R-010

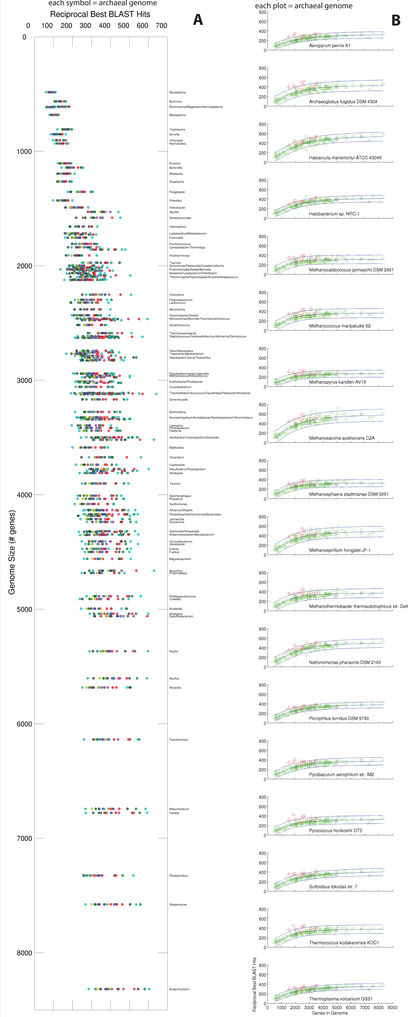
Vertical Descent or Lateral Transfer? Unravelling the Large Number of Whole-Genome Reciprocal BLAST Hits between Anaerobic, Thermophilic Bacteria and Archaea

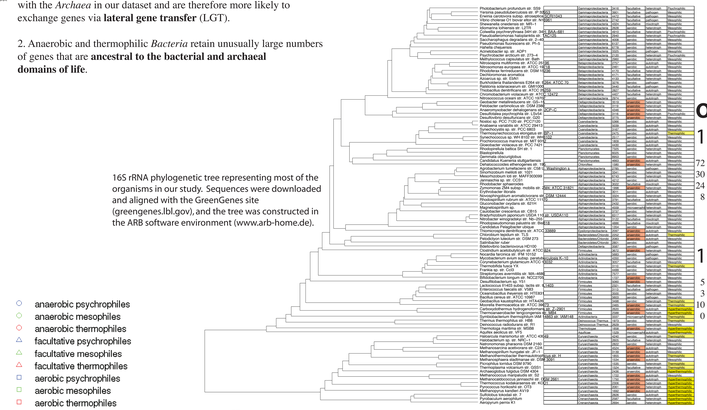
C. A. Fuchsman, W. J. Brazelton, R. E. Collins, M. C. Horner-Devine, and **G. Rocap,** College of Ocean and Fishery Sciences, University of Washington

School of Oceanography, University of Washington — Astrobiology, At the University of Washington

Contact Information:
Gabrielle Rocap — rocap@ocean.washington.edu
Billy Brazelton — braz@ocean.washington.edu
School of Oceanography, University of Washington, Seattle, WA 98195

Figure 1. Reciprocal Best BLAST Hits between Bacteria and Archaea





16S rRNA phylogenetic tree representing most of the organisms in our study. Sequences were downloaded and aligned with the GreenGenes site (greengenes.lbl.gov), and the tree was constructed in the ARB software environment (www.arb-home.de).

Abstract

Whole genome comparisons using reciprocal BLAST hits (RBHs) can provide a measure of the number of shared genes between two genomes, even across domains. Reciprocal BLAST, however, does not distinguish between genes that are shared because of evolutionary vertical descent and those shared due to lateral gene transfer. In a reciprocal BLAST dataset of pairwise comparisons among 132 Bacterial genomes and 18 Archaeal genomes, thermophilic and anaerobic Bacteria had significantly more reciprocal BLAST hits to Archaeal genomes than did other Bacteria when adjusted for genome size. Several lines of evidence indicate that the Last Common Ancestor may have evolved in a thermophilic, anaerobic environment -- environmental conditions in which many of the Archaea used in our dataset are found today. We initially tested the hypothesis that the genomic relationship was explainable by vertical descent (and not lateral gene transfer) by statistically comparing a distance matrix based on numbers of RBHs among Bacteria and Archaea to a 16S rRNA distance matrix, and found a significant correlation, though we also found a significant correlation between the RBH distance and a 'phenotypic' distance matrix generated using environmental parameters of temperature, oxygen requirements, and habitat. To further investigate these findings we sorted RBHs into clusters of orthologous groups (COGs) to determine which COGs were driving the genomic relationships. In general, COGs involved in 'core processes' (e.g. replication and transcription) better explained the relationships than did COGs considered to be more susceptible to lateral gene transfer (e.g coenzyme metabolism and transport), though important exceptions were evident. Phylogenetic analyses of individual genes from influential COGs that are more frequently shared between thermophilic or anaerobic Bacteria and Archaea will provide a higher resolution view of the combined influences of vertical descent and lateral gene transfer on microbial evolution.

1. Introduction

One way to identify orthologous genes in phylogenetically diverse genomes is by the reciprocal best BLAST hits (RBHs) between the two genomes (Snel *et al.*, 1999). In other words, a pair of orthologs can be operationally defined as two genes that are each others' best BLAST hit when performing a pairwise whole genome BLAST search. This method excludes paralogs by requiring the BLAST hits to be reciprocal, and it identifies orthologs in diverse genomes by comparing genomes pairwise rather than by querying a multiple-genome database. Furthermore, the pairwise comparison is less susceptible to taxonomic biases in the sequence databases.

Our previous study (Fuchsman and Rocap, 2006) reported that the number of RBHs between bacterial and archaeal genomes increases linearly with the number of genes in the bacterial genome until a plateau at approximately 4,000 genes (**Figure 1**). When the overall plot is partitioned into multiple plots each containing only one archaeal genome (the right-hand column), it is clear that anaerobic and thermophilic Bacteria are often above the curve representing the average number of RBHs for a bacterial genome of given size. In other words:

Anaerobic and thermophilic *Bacteria* often have more reciprocal best BLAST hits (RBHs) with *Archaea* than would be expected from their genome size.

This study explores ecological and evolutionary explanations of this observation. Specifically, we investigate two alternative hypotheses:

1. Anaerobic and thermophilic *Bacteria* inhabit similar environments with the *Archaea* in our dataset and are therefore more likely to exchange genes via **lateral gene transfer** (LGT).

2. Anaerobic and thermophilic *Bacteria* retain unusually large numbers of genes that are **ancestral to the bacterial and archaeal domains of life**.

2. Adjusting for Genome Size

To further explore the relationship between anaerobic, thermophilic Bacteria and Archaea, we need to quantitatively account for genome size when considering reciprocal best BLAST hits (RBHs).

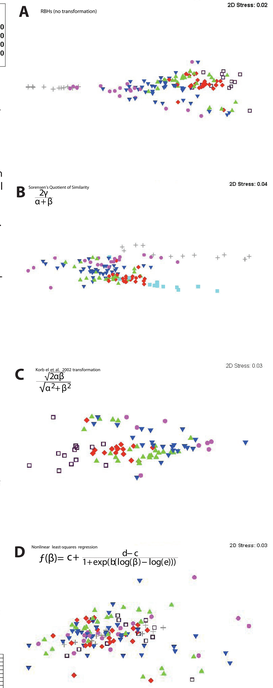
Figure 2A) The primary determinant of number of RBHs is the number of genes in the bacterial genome. Genome size class of each bacterium is shown here in a non-metric MDS plot of RBHs of each bacterial genome with the total set of archaeal genomes.

Figure 2B&C) Neither of two published methods satisfactorily removed the genome size artifact. Of the various methods put forth to account for genome size, we tested two for their ability to remove the genome size signal from the dataset: Sorensen's quotient of similarity (B) and the weighted average transformation of Korbel et al., 2002 (C).

The number of RBHs can be approximated by a 4-parameter logistic function. The number of RBHs between genomes increases to a plateau at about 4000 genes (**Figure 1A**) in a manner that can be approximated by a 4-parameter logistic function. For each archaeal genome, we performed nonlinear least-squares regressions of the number of RBHs between that archaeon and the total set of bacteria (shown in **Figure 1B**). From this best fit line we calculated the mean number of RBHs for each archaeon and a bacterial genome of arbitrary size (as well as 95% confidence intervals around that mean). Subtracting the mean from the RBHs we obtained the number of 'residual' RBHs for each bacterial genome against each archaeal genome.

Figure 2D) The 'residuals' method successfully removes the signal from bacterial genome size. The non-metric MDS plot shows no influence of bacterial genome size on RBHs, which was also statistically corroborated using an ANOSIM (Table 1).

Key for equations:
α is the number of genes in the archaeal genome, β is the number of genes in the bacterial genome, and γ is the number of RBHs between them.



Our dataset:

134 bacterial genomes:

72 aerobic	118 mesophilic	55 heterotrophic
30 facultative	7 thermophilic	48 pathogenic
24 anaerobic	5 psychrophilic	21 autotrophic
8 microaerophilic	4 hyperthermophilic	6 endosymbiotic
		4 mixotrophic

18 archaeal genomes:

5 aerobic	6 mesophilic	9 heterotrophic
3 facultative	4 thermophilic	0 pathogenic
10 anaerobic	0 pyschrophilic	9 autotrophic
0 microaerophilic	8 hyperthermophilic	0 endosymbiotic
		0 mixotrophic

3. Adjusting for Bacteria-Bacteria RBHs

Due to the predominance of bacterial genomes (134) over archaeal genomes (18) in our study, the Bacteria-Bacteria RBHs swamp any signal from Bacteria-Archaea RBHs in our analyses. Since we are interested in examining RBHs across the bacterial and archaeal domains, we need to create a matrix containing only Bacteria-Archaea RBHs (**Figure 3**).

	Archaea 1	Archaea 2	Archaea 3
Bacteria 1	282	399	355
Bacteria 2	330	454	497
Bacteria 3	283	410	474

RBHs between Bacteria and Archaea

	Archaea 1	Archaea 2	Archaea 3
Bacteria 1	23.85	156.83	1.06
Bacteria 2	-60.51	79.48	10.70
Bacteria 3	-101.14	41.84	-5.93

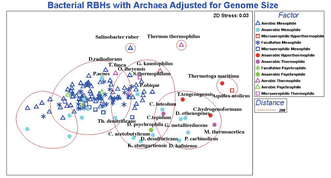
Residual RBHs (adjusted for genome size)

	Bacteria 1	Bacteria 2	Bacteria 3
Bacteria 1	0		
Bacteria 2	387.12	0	
Bacteria 3	121.41	489.96	0

Bacteria-Bacteria Euclidean distance matrix representing RBHs with Archaea

Figure 3. Generation of distance matrix based on bacterial residual RBHs with Archaea

Figure 4. Bacterial residual RBHs with Archaea visualized on a non-metric MDS plot.



These RBHs have been corrected for genome size and only include Bacteria-Archaea RBHs. The unusual relationship between anaerobes and thermophiles (our initial observation in Figure 1) is now evident.

Table 1. ANOSIM significance and statistics of bacterial phenotypic groupings

	Bacteria-Archaea residuals		Bacteria-Bacteria residuals	
	p	R	p	R
Oxygen Respiration (overall ANOSIM)	0.003	0.135	not signif	
Aerobic vs Anaerobic	0.001	0.392		
Anaerobic vs Facultative	0.001	0.345		
Growth Temperature (overall ANOSIM)	0.025	0.230	not signif	
Mesophilic vs Hyperthermophilic	0.005	0.423		
Mesophilic vs Thermophilic	0.004	0.424		
Trophic Level (overall ANOSIM)	0.034	0.065	not signif	
Pathogen vs Autotroph	0.001	0.274		
Pathogen vs Heterotroph	0.018	0.038		
# Genes in Bacterial Genome	not signif		0.001	0.091
Bacterial Phylum	0.001	0.188	0.001	0.604

An analysis of similarity (ANOSIM) test was conducted to determine whether the patterns seen in the MDS plot (Figure 5) are statistically significant. Boxes highlighted in yellow indicate that the grouping is statistically significant. The results show that the matrix consisting only of Bacteria-Archaea residual RBHs clearly groups bacteria based on their oxygen respiration abilities and their optimum growth temperature. **In short, anaerobic and thermophilic bacteria are statistically different from aerobes and mesophiles based on their residual RBHs with Archaea.**

4. Vertical Descent or Lateral Transfer?

Are the reciprocal best BLAST hits (RBHs) between anaerobic, thermophilic Bacteria and Archaea due to conservation of ancestral genes or to lateral transfer?

If the number of RBHs between Bacteria and Archaea is simply a reflection of their phylogenetic distance, then the residual RBH distance matrix generated above should strongly correlate to a distance matrix of their 16S rRNA sequences.

Alternatively, if the number of RBHs is greatly affected by lateral transfer, then the RBH distance matrices should strongly correlate to a distance matrix representing the organisms' opportunities to exchange genetic material. We have attempted to construct such a matrix using phenotypic characteristics which allow organisms to inhabit similar environments (e.g., namely, their optimum growth temperature and their ability to respire oxygen (aerobic, microaerophilic, facultative, or anaerobic). The matrix correlations were calculated with a Mantel test.

Table 2. Correlations among RBHs, 16S rRNA phylogeny, and phenotypes using Mantel tests.

	Matrix A	Matrix B	R	Matrix C	R
Matrices include Bacteria-Bacteria RBHs and 16S rRNA phylogenetic distances	16S rRNA	phenotype	0.324		
	RBHs	16S rRNA	0.830	phenotype	0.808
Matrices include only Bacteria-Archaea RBHs and 16S rRNA phylogenetic distances	RBHs	phenotype	0.411	16S rRNA	0.269
	16S	phenotype	0.290		
	residual RBHs	16S rRNA	0.311	phenotype	0.210
	residual RBHs	phenotype	0.458	16S rRNA	0.380

An R value of 1 = perfect correlation. All correlations were highly significant ($p < 0.002$)

Conclusion from Mantel Tests:
1. Bacteria-Archaea RBHs are less correlated to 16S rRNA phylogenetic distance than are RBHs among Bacteria.
2. Bacteria-Archaea RBHs are more strongly correlated to phenotypes (growth temp & oxygen respiration).

Therefore: Lateral transfer appears to be a significant cause of RBHs between Bacteria and Archaea.

5. Exploring vertical descent vs. lateral transfer in specific functional groups of proteins (COGs)

The same ANOSIM and Mantel analyses performed on the whole dataset in sections 3 and 4 were applied to the dataset partitioned into separate clusters of orthologous groups (COGs) representing functional categories of proteins.

Residual RBHs in some COGs show statistically significant different patterns for Bacteria grouped by: oxygen respiration... or growth temperature... or trophic level

Table 3. ANOSIM statistics for grouping of Bacteria by phenotype based on their RBHs with Archaea in each COG. Only R values that were significant ($p < 0.05$) are listed.

COG	Oxygen Respiration (overall ANOSIM)	Aerobic vs Anaerobic	Anaerobic vs Facultative	Aerobic vs Facultative	Growth Temperature (overall ANOSIM)	Mesophilic vs Hyperthermophilic	Mesophilic vs Thermophilic	Hyperthermophilic vs Thermophilic	Trophic Level (overall ANOSIM)	Pathogen vs Autotroph	Pathogen vs Heterotroph	Autotroph vs Heterotroph
Translation	0.152	0.286	0.167		0.226		0.439					
Replication	0.164	0.356	0.237		0.215	0.522	0.341					
Cell Wall									0.111	0.383		0.268
Posttrans. Mod.	0.145	0.283	0.092	0.097								
Energy	0.351	0.648	0.264	0.173								
Carbohydrates					0.241	0.600	0.248					
Aminoacids												
nucleotide												
Coenzyme												
Lipids												
Inorganic Ions	0.134	0.275	0.221	0.090								
not in COGs	0.100	0.314	0.262		0.265	0.507	0.401				0.031	
2° Metabolites												
General	0.170	0.413	0.349		0.320	0.468	0.560		0.088	0.404	0.082	
Unknown	0.127	0.342	0.272		0.257	0.315	0.453		0.058	0.390		

For example, relative RBHs in the Replication COG show strong differences in mesophiles vs thermophiles:

Figure 5. Non-metric MDS plot visualizing the Bacteria-Bacteria matrix of residual RBHs with archaeal proteins in the Replication COG. ANOSIM statistics for this plot are in Table 3.

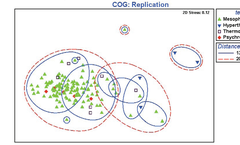
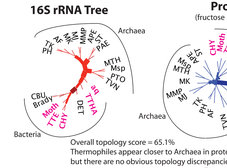


Table 4. Mantel test results for correlation of 16S rRNA phylogeny with a Bacteria-Bacteria matrix representing Archaeal residual RBHs in each COG.

Matrix A	Matrix B	p	R	Matrix C	p	R
16S	Nucleotide residuals	0.001	0.369	pheno	0.001	0.317
	Replication residuals	0.001	0.313		0.002	0.230
	NotinCOGs residuals	0.001	0.290		0.004	0.198
	Carbs residuals	0.002	0.289		0.001	0.219
	Translation residuals	0.001	0.228		0.003	0.187
	Inorgion residuals	0.003	0.196		0.017	0.125
	AminoAcid residuals	0.010	0.202		0.022	0.156
	Cell Wall residuals	0.035	0.101		0.032	0.097
	Energy residuals	0.007	0.164		0.169	0.046
	Coenzyme residuals	0.076	0.104		0.151	0.058
	Posttrans residuals	0.164	0.055		0.425	0.007
	Lipid residuals	0.360	-0.023		0.123	-0.051
	Transcription residuals	0.538	-0.005		0.092	-0.071
pheno	Energy residuals	0.001	0.413	16S	0.001	0.406
	NotinCOGs residuals	0.001	0.401		0.001	0.347
	Replication residuals	0.001	0.372		0.001	0.316
	Inorgion residuals	0.001	0.287		0.001	0.242
	Translation residuals	0.001	0.272		0.001	0.225
	Nucleotide residuals	0.001	0.264		0.001	0.177
	Carbs residuals	0.001	0.229		0.001	0.159
	Transcription residuals	0.001	0.214		0.001	0.225
	AminoAcid residuals	0.003	0.192		0.004	0.141
	Coenzyme residuals	0.001	0.173		0.001	0.188
	Posttrans residuals	0.001	0.168		0.001	0.159
	Lipid residuals	0.017	0.089		0.009	0.100
	Cell Wall residuals	0.241	0.028		0.500	-0.001

RBHs between 'housekeeping' proteins are correlated to 16S rRNA distance..

RBHs in most COGs show some correlation to a distance matrix representing phenotypes.



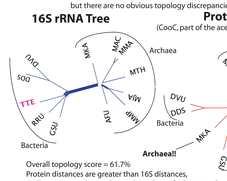


Figure 6. Topology comparison between two protein family trees and the corresponding 16S rRNA tree. Thermophilic Bacteria are highlighted in pink. A topology comparison score is generated for each pair by the algorithm of Nye et al. (2005). Blue branches are shorter than the corresponding red branches, and thicker branches have greater mismatch with the corresponding branch on the other tree.

6. Future direction: Testing lateral transfer by tree topology

The most definitive test of lateral gene transfer (LGT) is to compare the phylogeny of the gene of interest to that of a gene assumed to have experienced no lateral transfer, such as 16S rRNA. We are beginning to use this test on our dataset by constructing phylogenetic trees for each of the protein families comprised of reciprocal best BLAST hits (RBHs).

All 337,734 proteins which represent the 6,681,6331 RBHs in our study were clustered into 29,629 clusters according to their BLAST e-values by the MCL Markov clustering algorithm (Enright et al., 2002). Each cluster generated in this process is considered to be one protein family of orthologs. Clustal multiple sequence alignments of each cluster (we've made approx. 1400 so far) were converted to distance matrices and trees with Phylip's protdist and neighbor-joining algorithms.

Next, a 16S rRNA distance matrix containing the same set of species as represented in each protein tree was generated by pruning the overall 16S rRNA distance matrix.

The final (unfinished step) will be to compare the topology of each protein family tree with the corresponding 16S rRNA trees using the toplogy comparison tool of Nye et al. (2005). Two examples of the output of this procedure are shown in **Figure 6**.

Our goal is to use topology comparison with 16S rRNA phylogeny to identify individual proteins that appear to have been laterally transferred. **We can then ask to what degree these proteins are causing anaerobic, thermophilic Bacteria to have unusually high numbers of RBHs with Archaea.**

Acknowledgements

We would like to thank John Baross for inspiring discussions and Cedar McKay for computer technical assistance. We utilized publicly available software for the R Project of Statistical Computing and the BioPython package. Three of the authors (CAF, REC, WJB) were funded in part by NSF IGERT traineeships awarded to the UW Astrobiology program. Other funding sources include the NASA Astrobiology Institute and National Science Foundation.

References

Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families, Nucleic Acids Research 30(7):1575-1584 (2002).

Fuchsman, C.A. and G. Rocap (2006) Whole-genome reciprocal BLAST analysis reveals that Planctomycetes do not share an unusually large number of genes with Eukarya and Archaea. Appl Env Microbiol 72:6841-6844.

Korbel, J.O., B. Snel, M.A. Huynen, and P. Bork (2002) SHOT: a web server for the construction of genome phylogenies. Trends Genetics 18(3).

Nye, T.M.W., P. Lio, and W.R. Gilks (2005) A novel algorithm and web-based tool for comparing alternative phylogenetic trees. Bioinformatics.

Snel, B., P. Bork, and M.A. Huynen (1999) Genome phylogeny based on gene content. Nat Genet 21:108-110.