

1) **Start:** with N partial genomes and the FASTA files containing their genes (amino acid sequences). We also have metadata and annotations for each gene.

#	Name	Source	# Genes	Family	Partial
G1	AAA288-N07	Freshwater (LD12)	560	1	yes
G2	AAA028-D10	Freshwater (LD12)	478	1	yes
G3	AAA023-L09	Freshwater (LD12)	609	1	yes
G4	HIMB140	Marine (SAR11)	1091	2	yes
G5	HTCC9565	Marine (SAR11)	951	3	no
G6	HIMB083	Marine (SAR11)	980	3	no

LD12 genomes analysis pipeline

Lucas Sinclair, lucas.sinclair@me.com

Last modified:

Tue Nov 18 2014

All numbers are fake.
This document
describes a plan, it is
not scientific work.



2) **All against all BLAST:** Perform a sequence similarity search of all genes against all genes. With a e-value cutoff of 0.001

#	g1	g2	g3	g4	g5	g6	...
g1	-	1	89	0	112	1	...
g2	1	-	1	1	1	0	...
g3	45	1	-	1	44	1	...
g4	2	1	1	-	1	1	...
g5	87	0	12	1	-	1	...
g6	1	1	1	1	0	-	...
..

3) **Filter the BLAST hits:** Any hit with a minimum identity of 30% and a minimal coverage of 50%

#	g1	g2	g3	g4	g5	g6	...
g1	-	1	89	0	112	1	...
g2	1	-	1	1	1	0	...
g3	45	1	-	1	44	1	...
g4	2	1	1	-	1	1	...
g5	87	0	12	1	-	1	...
g6	1	1	1	1	0	-	...
.

4) **Compte synteny scores:**
Come up a matrix of synteny distances

#	g1	g2	g3	g4	g5	g6	.
g1	-	1	89	0	112	1	...
g2	1	-	1	1	1	0	...
g3	45	1	-	1	44	1	...
g4	2	1	1	-	1	1	...
g5	87	0	12	1	-	1	...
g6	1	1	1	1	0	-	...
.

5) **Merge both matrices:** Using a elitic-type weighing equation (if this slips is a circle we are just doing $0.5 * M1 + 0.5 * M2$)

#	g1	g2	g3	g4	g5	g6	...
g1	-	1	89	0	112	1	...
g2	1	-	1	1	1	0	...
g3	45	1	-	1	44	1	...
g4	2	1	1	-	1	1	...
g5	87	0	12	1	-	1	...
g6	1	1	1	1	0	-	...
.

6) **Cluster:** Using the MCL clustering, input the distance matrix and get a list of clusters of genes.

In a cluster do they all have the same annotations ?

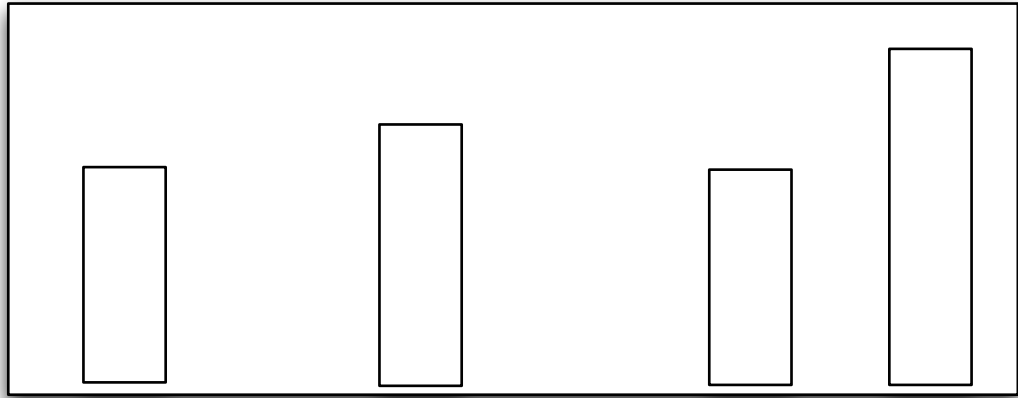
7) **Count within genomes:** Given each cluster, how many of its members are associated with each genome

8) **Choose some cluster that we like:** Have at least one representative in each family and no more than 29 genes in the cluster.

9) **Take only one gene per genome:** Since we are allowing some cluster to contain two or more genes pertaining to the same genome, we filter out extraneous genes, keeping one randomly.

10i) **With the genes of a cluster:** Align each protein with muscle, run gblocks on the result, run that into RaxML to make a tree.

Extra: Taking columns such C3 where we have only freshwater genomes, BLAST against "refseq+marine" database. And check how many hits are to SAR11, alphaproteobacteria, proteobacteria, archea ?



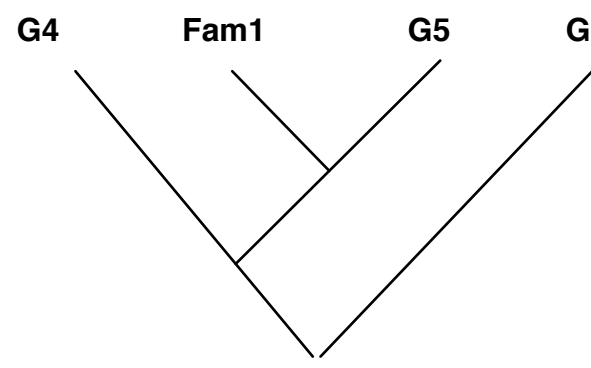
#	L29	S17	Rib3	RIb
G1	1	1	1	1
G2	1	1	0	1
G3	1	0	1	1
G4	1	1	1	1
G5	1	1	1	1
G6	1	1	1	1

1B) **Extract ribosomal genes:** based on the annotations. This will act as the reference tree, these are supposed to be the less likely genes to be exchanged.

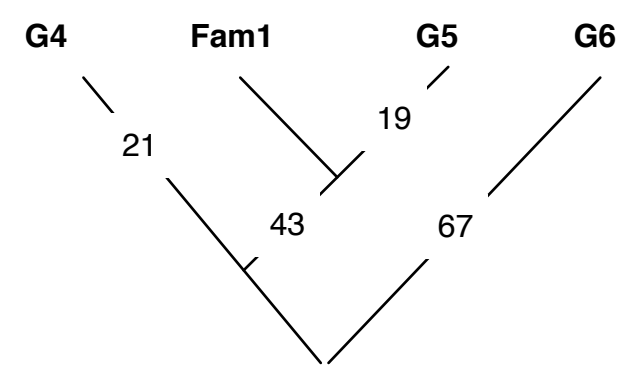
Multiple sequence aligement	
P1	KVFGRCELAAMKRHGLDNYRGYSLGNWVC----
P3	-----AAKFESNFNTQATNRNTDGSTDYGI--
P4	RWWCN--DGRTPGSRNLCN-----IPCSAL---IT
P5	IWWCAK-----KITPDGNGMNAWVAS--
G3	PRAA-----AAAERLWSNFF-----EGTFFIAA
G6	-----YVVWQEVFDNKKVKVRP-----DTIIQVE
G8	-----RLWPRAGAVAERLWSSNLTNTNIDFRL
H1	-----AAAM--
H2	--KVFGRCCELAAM-----
H4	--EVD-----FTCW-----KS-----NP---
H5	-----AGFRALLSAPW-----
H6	--QKT-----NRNTDGSTD--
H7	-----RWWCNDGRT-----PGS---

Using all ribosomal genes

2B) **Align each ribosomal protein cluster separately:** run gblocks on the result, then concatenate each cluster together into one file, run that into RaxML



3B) **Build a separte tree:** the tips can be mixed, if we didn't find a specific ribosomal protein in one given genome, we can merge it with another genome of the same family and indicate the family on the tree leaf



4B) **Build a separte tree:** Add the bootstrap. See dendropy.sumtrees

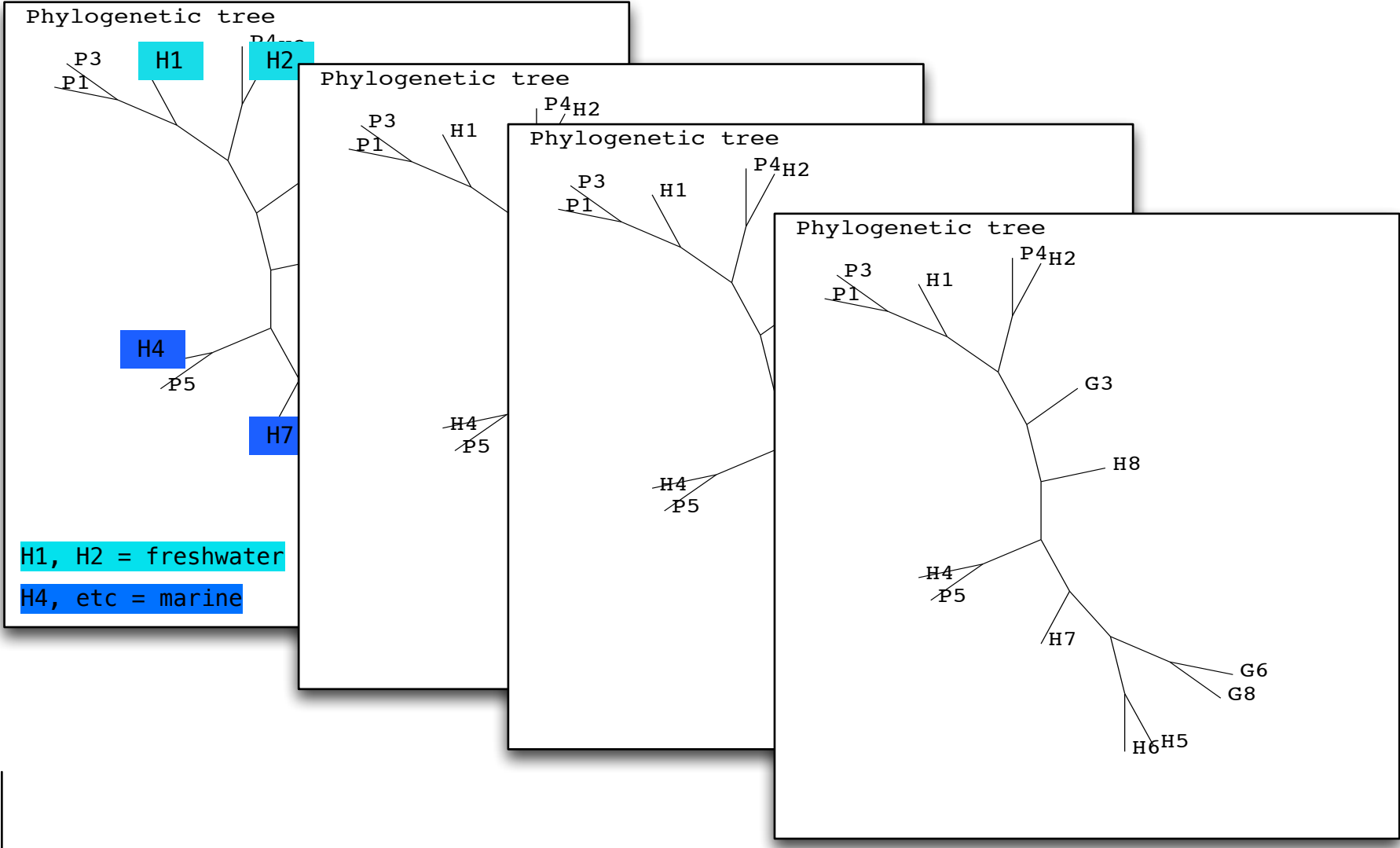
Finally: For every tree (from clusters) compare with the ribosomal tree.

For every tree compare with the reference tree: How well does the structure correspond ? How often do we find the same split between the families ? When we don't fin the same split, which families are miss-spiting ?

1) Check that on the cluster trees the families are grouping together, either kick the ones that don't out, or report how many times a certain split is not conserved (uncollapsible)

2) The trees that collapse nicely compare it with master tree setting the root at V and checking that the same branching patterns exists.

3) If there is a popular alternative branching pattern, draw it in the supplementary and add as numbers how many clusters had this exact alternative branching.



Bacteria Proteobacteria Family1 Family2