Phylogenetic Tree Branch Density Estimation & Tree-base OTU Grouping Revisit

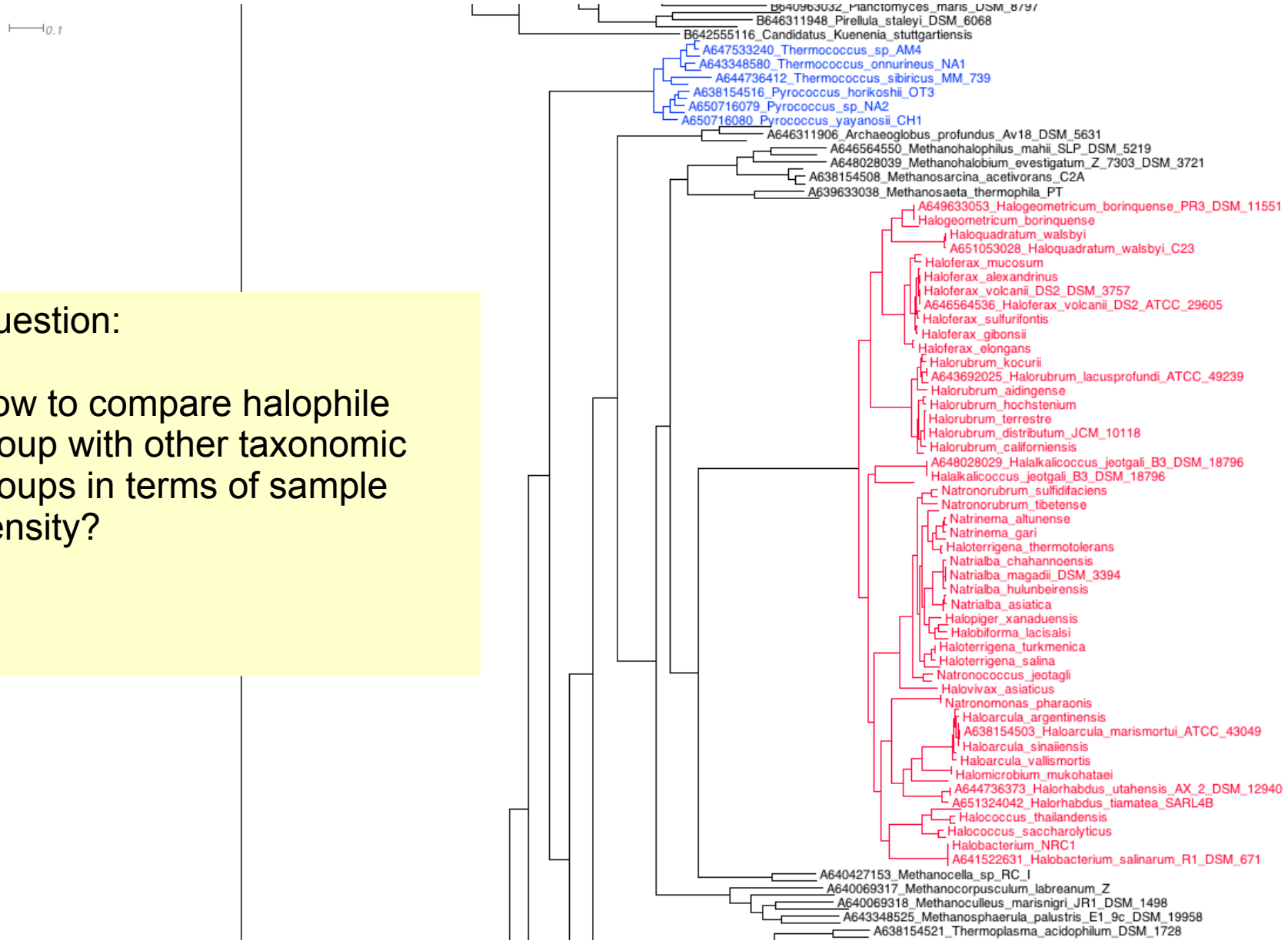
Dongying Wu

July, 2012

Fasttree of halophiles and all IMG bacteria and archaea based on the concatenated alignment of 40 markers

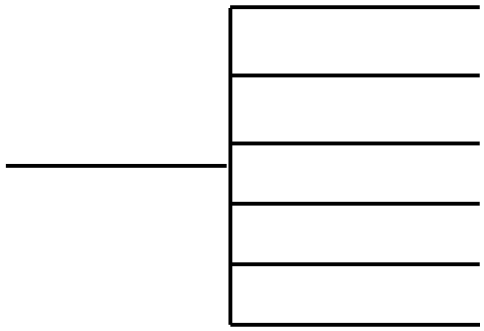
Question:

How to compare halophile group with other taxonomic groups in terms of sample density?

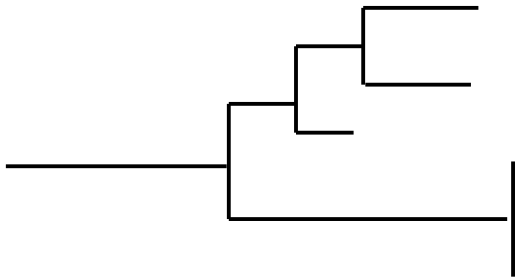


Estimate taxa density of a node in a tree

Taxa number as Density

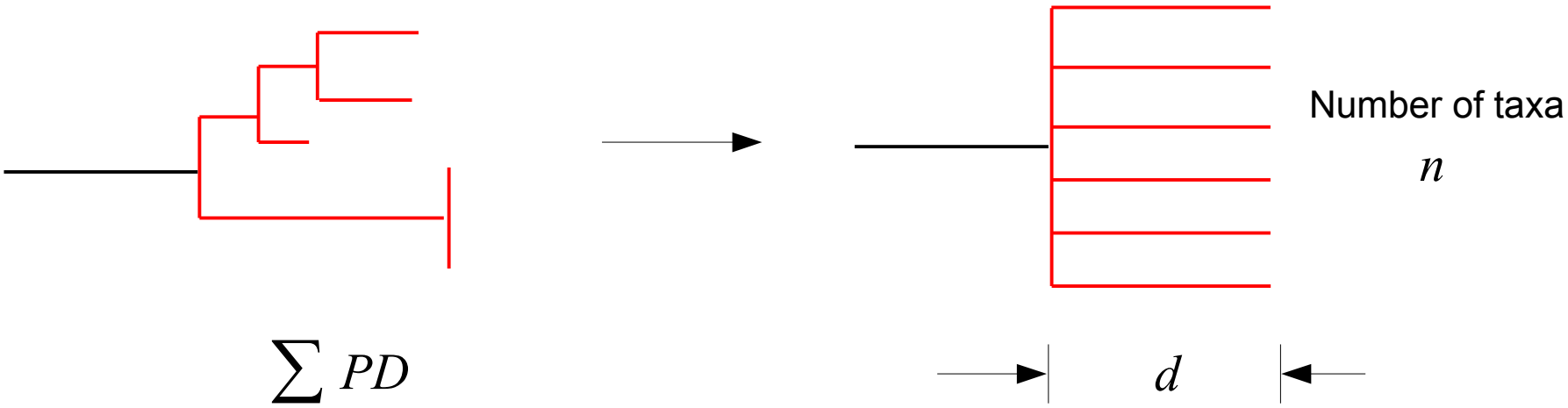


The underline assumption for taxa count as density



Density Calculation should take tree structure into account

Estimate taxa density of a node in a tree



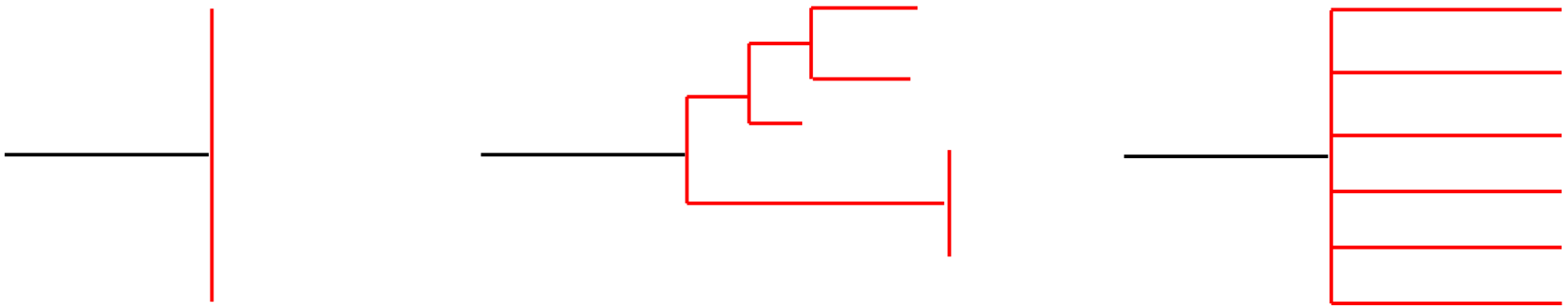
Average Depth of all the leaves to a node

$$Density = \frac{\sum PD}{n \times d} \times n$$

$$Density = \frac{\sum PD}{d}$$

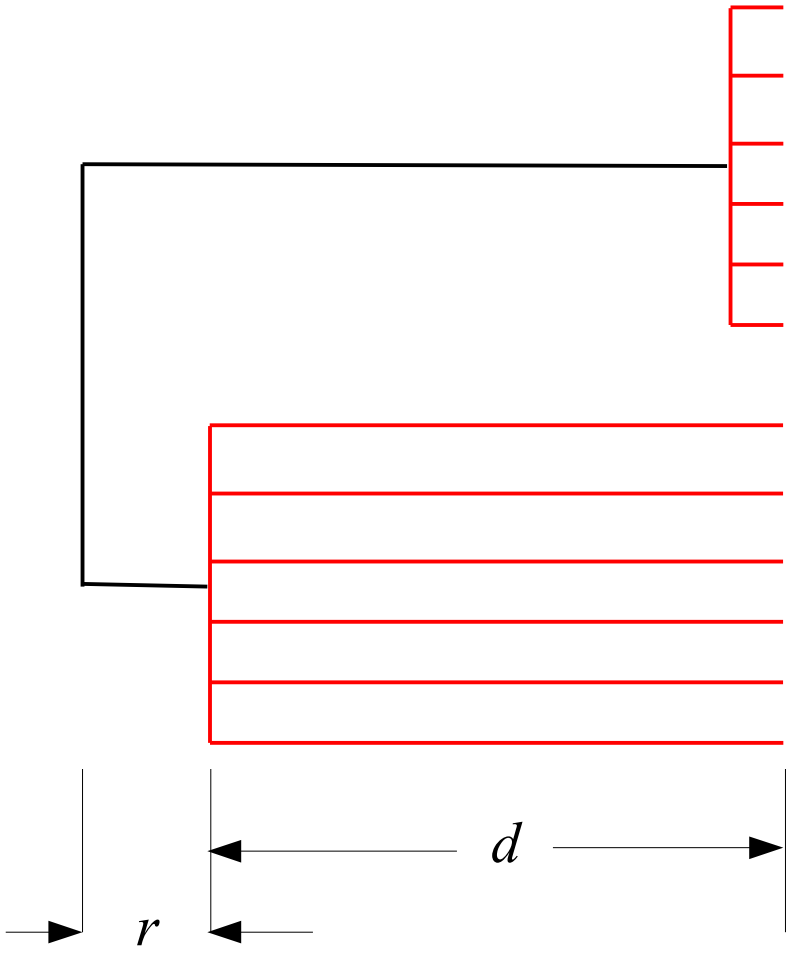
$$Density = \frac{\sum PD}{d}$$

For a node with n leaves
Density is a number between 1 and n



Density of 1

Density of n

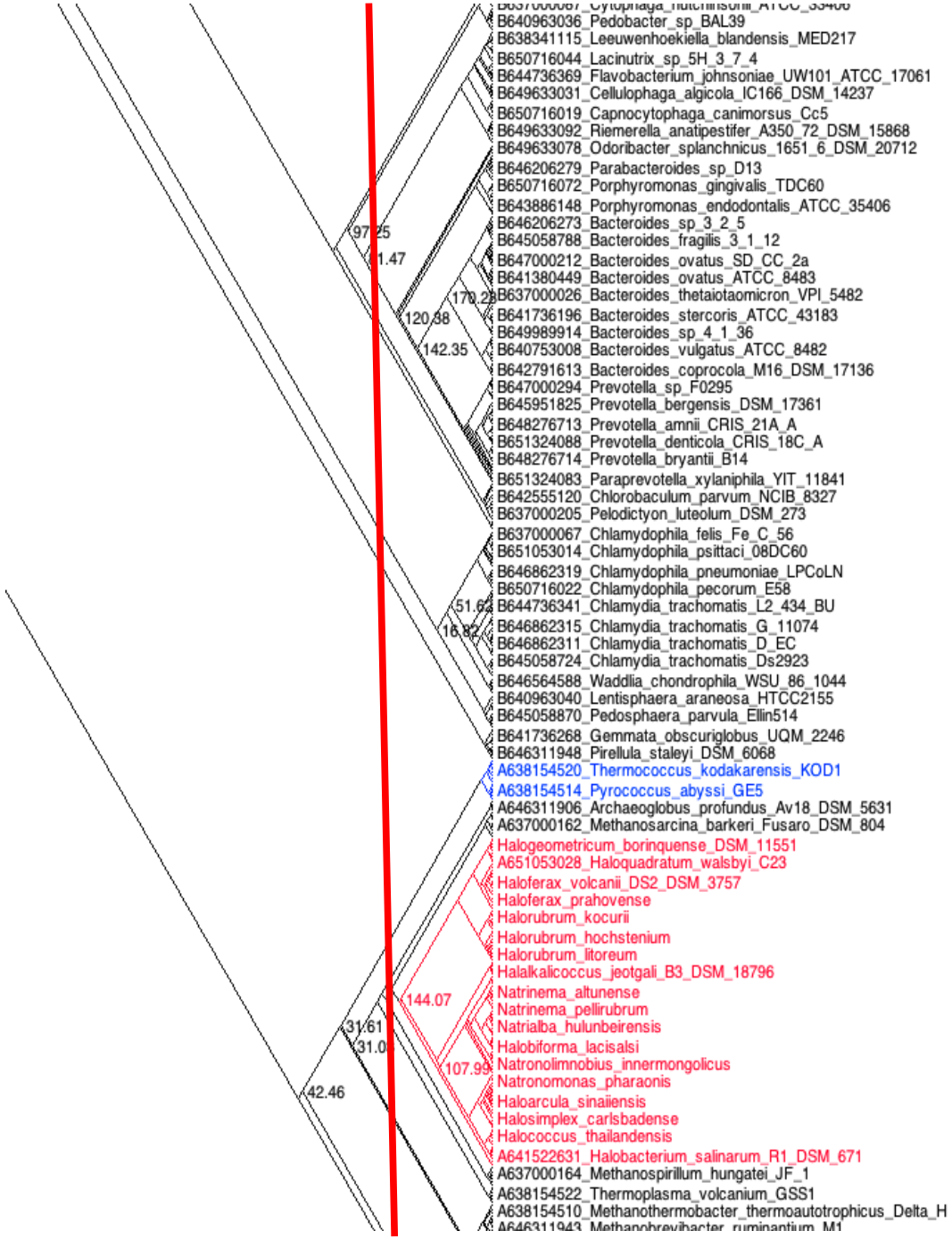


Density calculation need to take the relationship of root, node and leaves into account

$$Density = \frac{\sum PD}{d}$$

$$Relative\ Density = \frac{Density}{d/(r+d)}$$

$$Relative\ Density = \frac{(d+r)\times\sum PD}{d^2}$$



Halophile Group Relative Density: 144.07
Thermococcus/Pyrococcus Group: 69.79

Does density comparison between nodes make biological sense?

Tree-base OTU grouping revisit

Version 1: For each node, calculate the average distance of all the leaves and the standard deviation, and decide if a node defines an OTU

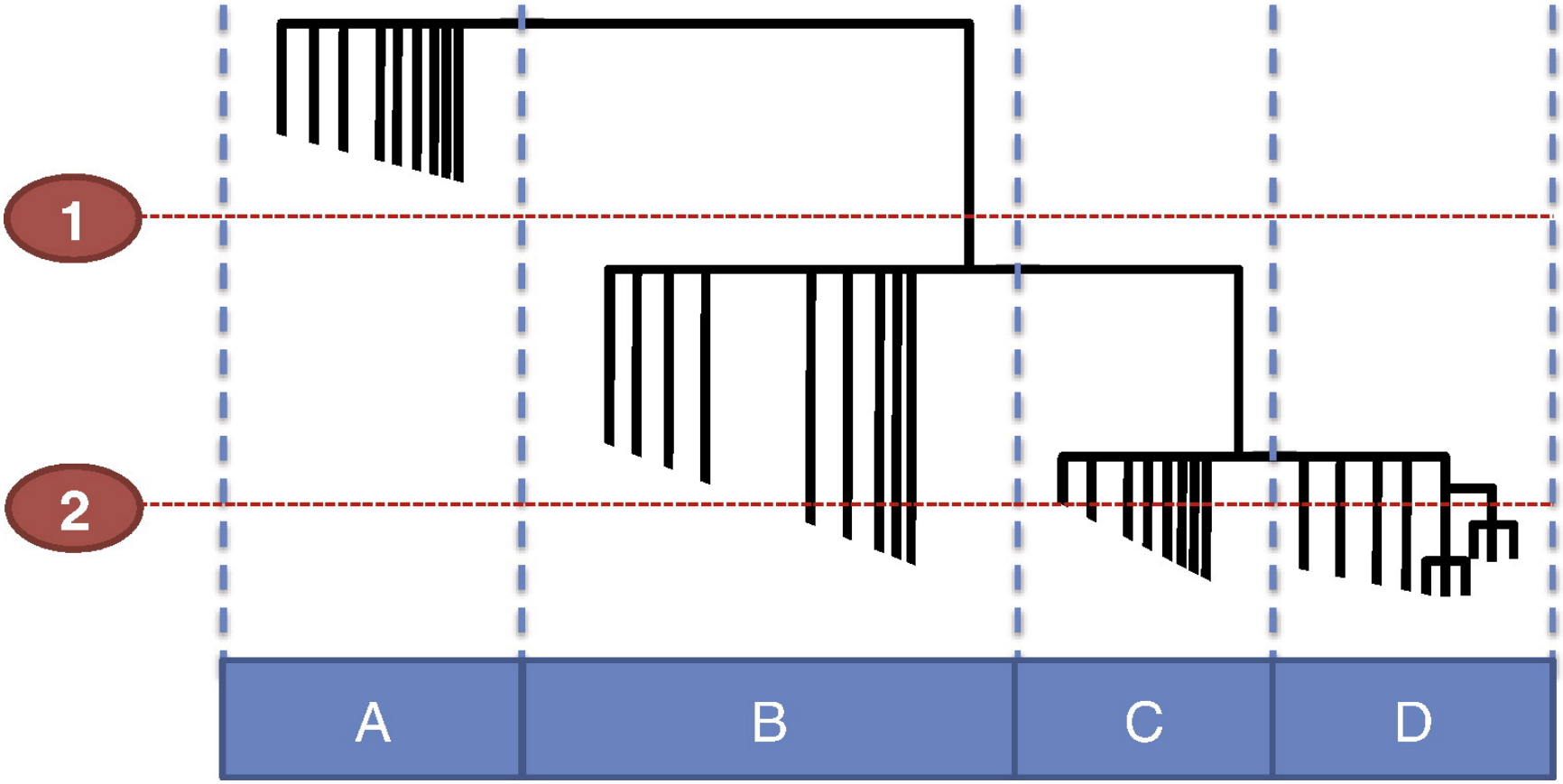
Problem: Slow, RAM demanding

Outside Progress:

1. Tree topology based OTU grouping (Dynamic Tree Cutting, D. Serre)

2. Tree branch-length based OTU grouping (PHYLOSEQ, P. McMurdie & S. Holmes)

Dynamic Tree Cutting



Genomics. 2011 Oct;98(4):253-9. Epub 2011 Apr 15.
A novel method for determining microflora composition using dynamic phylogenetic analysis of 16S ribosomal RNA deep sequencing data.

Chan ER, Hester J, Kalady M, Xiao H, Li X, Serre D.

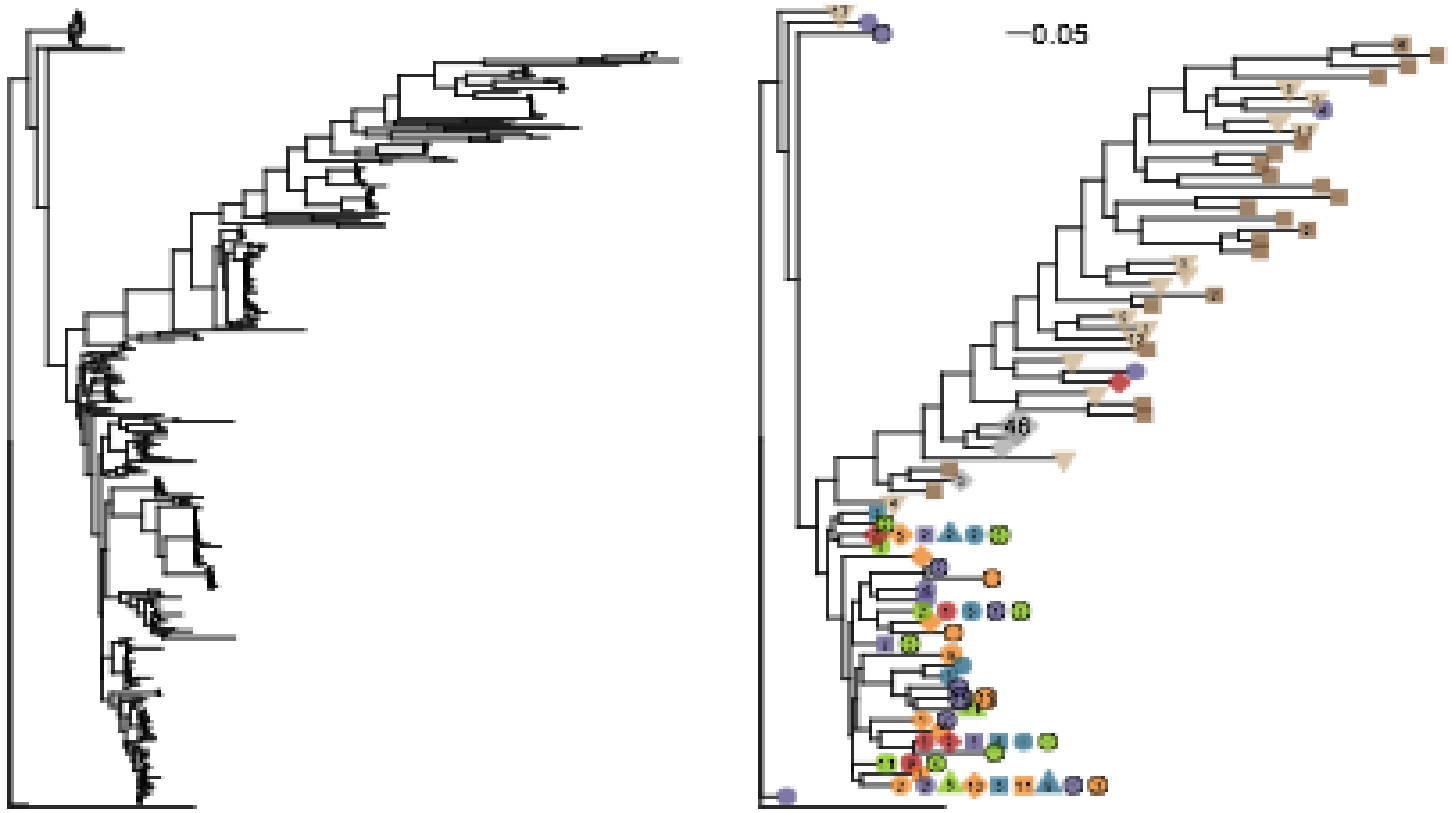
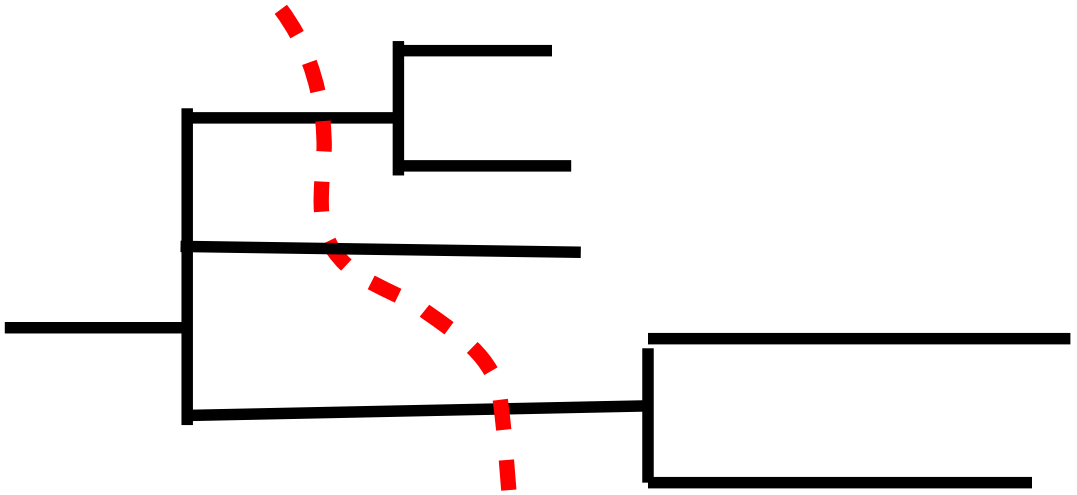


Fig. 6: Example of phylogenetic sequence data before and after basic clustering with `tipglom()` function. (Left) Standard phylogram produced using default plotting function and no OTU clustering. (Right) Annotated phylogram after OTU clustering with `tipglom()`. Different symbols next to each tip indicate different samples in which the OTU was observed. The number inside each symbol indicates the respective number of individuals of a given OTU were observed in each sample.

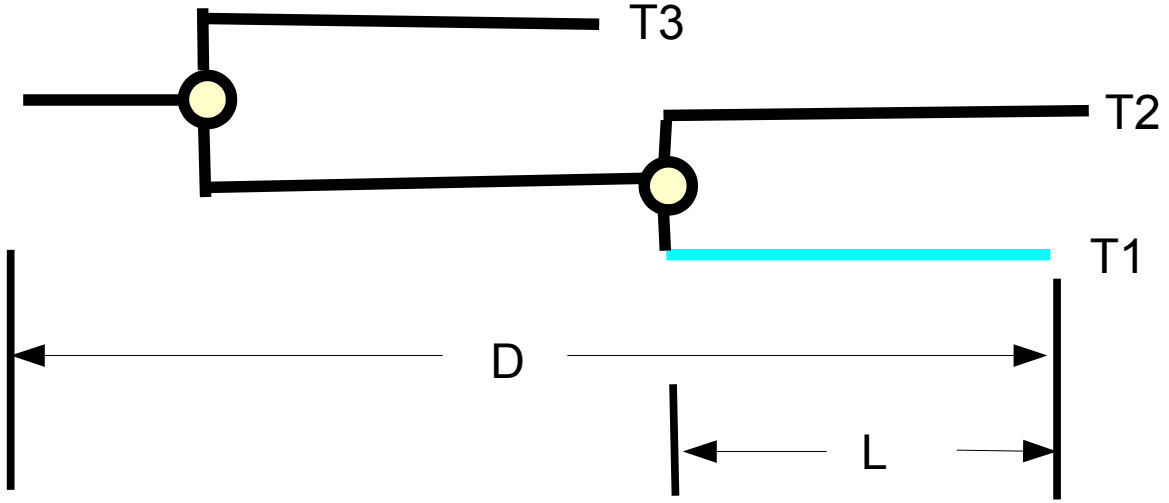
PHYLOSEQ, P. McMurdie & S. Holmes

Tree-based OTU classification

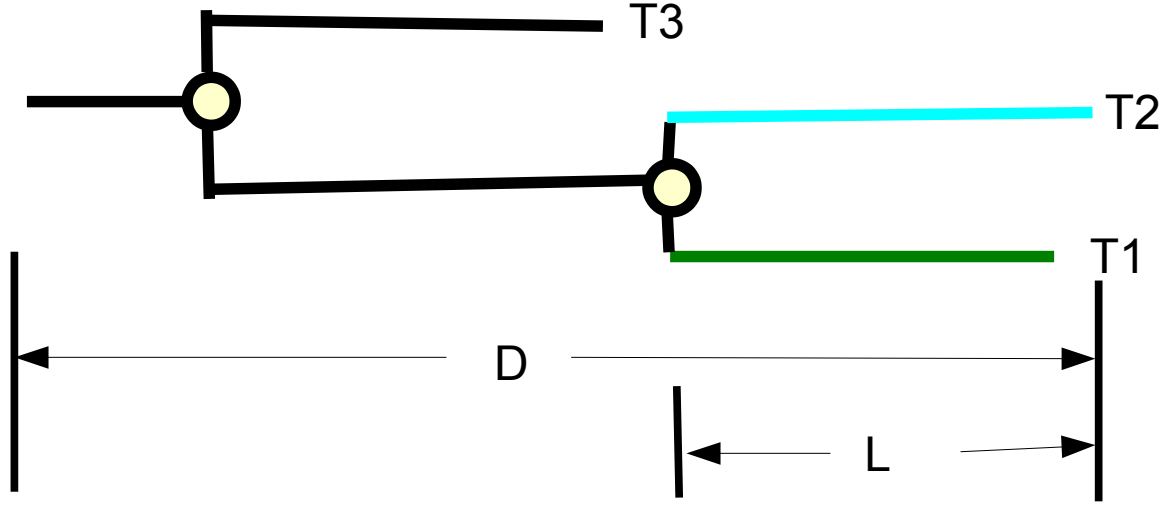


Rule 1: The classification should be based upon branch length rather than tree topology

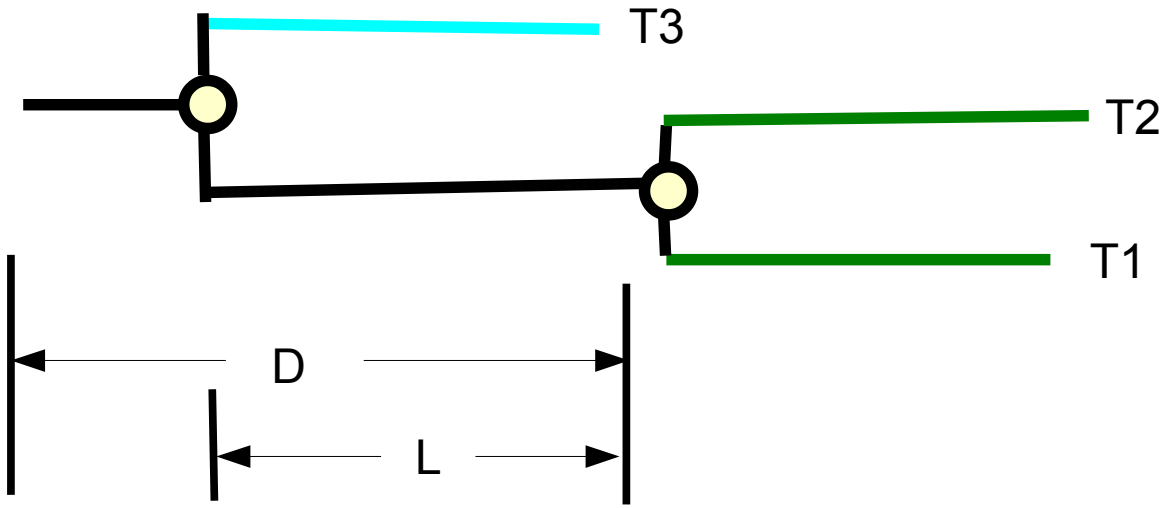
Rule 2: Dynamic cutting (branch length should be normalized to deal with different evolution rates fro different lineages)



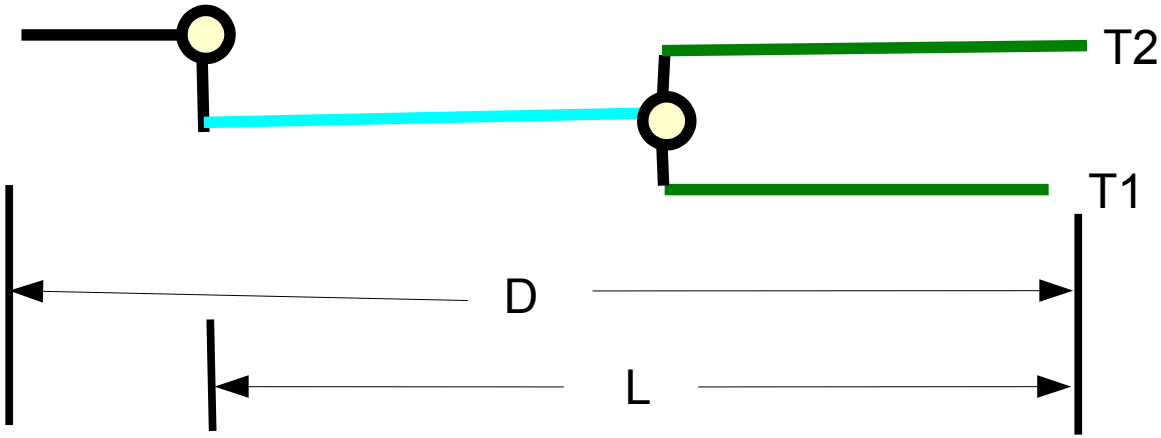
L/D >= cutoff ? No



L/D >= cutoff ? No

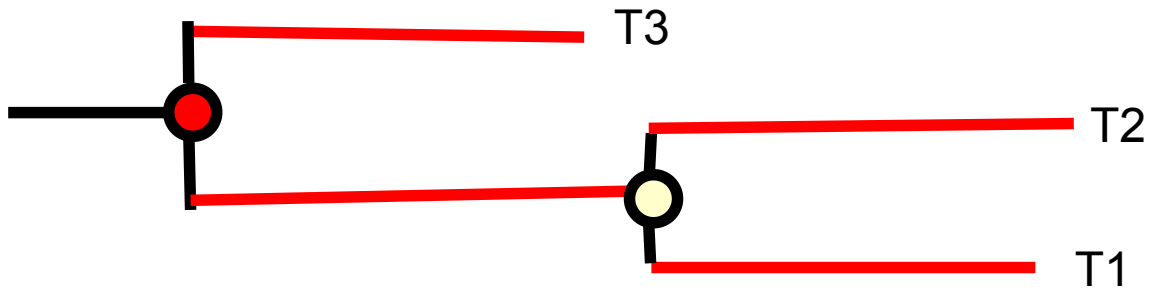


L/D >= cutoff ? Yes → (T3)

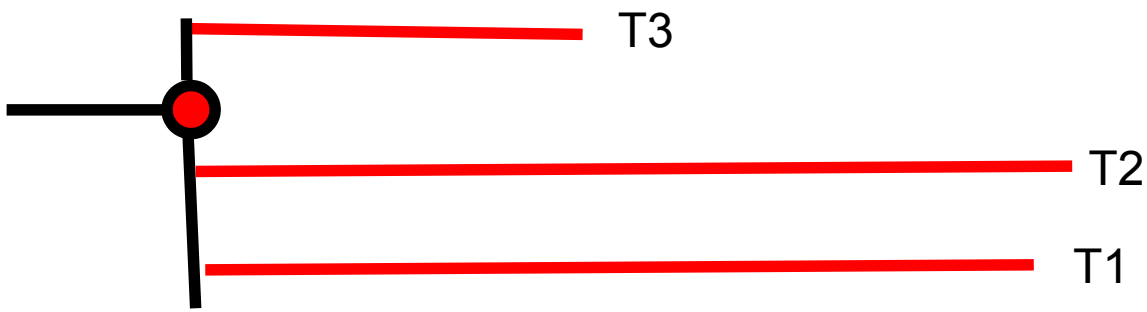


L/D >= cutoff ? Yes → (T1,T2)

How to calculate the average depth from tips to a node?

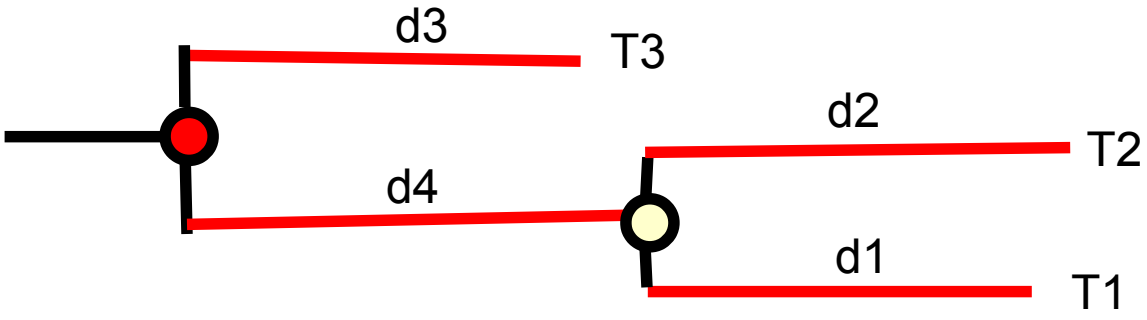


Naive Average:



Weighted Average:

Weight factor: The true contribution of a taxon to a node



T3 weight factor : 1

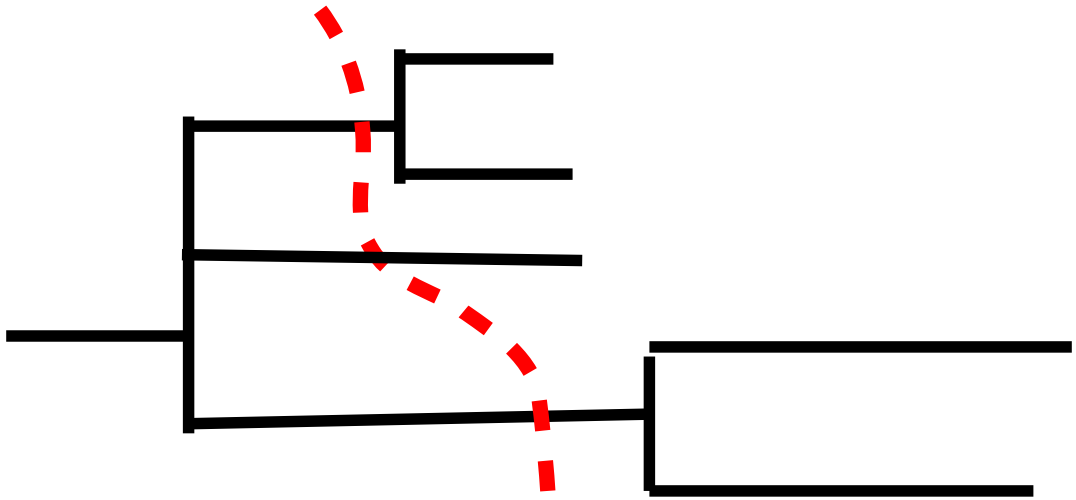
T2 weight factor : (d2+d4/2)/(d2+d4)

T1 weight factor : (d1+d4/2)/(d1+d4)

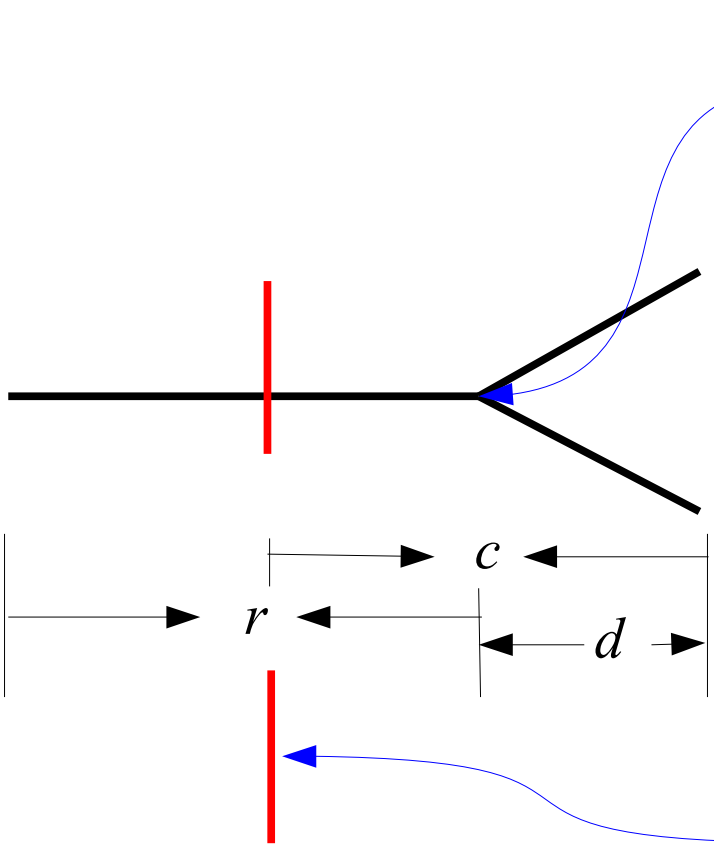
Speed up the OTU script:

Adept data structure for the rooted tree according to the following paper:

Syst Biol. 2006 Oct;55(5):769-73. Phylogenetic diversity within seconds.
Minh BQ, Klaere S, von Haeseler A.



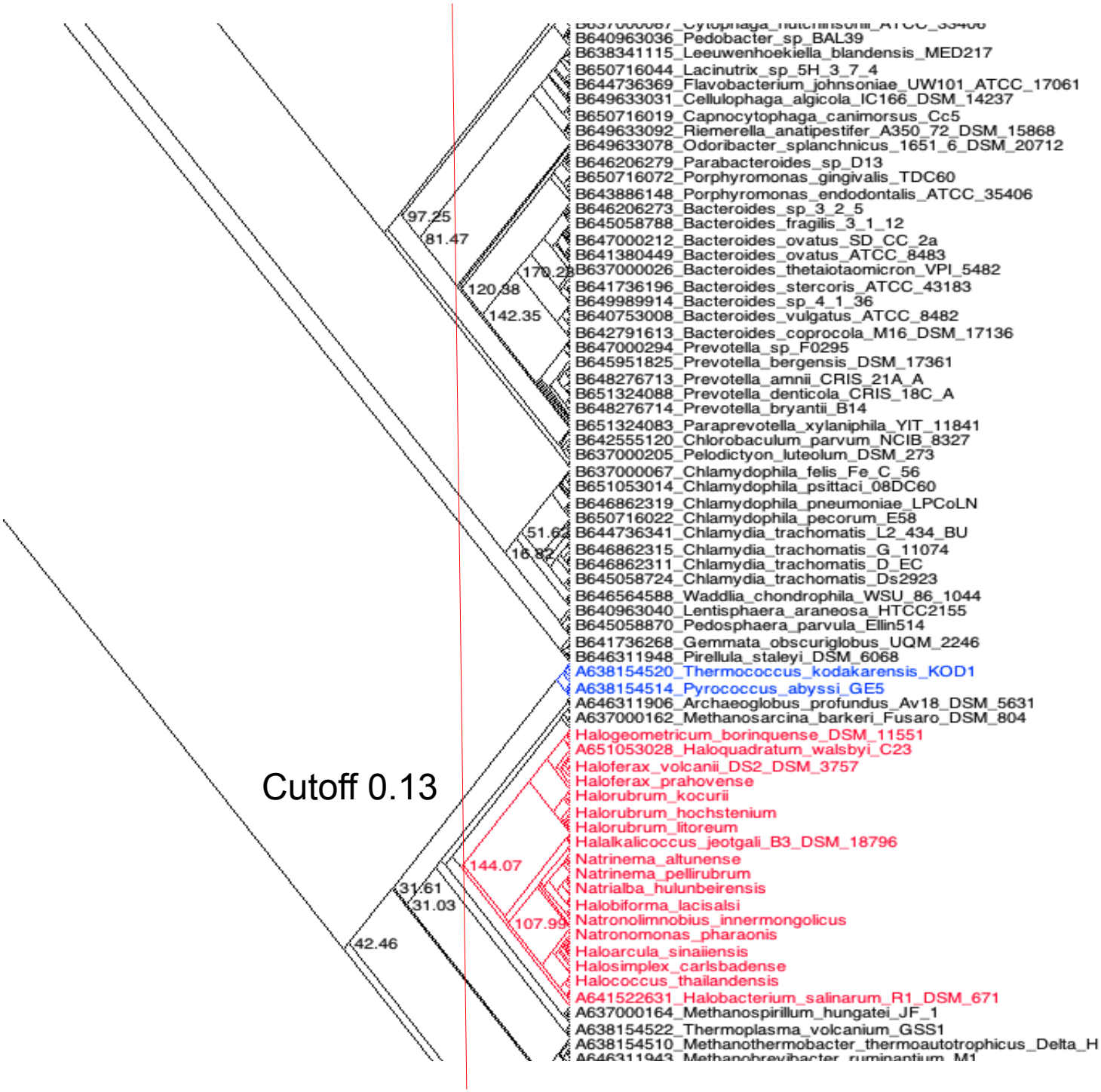
Node based relative densities are not comparable, densities need to be calculated at the cutoff line instead of the nodes that defines the OTUs

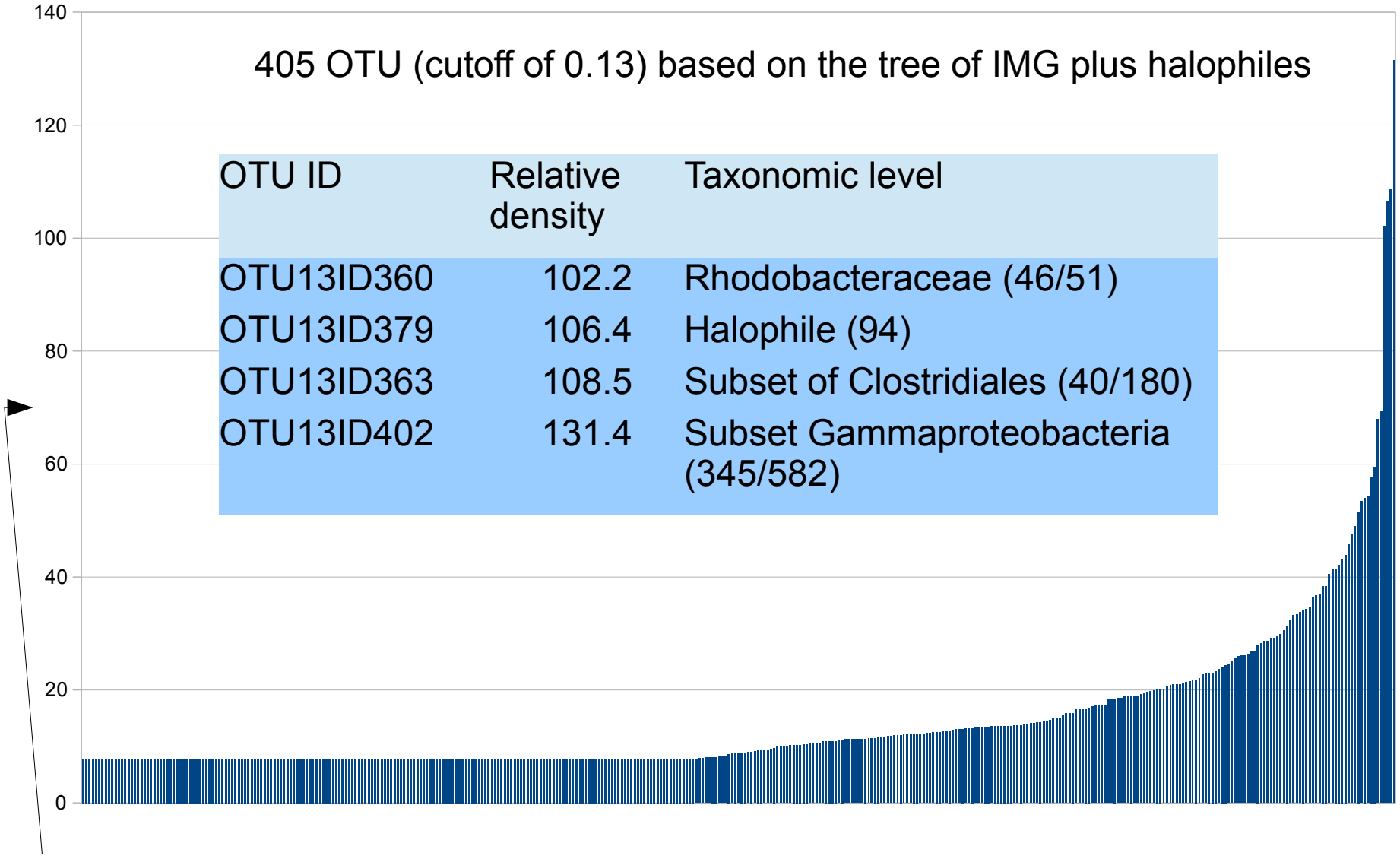


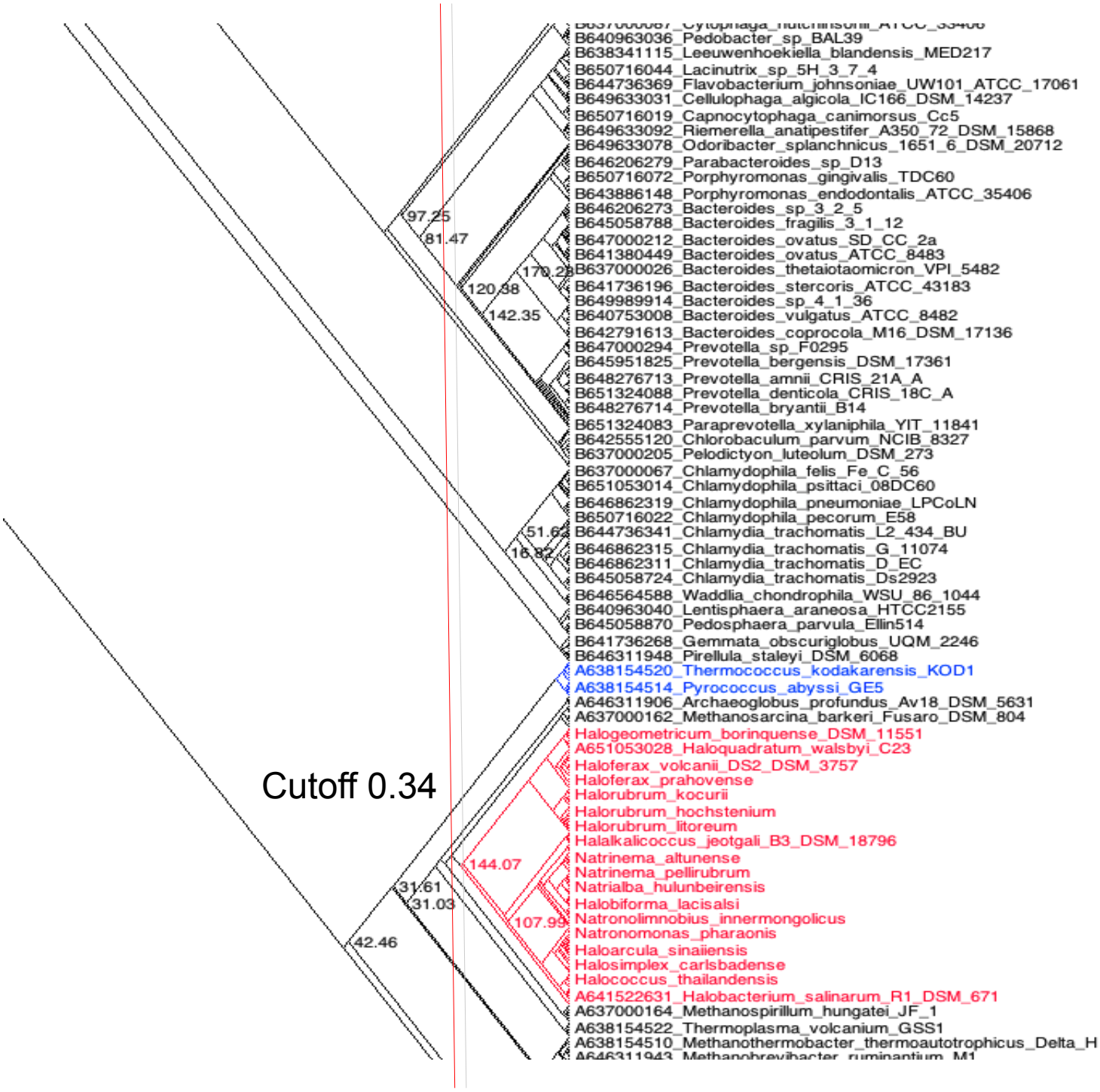
$$Relative\ Density = \frac{(d+r)\times\sum PD}{d^2}$$

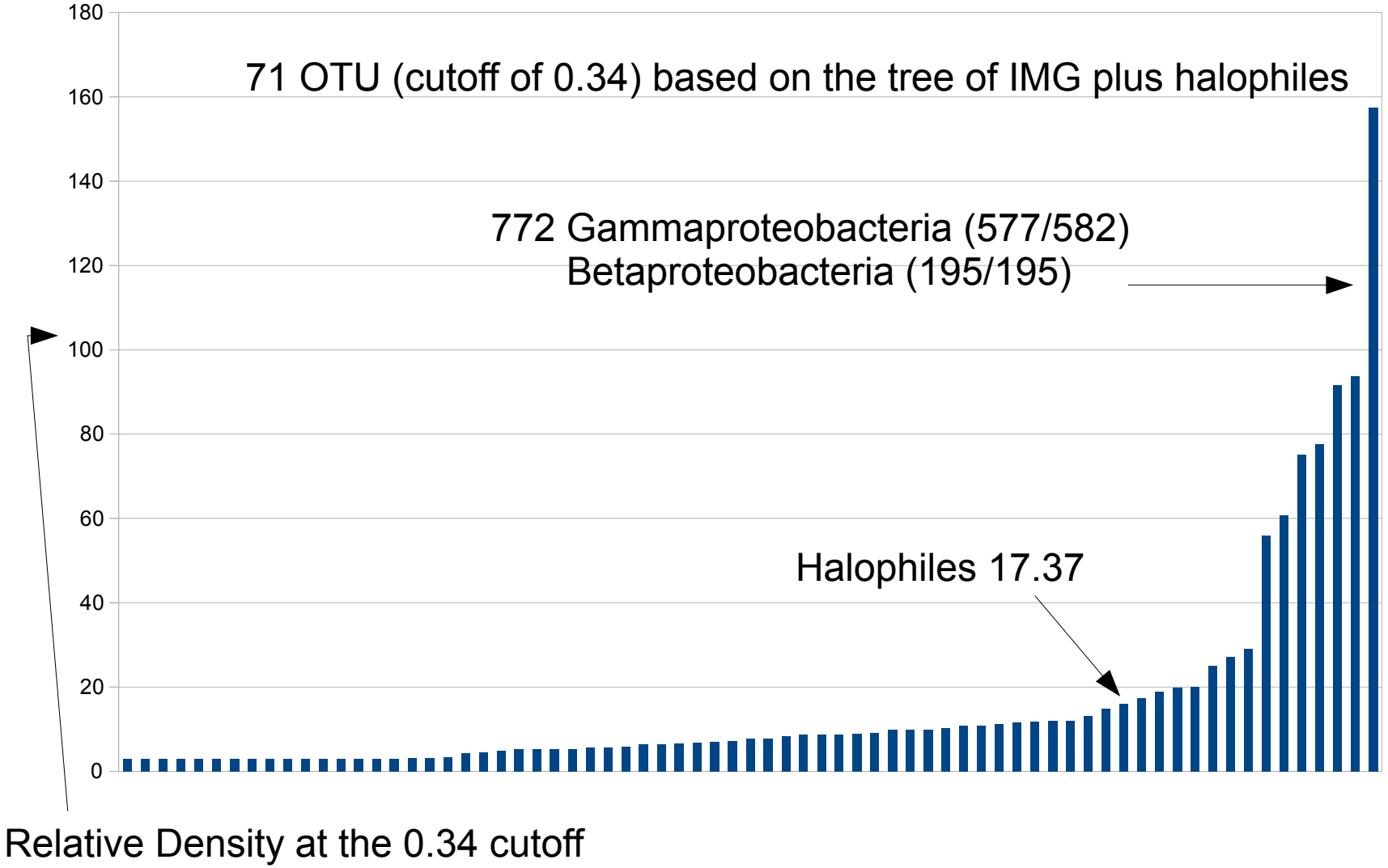
$$Relative\ Density = \frac{(d+r)\times(c-d+\sum PD)}{c^2}$$

Relative Density at a cutoff line


B640963036_Pedobacter_sp_BAL39
B638341115_Leeuwenhoekiella_blandensis_MED217
B650716044_Lacinutrix_sp_5H_3_7_4
B644736369_Flavobacterium_johnsoniae_UW101_ATCC_17061
B649633031_Cellulophaga_algicola_IC166_DSM_14237
B650716019_Capnocytophaga_canimorsus_Cc5
B649633092_Riemerella_anatipestifer_A350_72_DSM_15868
B649633078_Odoribacter_splanchnicus_1651_6_DSM_20712
B646206279_Parabacteroides_sp_D13
B650716072_Porphyromonas_gingivalis_TDC60
B643886148_Porphyromonas_endodontalis_ATCC_35406
B646206273_Bacteroides_sp_3_2_5
B645058788_Bacteroides_fragilis_3_1_12
B647000212_Bacteroides_ovatus_SD_CC_2a
B641380449_Bacteroides_ovatus_ATCC_8483
B637000026_Bacteroides_thetaiotaomicron_VPI_5482
B641736196_Bacteroides_stercoris_ATCC_43183
B649989914_Bacteroides_sp_4_1_36
B640753008_Bacteroides_vulgatus_ATCC_8482
B642791613_Bacteroides_coprocola_M16_DSM_17136
B647000294_Prevotella_sp_F0295
B645951825_Prevotella_bergensis_DSM_17361
B648276713_Prevotella_amnii_CRIS_21A_A
B651324088_Prevotella_denticola_CRIS_18C_A
B648276714_Prevotella_bryantii_B14
B651324083_Paraprevotella_xylaniphila_YIT_11841
B642555120_Chlorobaculum_parvum_NCIB_8327
B637000205_Pelodictyon_luteolum_DSM_273
B637000067_Chlamydophila_felis_Fe_C_56
B651053014_Chlamydophila_psittaci_08DC60
B646862319_Chlamydophila_pneumoniae_LPCoLN
B650716022_Chlamydophila_pecorum_E58
B644736341_Chlamydia_trachomatis_L2_434_BU
B646862315_Chlamydia_trachomatis_G_11074
B646862311_Chlamydia_trachomatis_D_EC
B645058724_Chlamydia_trachomatis_Ds2923
B646564588_Waddlia_chondrophila_WSU_86_1044
B640963040_Lentisphaera_araneosa_HTCC2155
B645058870_Pedosphaera_parvula_Ellin514
B641736268_Gemmata_obscuriglobus_UQM_2246
B646311948_Pirellula_staleyi_DSM_6068
A638154520_Thermococcus_kodakarensis_KOD1
A638154514_Pyrococcus_abyssi_GE5
A646311906_Archaeoglobus_profundus_Av18_DSM_5631
A637000162_Methanosarcina_barkeri_Fusaro_DSM_804
Halogeometricum_borinquense_DSM_11551
A651053028_Haloquadratum_walsbyi_C23
Haloferax_volcanii_DS2_DSM_3757
Haloferax_prahovense
Halorubrum_kocurii
Halorubrum_hochstenium
Halorubrum_litoreum
Halalkalicoccus_jeotgali_B3_DSM_18796
Natrinema_altunense
Natrinema_pellirubrum
Natrialba_hulunbeirensis
Halobiforma_lacisalsi
Natronolimnobius_innermongolicus
Natronomonas_pharaonis
Haloarcula_sinaiiensis
Halosimplex_carlsbadense
Halococcus_thailandensis
A641522631_Halobacterium_salinarum_R1_DSM_671
A637000164_Methanospirillum_hungatei_JF_1
A638154522_Thermoplasma_volcanium_GSS1
A638154510_Methanothermobacter_thermoautotrophicus_Delta_H
97.25
81.47
120.38
142.35
51.62
16.82
144.07
31.61
31.03
107.99
42.46
Cutoff 0.13


140
120
100
80
60
40
20
0
405 OTU (cutoff of 0.13) based on the tree of IMG plus halophiles
OTU ID
Relative density
Taxonomic level
OTU13ID360
102.2
Rhodobacteraceae (46/51)
OTU13ID379
106.4
Halophile (94)
OTU13ID363
108.5
Subset of Clostridiales (40/180)
OTU13ID402
131.4
Subset Gammaproteobacteria (345/582)
Relative Density at the 0.13 cutoff levelc


B640963036_Pedobacter_sp_BAL39
B638341115_Leeuwenhoekiella_blandensis_MED217
B650716044_Lacinutrix_sp_5H_3_7_4
B644736369_Flavobacterium_johnsoniae_UW101_ATCC_17061
B649633031_Cellulophaga_algicola_IC166_DSM_14237
B650716019_Capnocytophaga_canimorsus_Cc5
B649633092_Riemerella_anatipestifer_A350_72_DSM_15868
B649633078_Odoribacter_splanchnicus_1651_6_DSM_20712
B646206279_Parabacteroides_sp_D13
B650716072_Porphyromonas_gingivalis_TDC60
B643886148_Porphyromonas_endodontalis_ATCC_35406
B646206273_Bacteroides_sp_3_2_5
B645058788_Bacteroides_fragilis_3_1_12
B647000212_Bacteroides_ovatus_SD_CC_2a
B641380449_Bacteroides_ovatus_ATCC_8483
B637000026_Bacteroides_thetaiotaomicron_VPI_5482
B641736196_Bacteroides_stercoris_ATCC_43183
B649989914_Bacteroides_sp_4_1_36
B640753008_Bacteroides_vulgatus_ATCC_8482
B642791613_Bacteroides_coprocola_M16_DSM_17136
B647000294_Prevotella_sp_F0295
B645951825_Prevotella_bergensis_DSM_17361
B648276713_Prevotella_amnii_CRIS_21A_A
B651324088_Prevotella_denticola_CRIS_18C_A
B648276714_Prevotella_bryantii_B14
B651324083_Paraprevotella_xylaniphila_YIT_11841
B642555120_Chlorobaculum_parvum_NCIB_8327
B637000205_Pelodictyon_luteolum_DSM_273
B637000067_Chlamydophila_felis_Fe_C_56
B651053014_Chlamydophila_psittaci_08DC60
B646862319_Chlamydophila_pneumoniae_LPCoLN
B650716022_Chlamydophila_pecorum_E58
B644736341_Chlamydia_trachomatis_L2_434_BU
B646862315_Chlamydia_trachomatis_G_11074
B646862311_Chlamydia_trachomatis_D_EC
B645058724_Chlamydia_trachomatis_Ds2923
B646564588_Waddlia_chondrophila_WSU_86_1044
B640963040_Lentisphaera_araneosa_HTCC2155
B645058870_Pedosphaera_parvula_Ellin514
B641736268_Gemmata_obscuriglobus_UQM_2246
B646311948_Pirellula_staleyi_DSM_6068
A638154520_Thermococcus_kodakarensis_KOD1
A638154514_Pyrococcus_abyssi_GE5
A646311906_Archaeoglobus_profundus_Av18_DSM_5631
A637000162_Methanosarcina_barkeri_Fusaro_DSM_804
Halogeometricum_borinquense_DSM_11551
A651053028_Haloquadratum_walsbyi_C23
Haloferax_volcanii_DS2_DSM_3757
Haloferax_prahovense
Halorubrum_kocurii
Halorubrum_hochstenium
Halorubrum_litoreum
Halalkalicoccus_jeotgali_B3_DSM_18796
Natrinema_altunense
Natrinema_pellirubrum
Natrialba_hulunbeirensis
Halobiforma_lacisalsi
Natronolimnobius_innermongolicus
Natronomonas_pharaonis
Haloarcula_sinaiiensis
Halosimplex_carlsbadense
Halococcus_thailandensis
A641522631_Halobacterium_salinarum_R1_DSM_671
A637000164_Methanospirillum_hungatei_JF_1
A638154522_Thermoplasma_volcanium_GSS1
A638154510_Methanothermobacter_thermoautotrophicus_Delta_H
97.25
81.47
120.38
142.35
51.62
16.82
144.07
31.61
31.03
107.99
42.46
Cutoff 0.34


180
160
140
120
100
80
60
40
20
0
71 OTU (cutoff of 0.34) based on the tree of IMG plus halophiles
772 Gammaproteobacteria (577/582)
Betaproteobacteria (195/195)
Halophiles 17.37
Relative Density at the 0.34 cutoff

Summary of the halophile sample density study:

Based on a fasttree built from all IMG bacteria/archea with halophile samples, halophile group has a relative density of 106.4 at the 87% relative depth level from the root. It is one of the top 4 best sampled group at this level. From the level of 87% to 66% towards the root, the relative sample density decreases to 17.4 because no more halophile genomes are added.

Test Run of OTU grouping on Merlot: compute-5-10

1. OTU grouping

Input: Greengenes fasttree of 399,817 taxa (mid point rooted)
Cutoff: normalized branch length of 0.10 from the tip

Naive depth calculation:
Start: Mon Jul 9 12:35:21 PDT 2012
End: Mon Jul 9 12:39:49 PDT 2012

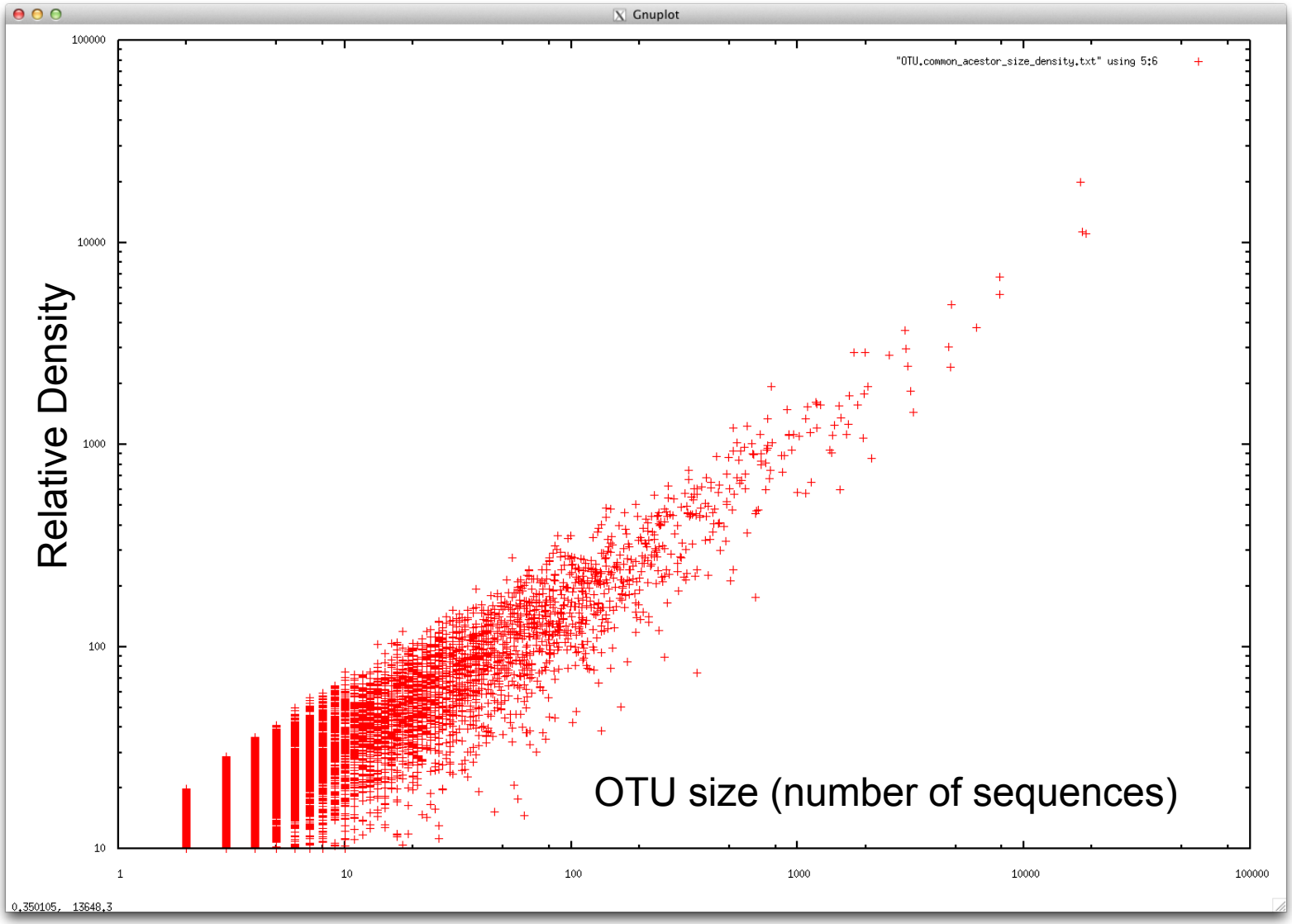
Weighted depth calculation:
Tue Jul 10 15:33:01 PDT 2012
Tue Jul 10 15:37:29 PDT 2012

399,817 sequences have been grouped into 52,345 OTUs by naïve average depth method
Using the same cutoff, they are grouped into 69,339 OTUs by weighted average depth method

2.Calculate Density at the cutoff of 0.10 from tip for all the OTU groups

Relative density of OTUs at the 0.10 relative depth level from the tips

Density	OTU count	ssu-rRNA count
10(one seq per OTU)	38781	39453
11-20	7483	21084
21-50	4129	34705
51-100	1112	32180
101-200	477	34626
201-500	254	50525
501-1000	61	37722
1001-2000	33	43390
2001-5000	10	35688
5001-10000	2	15641
10000-20000	3	54803


Gnuplot
"OTU.common_acestor_size_density.txt" using 5:6
Relative Density
OTU size (number of sequences)
100000
10000
1000
100
10
1
10
100
1000
10000
100000
0.350105, 13648.3

The most densely sample OTUs (0.10 cutoff)

OTU52330 ssu-rRNA count:17842 Relative Density:19864.4

201 Unclassified:OTU
10601 k__Bacteria:p__Firmicutes:c__Bacilli:o__Bacillales:f__Staphylococcaceae:g__Staphylococcus:s__Staphylococcus:OTU
7037 k__Bacteria:p__Firmicutes:c__Bacilli:o__Bacillales:f__Staphylococcaceae:g__Staphylococcus:Unclassified:OTU

OTU52315 ssu-rRNA count:18153 Relative Density:11353.3

1190 Unclassified:OTU
16376 k__Bacteria:p__Actinobacteria:c__Actinobacteria:o__Actinomycetales:f__Propionibacteriaceae:g__Propionibacterium:s__Propionibacterium:OTU
587 k__Bacteria:p__Actinobacteria:c__Actinobacteria:o__Actinomycetales:f__Propionibacteriaceae:g__Propionibacterium:Unclassified:OTU

OTU52336 ssu-rRNA count:18808 Relative Density:11097.1

9 k__Bacteria:p__Actinobacteria:c__Actinobacteria:o__Actinomycetales:f__Corynebacteriaceae:g__Corynebacterium:s__Brevibacterium:OTU
469 k__Bacteria:p__Actinobacteria:c__Actinobacteria:o__Actinomycetales:f__Corynebacteriaceae:g__Corynebacterium:s__Corynebacterium:OTU
14561 k__Bacteria:p__Actinobacteria:c__Actinobacteria:o__Actinomycetales:f__Corynebacteriaceae:g__Corynebacterium:Unclassified:OTU
3768 Unclassified:OTU

** not all greengenes entries have Hugenholtz taxonomy assignments*

Future direction: An Unifrac alternative

1. OTU grouping reference trees at different cutoff level (**multi-level**)
2. Map metagenomic sequences on the trees
3. Calculate the relative density for the metagenomic subtrees
4. generate density matrix and compare different samples (**intuitive**)