

MCB 5472

Gene Families, Super Trees and Super Matrices

Peter Gogarten

Office: *BSP 404*

phone: *860 486-4061,*

Email:

gogarten@uconn.edu

Automated Assembly of Gene Families Using BranchClust

J. Peter Gogarten

University of Connecticut
Dept. of Molecular and Cell Biol.

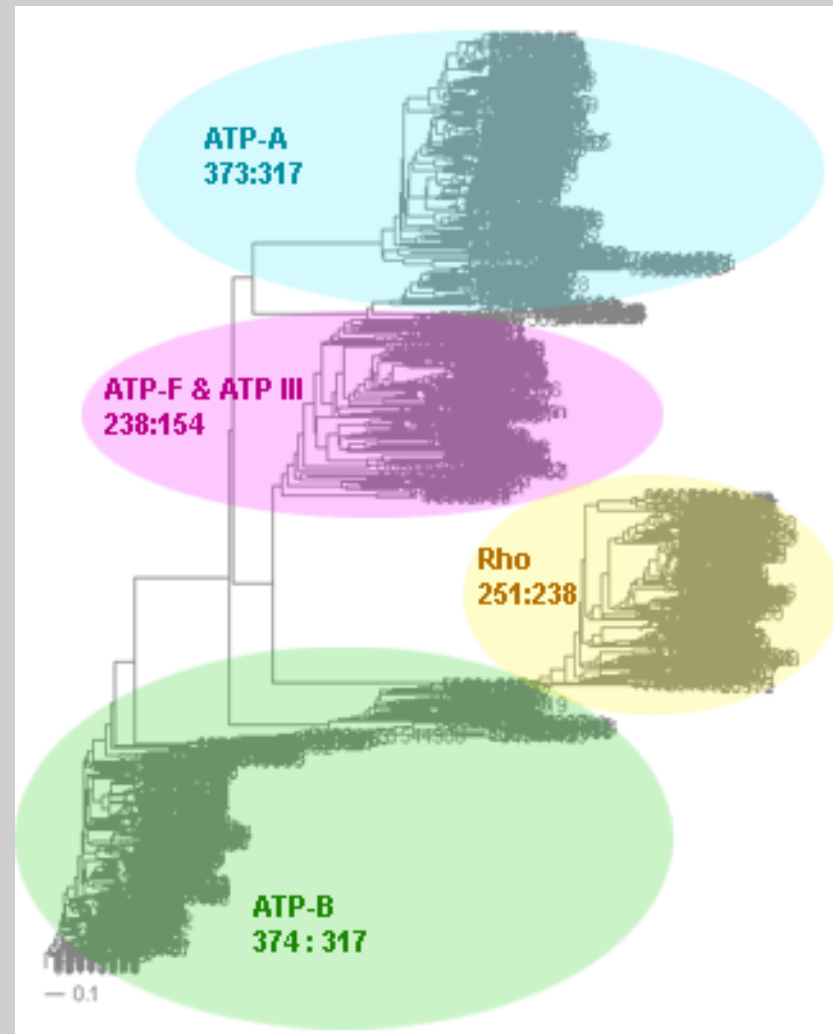
Collaborators:

Maria Poptsova (UConn)

Fenglou Mao (UGA)

Funded through the
Edmond J. Safrá Bioinformatics Program,
Fulbright Fellowship,
NASA Exobiology Program,
NSF Assembling the Tree of Life Programm and
NASA Applied Information Systems Research Program

Workshop at Te Aviv University, November 29th, 2009.



to use other genomes:

- The easiest source for other genomes is via anonymous ftp from <ftp.ncbi.nlm.nih.gov>
Genomes are in the subfolder genomes.
Bacterial and Archaeal genomes are in the subfolder Bacteria
- For use with BranchClust you want to retrieve the .faa files from the folders of the individual organisms (in case there are multiple .faa files, download them all and copy them into a single file).
- Copy the genomes into the fasta folder in directory where the branchclust scripts are.
- To create a table that links GI numbers to genomes run `perl extract_gi_numbers.pl` or `qsub extract_gi_numbers.sh`

If you use other genomes you will need to generate a file that contains assignments between name of the ORF and the name of the genome. This file should be called `gi_numbers.out`

If your genomes follow the JGI convention, every ORF starts with a four letters designating the species followed by 4 numbers identifying the particular ORF. In this case the file `gi_numbers.out` should look as follows. It should be straight forward to create this file by hand 😊

```
Thermotoga maritima | Tmar.....  
Thermotoga naphthophila | Tnap.....  
Thermotoga neapolitana | Tnea.....  
Thermotoga petrophila | Tpet.....  
Thermotoga sp. RQ2 | TRQ2.....
```

If your genomes conform to the NCBI *.faa convention, put the genomes into a subdirectory called fasta, and run the script extract_gi_numbers.pl in the parent directory. (Best is probably ~/workshop/test.)

The script should generate a log file and an output file called gi_numbers.out

```
Burkholderia phage Bcep781 |      2375.....      4783.....      1179.....
Enterobacteria phage K1F | 7711.....
Enterobacteria phage N4 | 1199.....
Enterobacteria phage P22 | 5123.....      9635...      1271.....
      193433..
Enterobacteria phage RB43 |      6639.....
Enterobacteria phage T1 | 4568.....
Enterobacteria phage T3 | 1757.....
Enterobacteria phage T5 | 4640.....
Enterobacteria phage T7 | 9627...
Kluyvera phage Kvp1 |      2126.....
Lactobacillus phage phiAT3 |      4869.....
Lactobacillus prophage Lj965 |      4117.....
Lactococcus phage r1t |      2345.....
Lactococcus phage sk1 |      9629...      193434..
Mycobacterium phage Bxz2 | 29566...
```

IF YOU USE GENOMES WITH NCBI ANNOTATION LINES, YOU NEED TO USE THE SCRIPTS CALLED BY `do_all_GI.sh` !!

(Sorry, in its present form this version does not allow to filter the E-values in the parsing of the blast searches. This means that you need to select a reasonable E-value in your initial blast searches.

If you want to use an E-value cut-off of 10^{-20} , you need to edit the `do_blast.pl` script!

If you use the JGI format, you can use the `parse_blast_cutoff_Thermotoga.pl` script to change the E-value, i.e, you don't have to re-run all of the blast searches.).

Create super families, alignments and trees

```
vi do_blast.pl
# to see what the parameters are doing type blastall or
# blastall | more at the commandline.
# If you move this to a different computer you might need to change a 2 to
a 1
```

```
vi parse_blast_cutoff_thermotoga.pl
# change bioperl directory; change cutoff E-value
# the script as written uses the bioperl library in my home directory
# Note: if using closely related genomes, you can cut back on the
# size of the superfamilies by using a smaller E-value
# (if you genomes have normal GI numbers, use
# vi parse_blast_cutoff1.pl)
```

```
# check output:
more parsed/all_vs_all.parsed #### type q to leave more
more parsed/all_vs_all.parsed | wc -l
# checks for number of lines=super families output
```

Super Families to Trees

- `perl parse_superfamilies_singlelink.pl 1`
1 gives the minimum size of the superfamily
- `perl prepare_fa_thermotoga.pl parsed/
all_vs_all.fam`
Creates a multiple fasta file for each superfamily
- `perl do_clustalw_aln.pl`
aligns sequences using clustalw
- `perl do_clustalw_dist_kimura.pl`
calculates trees using Kimura distances for all families in fa
#trees stored in trees Check #1, 106, 1027, 111
- `perl prepare_trees.pl`
reformats trees

Branchclust

```
perl branchclust_all_thermotoga.pl 2  
# Parameter 2 (MANY) says that a family needs to have  
# at least 2 members.
```

```
make_clusterlist.sh  
# runs perl make_fam_list_inpar.pl 5 4 0  
# results in test called families_inpar_5_4_0.list  
# 5: number of genomes;  
# 4: number of genomes in cluster ;  
# 0: number of inparalogs  
# (a 1 returns all the families with exactly 1 inparalog)  
# you could add additional lines to the shell script:  
# perl make_fam_list_inpar.pl 5 4 1
```

Process Branchclust output

```
perl names_for_cluster_all.pl
```

```
# (Parses clusters and attaches names.
```

```
# Results in sub directory clusters. List in test)
```

```
perl summary.pl
```

```
# (makes list of number of complete and incomplete families
```

```
# file is stored in test)
```

```
perl detailed_summary_dashes.pl
```

```
# (result in test: detailed_summary.out - can be used in Excel)
```

```
perl prepare_bcfam_thermotoga.pl families_inpar_5_4_0.list #
```

```
(writes multiple fasta files into bcfam subdirectory.
```

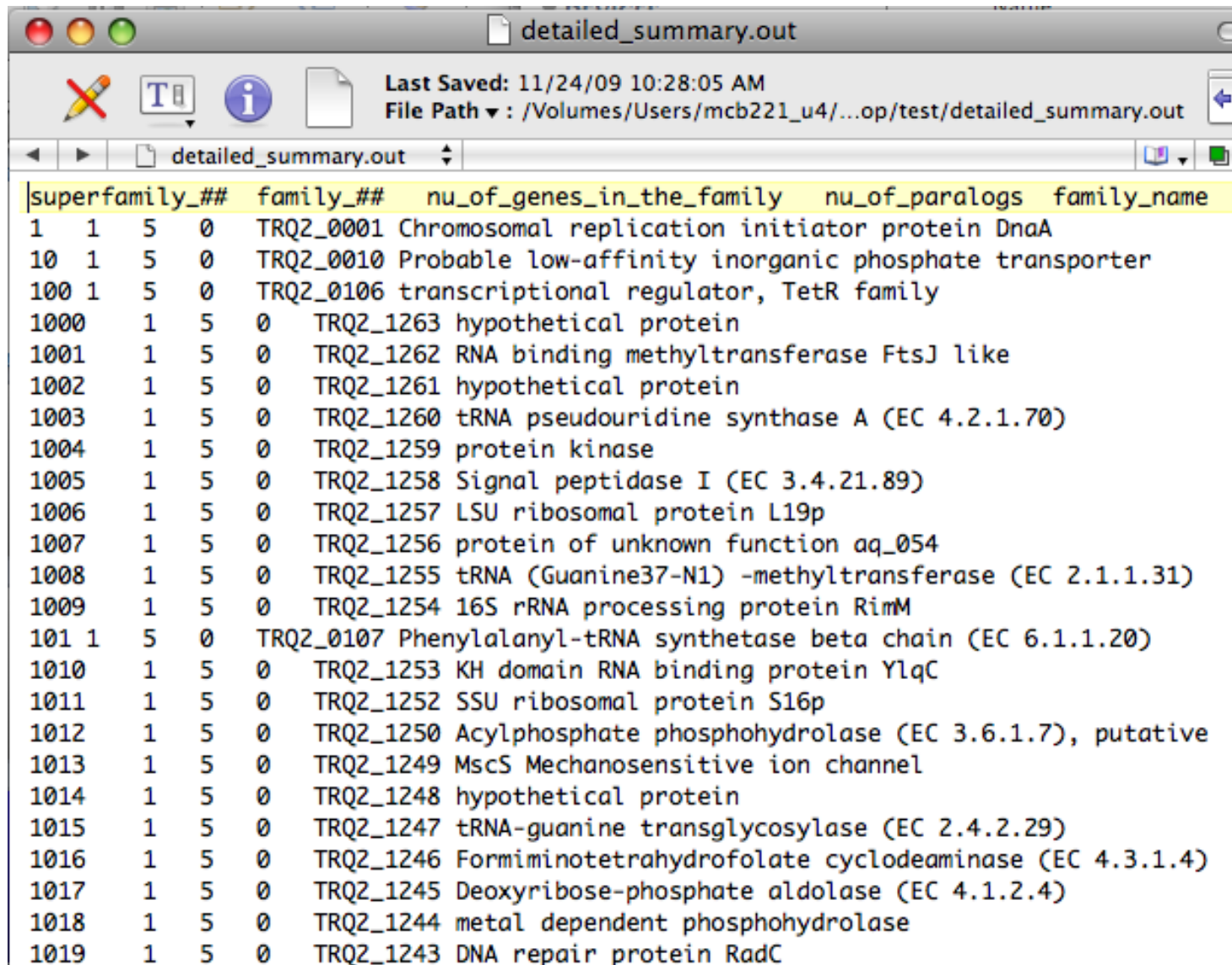
```
# Can be used for alignment and phylogenetic reconstruction)
```

Summary Output

done with many = 3 and
E-value cut-off of 10^{-25}

- complete: 1564
- incomplete: 248
- total: 1812
- ----- details -----
- incomplete 4: 87
- incomplete 3: 53
- incomplete 2: 66
- incomplete 1: 42

Detailed Summary in Text Wrangler



The screenshot shows a window titled "detailed_summary.out" in a text editor. The window includes a menu bar with icons for editing, text, information, and file operations. The status bar at the top indicates "Last Saved: 11/24/09 10:28:05 AM" and "File Path: /Volumes/Users/mcb221_u4/...op/test/detailed_summary.out". The main content area displays a table with the following data:

superfamily_##	family_##	nu_of_genes_in_the_family	nu_of_paralogs	family_name
1	1	5	0	TRQ2_0001 Chromosomal replication initiator protein DnaA
10	1	5	0	TRQ2_0010 Probable low-affinity inorganic phosphate transporter
100	1	5	0	TRQ2_0106 transcriptional regulator, TetR family
1000	1	5	0	TRQ2_1263 hypothetical protein
1001	1	5	0	TRQ2_1262 RNA binding methyltransferase FtsJ like
1002	1	5	0	TRQ2_1261 hypothetical protein
1003	1	5	0	TRQ2_1260 tRNA pseudouridine synthase A (EC 4.2.1.70)
1004	1	5	0	TRQ2_1259 protein kinase
1005	1	5	0	TRQ2_1258 Signal peptidase I (EC 3.4.21.89)
1006	1	5	0	TRQ2_1257 LSU ribosomal protein L19p
1007	1	5	0	TRQ2_1256 protein of unknown function aq_054
1008	1	5	0	TRQ2_1255 tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)
1009	1	5	0	TRQ2_1254 16S rRNA processing protein RimM
101	1	5	0	TRQ2_0107 Phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20)
1010	1	5	0	TRQ2_1253 KH domain RNA binding protein YlqC
1011	1	5	0	TRQ2_1252 SSU ribosomal protein S16p
1012	1	5	0	TRQ2_1250 Acylphosphate phosphohydrolase (EC 3.6.1.7), putative
1013	1	5	0	TRQ2_1249 MscS Mechanosensitive ion channel
1014	1	5	0	TRQ2_1248 hypothetical protein
1015	1	5	0	TRQ2_1247 tRNA-guanine transglycosylase (EC 2.4.2.29)
1016	1	5	0	TRQ2_1246 Formiminotetrahydrofolate cyclodeaminase (EC 4.3.1.4)
1017	1	5	0	TRQ2_1245 Deoxyribose-phosphate aldolase (EC 4.1.2.4)
1018	1	5	0	TRQ2_1244 metal dependent phosphohydrolase
1019	1	5	0	TRQ2_1243 DNA repair protein RadC

Detailed Summary in Excel

- copy detailed summary out onto your computer
- In EXEL Menu: Data -> get external data -> import text file -> in English version use defaults for other options.
- In EXEL Menu: Data -> sort -> sort by “superfamily number”-> if asked, check expand selection
- Scrolling down the list, search for a superfamily that was broken down into many families.

Do the families that were part of a superfamily have similar annotation lines?

How many of the families were complete?

Do any have inparalogs? Take note of a few super families.

superfamily_##	ily_##	he_fa mily	nu_of_p aralogs	family_name
129	51	2	0	Tnea_0520 Inositol transport system ATP-binding protein
129	52	2	0	TRQ2_1091 oligopeptide ABC transporter, ATP-binding protein
129	53	1	0	Tnea_0642 ABC transporter related
129	54	1	0	Tnap_0004 oligopeptide/dipeptide ABC transporter, ATPase subunit
129	55	5	0	TRQ2_0766 ABC transporter related
129	56	4	0	Tpet_0504 sugar ABC transporter, ATP-binding protein
129	57	5	0	TRQ2_0228 ABC transporter related
129	58	5	0	TRQ2_0461 ABC transporter related
129	59	5	0	TRQ2_0594 ABC transporter related
129	60	1	0	Tnap_0003 oligopeptide/dipeptide ABC transporter, ATPase subunit
129	61	5	0	TRQ2_1593 Phosphate transport ATP-binding protein PstB (TC 3.A.1.7.1)
129	62	1	0	Tnea_0524 ABC transporter related
130	1	5	0	TRQ2_0139 Putative preQ0 transporter
131	1	5	0	TRQ2_0140 NADPH dependent preQ0 reductase
132	1	5	0	TRQ2_0141 Phosphomethylpyrimidine kinase (EC 2.7.4.7) / Thiamin-phosphate synt

clusters/clusters_NNN.out.names

- Check a superfamily of your choice.
Within a family, are all the annotation lines uniform?
- Within this report, if there are inparalogs, one is listed as a family member, the other one as inparalog. This is an arbitrary choice, both inparalogs from the same genome should be considered as being part of the family.
- Out of cluster paralogs are paralogs that did not make it into a cluster with “many” genomes.

```
COMPLETE: 5
```

```
----- CLUSTER -----
```

```
>lclITnea_1049 ABC transporter related [Thermotoga neapolitana]  
>lclITRQ2_0990 ABC transporter related [Thermotoga sp. RQ2]  
>lclITnea_1896 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.  
>lclITmar_1872 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.  
>lclITpet_1811 ABC transporter related [Thermotoga petrophila]  
>lclITnap_1536 ABC transporter related [Thermotoga naphthophila]
```

```
----- FAMILY -----
```

```
>lclITmar_1872 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.  
>lclITnap_1536 ABC transporter related [Thermotoga naphthophila]  
>lclITnea_1049 ABC transporter related [Thermotoga neapolitana]  
>lclITpet_1811 ABC transporter related [Thermotoga petrophila]  
>lclITRQ2_0990 ABC transporter related [Thermotoga sp. RQ2]
```

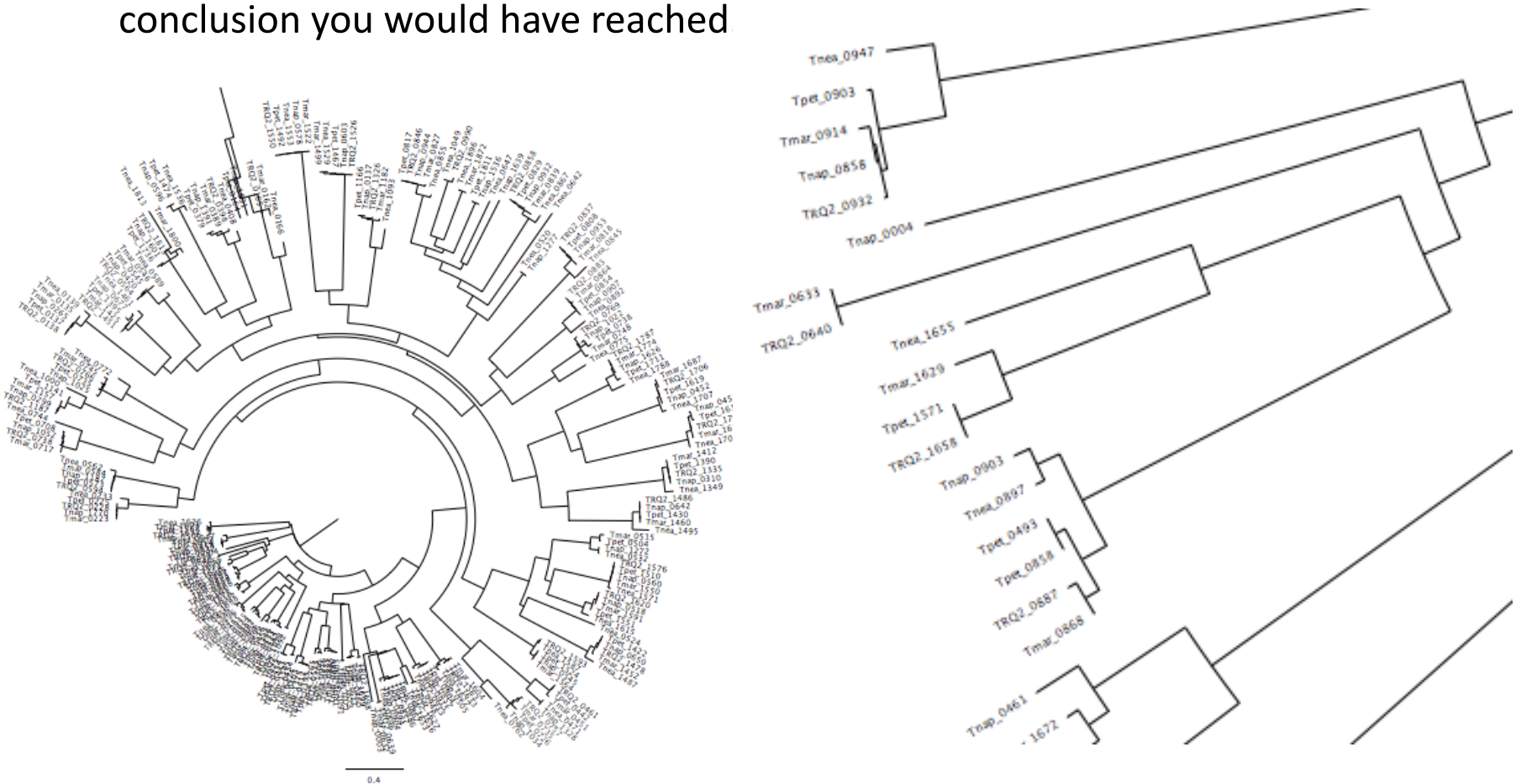
```
COMPLETE: 5
```

```
>>>> IN-PARALOGS -----
```

```
>lclITnea_1896 Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.
```

trees/fam_XYZ.tre

- Check the tree for a superfamily of your choice. Copy the file to your computer and open it in TreeView, NJPLOT, or FigTree (check with your neighbor on which program works).
- For at least one cluster, in the tree, check if branchclust came to the same conclusion you would have reached



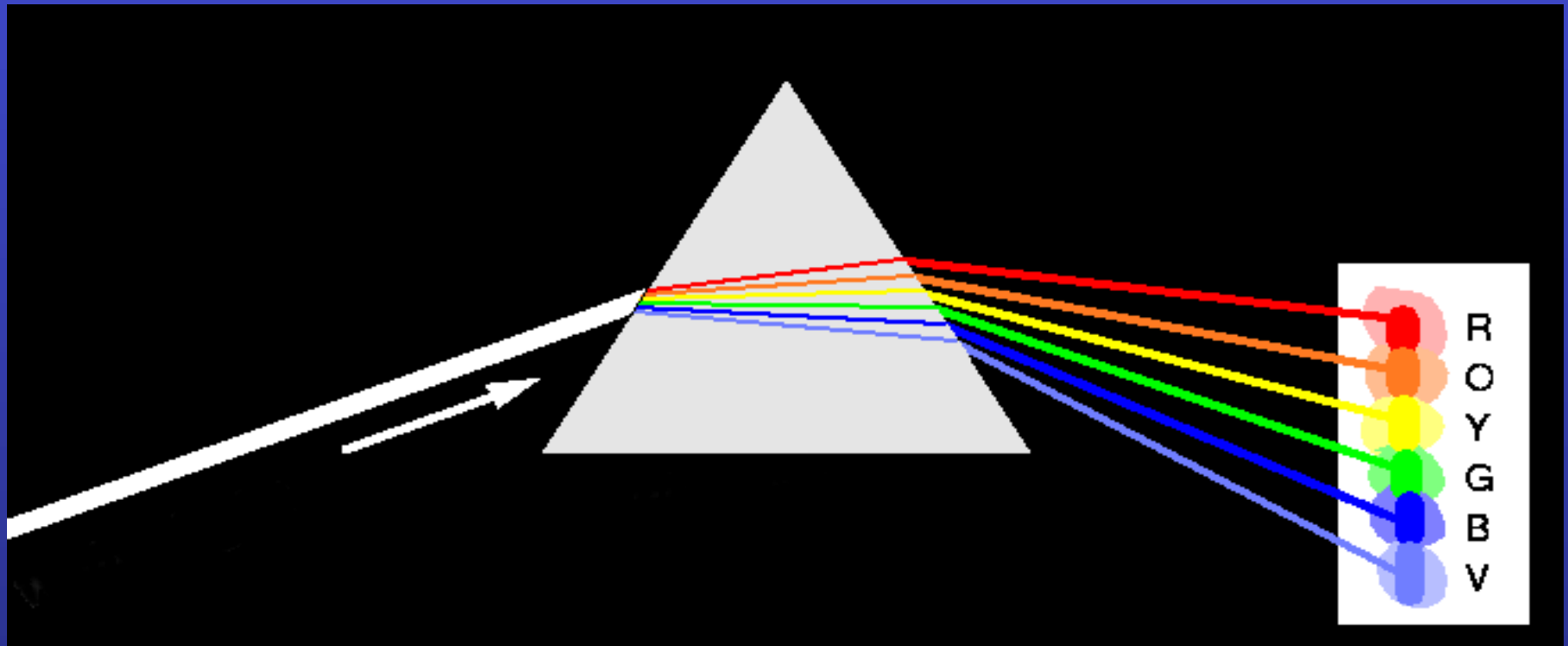

```
prepare_bcfam_thermotoga.pl  
families_inpar_5_4_0.list
```

The script `prepare_bcfam_thermotoga.pl` takes a list of families (created by `make_fam_list_inpar.pl`) and for each family retrieves the fasta sequences from the combined genome databank and stores the sequences in the BCfam folder, one multiple sequence file per family.

One possibility for further evaluation is to take multiple sequence files, align the sequences and perform a phylogenetic reconstruction (including bootstrap analysis) using programs like [phymml](#) or [Raxml](#).

The resulting trees can be analyzed by decomposition and supertree approaches.

Decomposition of Phylogenetic Data



Phylogenetic information present in genomes

Break information into small quanta of information (bipartitions or embedded quartets)

Analyze spectra to detect transferred genes and plurality consensus.

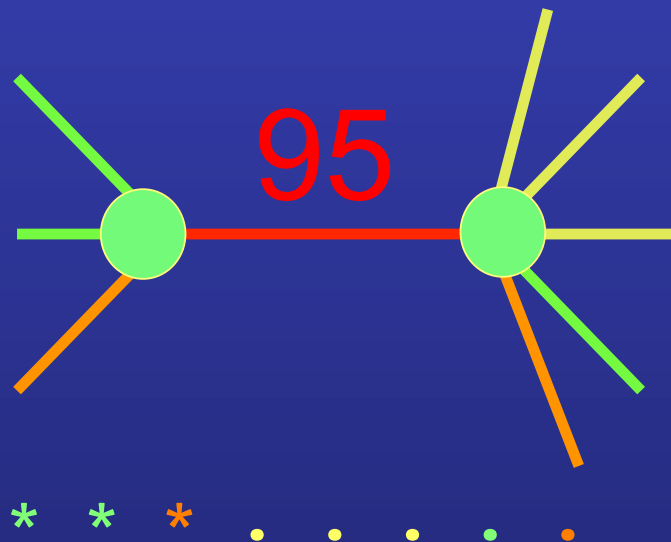
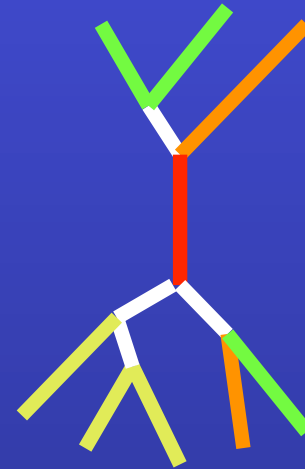


TOOLS TO ANALYZE
PHYLOGENETIC INFORMATION
FROM MULTIPLE GENES IN
GENOMES:

Bipartition Spectra (Lento Plots)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition (or split) – a division of a phylogenetic tree into two parts that are connected by a single branch. It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.



Yellow vs Rest

* * * . . . * *

compatible to illustrated
bipartition

Orange vs Rest

. . * . . . *

incompatible to illustrated
bipartition

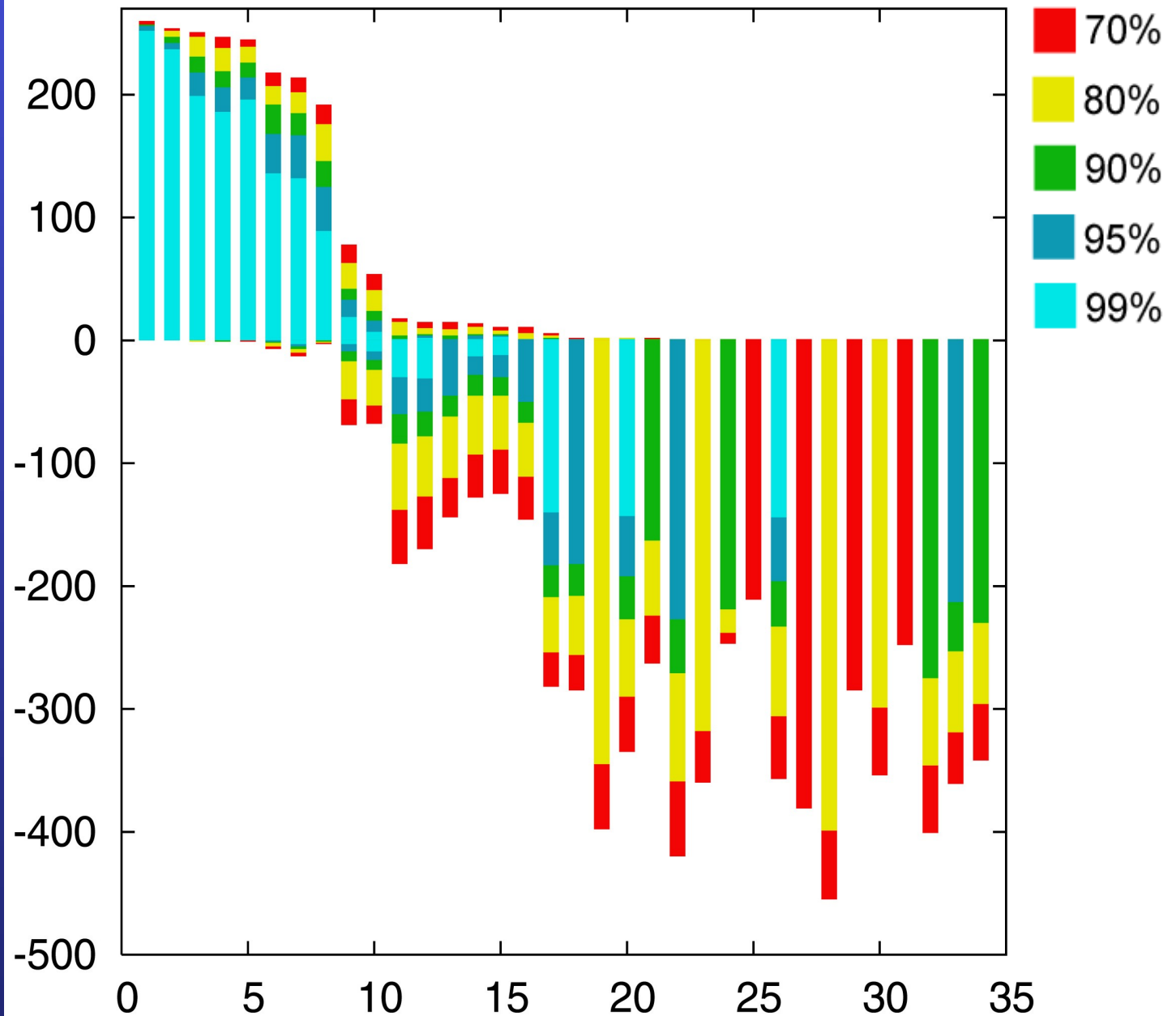
“Lento”-plot of 34 supported bipartitions (out of 4082 possible)

13 gamma-proteobacterial genomes

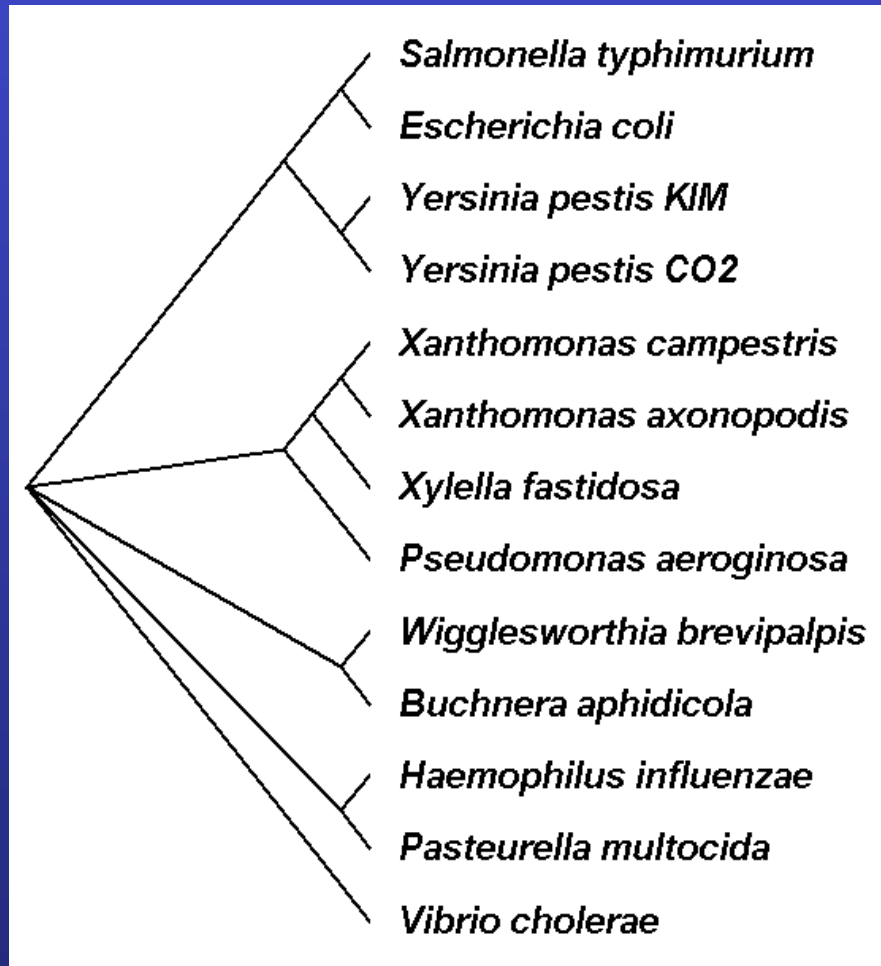
(258 putative orthologs):

- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

There are **13,749,310,575** possible unrooted tree topologies for 13 genomes

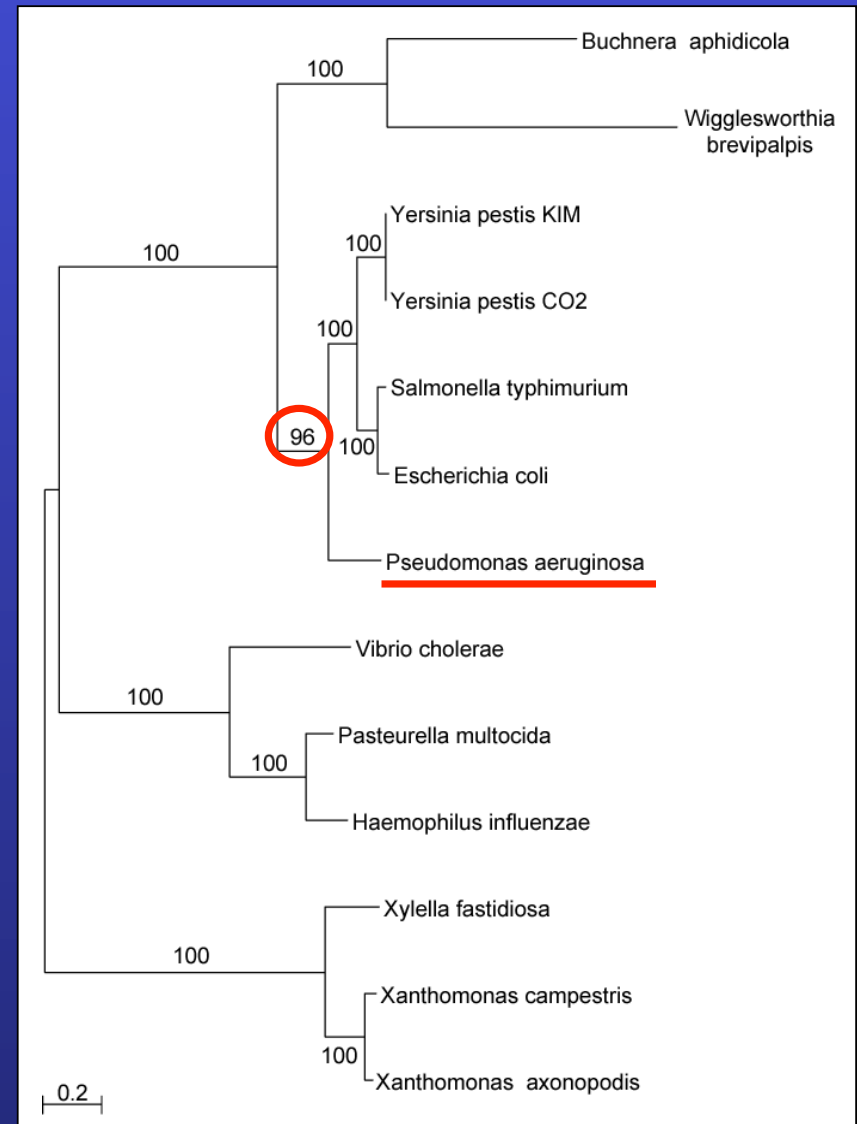


Consensus clusters of eight significantly supported bipartitions



only 258 genes analyzed

Phylogeny of putatively transferred gene (virulence factor homologs (mviN))

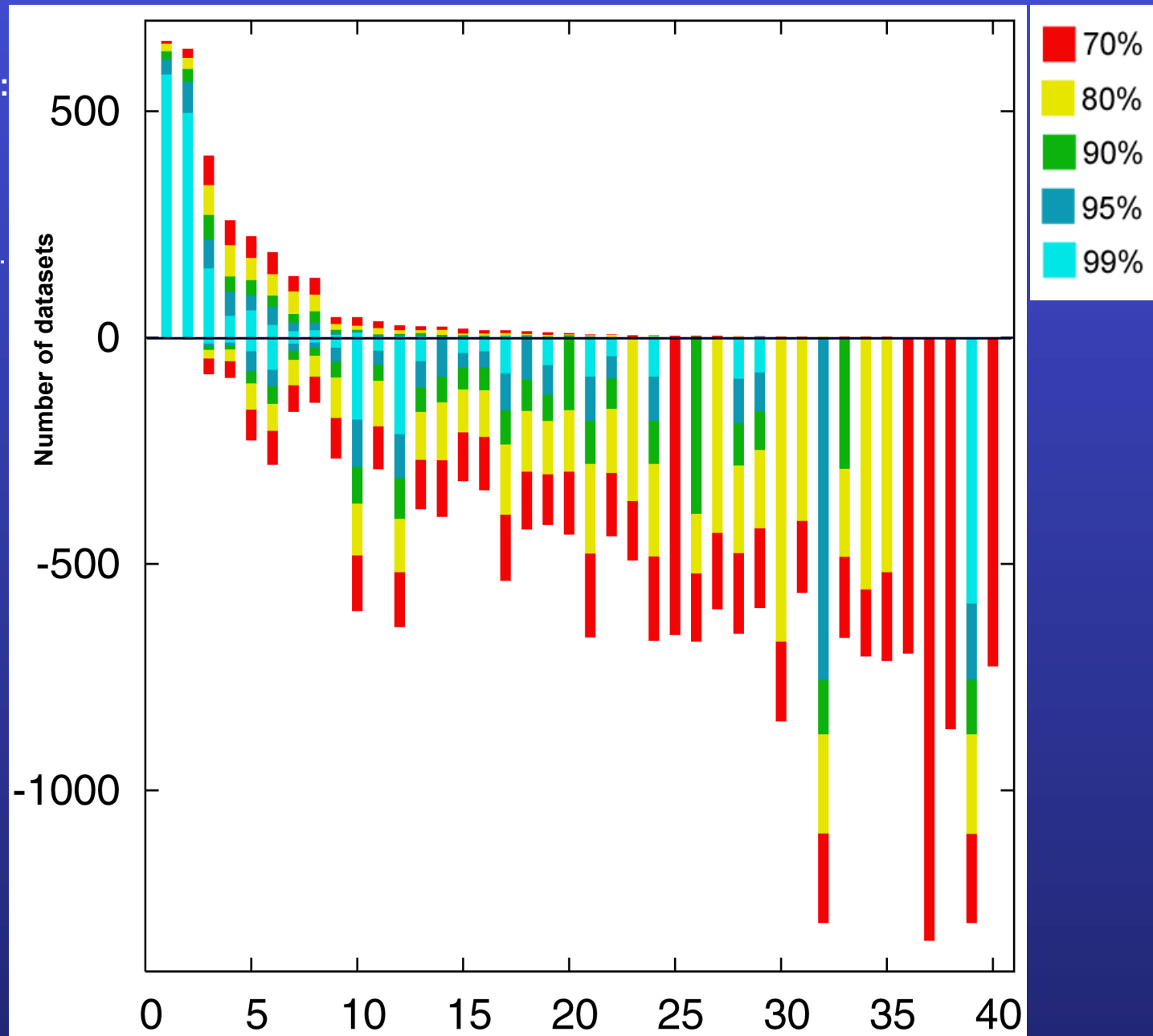


“Lento”-plot of supported bipartitions (out of 501 possible)

10 cyanobacteria:

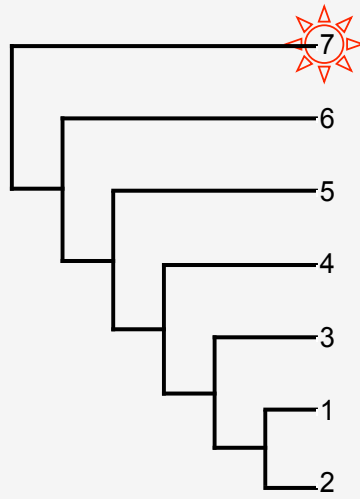
- *Anabaena*
- *Trichodesmium*
- *Synechocystis* sp.
- *Prochlorococcus marinus* (3 strains)
- Marine *Synechococcus*
- *Thermosynechococcus elongatus*
- *Gloeobacter*
- *Nostoc punctioforme*

Based on 678 sets of orthologous genes



PROBLEMS WITH BIPARTITIONS (CONT.)

- Q₁={
 4 5 6 7
 1 5 6 7
 2 5 6 7
 3 5 6 7
 3 4 6 7
 1 4 6 7
 2 4 6 7
 2 3 6 7
 1 3 6 7
 1 2 6 7
 1 2 3 7
 1 2 4 7
 1 3 4 7
 2 3 4 7
 2 3 5 7
 1 3 5 7
 1 2 5 7
 1 4 5 7
 2 4 5 7
 3 4 5 7
 3 4 5 6
 1 4 5 6
 2 4 5 6
 2 3 5 6
 1 3 5 6
 1 2 5 6
 1 2 3 6
 1 2 4 6
 1 3 4 6
 2 3 4 6
 2 3 4 5
 1 3 4 5
 1 2 4 5
 1 2 3 5
 1 2 3 4}



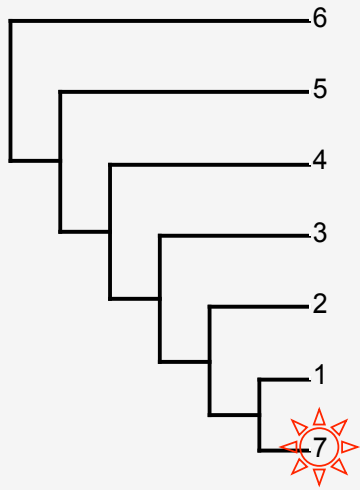
- B₁={
 **
 ***
 ****
 ***** ..
 }
 bipartitions

embedded quartets

A single rogue sequence that moves from one end of a Hennigian comb to the other changes all bipartition

supported quartets
 $Q_1 \cap Q_2 =$
 {3 4 5 6, 1 4 5 6, 2 4 5 6, 2 3 5 6,
 1 3 5 6, 1 2 5 6, 1 2 3 6, 1 2 4 6,
 1 3 4 6, 2 3 4 6, 2 3 4 5, 1 3 4 5,
 1 2 4 5, 1 2 3 5, 1 2 3 4}

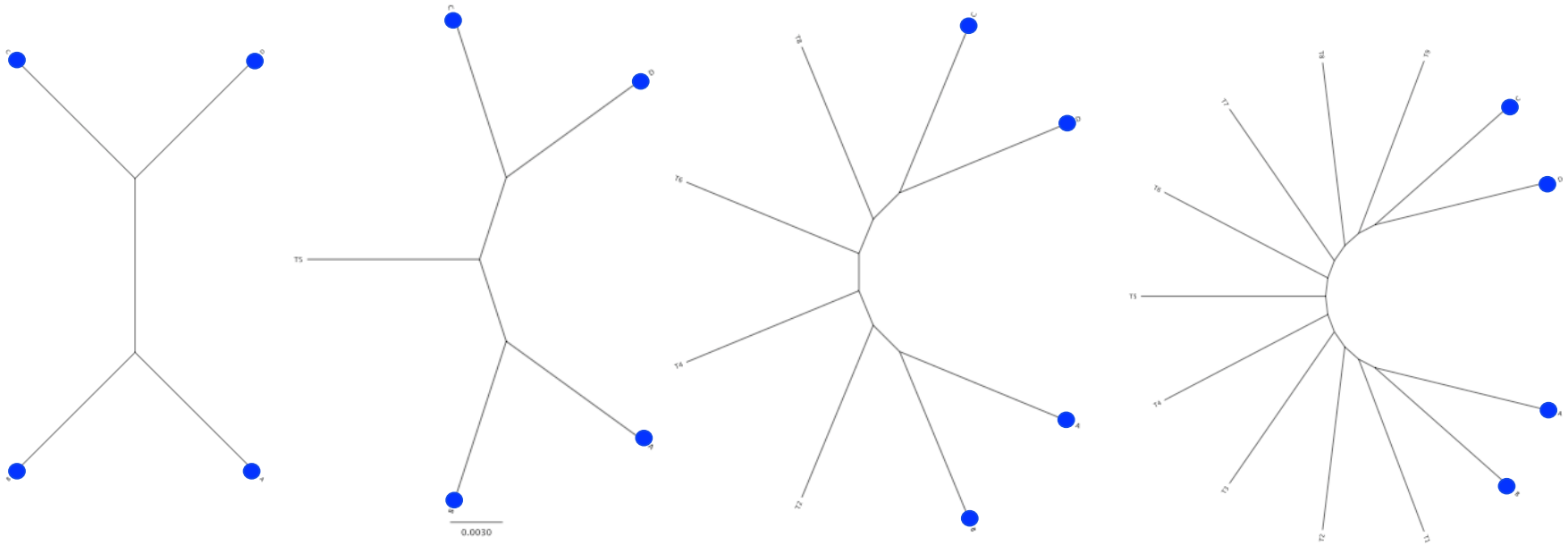
supported bipartitions:
 $B_1 \cap B_2 = \emptyset$



- B₂={
 * *
 ** *
 *** *
 **** *
 }
 bipartitions

- Q₂={
 3 4 5 6
 1 4 5 6
 7 4 5 6
 2 4 5 6
 2 3 5 6
 1 3 5 6
 7 3 5 6
 7 2 5 6
 1 2 5 6
 1 7 5 6
 1 7 2 6
 1 7 3 6
 1 2 3 6
 7 2 3 6
 7 2 4 6
 1 2 4 6
 1 7 4 6
 1 3 4 6
 7 3 4 6
 2 3 4 6
 2 3 4 5
 1 3 4 5
 7 3 4 5
 7 2 4 5
 1 2 4 5
 1 7 4 5
 1 7 2 5
 1 7 3 5
 1 2 3 5
 7 2 3 5
 7 2 3 4
 1 2 3 4
 1 7 3 4
 1 7 2 4
 1 7 2 3}

Decay of bipartition support with number of OTUs

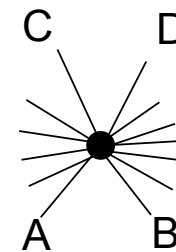
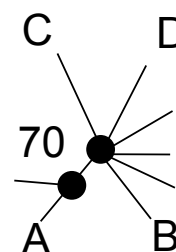
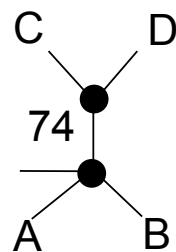
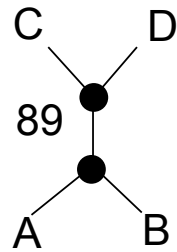


Phylogenies used for simulation

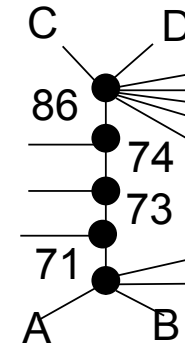
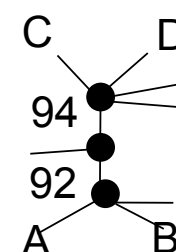
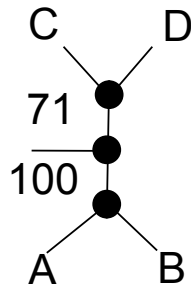
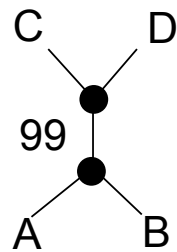
Example for decay of bipartition support with number of OTUs

Sequence lengths

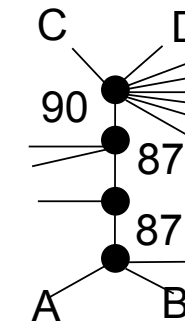
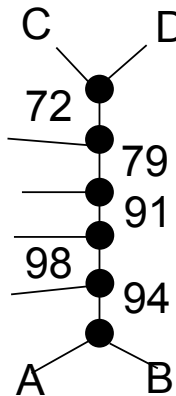
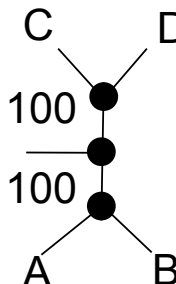
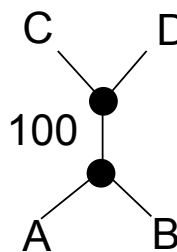
200



500

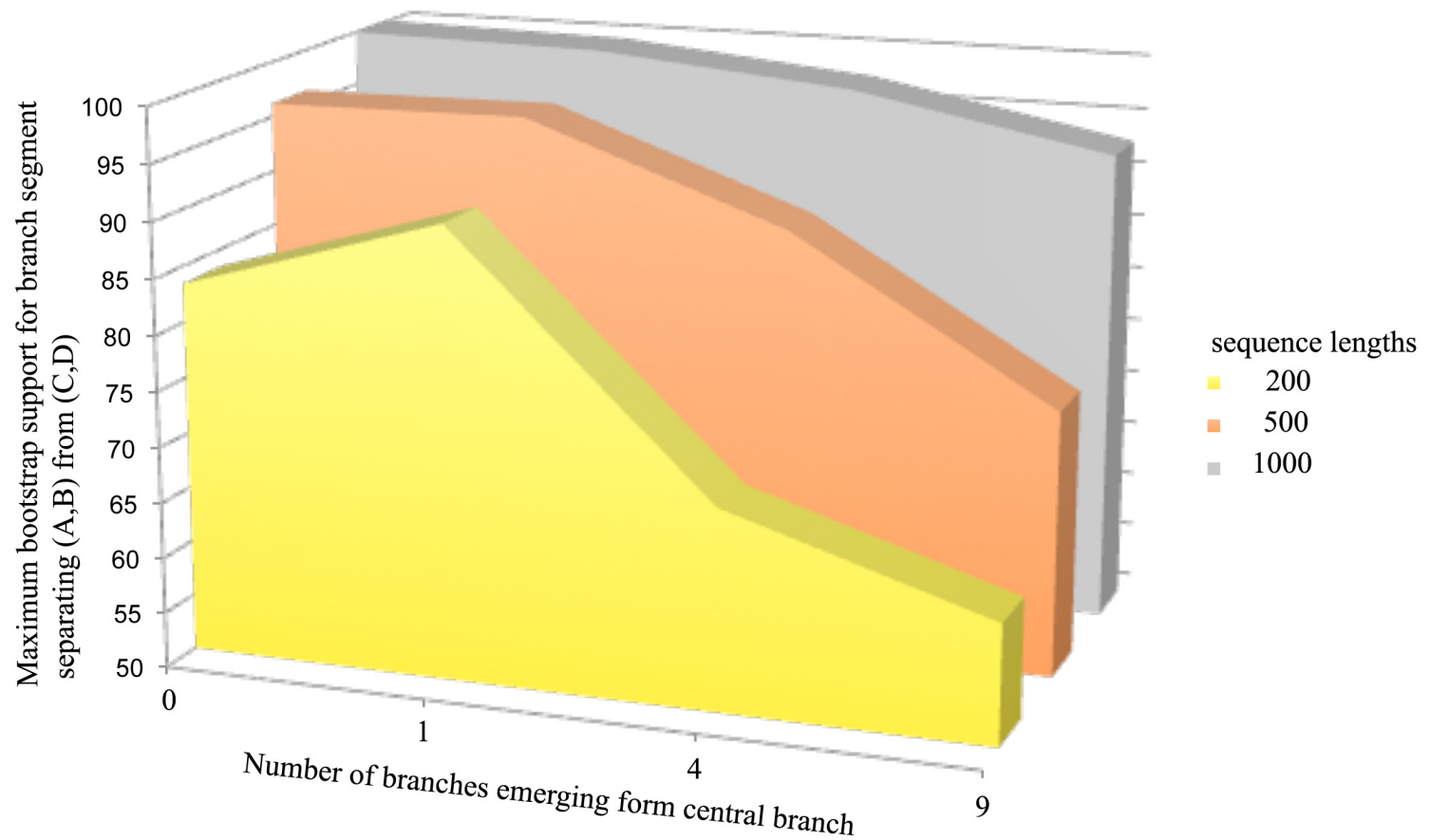


1000



Only branches with better than 70% bootstrap support are shown

Decay of bipartition support with number of OTUs



Each value is the average of 10 simulations using seq-gen.
Simulated sequences were evaluated using PHYML.
Model for simulation and evaluation WAG + $\Gamma(\alpha=1, 4 \text{ rate categories})$

Bipartition Paradox:

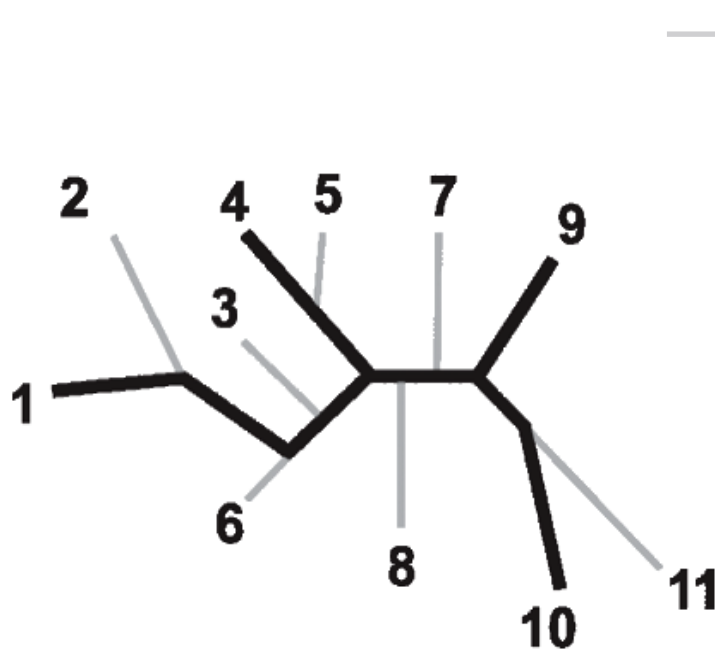
- The more sequences are added, the lower the support for bipartitions that include all sequences. The more data one uses, the lower the bootstrap support values become.
- This paradox disappears when only embedded splits for 4 sequences are considered.



TOOLS TO ANALYZE
PHYLOGENETIC INFORMATION
FROM MULTIPLE GENES IN
GENOMES:

QUARTET DECOMPOSITION

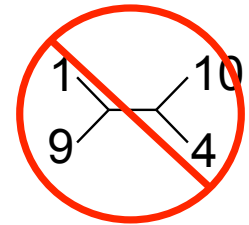
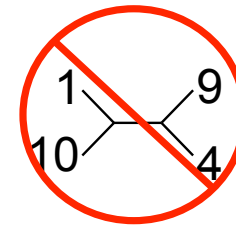
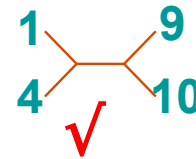
Bootstrap support values for embedded quartets



— + — : tree calculated from one pseudo-sample generated by bootstrapping from an alignment of one gene family present in 11 genomes

— : embedded quartet for genomes 1, 4, 9, and 10 .

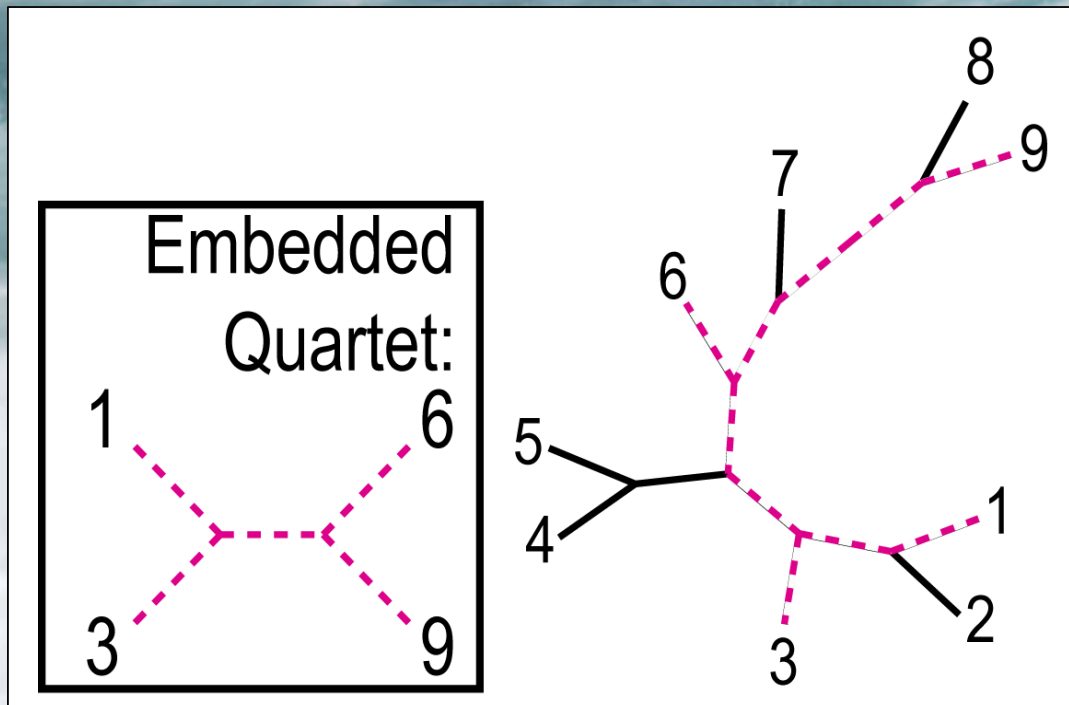
This bootstrap sample supports the topology ((1,4),9,10).



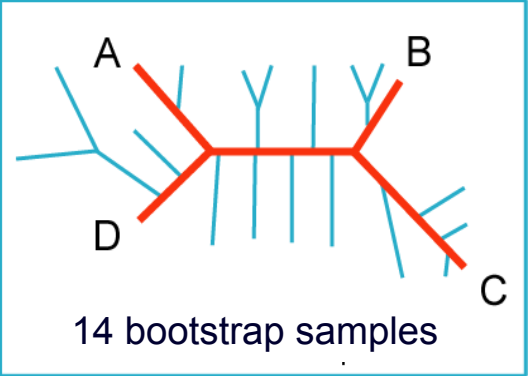
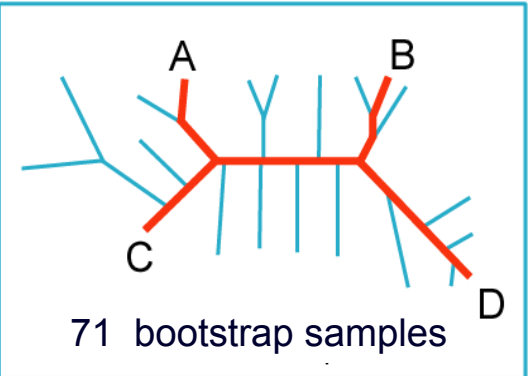
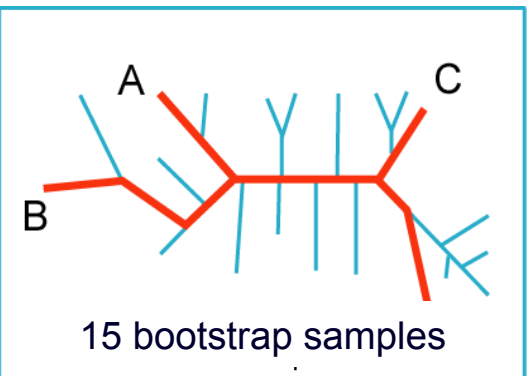
Quartet spectral analyses of genomes iterates over three loops:

- Repeat for all bootstrap samples.
- Repeat for all possible embedded quartets.
- Repeat for all gene families.

QUARTET DECOMPOSITION METHOD



- Quartet is a smallest unit of phylogenetic information
- Each quartet is associated with only three unrooted tree topologies
- Support for different quartet topologies can be summarized for all gene families



BOOTSTRAP SUPPORT VALUE VECTOR:

(15, 71, 14)

Illustration of one component of a quartet spectral analyses

analyses Summary of phylogenetic information for one genome quartet for all gene families

Total number of gene families containing the species quartet

Number of gene families supporting the same topology as the plurality (colored according to bootstrap support level)

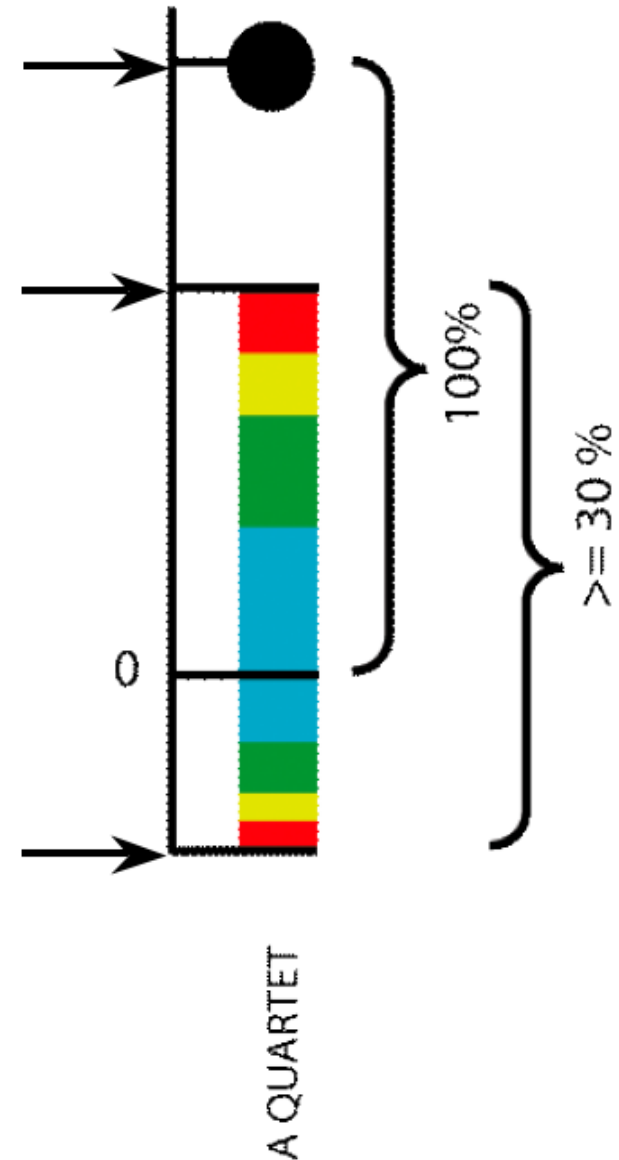
80%

90%

95%

99%

Number of gene families supporting one of the two alternative quartet topologies



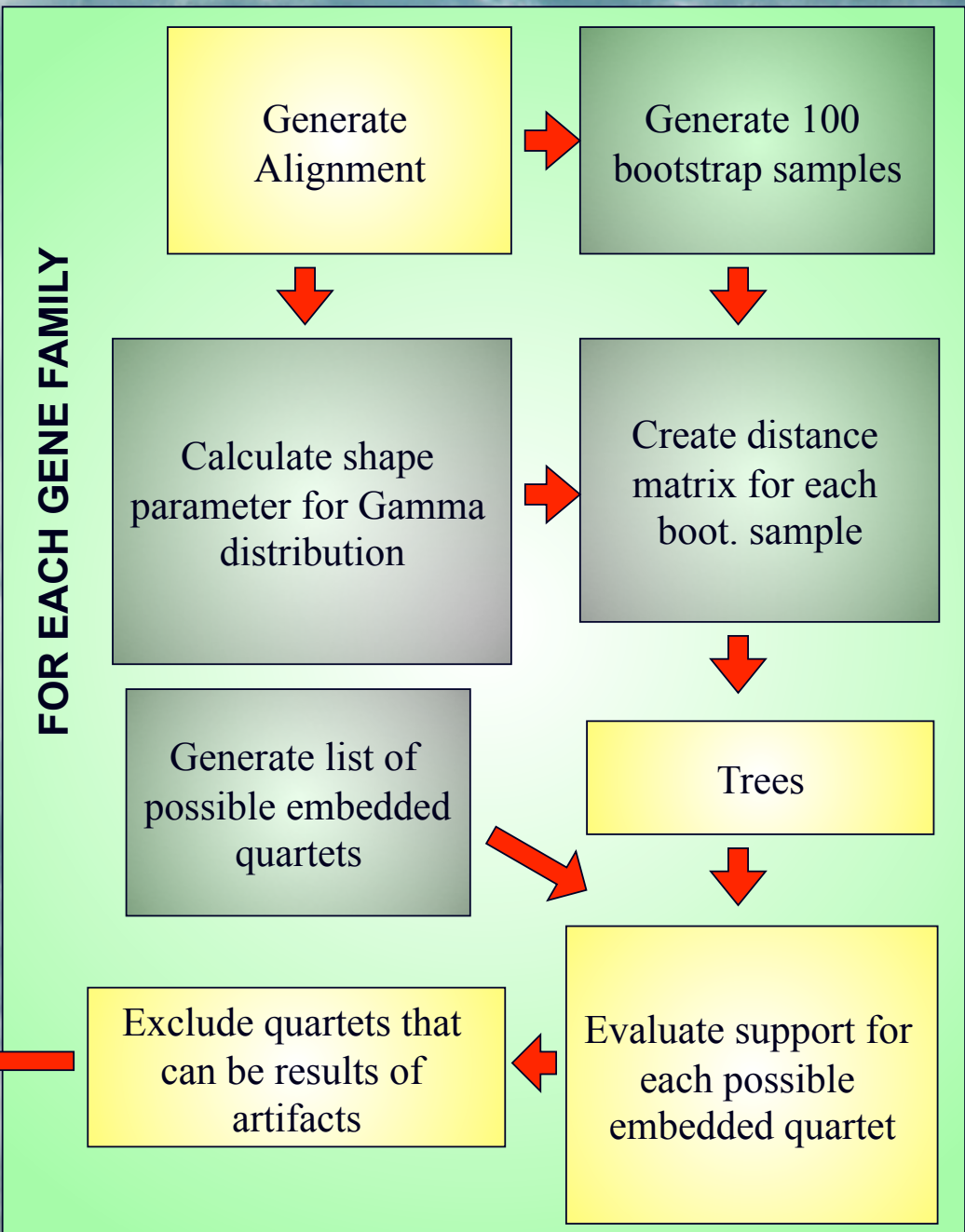
QUARTET DECOMPOSITION ANALYSES: DATA FLOW

N completely sequenced genomes
(a.a. sequences)

Detect gene families

Families with missing data are considered

Visualize support for embedded quartets



Generate Alignment

Generate 100 bootstrap samples

Calculate shape parameter for Gamma distribution

Create distance matrix for each boot. sample

Generate list of possible embedded quartets

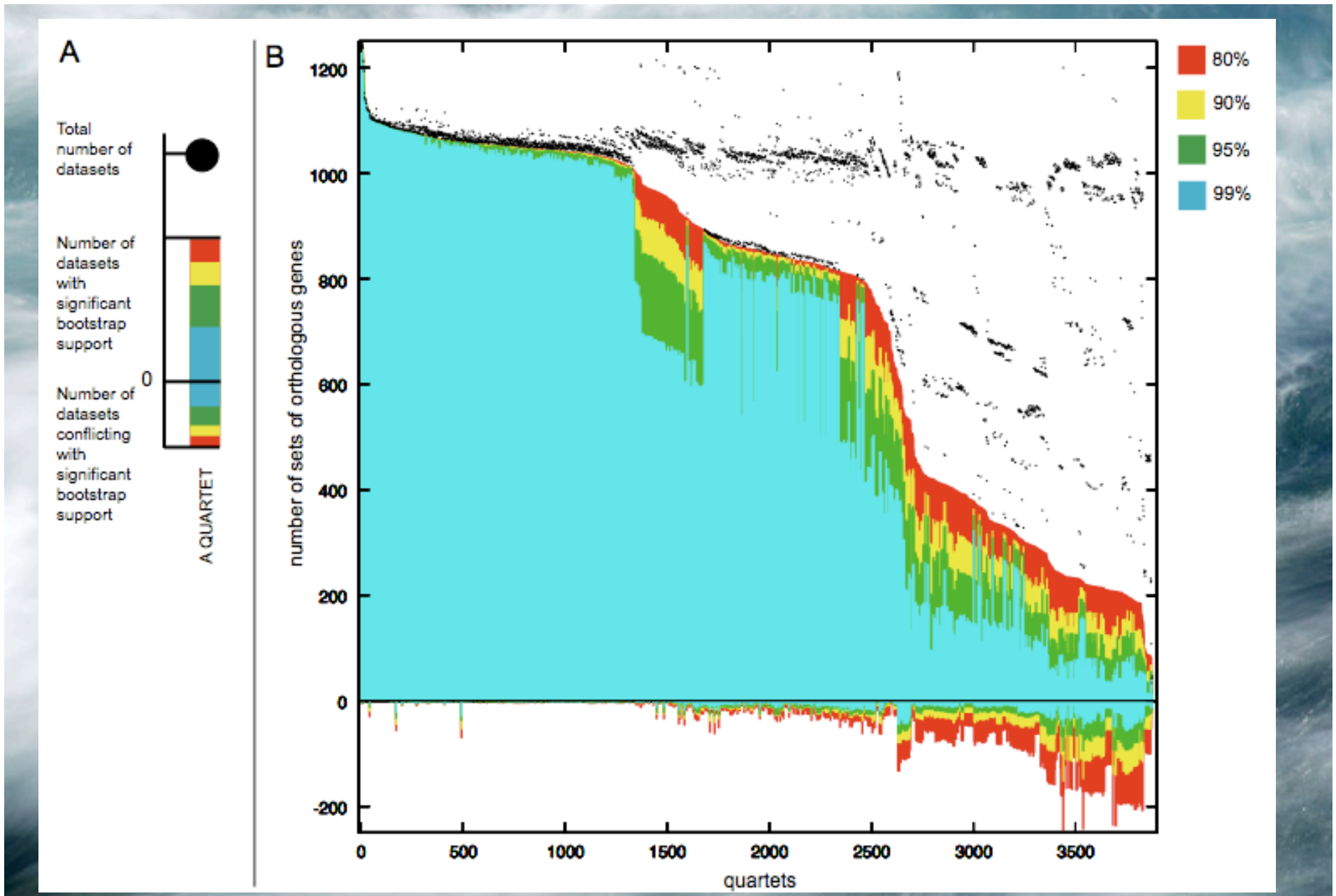
Trees

Exclude quartets that can be results of artifacts

Evaluate support for each possible embedded quartet

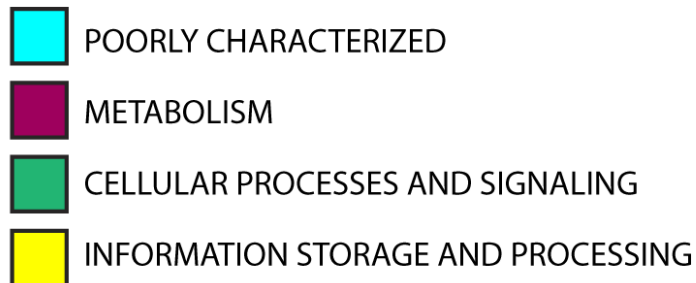
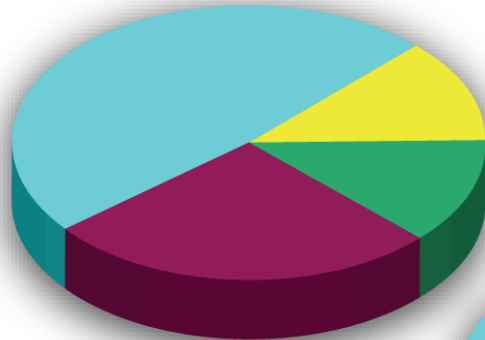
Other POSITIVE THOUGHTS ABOUT THE METHOD

1. No assumption that all genes in a genome have the same phylogenetic history.
2. The total number of quartets is much smaller than number of tree topologies, which makes it possible to evaluate all quartets.
3. Gene families present only in few analyzed genomes can be included in the analyses
4. Phylogenetic signal can be divided into consensus supported by the plurality of gene families and the conflicting signal.
5. Allows us to partition analyzed genomes according to some scenario (e.g., grouping by ecology) and retrieve gene families that support or conflict it.



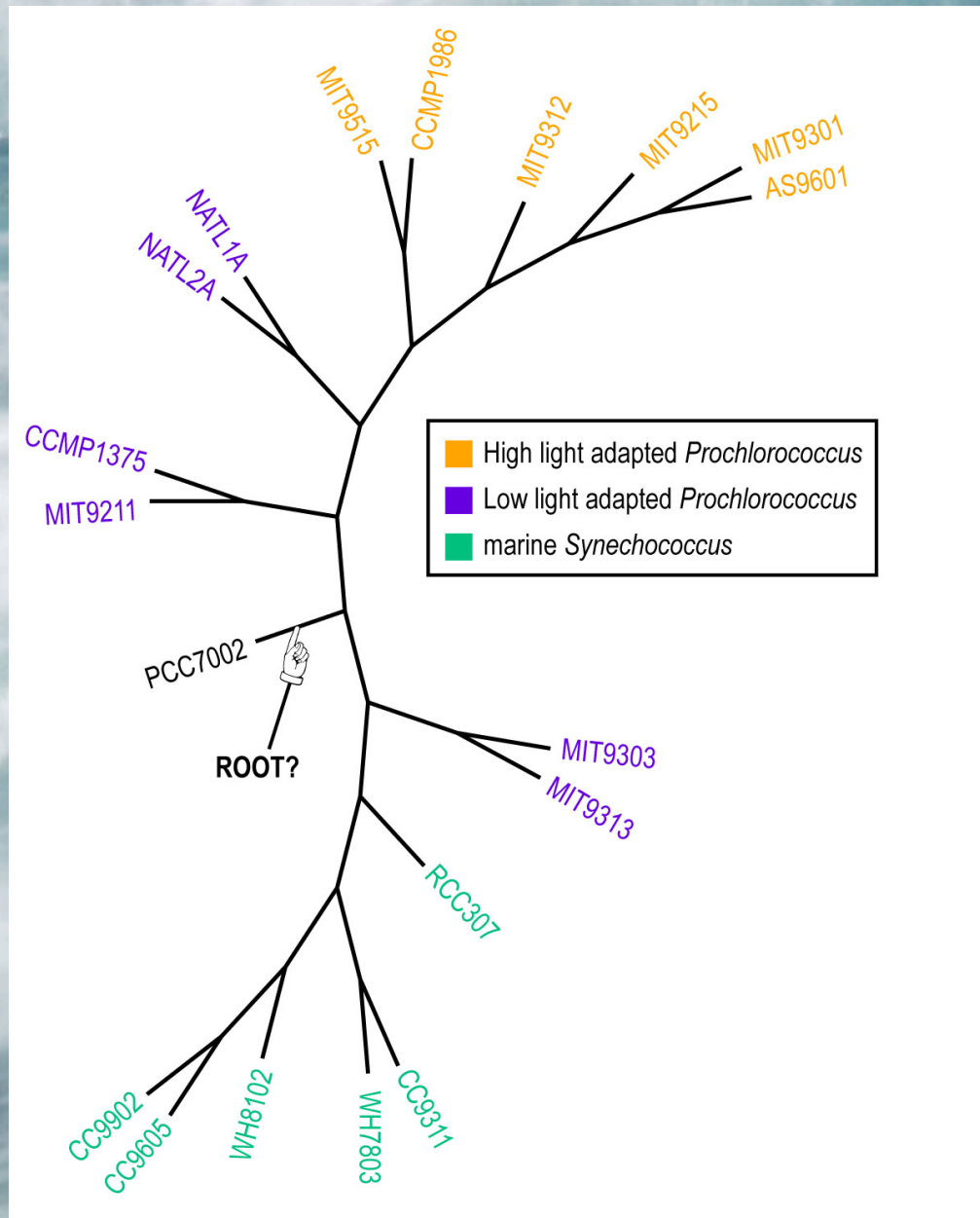
Quartet decomposition analysis of 19 *Prochlorococcus* and marine *Synechococcus* genomes. Quartets with a very short internal branch or very long external branches as well those resolved by less than 30% of gene families were excluded from the analyses to minimize artifacts of phylogenetic reconstruction.

1812
gene
families



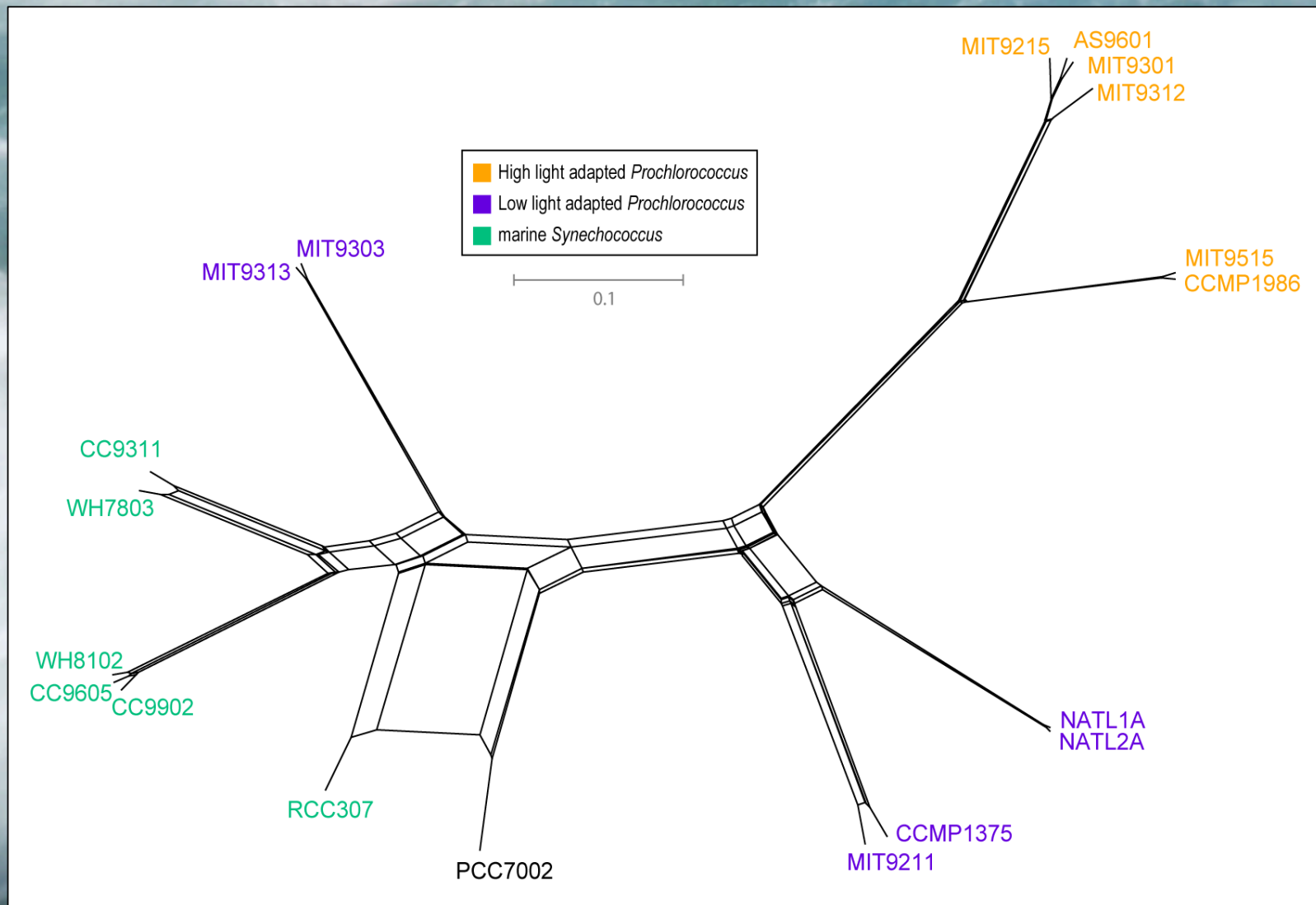
932 gene families
in conflict with
plurality

Figure 8. Distribution of gene families without paralogs across functional categories. The four super-categories are defined by COG database. Notably, genes of informational storage and processing are represented in equal proportions in genes in conflict with plurality as compared to all 1812 gene families, which contradicts complexity hypothesis {Jain, 1999 #31}. Metabolic genes appear to be overrepresented in the gene family pool which conflicts with plurality.



Plurality consensus calculated as supertree (MRP) from quartets in the plurality topology.

NeighborNet (calculated with SplitsTree 4.0)



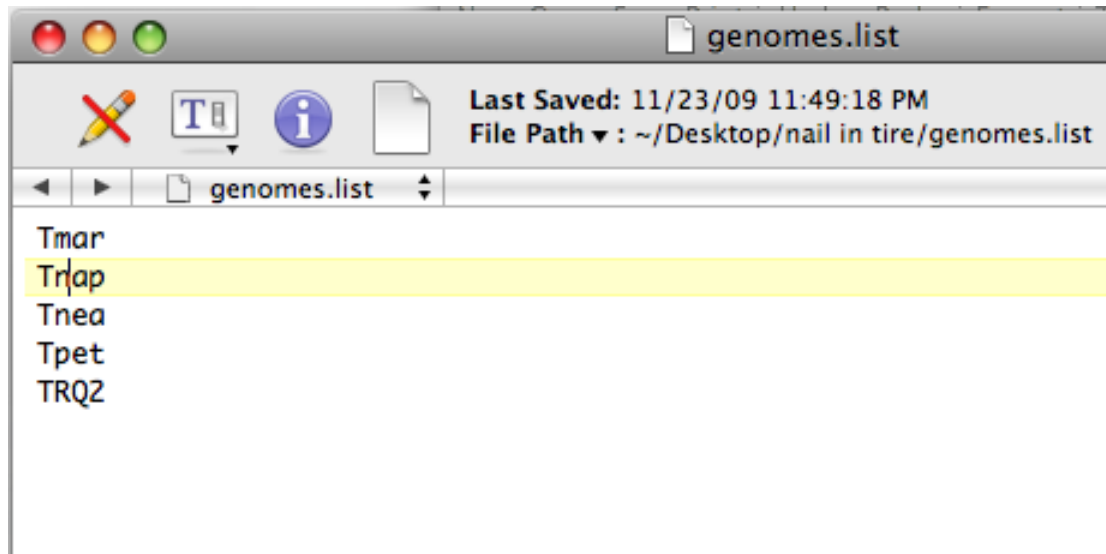
Plurality neighbor-net calculated as supertree (from the MRP matrix using SplitsTree 4.0) from all quartets significantly supported by all individual gene families (1812) without in-paralogs.

The Quartet Decomposition Server

<http://csbl1.bmb.uga.edu/QD/phytree.php>

Input A):

a file listing the names of genomes: E.g.:

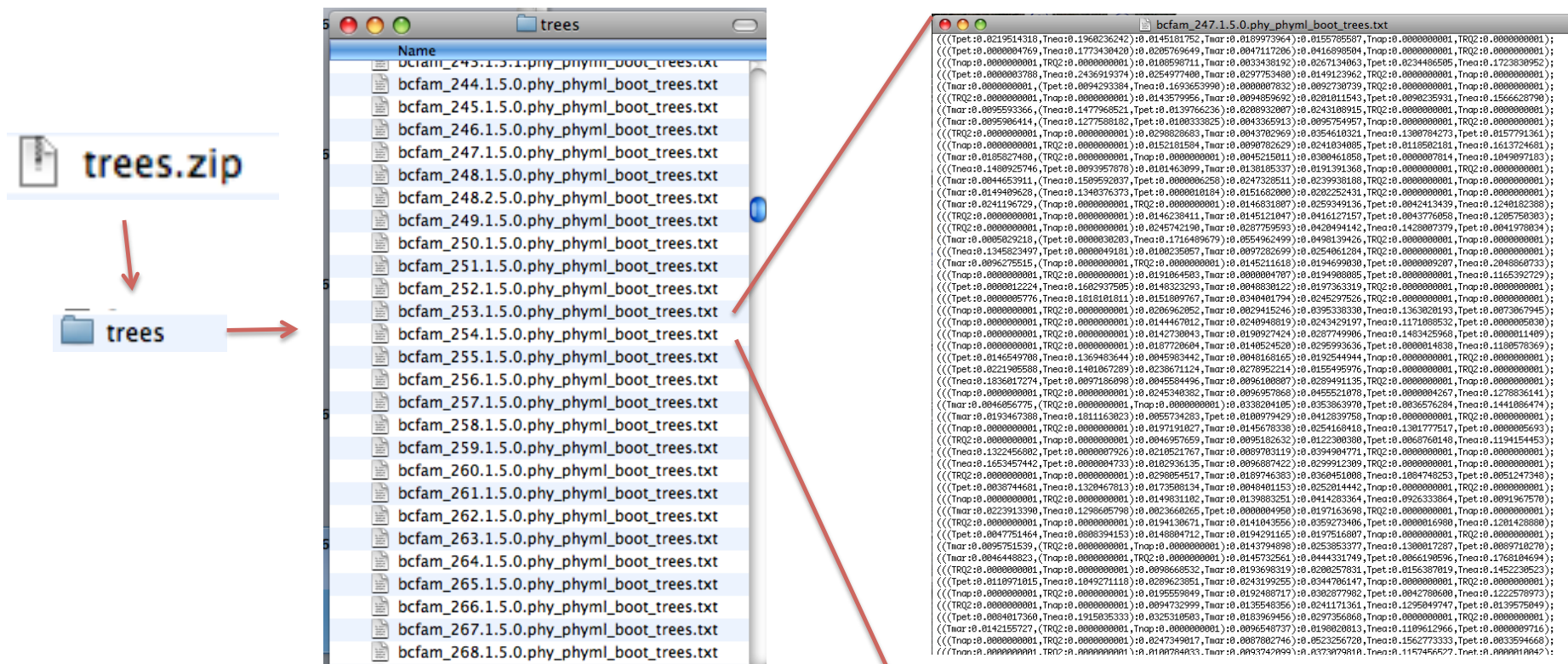


The Quartet Decomposition Server

<http://csbl1.bmb.uga.edu/QD/phytree.php>

Input B):

An Archive of files where every file contains all the trees that resulted from a bootstrap analysis of one gene family:



One file per family

100 trees per file

The Quartet Decomposition Server

<http://csbl1.bmb.uga.edu/QD/phytree.php>

Trees from the bootstrap samples should contain branch lengths, but the name for each sequence should be translated to the genome name, using the names in the genome list. See the following three trees in Newick notation for an example:

```
((Tnea:0.1559823230,Tpet:0.0072068797):  
0.0287486818,Tmar:0.0046676053):0.0407339037,Tnap:  
0.0000000001,TRQ2:0.0000000001);  
(((Tpet:0.0219514318,Tnea:0.1960236242):  
0.0145181752,Tmar:0.0189973964):0.0155785587,Tnap:  
0.0000000001,TRQ2:0.0000000001);  
(((Tpet:0.0000004769,Tnea:0.1773430420):  
0.0205769649,Tmar:0.0047117206):0.0416898504,Tnap:  
0.0000000001,TRQ2:0.0000000001);
```


The spectrum

<http://csbl1.bmb.uga.edu/QD/jobstatus.php?jobid=QDSgArf2&source=0&resolve=0&support=0>

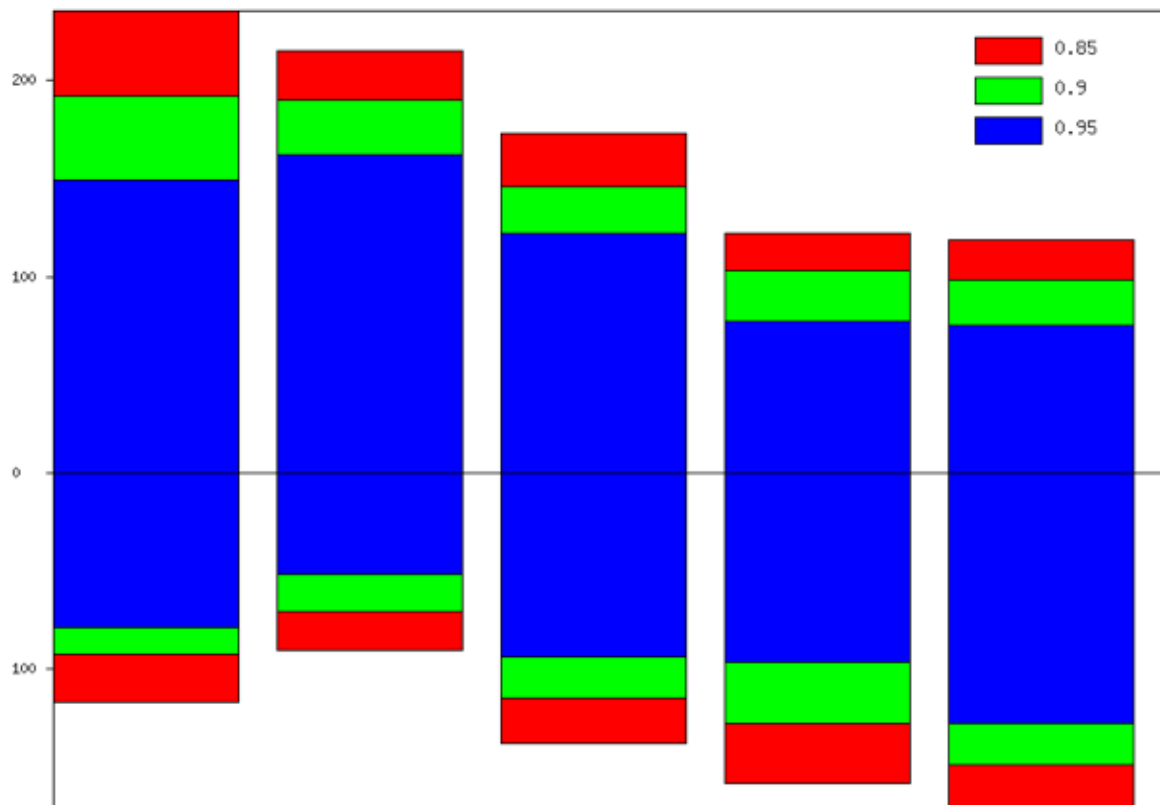
Quartet Decomposition

Quartet Decomposition Spectrum for job: **QDSgArf2**

Download quartets with at least % bootstrap support value in at least gene families

Download quartets with bootstrap support value threshold %

Remove quartets resolved in less than % gene families with at least % bootstrap support value



good and bad quartets

Quartet Decomposition

Good quartets with bootstrap support value > 0.9

[Download](#) as newick trees

Quartet ID	Gene Family Numbers	Quartet Topology
1	192	((Tmar,Tnea),(Tnap,Tpet));
4	98	((Tmar,Tnea),(Tnap,TRQ2));
8	190	((Tmar,TRQ2),(Tnap,Tpet));
9	103	((Tmar,Tnea),(Tpet,TRQ2));
13	146	((Tnap,Tpet),(Tnea,TRQ2));

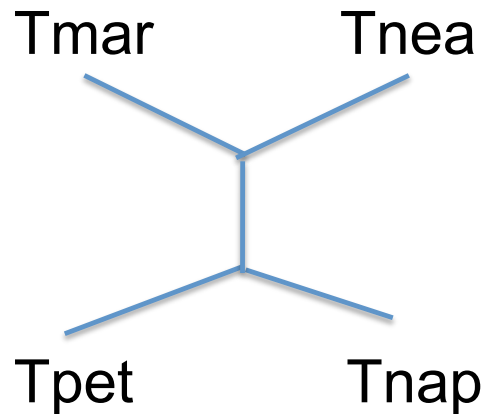
Quartet Decomposition

Bad quartets with bootstrap support value > 0.9

[Download](#) as newick trees

Quartet ID	Gene Family Numbers	Quartet Topology
0	38	((Tmar,Tnap),(Tnea,Tpet));
2	55	((Tmar,Tpet),(Tnap,Tnea));
3	64	((Tmar,Tnap),(Tnea,TRQ2));
5	85	((Tmar,TRQ2),(Tnap,Tnea));
6	46	((Tmar,Tnap),(Tpet,TRQ2));
7	25	((Tmar,Tpet),(Tnap,TRQ2));
10	57	((Tmar,Tpet),(Tnea,TRQ2));
11	71	((Tmar,TRQ2),(Tnea,Tpet));
12	66	((Tnap,Tnea),(Tpet,TRQ2));
14	49	((Tnap,TRQ2),(Tnea,Tpet));

Quartets -> Matrix Representation Using Parsimony



matrix	
TRQ2	??
Tmar	10
Tnap	01
Tnea	10
Tpet	01

Quartet Decomposition

Good quartets with bootstrap support value > 0.9
[Download](#) as newick trees

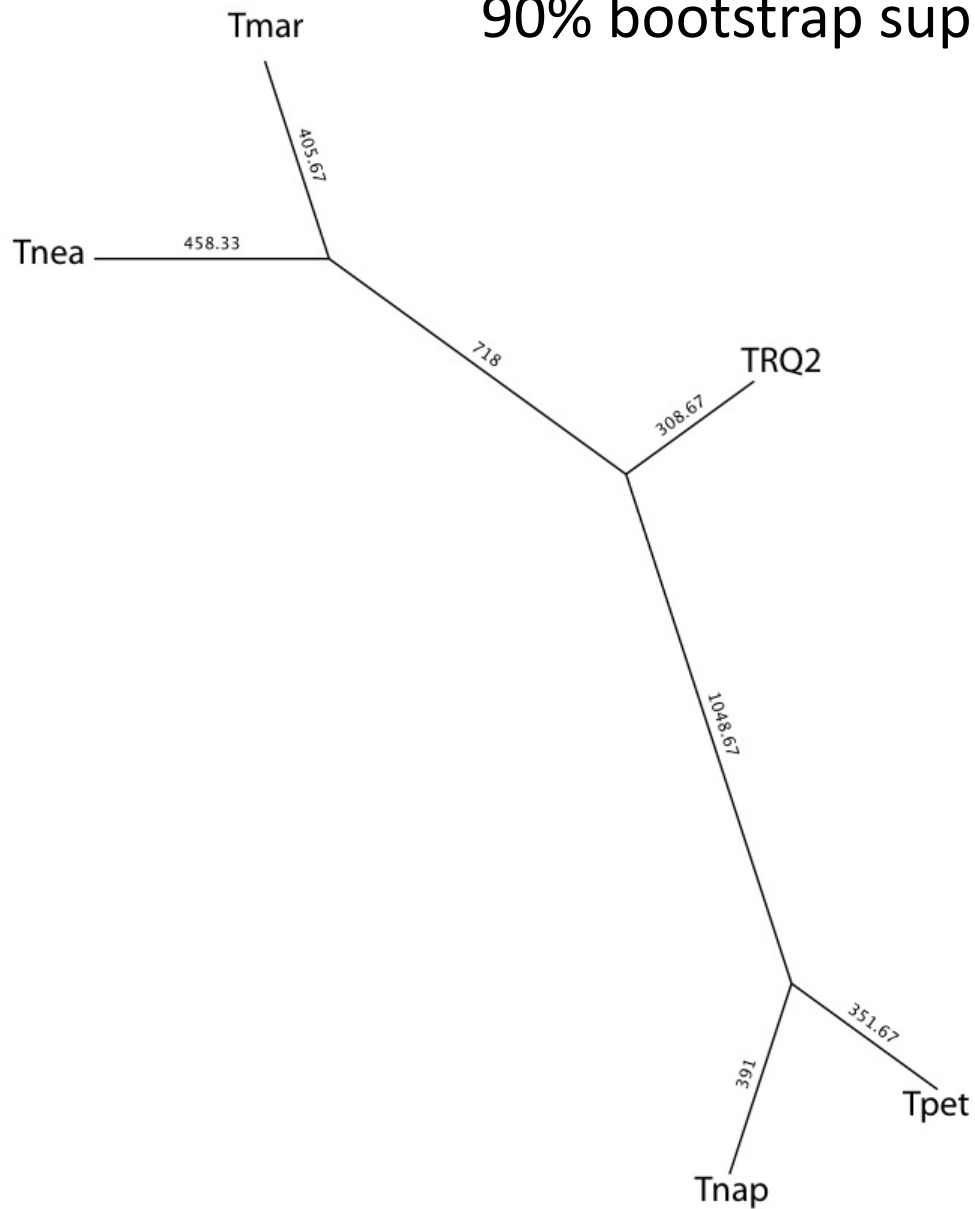
Quartet ID	Gene Family Numbers	Quartet Topology
1	192	((Tmar,Tnea),(Tnap,Tpet));
4	98	((Tmar,Tnea),(Tnap,TRQ2));
8	190	((Tmar,TRQ2),(Tnap,Tpet));
9	103	((Tmar,Tnea),(Tpet,TRQ2));
13	146	((Tnap,Tpet),(Tnea,TRQ2));



	5	2570
TRQ2	????????????????????????????????	10101010101010101010
Tmar	10101010101010101010101010101010	????????????????????
Tnap	01010101010101010101010101010101	10101010101010101010
Tnea	10101010101010101010101010101010	01010101010101010101
Tpet	01010101010101010101010101010101	01010101010101010101

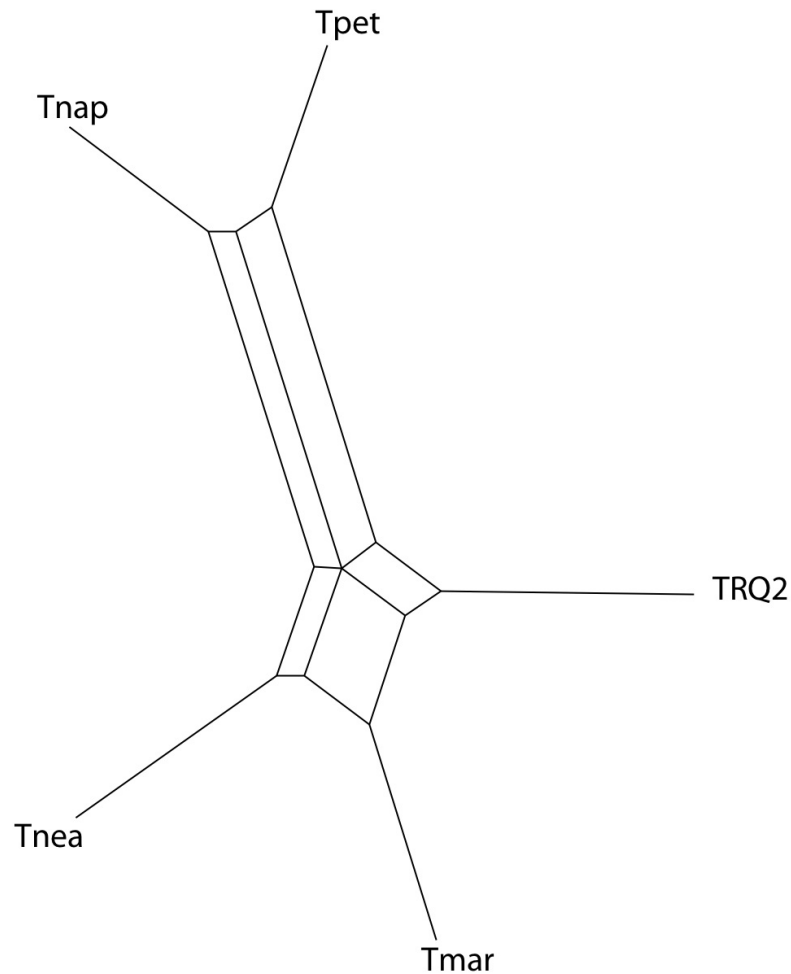
Most Parsimonious Tree (MRP)

Using all Quartets from all Gene Families that have more than 90% bootstrap support

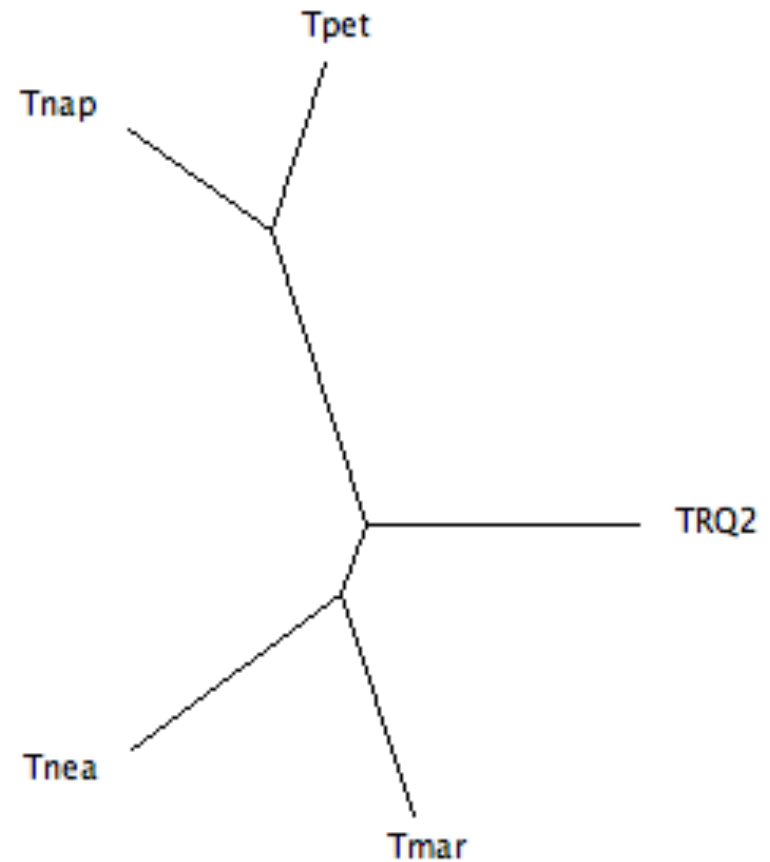


Splits Tree Representation

Using all Quartets from all Gene Families that have more than 90% bootstrap support

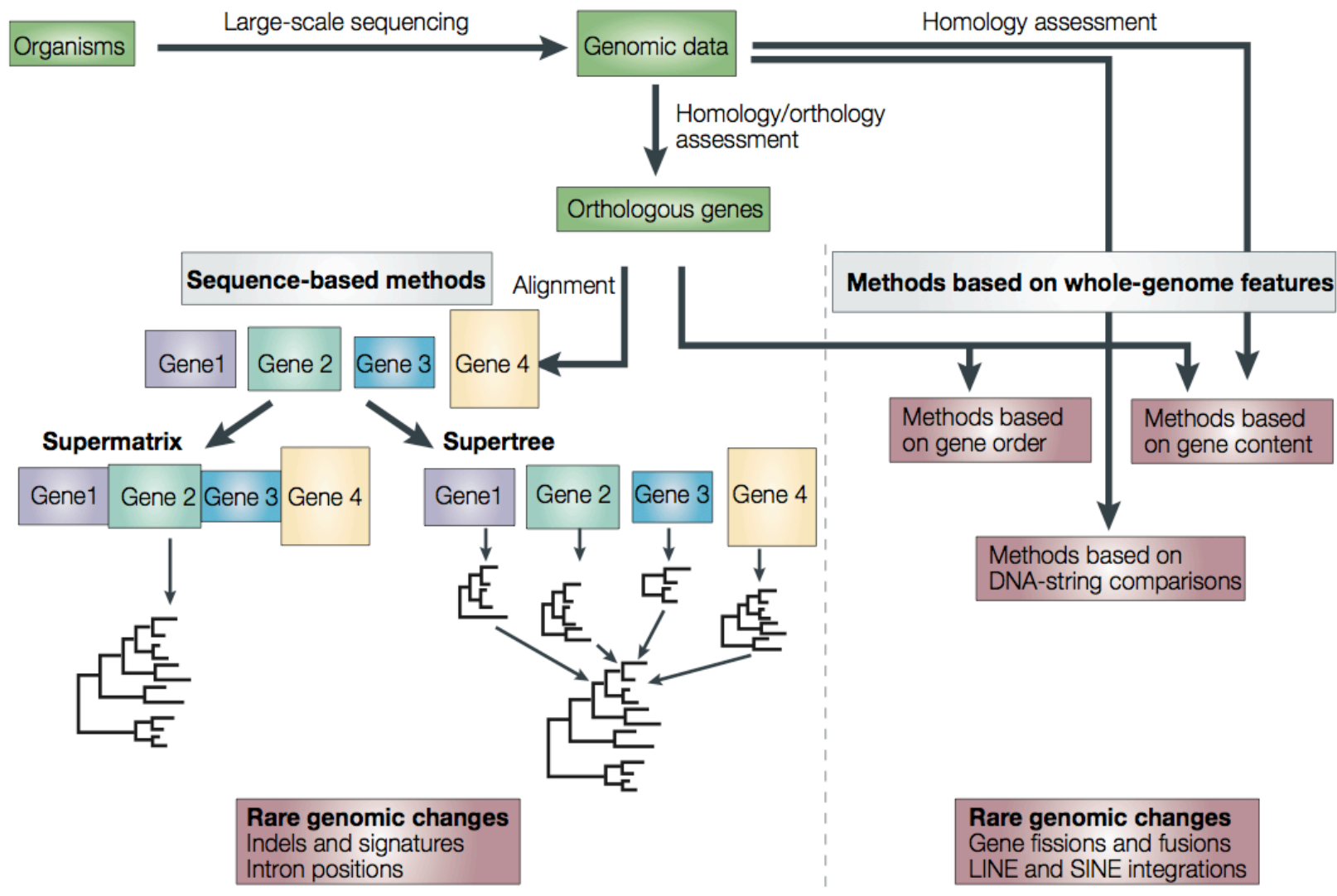


Split Decomposition tree
from uncorrected P distances



NJ tree
from uncorrected P distances

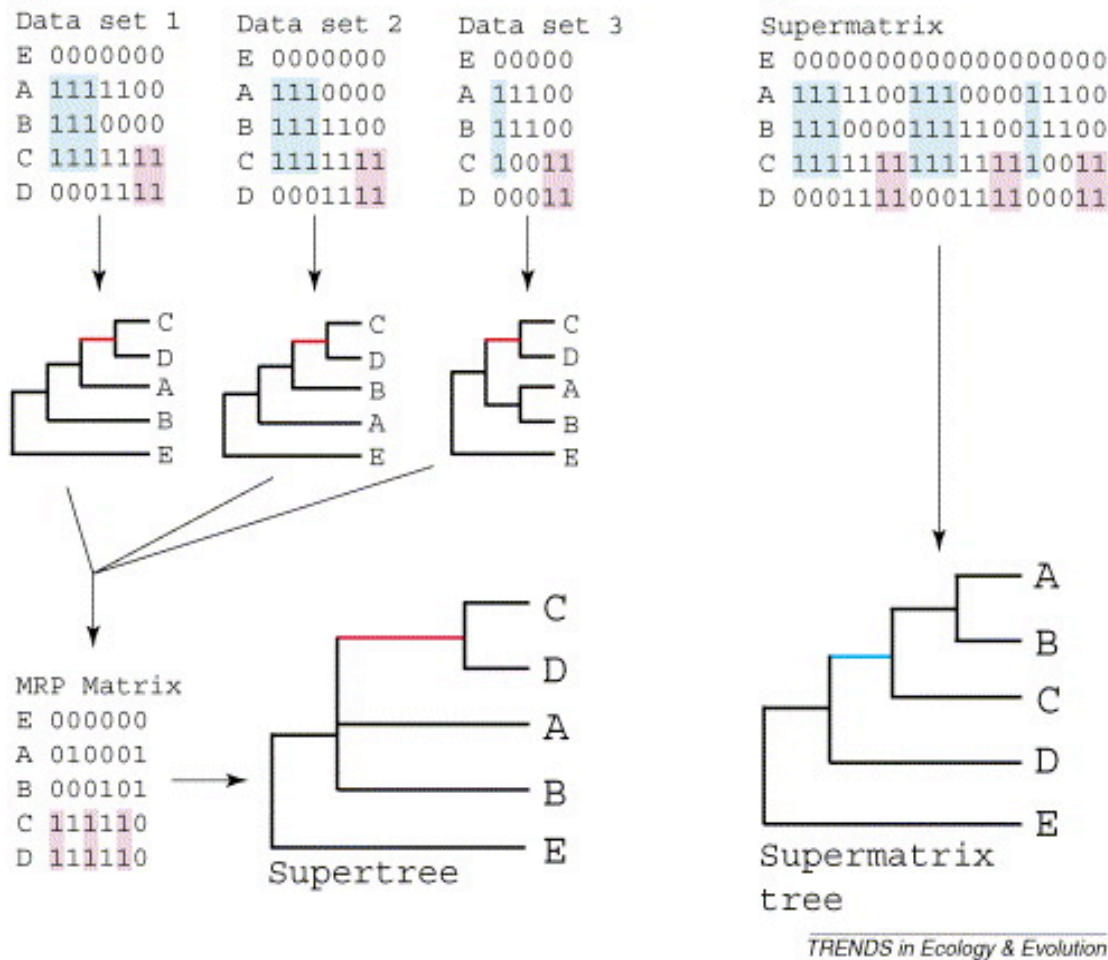
Box 2 | Methods of phylogenomic inference



The flowchart shows steps in the inference of evolutionary trees from genomic data. Genomic information is obtained by large-scale DNA sequencing. In general, sets of orthologous genes are then assembled from specific sets of species for phylogenetic analysis. This homology or orthology assessment is a crucial step that is almost always based on simple similarity comparisons (for example, **BLAST**¹⁵⁸ searches). Most methods used for the subsequent reconstruction of phylogenetic trees are either sequence-based or are based on whole-genome features.

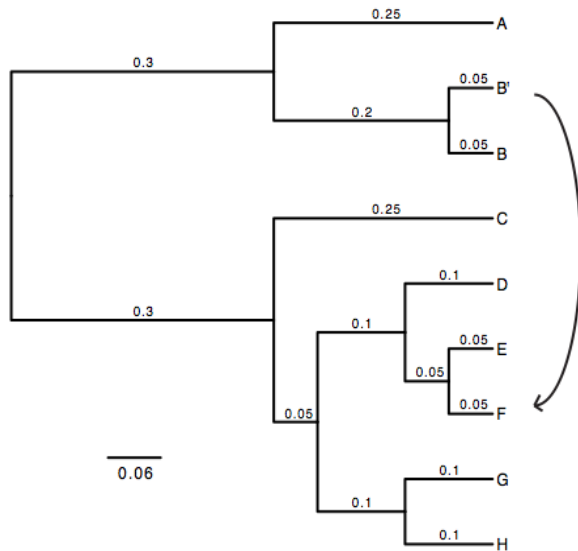
From:
 Delsuc F, Brinkmann H, Philippe H.
 Phylogenomics and the reconstruction of the tree of life.
 Nat Rev Genet. 2005 May;6(5):361-75.

Supertree vs. Supermatrix



From:
 Alan de Queiroz John Gatesy:
 The supermatrix approach to systematics
 Trends Ecol Evol. 2007 Jan;22(1):34-41

Schematic of MRP supertree (left) and parsimony supermatrix (right) approaches to the analysis of three data sets. Clade C+D is supported by all three separate data sets, but not by the supermatrix. Synapomorphies for clade C+D are highlighted in pink. Clade A+B+C is not supported by separate analyses of the three data sets, but is supported by the supermatrix. Synapomorphies for clade A+B+C are highlighted in blue. E is the outgroup used to root the tree.



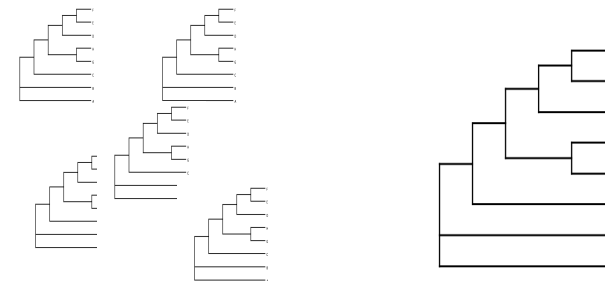
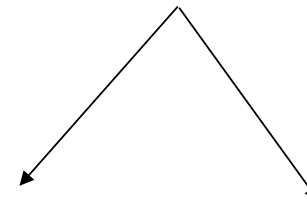
A) Template tree

```

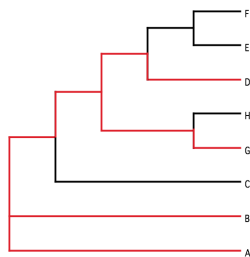
A  IYQILLVNNSSLSTVNWALGQDEDELETQTKTAPLDMSPITKI KAVQDVGEYALFNAENAG
B  ILQILLVNLSSLSTVKNWHLGQDEDELETQTESAFLDMTFVNKIEAVQDVGEYVVFNAENAW
C  ICQILMVNLSSFSTVSWQLAQDEDELETQGGLLDMRFITKVTTQQDVAEYPLFNAENAI
D  ICAILMINVSALSTVYWKLAQDEDELETQTSGLFSLMRPMAKIATQQDVGEYSLFNAKNTV
E  ICLILLINTSAESTVNWRLTQDEDELETQGGFFLSMRPMTKIRTRQDVGEYSLFNAKNTV
F  ICAILLINTSAHSTVNWSLTQDEDELETQGGCFLSMRPMTKIRTRQDVGEYSLFNAKNTV
G  ICAILPINASATSTVDWTLKQDEDELETQGGFFLEMRPMTKIRTRQDVAEYLLFNAENAS
H  ICAILLINASALSTVNWHLQDEDELETQGGFFLEMRPMTKIRTRQDVAEYSLFNAENAT
  *  ** : * * : *** * * ***** : * . * * : : : *** . ** : *** : * :

```

B) Generate 100 datasets using Evolver with certain amount of HGTs



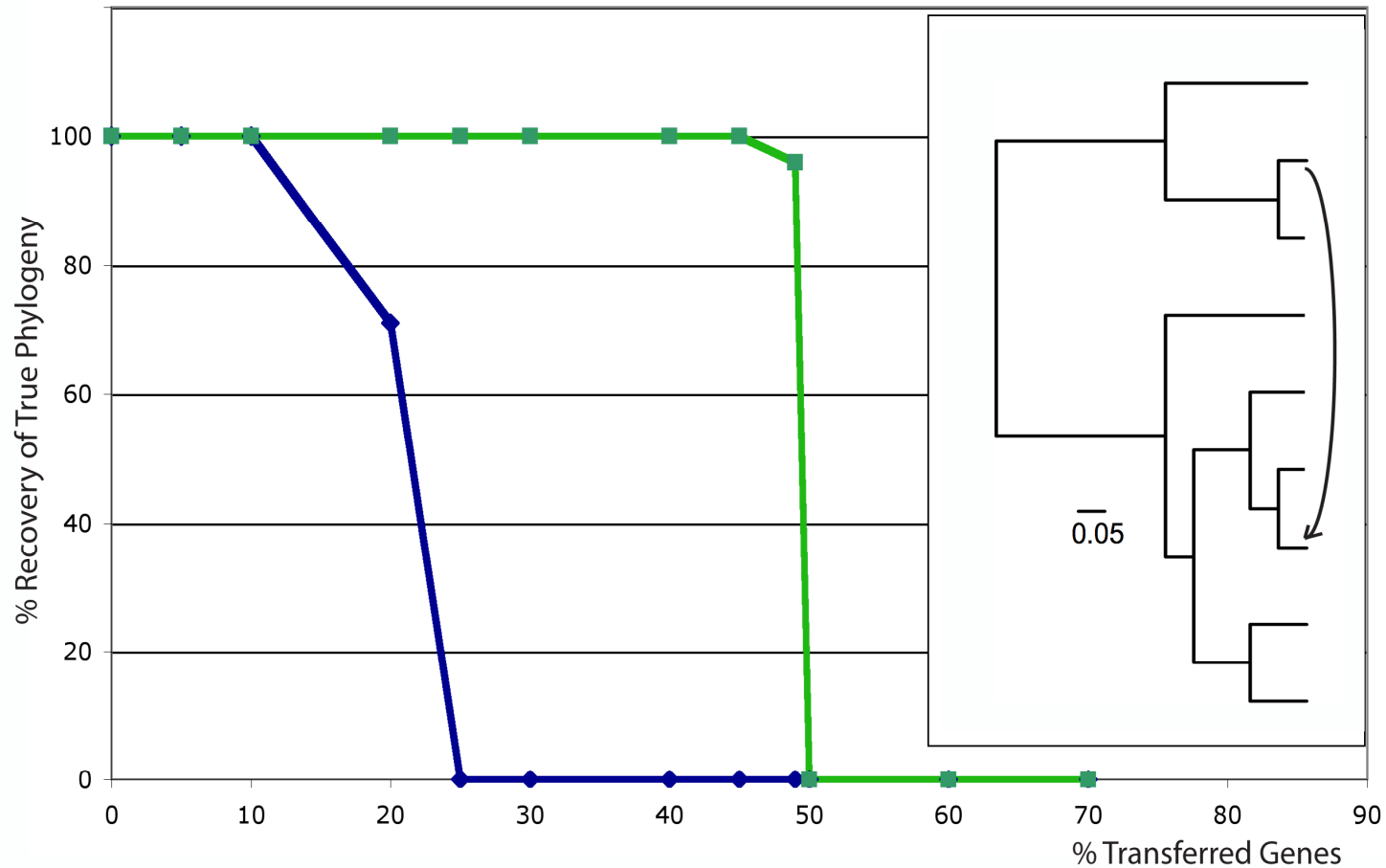
C) Calculate 1 tree using the concatenated dataset or 100 individual trees



D) Calculate Quartet based tree using Quartet Suite

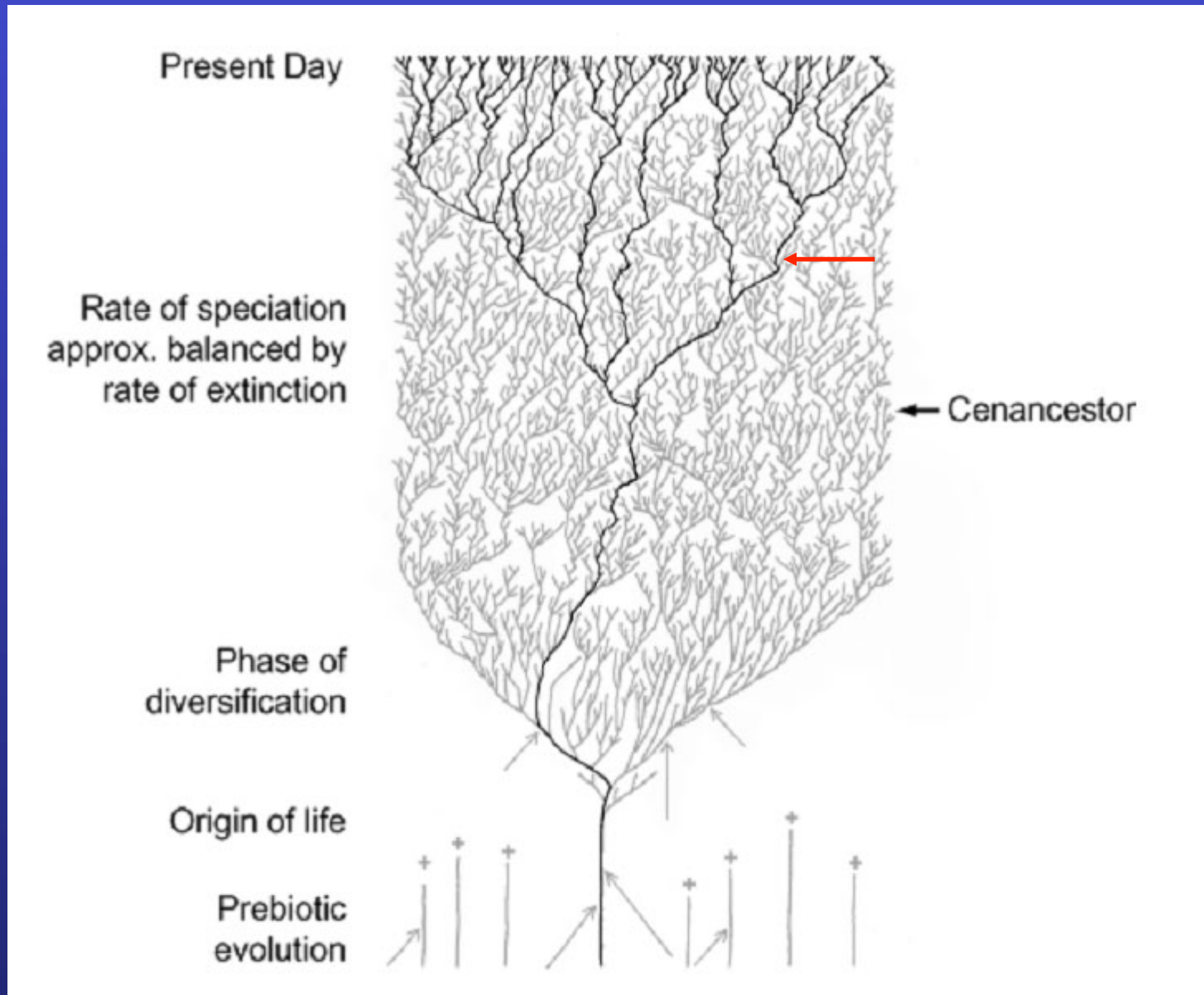
Repeated 100 times...

Supermatrix versus Quartet based Supertree



inset: simulated phylogeny

The Coral of Life (Darwin)



Coalescence – the process of tracing lineages backwards in time to their common ancestors. Every two extant lineages coalesce to their most recent common ancestor. Eventually, all lineages coalesce to the cenancestor.

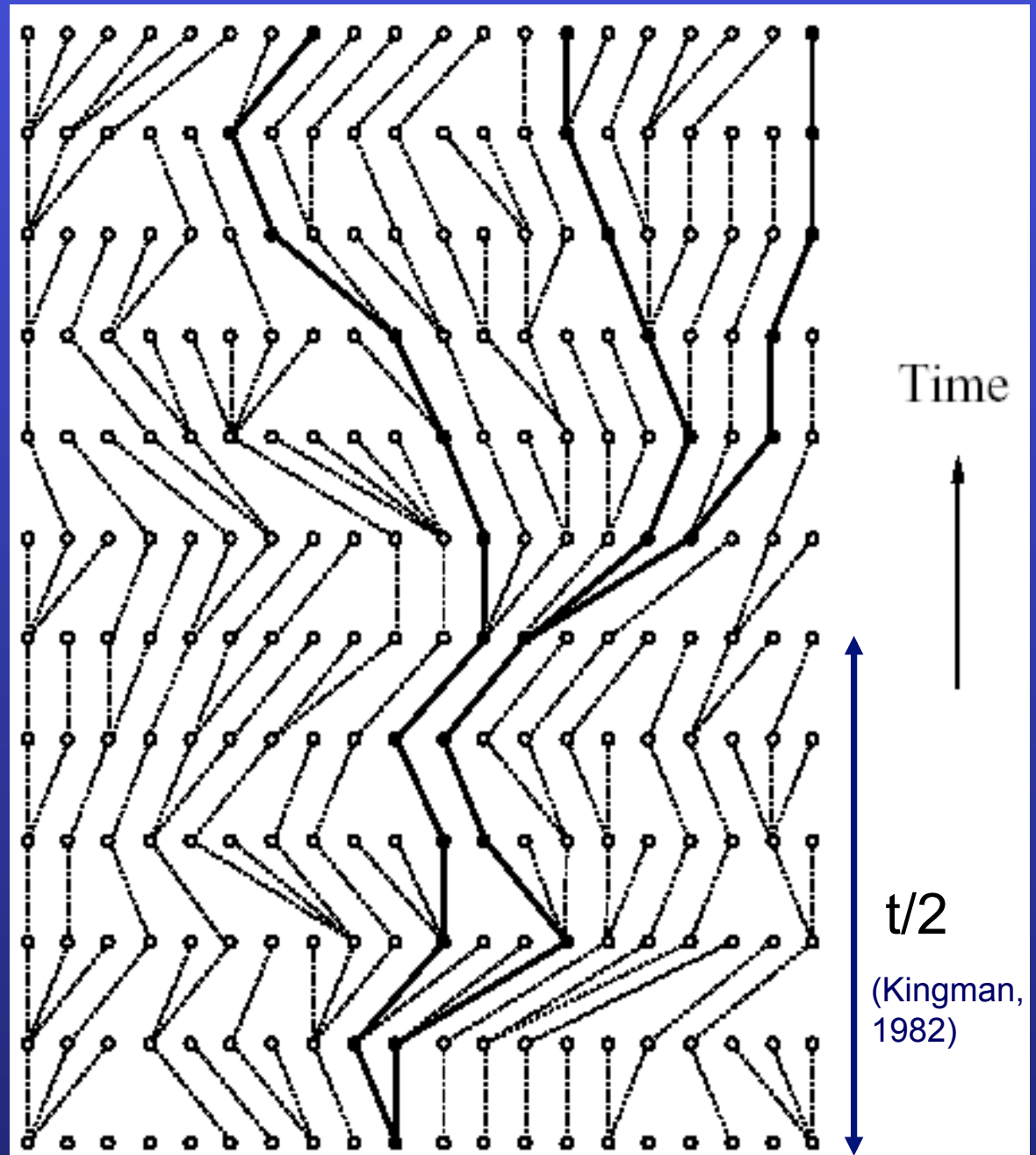
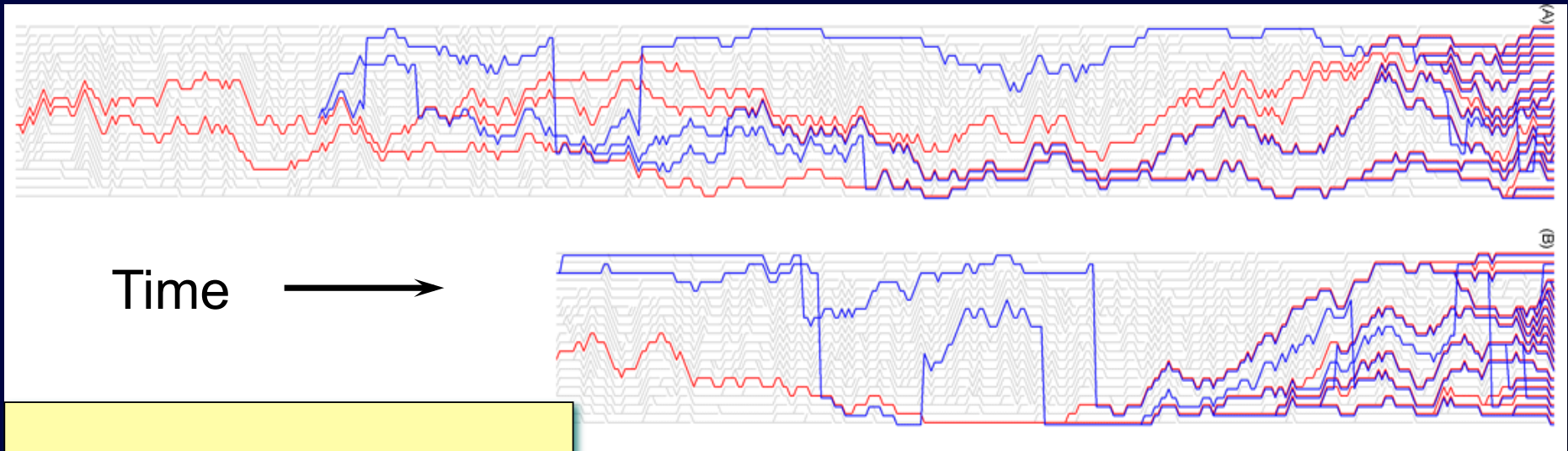


Illustration is from J. Felsenstein, "Inferring Phylogenies", Sinauer, 2003

Coalescence of **ORGANISMAL** and **MOLECULAR** Lineages



•20 lineages

•One extinction and one speciation event per generation

•One horizontal transfer event once in 10 generations (i.e., speciation events)

RED: organismal lineages (no HGT)

BLUE: molecular lineages (with HGT)

GRAY: extinct lineages

RESULTS:

•Most recent common ancestors are different for organismal and molecular phylogenies

•Different coalescence times

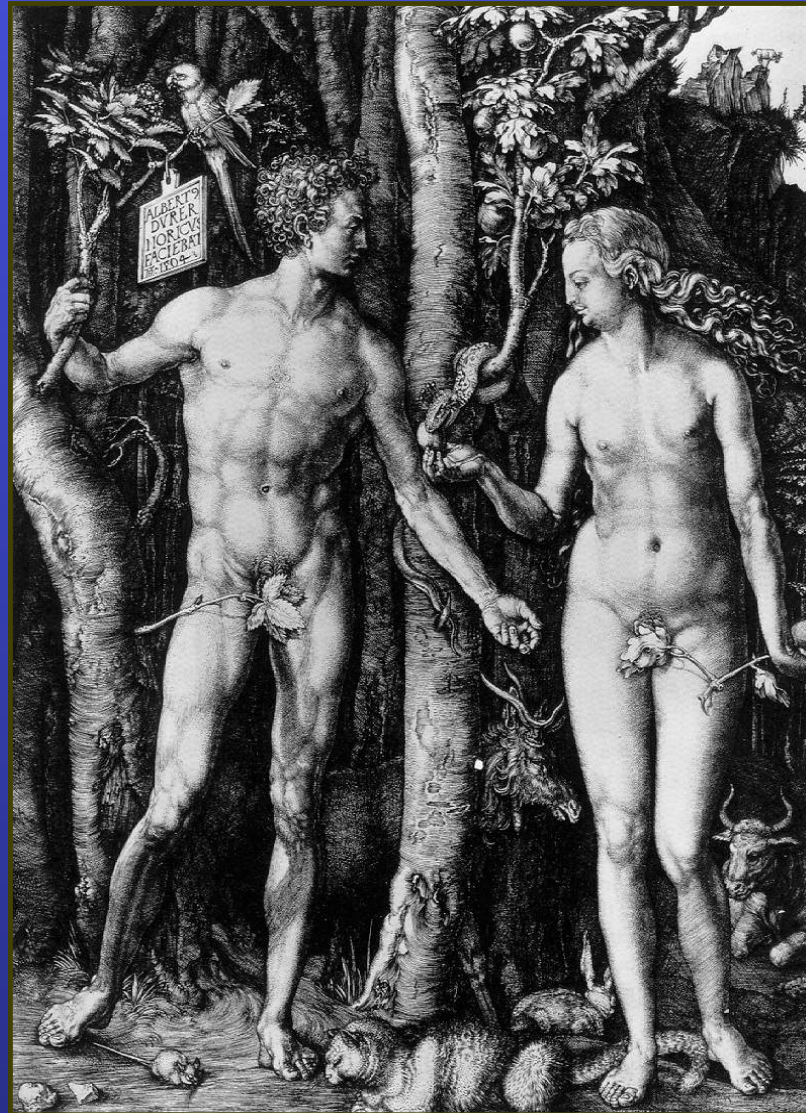
•Long coalescence time for the last two lineages

Y chromosome Adam

Lived approximately 50,000 years ago

Thomson, R. *et al.* (2000)
Proc Natl Acad Sci U S A 97, 7360-5

Underhill, P.A. *et al.* (2000)
Nat Genet 26, 358-61



Albrecht Dürer, The Fall of Man, 1504

Adam and Eve never met ☹

The same is true for ancestral rRNAs, EF, SRP, ATPases!

Mitochondrial Eve

Lived 166,000-249,000 years ago

Cann, R.L. *et al.* (1987) *Nature* 325, 31-6

Vigilant, L. *et al.* (1991) *Science* 253, 1503-7

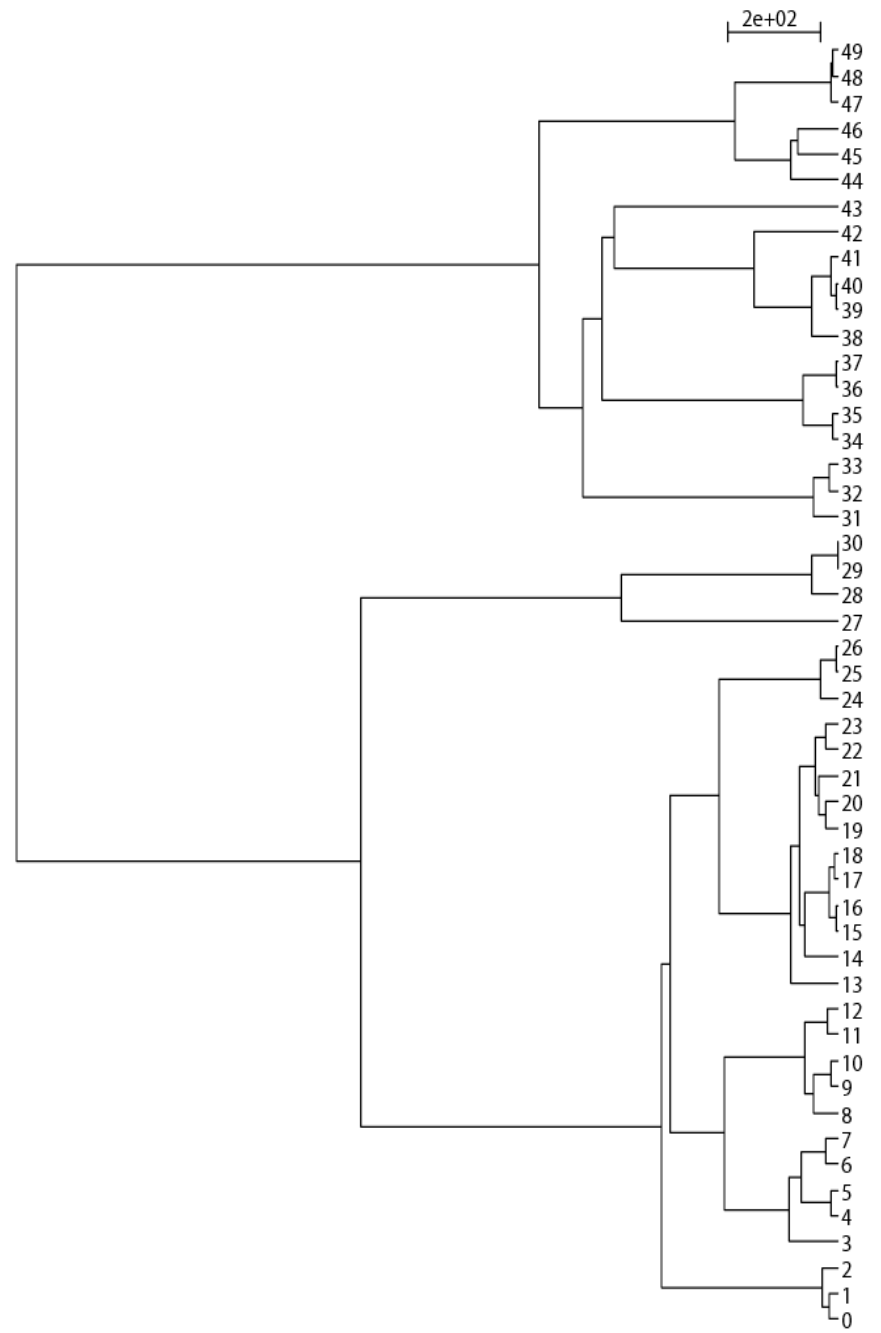
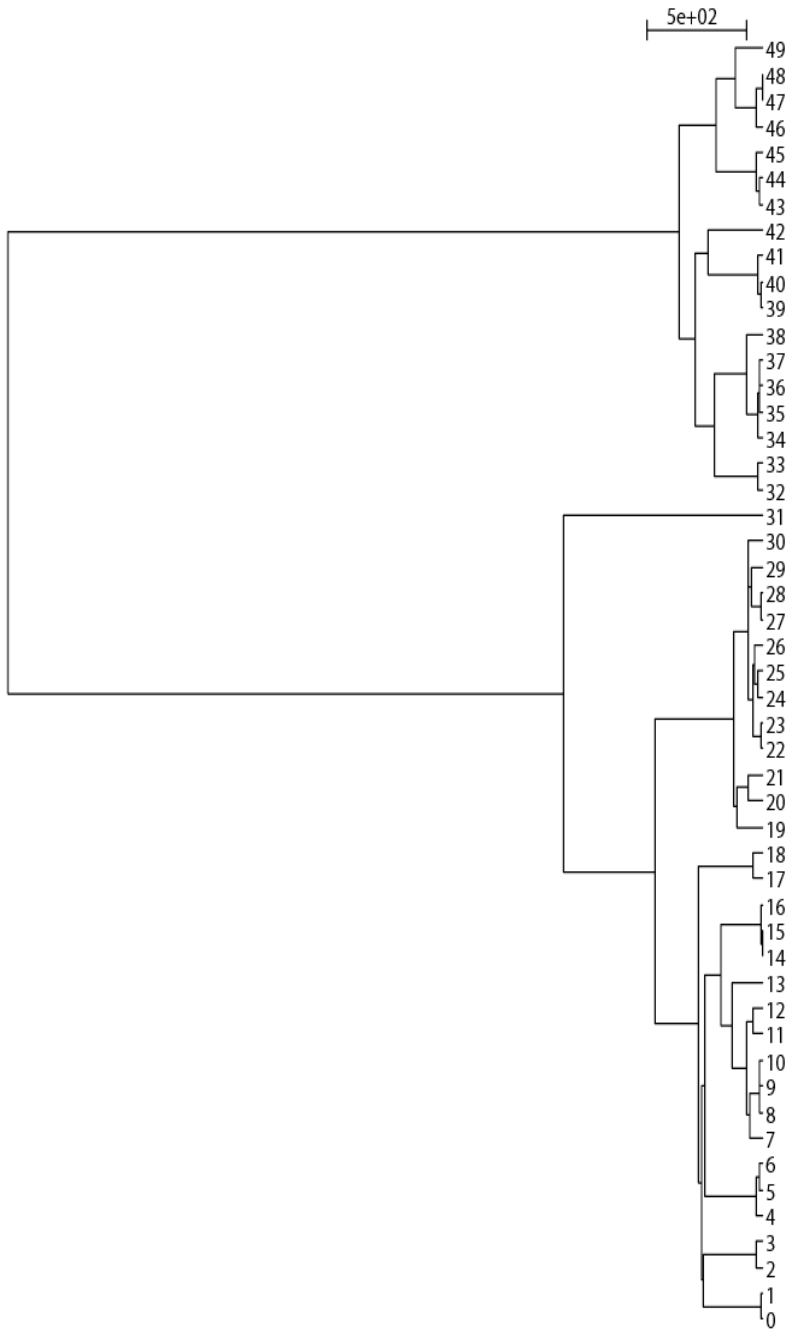
The most recent common ancestor in pedigrees:
(aside the most recent common ancestor of all humans, i.e. the person found in all pedigrees of now existing human was estimated to have lived only a few thousand years ago. (About 4500 years BP under a realistic model for migration and non random mating)

see D.L. Rohde, S. Olson, J.T. Chang, Nature 431(7008), 562–566 (2004)

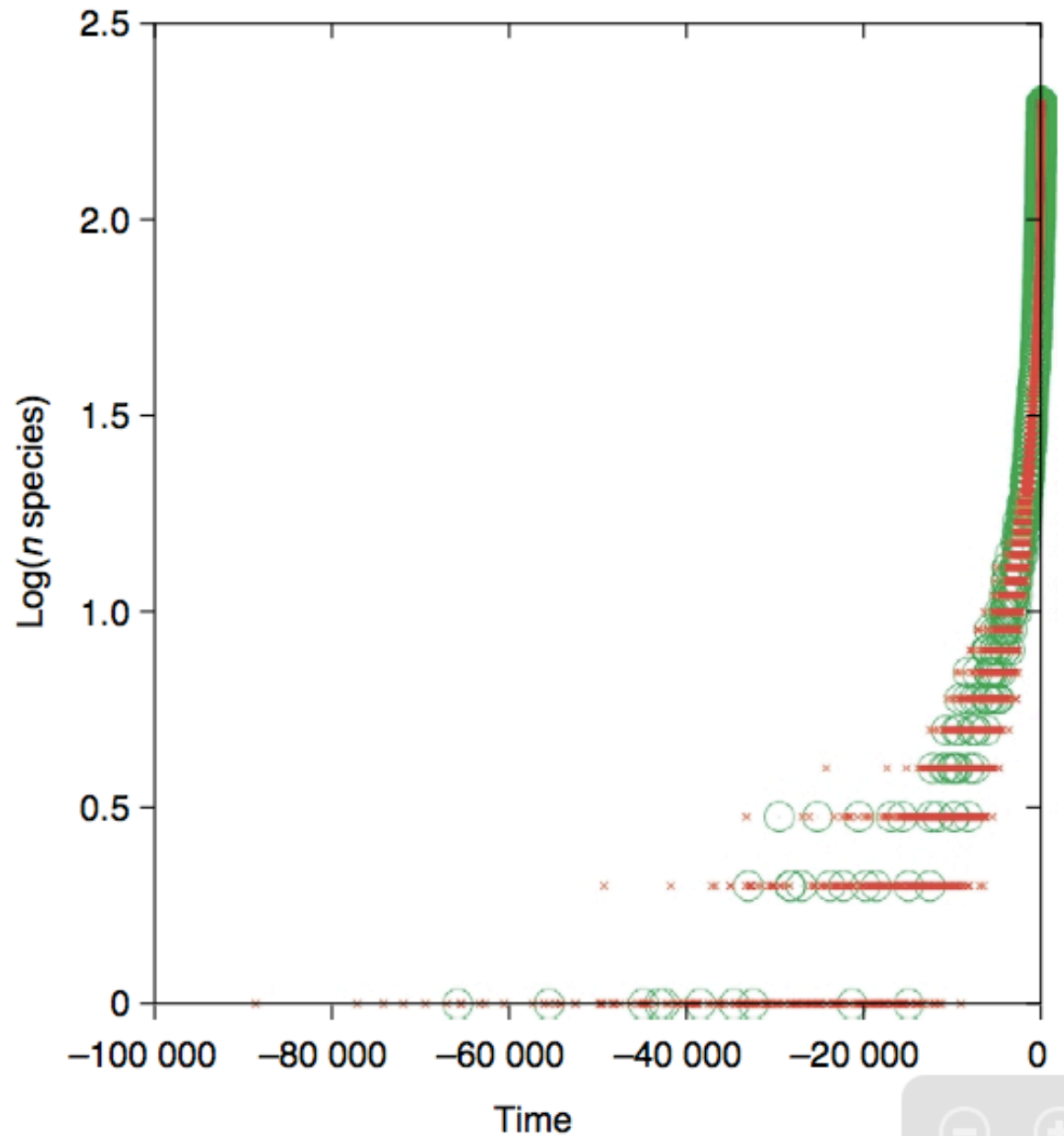
Did this genealogical MRCA contribute any genes to your genome?

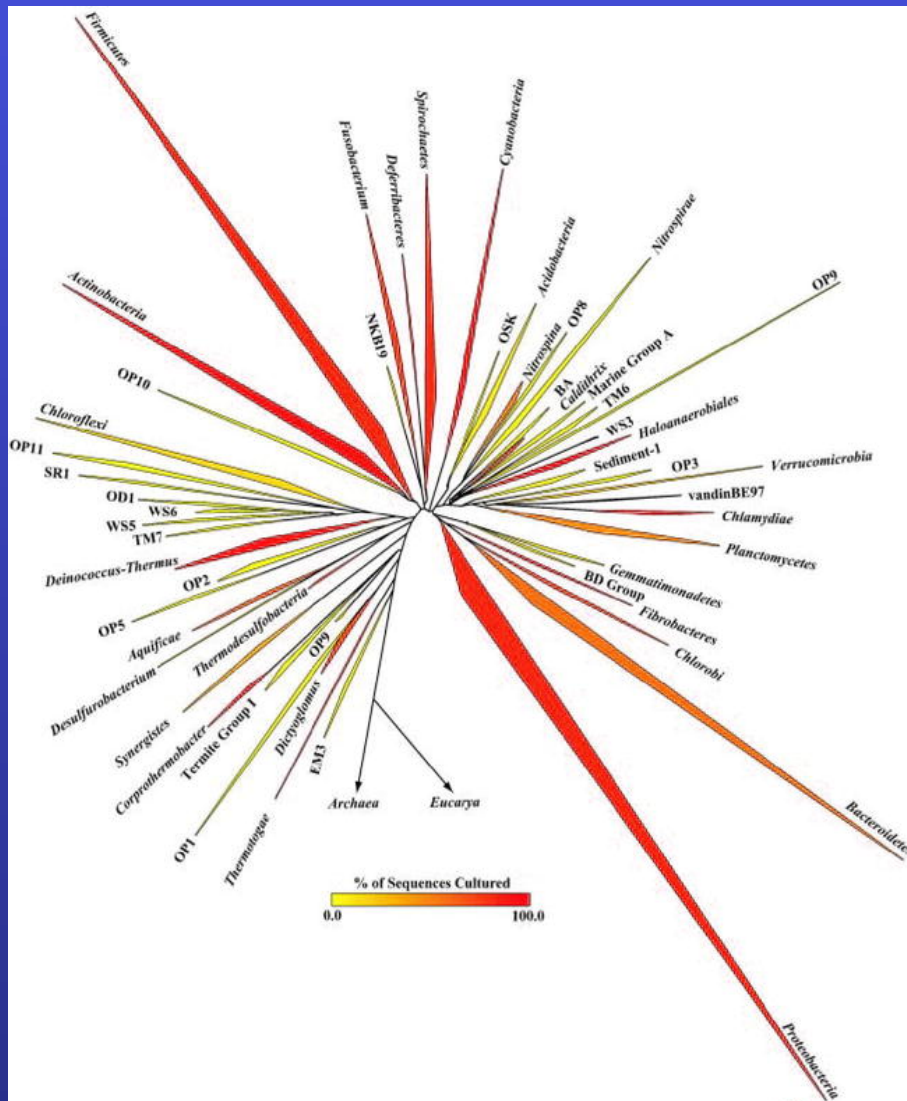
(Provide a back of the envelope calculation on how many nucleotides in your genome were contributed by this MRCA).

EXTANT LINEAGES FOR THE SIMULATIONS OF 50 LINEAGES



Lineages through time plot for simulated data, 200 species per generation. Data from 10 independent simulations of organismal evolution are shown in green, and for each organismal simulation 25 simulations of gene evolution were performed [one horizontal gene transfer (HGT) event per 10 generations] and are shown in red.





- The deviation from the “long branches at the base” pattern could be due to
- under sampling
 - an actual radiation
 - due to an invention that was not transferred
 - following a mass extinction



Bacterial 16S rRNA based phylogeny
 (from P. D. Schloss and J. Handelsman,
 Microbiology and Molecular Biology Reviews,
 December 2004.)