

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:

Fri Oct 10 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	-	...
..

#	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.

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	-	...
.

Multiple sequence aligement	
P1	KVFGRCELAAAMKRHGLDNYRGYSLGNWVC----
P3	-----AAKFESNFNTQATNRNTDGSTDYGI--
P4	RWWCN--DGRTPGSRNLCN-----IPCSAL---IT
P5	IWWCAK-----KITPDGNGMNAWVAS--
G3	PRAA-----AAAERLWSNFF-----EGTFFIAA
G6	-----YVVWQEVFDNKVKVVRP-----DTIIQVE
G8	-----RLWPRAGAVAERLWSSNLTNTIDFRL
H1	-----AAAM--
H2	--KVFGRCELAAAM-----
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

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 the 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	-	...
.

Clusters	
C1	= g1+g3+g5
C2	= g2+g4
C3	= g6+g7+g8+g9
ETC...	

#	C1	C2	C3	C4
G1	1	1	1	1
G2	1	1	1	1
G3	1	1	1	1
G4	1	2	0	2
G5	1	1	0	3
G6	2	0	0	3

#	C1	C2
G1	1	1
G2	1	1
G3	1	1
G4	1	2
G5	1	1
G6	2	0

#	C1	C2
G1	1	1
G2	1	1
G3	1	1
G4	1	1
G5	1	1
G6	1	0

Current bottleneck

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

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

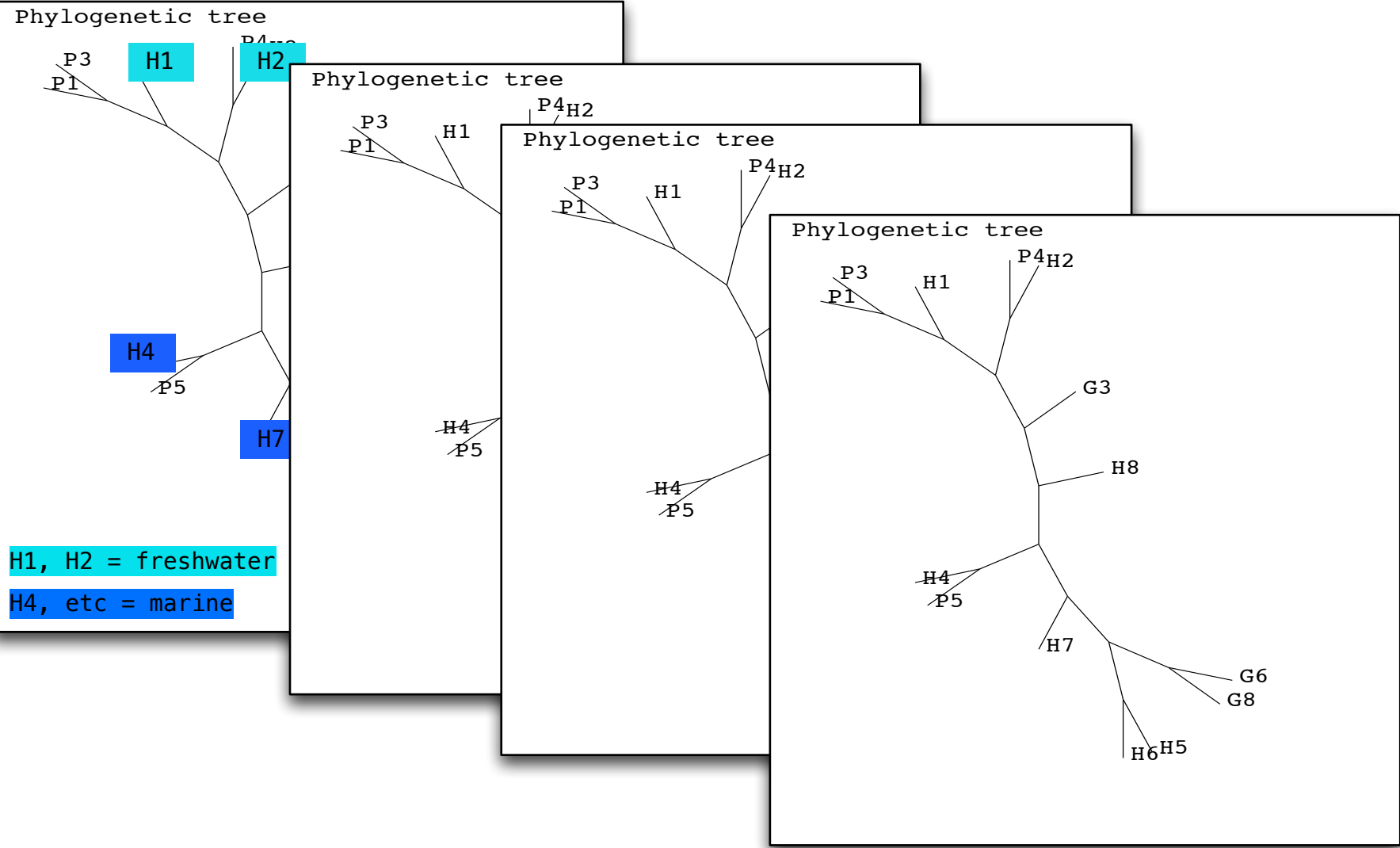
8) **Choose some cluster that we like:** using a simple scoring method that penalizes partial genomes less, we pick the 50 clusters that most likely are "single-copy" genes

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

Check that the families group together. Report those that don't.

Taking columns such C3: where a protein family exists in the freshwater, BLAST against "refseq+marine" database. How many hits are to SAR11, alphaproteobacteria, proteobacteria, archaea.



Final) For every tree (from clusters) compare with the ribosomal tree: How well does the structure correspond. How often do we find the same split between the families.

3B) **Build a sepearte 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