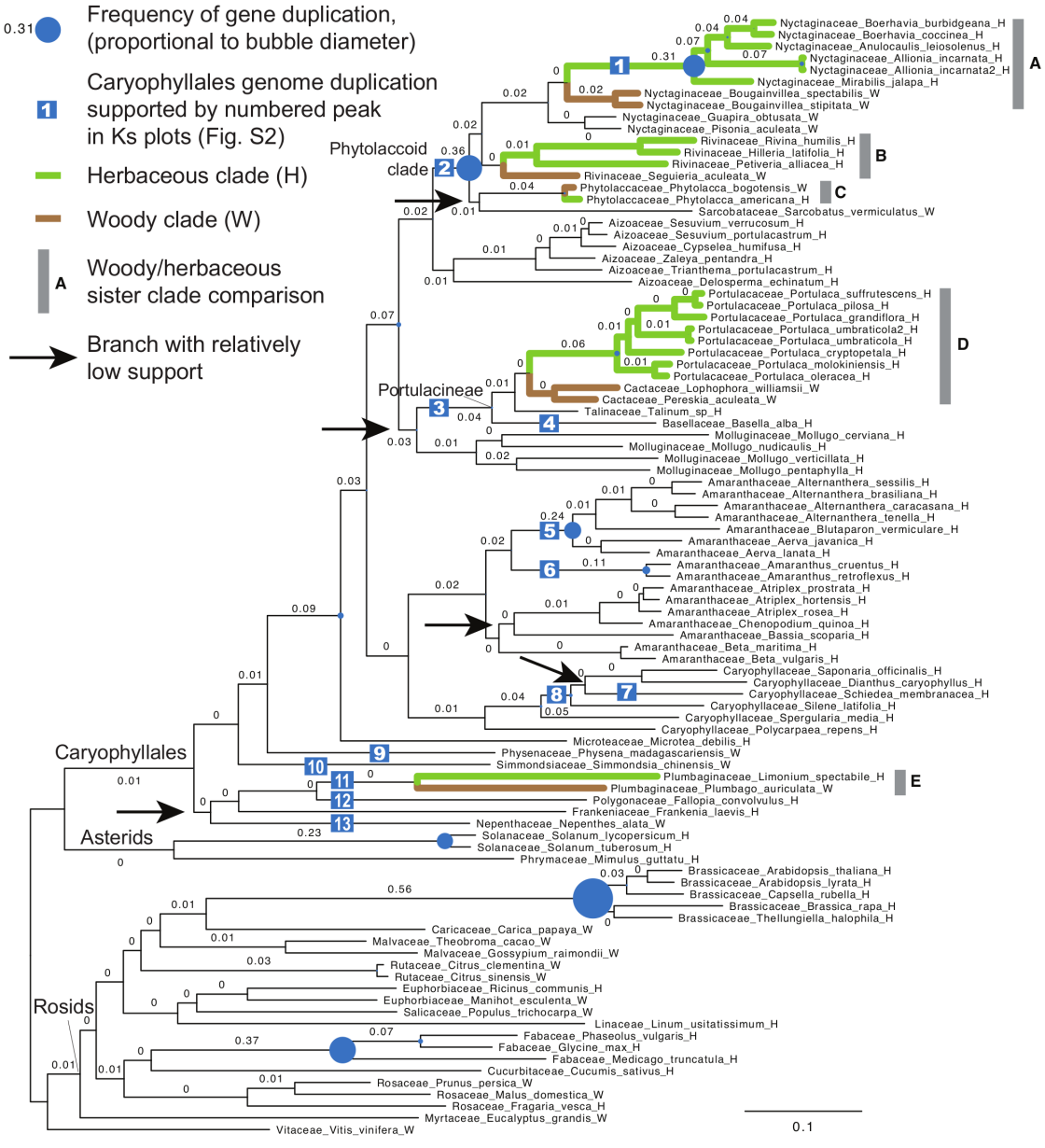


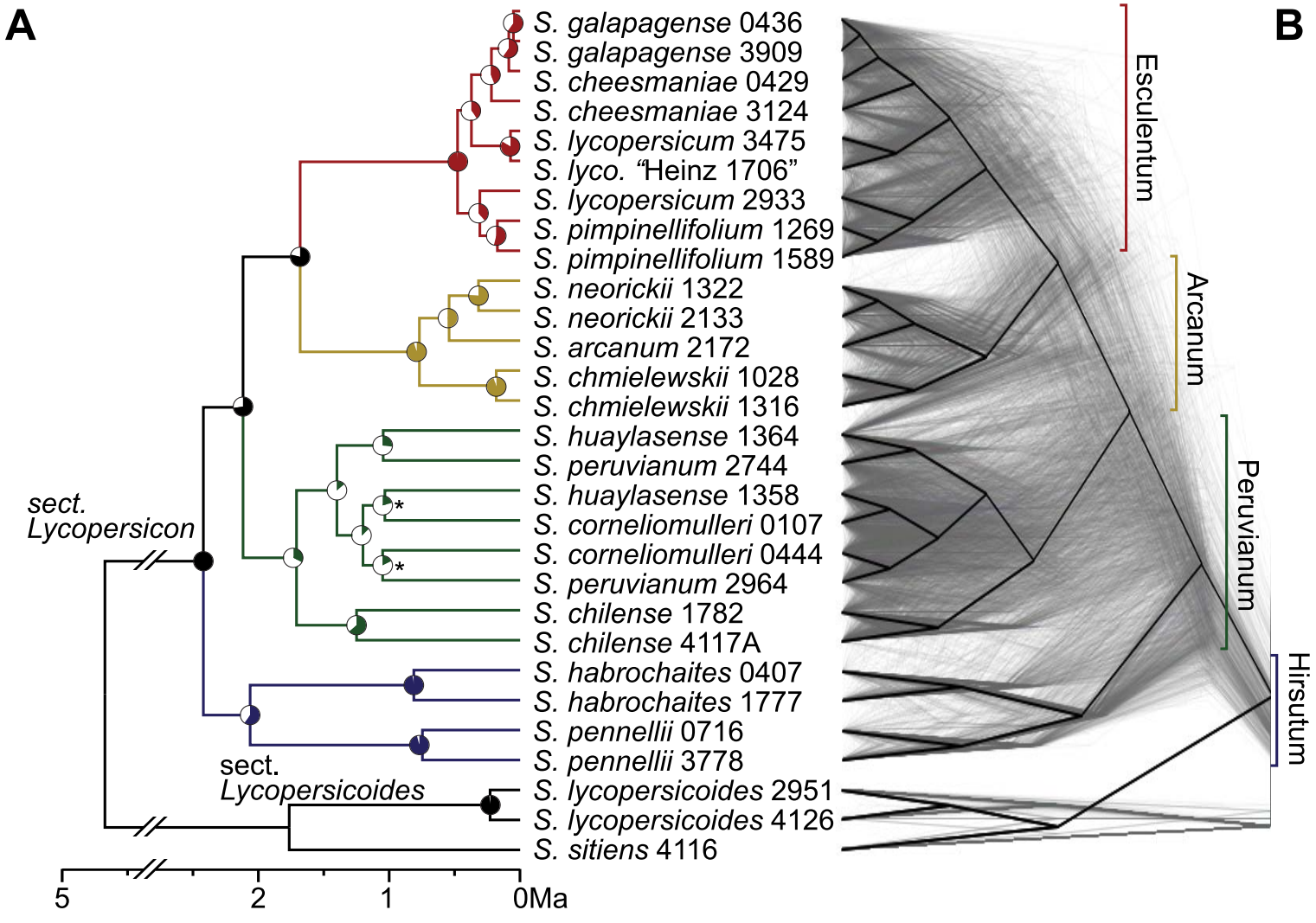

Synonymous rate contrast
Non-synonymous rate contrast
A
B
C
D
E
Counts

Gene duplications






0.31
Frequency of gene duplication, (proportional to bubble diameter)
1
Caryophyllales genome duplication supported by numbered peak in Ks plots (Fig. S2)
Herbaceous clade (H)
Woody clade (W)
A
Woody/herbaceous sister clade comparison
Branch with relatively low support
Phytolaccoid clade
Portulacineae
Caryophyllales
Asterids
Rosids
Nyctaginaceae_Boerhavia_burbidgeana_H
Nyctaginaceae_Boerhavia_coccinea_H
Nyctaginaceae_Anulocaulis_leiosolenus_H
Nyctaginaceae_Allionia_incarnata_H
Nyctaginaceae_Allionia_incarnata2_H
Nyctaginaceae_Mirabilis_jalapa_H
Nyctaginaceae_Bougainvillea_spectabilis_W
Nyctaginaceae_Bougainvillea_stipitata_W
Nyctaginaceae_Guapira_obtusata_W
Nyctaginaceae_Pisonia_aculeata_W
Rivinaceae_Rivina_humilis_H
Rivinaceae_Hilleria_latifolia_H
Rivinaceae_Petiveria_alliacea_H
Rivinaceae_Seguieria_aculeata_W
Phytolaccaceae_Phytolacca_bogotensis_W
Phytolaccaceae_Phytolacca_americana_H
Sarcobataceae_Sarcobatus_vermiculatus_W
Aizoaceae_Sesuvium_verrucosum_H
Aizoaceae_Sesuvium_portulacastrum_H
Aizoaceae_Cypselea_humifusa_H
Aizoaceae_Zaleya_pentandra_H
Aizoaceae_Trianthema_portulacastrum_H
Aizoaceae_Delosperma_echinatum_H
Portulacaceae_Portulaca_suffrutescens_H
Portulacaceae_Portulaca_pilosa_H
Portulacaceae_Portulaca_grandiflora_H
Portulacaceae_Portulaca_umbraticola2_H
Portulacaceae_Portulaca_umbraticola_H
Portulacaceae_Portulaca_cryptopetala_H
Portulacaceae_Portulaca_molokiniensis_H
Portulacaceae_Portulaca_oleracea_H
Cactaceae_Lophophora_williamsii_W
Cactaceae_Pereskia_aculeata_W
Talinaceae_Talinum_sp_H
Basellaceae_Basella_alba_H
Molluginaceae_Mollugo_cerviana_H
Molluginaceae_Mollugo_nudicaulis_H
Molluginaceae_Mollugo_verticillata_H
Molluginaceae_Mollugo_pentaphylla_H
Amaranthaceae_Alternanthera_sessilis_H
Amaranthaceae_Alternanthera_brasiliana_H
Amaranthaceae_Alternanthera_caracasana_H
Amaranthaceae_Alternanthera_tenella_H
Amaranthaceae_Blutaparon_vermiculare_H
Amaranthaceae_Aerva_javanica_H
Amaranthaceae_Aerva_lanata_H
Amaranthaceae_Amaranthus_cruentus_H
Amaranthaceae_Amaranthus_retroflexus_H
Amaranthaceae_Atriplex_prostrata_H
Amaranthaceae_Atriplex_hortensis_H
Amaranthaceae_Atriplex_rosea_H
Amaranthaceae_Chenopodium_quinoa_H
Amaranthaceae_Bassia_scoparia_H
Amaranthaceae_Beta_maritima_H
Amaranthaceae_Beta_vulgaris_H
Caryophyllaceae_Saponaria_officinalis_H
Caryophyllaceae_Dianthus_caryophyllus_H
Caryophyllaceae_Schiedea_membranacea_H
Caryophyllaceae_Silene_latifolia_H
Caryophyllaceae_Spergularia_media_H
Caryophyllaceae_Polycarpaea_repens_H
Microteaceae_Microtea_debilis_H
Physenaceae_Physena_madagascariensis_W
Simmondsiaceae_Simmondsia_chinensis_W
Plumbaginaceae_Limonium_spectabile_H
Plumbaginaceae_Plumbago_auriculata_W
Polygonaceae_Fallopia_convolvulus_H
Frankeniaceae_Frankenia_laevis_H
Nepenthaceae_Nepenthes_alata_W
Solanaceae_Solanum_lycopersicum_H
Solanaceae_Solanum_tuberosum_H
Phrymaceae_Mimulus_guttatu_H
Brassicaceae_Arabidopsis_thaliana_H
Brassicaceae_Arabidopsis_lyrata_H
Brassicaceae_Capsella_rubella_H
Brassicaceae_Brassica_rapa_H
Brassicaceae_Thellungiella_halophila_H
Caricaceae_Carica_papaya_W
Malvaceae_Theobroma_cacao_W
Malvaceae_Gossypium_raimondii_W
Rutaceae_Citrus_clementina_W
Rutaceae_Citrus_sinensis_W
Euphorbiaceae_Ricinus_communis_H
Euphorbiaceae_Manihot_esculenta_W
Salicaceae_Populus_trichocarpa_W
Linaceae_Linum_usitatissimum_H
Fabaceae_Phaseolus_vulgaris_H
Fabaceae_Glycine_max_H
Fabaceae_Medicago_truncatula_H
Cucurbitaceae_Cucumis_sativus_H
Rosaceae_Prunus_persica_W
Rosaceae_Malus_domestica_W
Rosaceae_Fragaria_vesca_H
Myrtaceae_Eucalyptus_grandis_W
Vitaceae_Vitis_vinifera_W
A
B
C
D
E
0.1



Pease et al. 2016 *PLoS Biol.*

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 (na "ive):

- 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