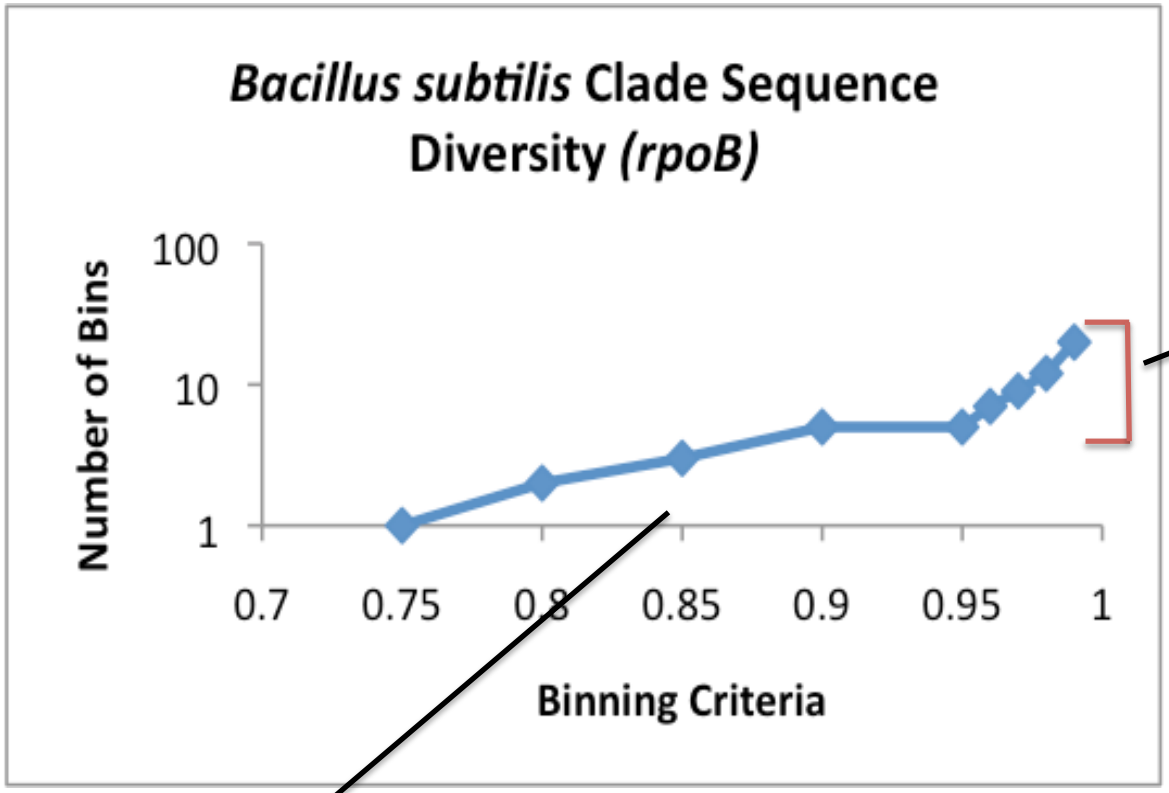
Anatomy of a Clade Sequence Diversity Curve (Interpretations from Stable Ecotype Model)

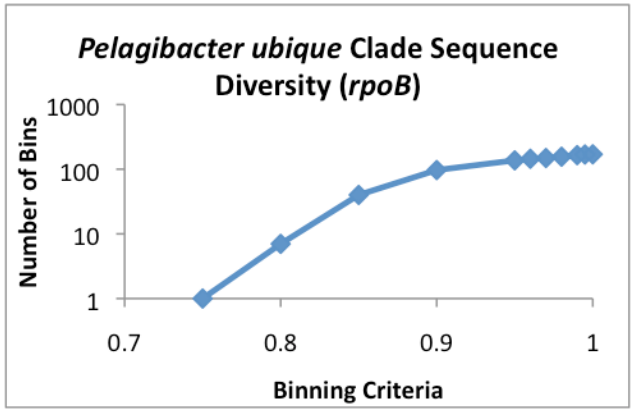


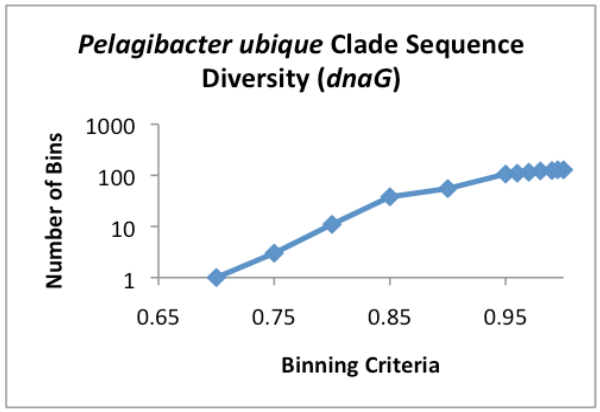
Flare-up at high identity criteria levels:
Represents temporary diversity that will be purged by periodic selection. Slope and position of flare-up helps ES determine the rate of periodic selection.

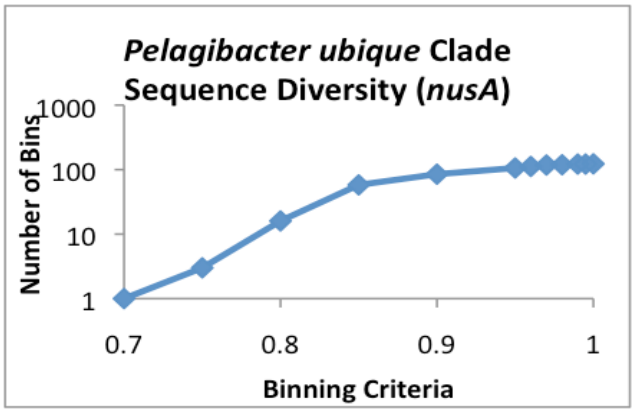
Section of (approximate) log-linear increase: Represents the net accumulation of ecotypes in the clade over time. Helps ES estimate the rate of ecotype formation and the number of ecotypes in the clade.

Clade Sequence Diversity Curves for *Pelagibacter ubique* sequences from GOS Data

Three different genes. No flare-ups.







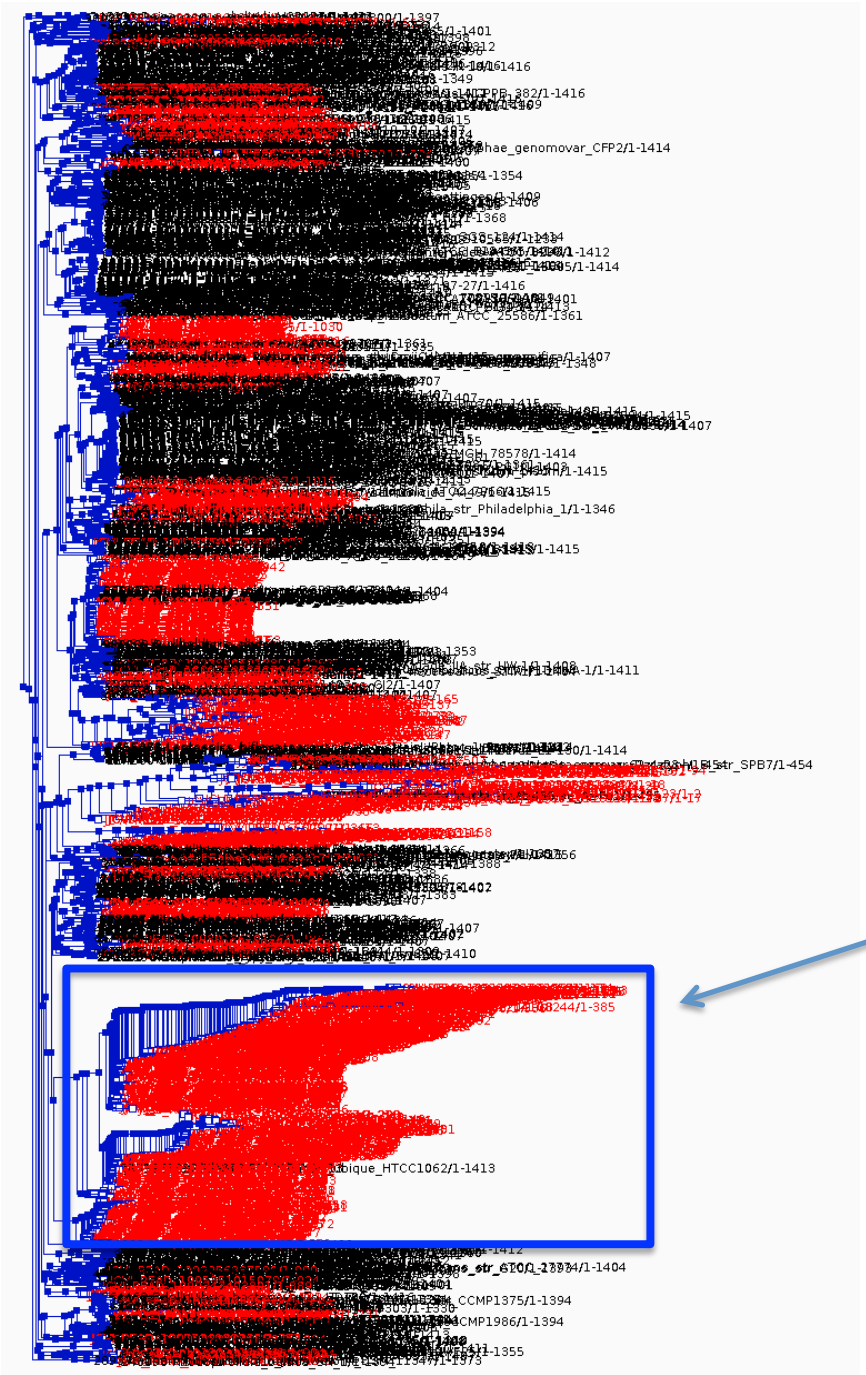
Difference in diversity signature is reflected in ES Parameter Solution Estimates:

					npop			omega			sigma		
Taxon	**Gene**	**# Seqs**	**Length**	**ML**	**lower**	**upper**	**ML**	**lower**	**upper**	**ML**	**lower**	**upper**
B. simplex	*gapA*	118	635	11	3	40	0.2	0.042	0.44	3.55	0.24	>100
	rpoB	130	871	34	9	79	0.43	0.2	0.93	54.16	1.14	>100
	uvrA	121	614	10	3	65	0.29	0.061	0.63	1.28	0.084	>100
B. sub/B. lich	*gapA*	77	426	10	6	20	0.083	0.038	0.18	2.11	0.44	31.33
	gyrA	88	539	20	13	27	0.049	0.036	0.066	2.49	0.12	6.25
	rpoB	88	509	12	8	19	0.1	0.05	0.158	8.47	1.34	>100
P. ubiq. (SC1)	*rpoB*	175	570	31	1	175	0.043	-	-	0.01	-	-
P. ubiq. (SC2)	*rpoB*	144	660	144	2	144	0.059	-	-	0.025	-	-
P. ubiq. (SC2)	*rpoB*	91	747	2	2	91	0.056	-	-	0.01	-	-
P. ubique	*dnaG*	136	471	136	35	136	0.069	0.047	0.069	0.047	0.001	>100
P. ubique	*nusA*	126	816	126	2	126	0.035	0.035	0.035	0.017	8E-04	9.26

Sigma:
Rate of Periodic Selection.
Estimates ~100x Lower than for *Bacillus* taxa

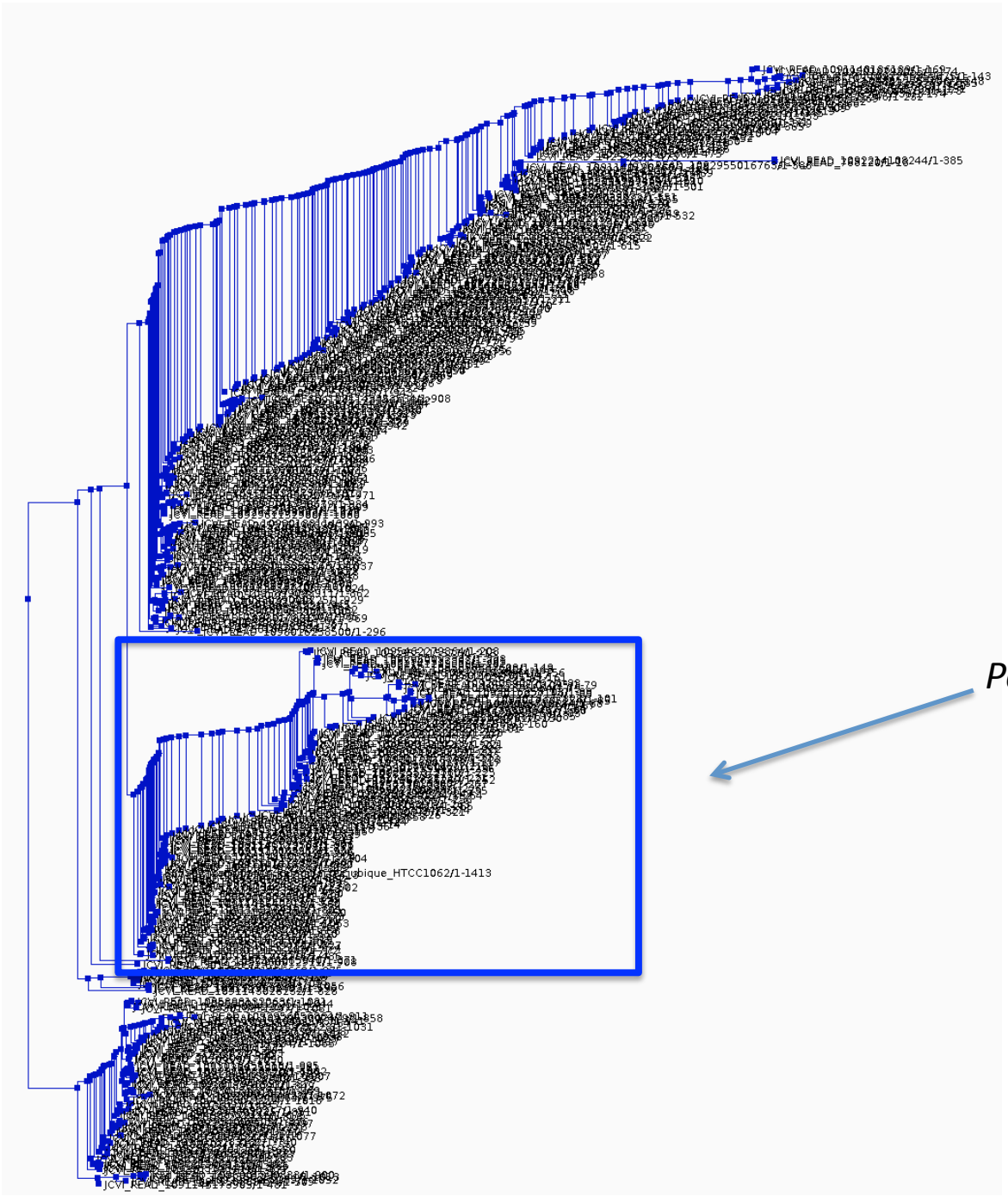
Npop: Number of Ecotypes. Not getting any sort of precise estimate. 95% CI's range from 1 ecotype to each sequence representing it's own ecotype.

Omega: Rate of Ecotype Formation. Estimates Comparable to *Bacillus* taxa.



Phylogeny of ComboDB 16s Markers (black) + GOS reads that were BLAST hits to *Pelagibacter* 16s with >99% sequence identity (red).

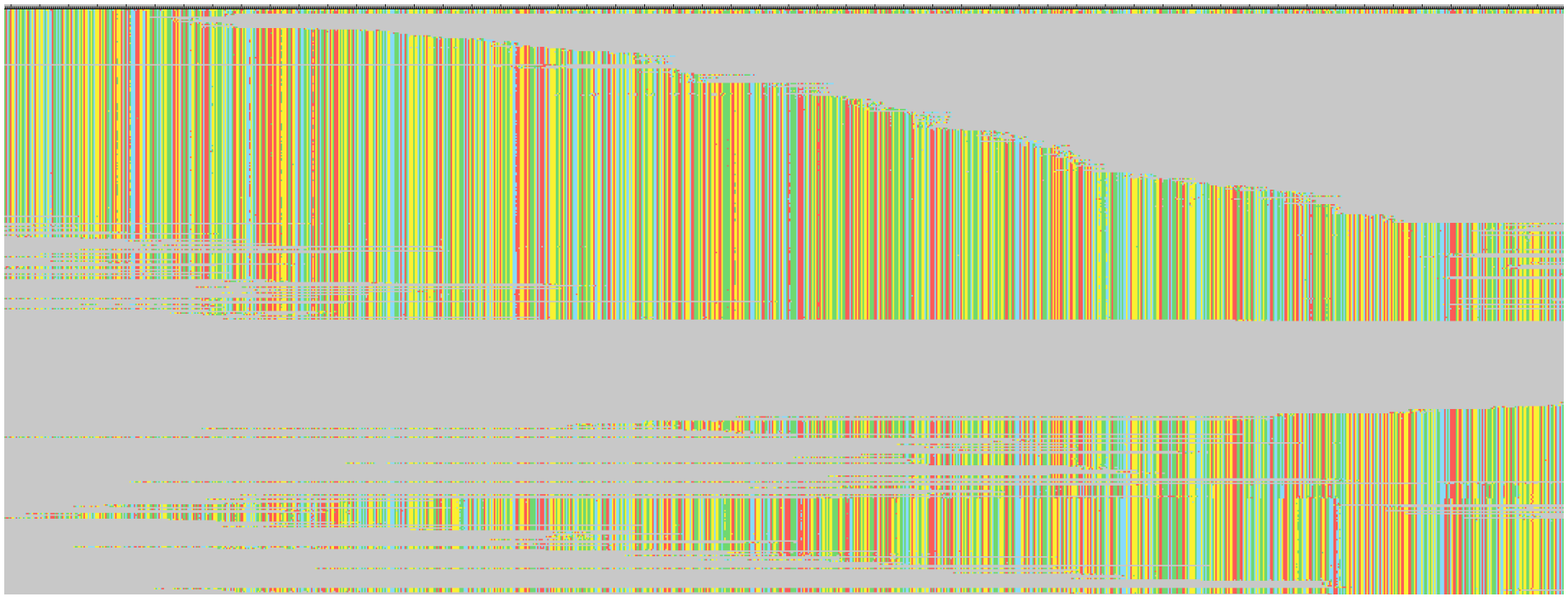
Pelagibacter Clade



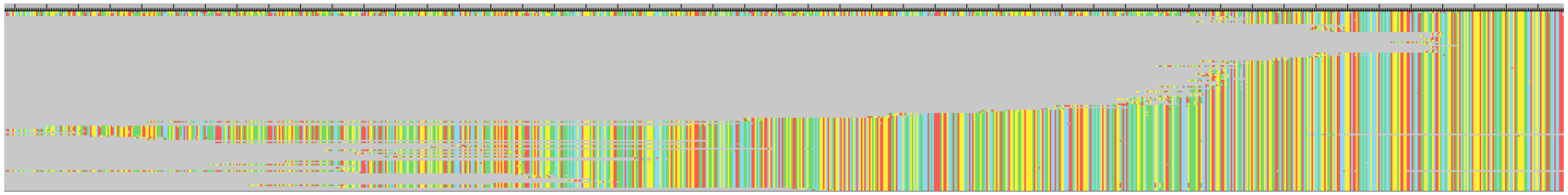
Same Tree, but zoomed in on *Pelagibacter* clade

Pelagibacter Subclade

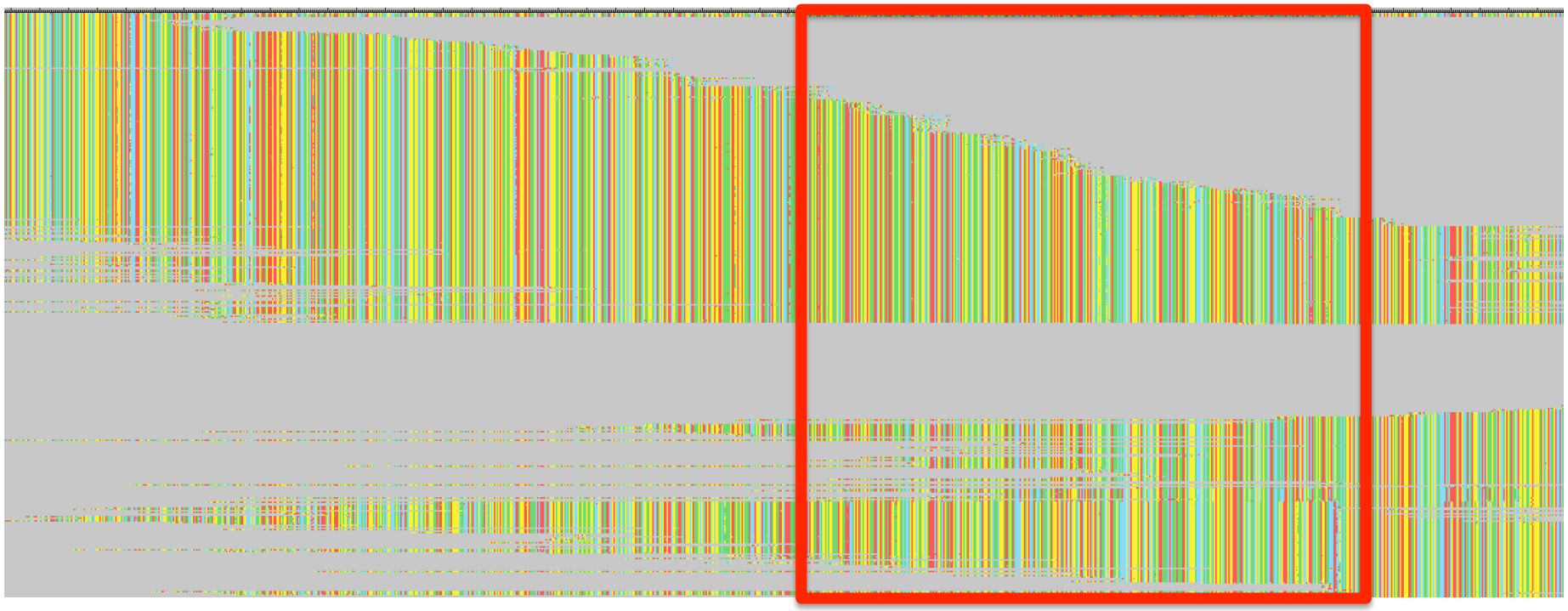
Alignment of whole *Pelagibacter* clade

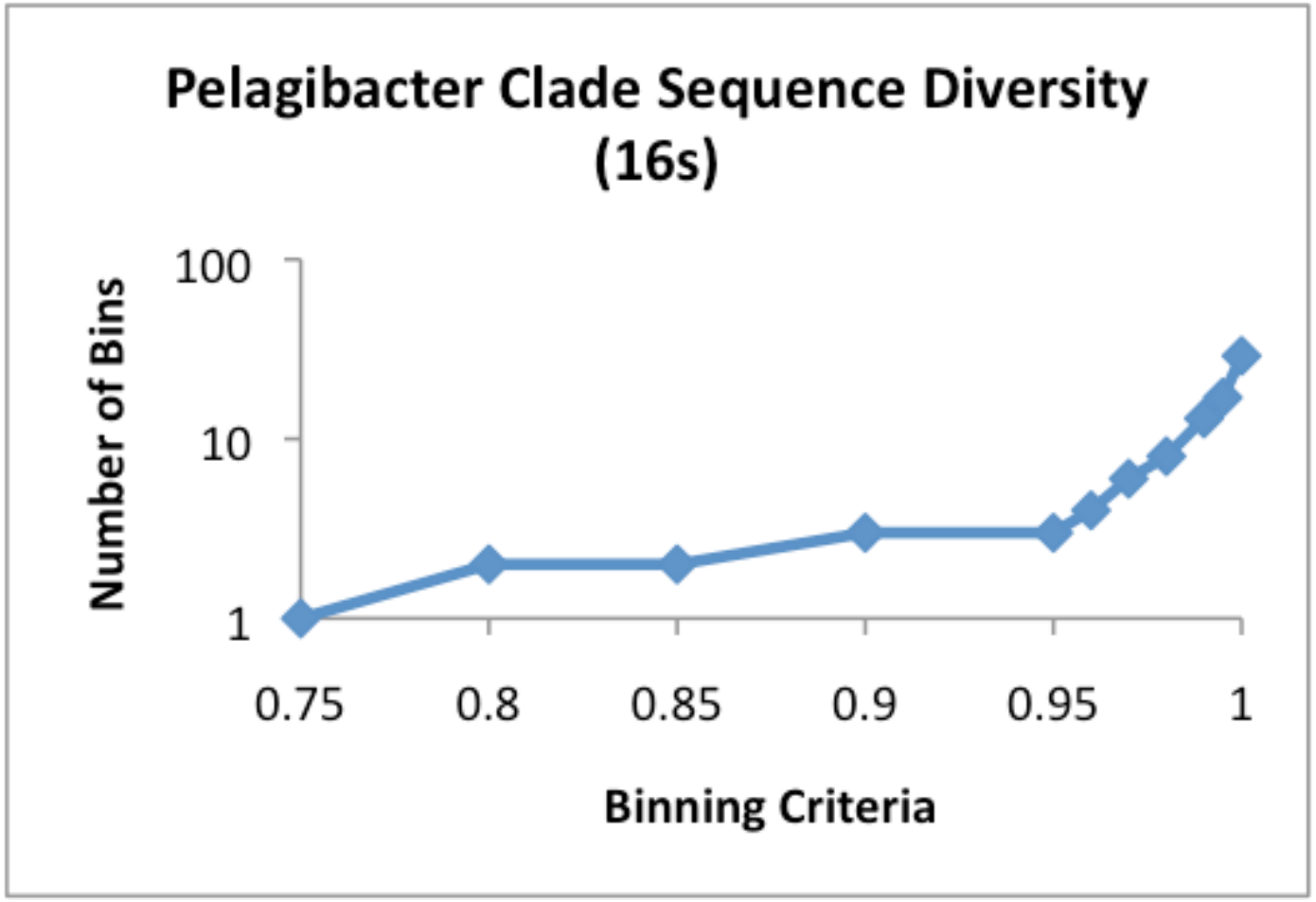


Alignment of just *Pelagibacter* subclade



Cut out a section (~500bp) with decent representation from both.





						npop			omega			sigma	
Taxon	**Gene**	**# Seqs**	**Length**	**Criteria**	**ML**	**lower**	**upper**	**ML**	**lower**	**upper**	**ML**	**lower**	**upper**
P. ubique	16s	113	479	2x	12	3	40	0.074	0.003	0.17	3.31	0.0071	>100

So why can I get 16s to work, but not the protein-coding genes?

1) Is it a sampling issue?
   a. Depth of sampling. All 16s "types" are well represented in the sample, but each rpoB "type" has only one (or a few) representatives, so they appear to be much more distantly related.
      i. GOS: Is there enough sampling at each location to be able resolve ecotype-scale diversity using protein-coding genes?
      ii. Methods: Is the sampling there, but I'm having to cut too many sequences out of each alignment?
   b. Potential Solutions.
      i. Try a different taxon (*Prochlorococcus*?). Same environment, same sampling, Do we see the same issue with protein-coding genes? Could help rule out sampling issues.

2) Is it an issue with Ecotype Simulation?
   a. ES assumes homogeneity of rates (of periodic selection and ecotype formation) over time within the sampled clade. Maybe we're backing up too far, and need to zoom in on more closely related groups. (is 16s the best we can do for now for using ES on *Pelagibacter*?)
   b. The clade sequence diversity curve as currently set up puts extra weight on recent substitutions. Perhaps adding more point to the curve would allow ES to get better estimates.