Authors' pre-print version.

Manuscript published in MBE-13-0466.R1

http://mbe.oxfordjournals.org/cgi/content/abstract/mst145?ijkey=XWEzZRwAu70saND&keytype=ref

Title:

Reconstruction of Ancestral 16S rRNA Reveals Mutation Bias in the Evolution of Optimal Growth Temperature in the Thermotogae Phylum

Authors: Anna G. Green[1], Kristen S. Swithers[1], Jan F. Gogarten[2], J. Peter Gogarten[1]

1. Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Rd Unit 3125, Storrs, Connecticut, 06269, United States of America

2. Department of Biology, McGill University, 1205 Docteur Penfield, Montreal, Quebec, H3A 1B1, Canada

Corresponding Author:

J. Peter Gogarten
Department of Molecular and Cell Biology
University of Connecticut
91 North Eagleville Rd, Unit 3125
Storrs, CT 06269
United States of America
jpgogarten@gmail.com and gogarten@uconn.edu
Phone: 860-486-4061

Email Addresses:

Anna G. Green:	*anna.g.green@gmail.com*
Jan F. Gogarten:	*jan.gogarten@gmail.com*
Kristen S. Swithers:	*kswithers@gmail.com*
J. Peter Gogarten:	*jpgogarten@gmail.com*

Abstract

Optimal growth temperature is a complex trait involving many cellular components, and its physiology is not yet fully understood. Evolution of continuous characters, such as optimal growth temperature, is often modeled as a one-dimensional random walk, but such a model may be an oversimplification given the complex processes underlying the evolution of continuous characters. Recent papers have used ancestral sequence reconstruction to infer the optimal growth temperature of ancient organisms from the guanine and cytosine content of the stem regions of ribosomal RNA, allowing inferences about the evolution of optimal growth temperature. Here, we investigate the optimal growth temperature of the bacterial phylum Thermotogae. Ancestral sequence reconstruction using a non-homogeneous model was used to reconstruct the stem guanine and cytosine content of 16S rRNA sequences. We compare this sequence reconstruction method to other ancestral character reconstruction methods, and show that sequence reconstruction generates smaller confidence intervals and different ancestral values than other reconstruction methods. Unbiased random walk simulation indicate that the lower temperature members of the Thermotogales have been under directional selection, however, when a simulation is performed that takes possible mutations into account, it is the high temperature lineages that are, in fact, under directional selection. We find that the evolution of Thermotogales optimal growth temperatures is best fit by a biased random walk model. These findings suggest that it may be 'easier' to evolve from a high optimal growth temperature to a lower one than vice versa.

The evolution of continuous phenotypic traits is often modeled as a simple one-dimensional random walk. Under this model, the value of a trait may increase or decrease in defined, distinct time intervals, and the magnitude of change is constant with each step (Felsenstein 1985; Gingerich 1993). With increasing number of time steps the rate of net change in the trait decreases predictably (Gingerich 1993). Modeling trait evolution as a random walk provides a useful null hypothesis against which to test a dataset, to determine whether a trait is evolving according to a deterministic process, such as directional selection. If the rate of change does not decrease over the sum of the time intervals, this is grounds to reject the null hypothesis that evolution of this particular trait proceeds via a random walk. However, even though the evolution of a particular trait appears to be random, this does not mean that the processes underlying the evolution of that particular trait are truly random (Gingerich 1993). While trait evolution may appear to proceed by a random walk, there may be underlying factors that influence and even bias evolution. Given the wealth of molecular data now available, it may be time to move away from simple models like the unbiased random walk, which do not capture the processes underlying phenotypic trait evolution.

Optimal growth temperature (OGT) is an example of a continuous phenotypic trait with many complex underlying factors, and its evolution was suggested to proceed according to a random walk (Dahle et al. 2011). To date research has focused on many of the adaptations of biological macromolecules (protein, RNA and DNA) to extreme temperatures, but understanding of how these adaptations interact to produce a phenotype and how such phenotypes evolve is limited.

In thermophilic organisms, virtually every molecule must adapt in particular ways to remain stable and functional at high temperatures. RNA is likely to undergo 3' to 5' bond hydrolysis at high temperatures (Grosjean and Oshima 2007). Ribosomal and transfer RNAs of thermophiles have a characteristic high guanine-cytosine (GC) content, which increases the stability of secondary structures (Grosjean and Oshima 2007). Ribosomal RNA in particular displays a correlation between the stem GC content and the optimal growth temperature of the organism in which it is found (Galtier and Lobry 1997). DNA is prone to lose its helical structure and undergo depurination and depyrimidation at high temperatures (Grosjean and Oshima 2007). Thermophiles have adapted to these thermodynamic challenges by using small ligand binding and covalent modification of nucleic acids, generation of compact tertiary structures, and efficient DNA repair (Grosjean and Oshima 2007). The proteins of thermophiles are known to undergo a variety of adaptations, including certain amino acid biases that increase stability (Suhre and Claverie 2003; Zeldovich et al. 2007a), and an increase in binding affinity for certain metabolites (Massant 2007). In addition, certain metabolic intermediates may be unstable at high temperatures, leading thermophilic cells to compensate through increased production or sequestration of the metabolite in question (Massant 2007). Given the complex and numerous adaptations required to optimize a cell for growth at high temperatures, it is reasonable to assume that the evolution of this trait is influenced by many underlying factors.

An interesting model clade for the study of OGT evolution is the Thermotogae phylum. The Thermotogales, currently the only recognized order within the phylum Thermotogae, are an order of anaerobic bacteria whose ribosomal genes indicate a close relation to the Aquificae (Zhaxybayeva et al. 2009). However, in bacterial ribosomal phylogenies rooted with an archaeal

outgroup, the root is frequently based in the branch leading to the Aquificae, turning the Thermotogae into the second deepest branching lineage (Reysenbach et al. 2005). In contrast to ribosomal proteins and RNAs, the majority of genes in the genomes of the Thermotogales appear to have been acquired by horizontal gene transfer from Clostridia (Zhaxybayeva et al. 2009). OGTs within the Thermotogales range from 37 to 80 degrees Celsius. Previous work, based on five representatives of the order, suggested that the ancestor to the Thermotogales grew at a higher OGT than the extant members of the clade (Zhaxybayeva et al. 2009). A recent study by Dahle et al. (2011) suggests OGT in the Thermotogales evolved according to an unbiased random walk. This study was based on the pairwise comparisons of quantitative phenotypes of extant organisms and produced wide confidence intervals for predicted values that did not reject the random walk hypothesis. We propose that considering sequence data will allow for better resolution of evolutionary histories and processes.

Traditional techniques of ancestral character estimation rely on a single numeric values of the trait in the extant organisms, and use different methods of averaging these traits to estimate the ancestral value. Because they rely only on single numeric values and methods of averaging, they may not provide realistic reflections on evolutionary processes. Ancestral sequence reconstruction to infer ancestral characters has recently emerged as an alternative to these traditional techniques of ancestral character estimation, and has proved particularly useful for examining historic OGTs (Boussau et al. 2008; Galtier et al. 1999; Groussin and Gouy 2011; Hobbs et al. 2012; Zhaxybayeva et al. 2009). Instead of simply averaging extant values, ancestral state reconstruction takes into account all of the amino acids or nucleotides in a sequence, and can simulate their evolution using complex models. When reconstructing ancestral sequences from extant sequences of different composition, non-homogeneous substitution models are superior for ancestral state reconstruction (Boussau et al. 2008). To date, none of these studies have compared inferences about ancestral temperature obtained from sequence reconstruction to methods traditionally used to reconstruct ancestral character states in evolutionary studies, including Bayesian Markov models, Parsimony, and maximum likelihood.

While OGT evolution within the Thermotogales may conform to a pattern similar to a random walk Dahle et al. (Dahle et al. 2011), this does not mean that optimal growth temperature evolution is truly random or unbiased. *A priori*, an unbiased random walk appears unlikely, because proteins, DNA and RNA in thermophiles are so meticulously adapted to have highly stable structures that there are simply more mutations available that would disrupt these stable structures, and fewer that would continue to increase stability. Thus, we hypothesize that more mutations are available that can lower the optimal growth temperature of an organism than increase it, and this bias should be observable in the evolutionary history of a clade of thermophilic organisms.

Here we study the evolution of optimal growth temperature in the Thermotogales utilizing 30 representatives from the order (fig. 1). We used a known correlate of optimal growth temperature, the GC content of the stem regions of ribosomal RNA (Galtier and Lobry 1997); and reconstructed the stem GC content of the 16S rRNA molecule in the Thermotogales at every node in the tree. This allowed us to trace the evolution of thermophily within the clade, and to show that this trait is not evolving according to a random walk. We compare the inference of the ancestral state for optimal growth temperature from GC content of 16S rRNA to traditional

ancestral state reconstruction methods. We find that no traditional method agrees with the values obtained by ancestral sequence reconstruction, suggesting that sequence reconstruction may be able to reconstruct evolutionary patterns more accurately.

Results

Inference of OGT of Ancestral Sequences

OGT and stem GC content in the Thermotogae show a strong correlation when the phylogenetic signal is subtracted form the analysis using Felsenstein's (1985) method of phylogenetically independent contrasts (PIC) ($r = 0.377$, $p < 0.05$) (fig. 2a). This correlation is also detectable using traditional, non-phylogenetically independent methods, ($r = 0.845$, $p < 0.001$) (fig. 2b).

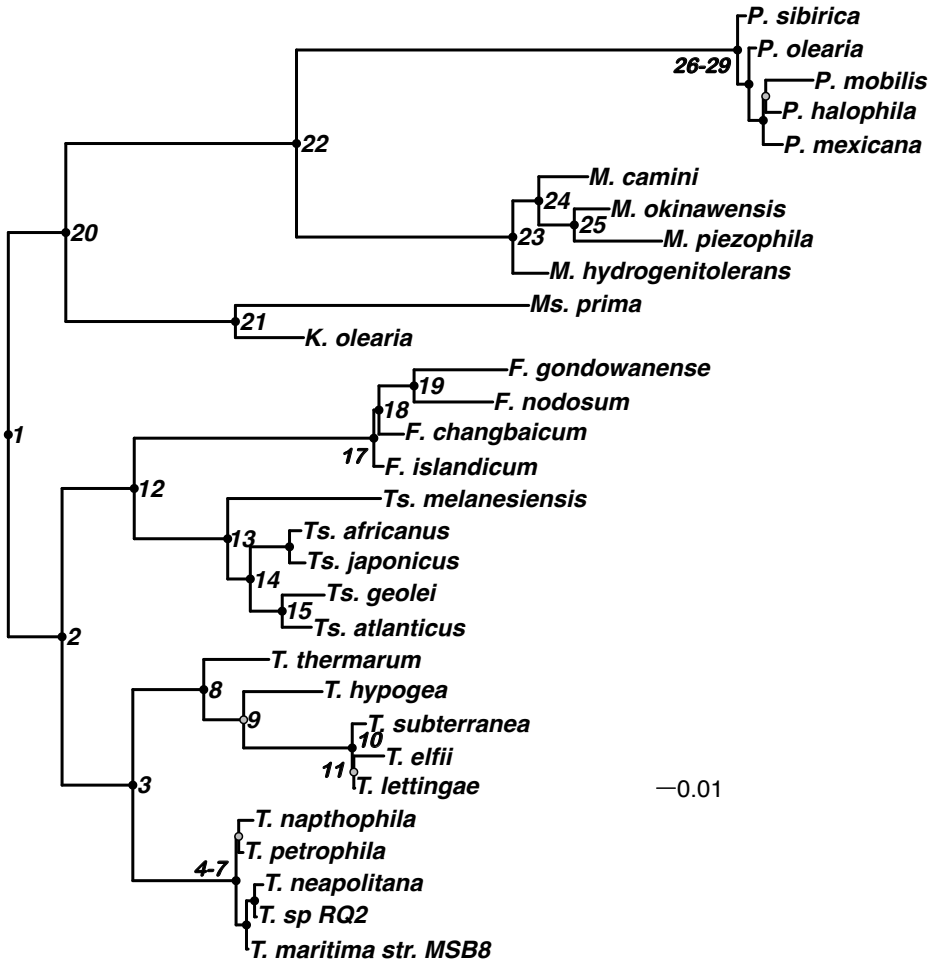

Figure 1: 16S rRNA tree produced using the GTR+G+I model in PhyML. The tree is rooted using 16S rRNAs from bacterial and archaeal genomes as the outgroup (see methods section). The black dots indicate nodes with greater than 75% bootstrap support. The grey dots indicate nodes with less than 75% bootstrap support. The dashed line indicates the position of the outgroup. The nodes are labeled corresponding to (tab. 1).

Table 1: Predicted ancestral OGTs for each node, with 95% confidence intervals, calculated from a linear regression between OGT and stem GC content of extant organisms

Node	Predicted OGT	Lower bound	Upper bound
1	75.93	72.70	79.17
2	76.79	73.38	80.20
3	77.60	74.02	81.18
4	77.87	74.23	81.51
5	78.12	74.43	81.81
6	77.87	74.23	81.51
7	77.87	74.23	81.51
8	75.42	72.28	78.55
9	74.19	71.29	77.10
10	69.12	66.90	71.34
11	68.87	66.67	71.07
12	74.97	71.92	78.01
13	71.49	69.02	73.96
14	68.96	66.75	71.16
15	69.23	67.01	71.46
16	70.01	67.72	72.31
17	70.29	67.97	72.62
18	70.11	67.80	72.41
19	71.80	69.28	74.31
20	76.18	72.90	79.47
21	71.16	68.73	73.59
22	66.10	63.96	68.23
23	59.51	56.65	62.36
24	59.20	56.29	62.11
25	61.20	58.62	63.78
26	53.35	49.22	57.48
27	54.30	50.39	58.21
28	53.26	49.12	57.41
29	52.98	48.77	57.19

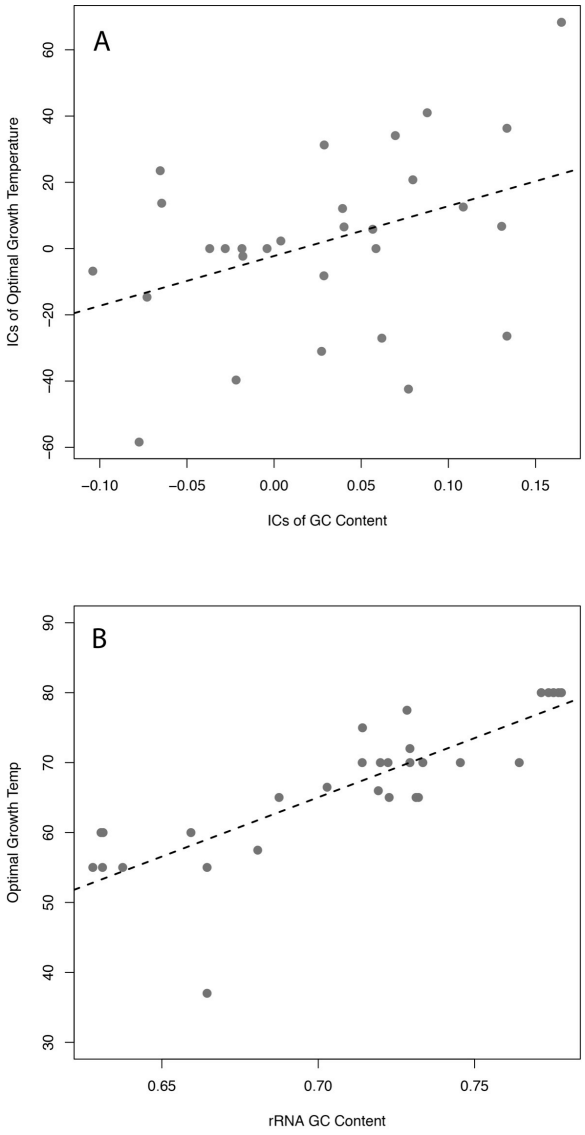

Figure 2: (a) The correlation between optimal growth temperature and stem GC content in the Thermotogae, when independent contrasts (IC's) are used, with regression line. r = 0.377, p = <0.05. Some points have negative values because independent contrasts takes the difference between existing phenotypes. (**b)** The correlation between optimal growth temperature and stem GC content in the Thermotogae, with regression line. r = 0.845, p = <0.001

We calculated a correlation line between the stem GC content and optimal growth temperature in the Thermotogae, $T_{opt} = -53.53 + 169.45 * C_{GC}$, where T_{opt} is the optimal growth temperature and C_{GC} is the stem GC content. We used this line to predict the optimal growth temperature of the last common ancestor of the Thermotogae, with an estimated optimal growth temperature of 76 ± 3.2 ° C. See (tab. 1) for complete list of ancestral OGTs, and (fig. 3) for graphical representation.

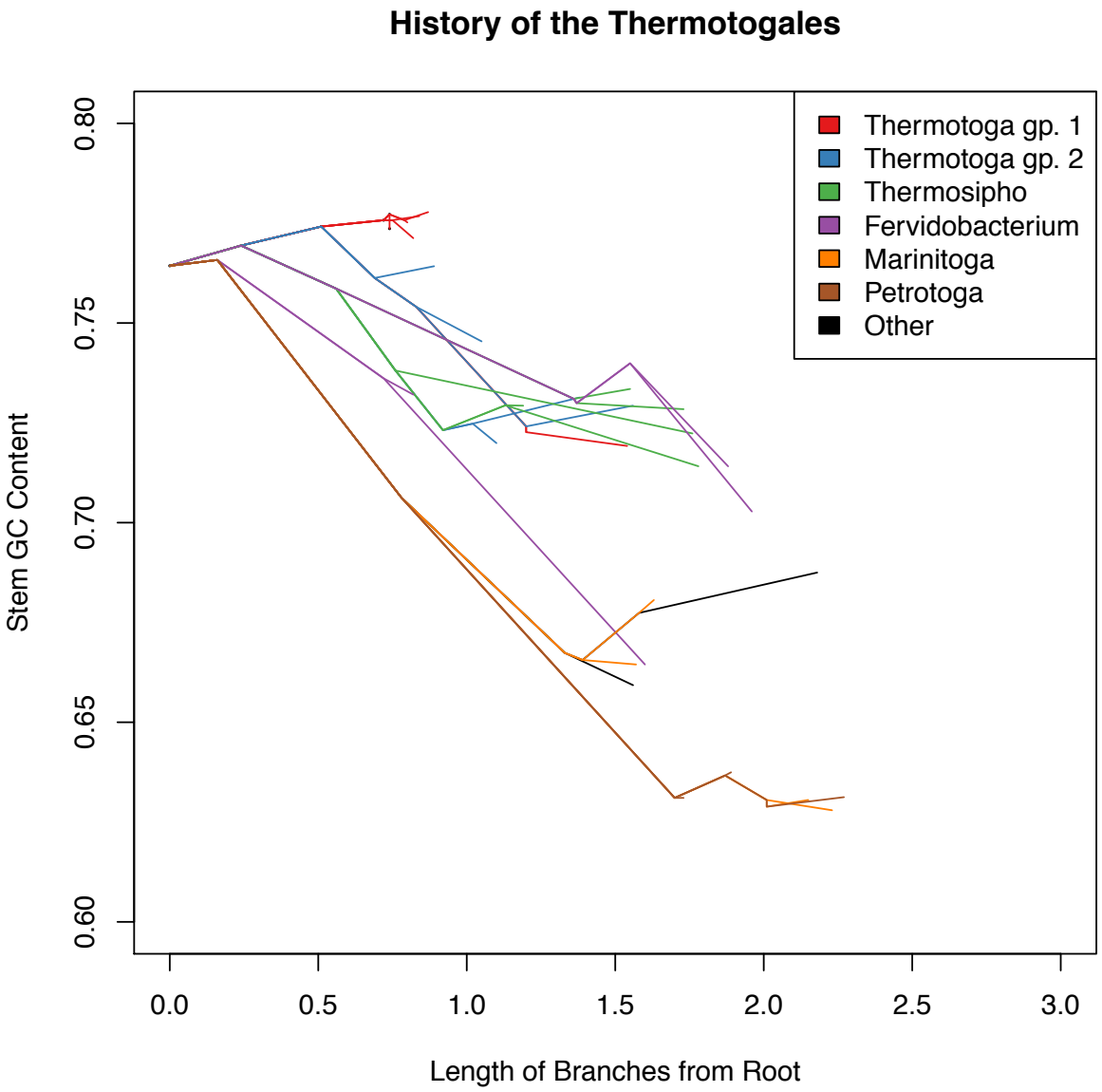

Figure 3: Evolution of the stem GC content in the Thermotogae lineage over time, which is represented by distance from the root. The different genera in the clade are labeled, "other" represent *Kosmotoga* and *Mesotoga*. From this graph we can see that the lineages with higher stem GC content are on shorter branches and have only increased their stem GC content slightly over time, whereas the lineages that have lower stem GC content are on much longer branches and have experienced dramatic decreases over time. Note that for this figure, the branch lengths were calculated from substitutions that did not affect the stem GC content.

Simulation of Changes to the T. maritima ribosome

To determine if it is more probable to accumulate mutations that increase or decrease GC content we simulated all possible point mutations to the *T. maritima* MSB8 ribosome, and

calculated the effects these mutations would have on stem GC content. The results are shown in (fig. 4 / tab. 2). This table indicates that there are a greater number of possible mutations that decrease the stem GC content (n=2279) than there are that increase (n=1120) or do not affect it (n=1278).

Table 2: Effect of simulated point mutations on stem GC content of *Thermotoga maritima MSB8* 16S rRNA.

	n	%	Mean change	Variance
Increase	1120	23.9	0.00118	7.19E-009
Neutral	1278	27.3	0	0
Decrease	2279	48.7	-0.0009938	8.14E-009

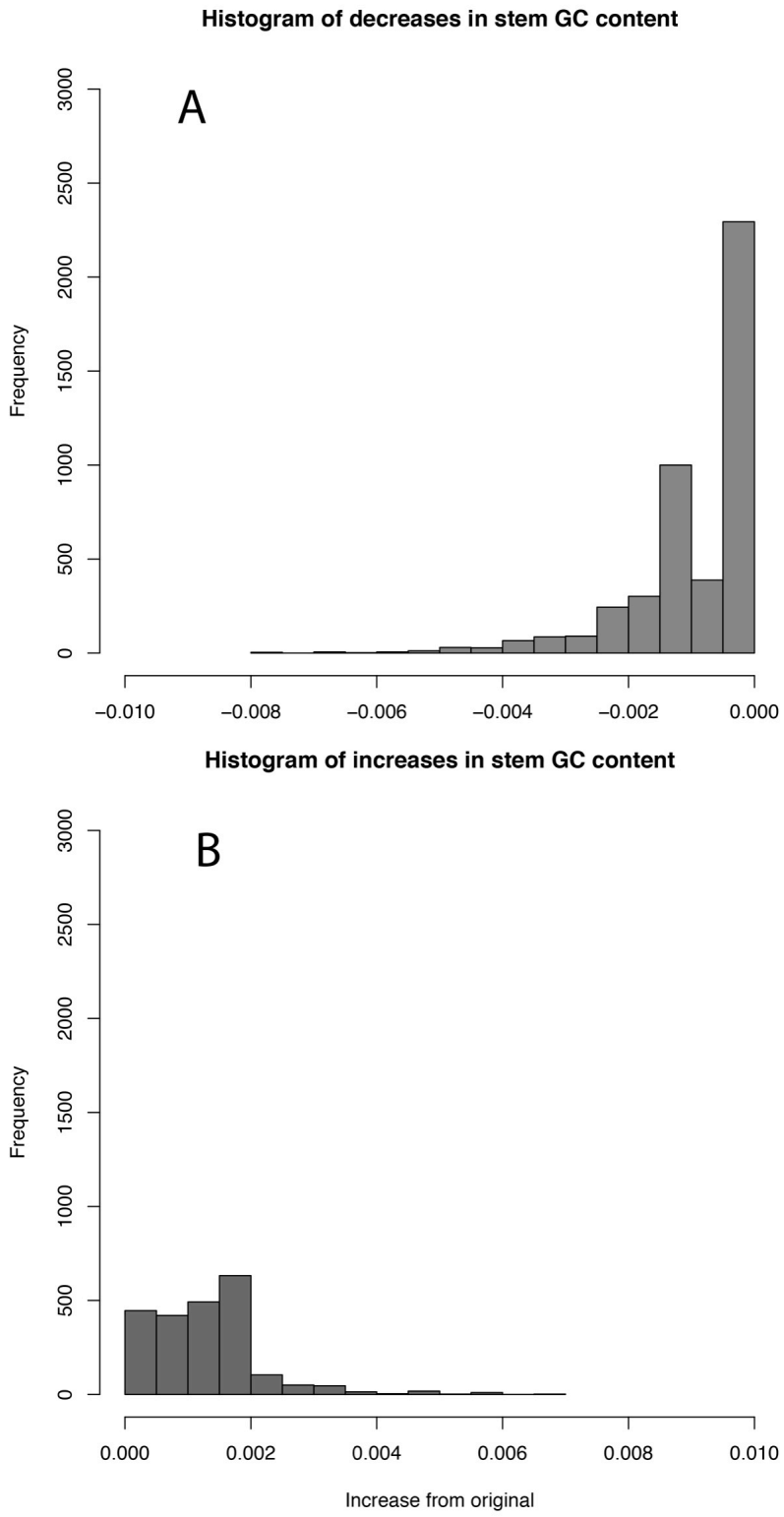

Figure 3: Histogram of the number of possible mutations that cause a decrease **(a)** or an increase **(b)** in optimal growth temperature. The histograms are shown with equal y axes and x axes of the same length to demonstrate the higher number of mutations that decrease the stem GC content.

Random Walk Simulation

We simulated evolution of stem GC content from the ancestral node of the tree, using a random walk simulation with 10,000 replicates, and used the simulated values to establish a 95% confidence interval of values produced by a random walk for each tip. Results are summarized in (tab. 3). None of the lineages fell above the 95% confidence interval. We found that the stem GC content in organisms *Petrotoga mexicana, Petrotoga halophila, Petrotoga olearia,* and *Petrotoga sibirica* was below the 95% confidence interval. Taken alone this could suggest directional evolution to decrease stem GC content, or a higher probability of mutations that decrease the stem GC content.

Levene's Test for Homogeneity of Variance was used to compare the variances of these two sets of possible changes, positive and negative. We found that there is a significant difference in the variances, with the variance of the negative group being greater ($p = <0.001$).

Table 3: Actual rRNA stem GC content of extant Thermotogales compared to results of an unbiased random walk simulation.

	Extant GC Content	2.5th percentile	97.5th percentile
Fervidobacterium changbaicum	0.728	0.649	0.881
Fervidobacterium gondowanense	0.703	0.645	0.887
Fervidobacterium islandicum	0.733	0.655	0.875
Fervidobacterium nodosum	0.714	0.645	0.884
Kosmotoga olearia	0.732	0.686	0.844
Marinitoga camini	0.664	0.655	0.876
Marinitoga hydrogenitolerans	0.659	0.656	0.874
Marinitoga okinawensis	0.681	0.653	0.875
Marinitoga piezophila	0.688	0.636	0.893
Mesotoga prima	0.664	0.654	0.875
Petrotoga halophila	**0.631**	0.634	0.895
Petrotoga mexicana	**0.628**	0.633	0.899
Petrotoga mobilis	0.631	0.631	0.898
Petrotoga olearia	**0.637**	0.642	0.887
Petrotoga sibirica	**0.631**	0.648	0.881
Thermosipho africanus	0.714	0.644	0.882
Thermosipho atlanticus	0.731	0.660	0.868
Thermosipho geolei	0.720	0.670	0.857
Thermosipho japonicus	0.729	0.668	0.862
Thermosipho melanesiensis	0.722	0.644	0.880
Thermotoga elfii	0.719	0.656	0.875
Thermotoga hypogea	0.745	0.673	0.856
Thermotoga lettingae	0.723	0.669	0.860
Thermotoga maritima MSB8	0.774	0.689	0.841
Thermotoga naphthophila	0.771	0.683	0.845
Thermotoga neapolitana	0.778	0.682	0.847
Thermotoga petrophila	0.777	0.682	0.845
Thermotoga sp. RQ2	0.775	0.684	0.843
Thermotoga subterranea	0.729	0.656	0.873
Thermotoga thermarum	0.764	0.680	0.848

Note**. --** Columns 2 and 3 show the upper and lower bounds of the 95% confidence interval of the unbiased random walk simulation. The bold values are the extant values that fall outside the confidence interval generated by the simulation.

Table 4: Results of evolutionary simulation of rRNA stem GC content using distribution of possible point mutations.

Organism	Extant GC Content	2.50%	97.50%
Fervidobacterium changbaicum	0.728	0.659	0.741
Fervidobacterium gondowanense	0.703	0.636	0.73
Fervidobacterium islandicum	0.733	0.663	0.743
Fervidobacterium nodosum	0.714	0.64	0.731
Kosmotoga olearia	0.732	0.692	0.757
Marinitoga camini	0.664	0.634	0.726
Marinitoga hydrogenitolerans	0.659	0.64	0.729
Marinitoga okinawensis	0.681	0.63	0.724
Marinitoga piezophila	0.688	0.621	0.718
Mesotoga prima	0.664	0.642	0.732
Petrotoga halophila	0.631	0.596	0.703
Petrotoga mexicana	0.628	0.595	0.705
Petrotoga mobilis	0.631	0.589	0.7
Petrotoga olearia	0.637	0.601	0.706
Petrotoga sibirica	0.631	0.602	0.707
Thermosipho africanus	0.714	0.682	0.752
Thermosipho atlanticus	0.731	0.677	0.75
Thermosipho geolei	0.72	0.675	0.748
Thermosipho japonicus	0.729	0.68	0.751
Thermosipho melanesiensis	0.722	0.663	0.743
Thermotoga elfii	0.719	0.663	0.743
Thermotoga hypogea	0.745	0.677	0.749
Thermotoga lettingae	0.723	0.669	0.746
Thermotoga maritima MSB8	**0.774**	0.698	0.759
Thermotoga naphthophila	**0.771**	0.696	0.758
Thermotoga neapolitana	**0.778**	0.694	0.757
Thermotoga petrophila	**0.777**	0.699	0.759
Thermotoga sp. RQ2	**0.775**	0.695	0.758
Thermotoga subterranea	0.729	0.666	0.745
Thermotoga thermarum	**0.764**	0.689	0.755

Note. - This table shows the extant GC content along with upper and lower bounds of the 95% confidence interval of simulated values. Extant values that are above or below the 95% confidence interval are bolded.

Table 5: Comparison of ancestral state reconstruction methods.

Method	# CI that overlap	G [f]	P [f]	Mean CI size [g]	T [h]	Df [h]	P [h]
REML[a]	29	2.975	0.08456	0.0279	1.0243	28.584	0.3134
SCP[b]	21	15.9239	6.59E-05	0.021	-6.627	51.065	2.00E-08
PIC[c]	29	2.975	0.08456	0.9732	8.9922	28.001	9.51E-10
GLSB[d]	28	0.1642	0.6853	0.3347	8.1825	28.004	6.59E-09
GLSG[e]	28	0.1642	0.6853	1.1825	19.2598	28.002	2.20E-16

Note. The following models were tested (see text for details):

[a] Brownian motion with restricted maximum likelihood (REML)

[b] squared change parsimony (SCP)

[c] phylogenetically independent contrasts (PIC)

[d] generalized least squares with a Brownian motion model (GLSB)

[e] Grafen model (GLSG)

[f] The overlap of confidence intervals generated by the ancestral state reconstruction methods with those generated by ancestral sequence reconstruction was examined using a G-test. Columns 2 and 3 show the G-test statistic and p-value (a low p-value indicates that the tested confidence intervals are unlikely to have come from the same distribution).

[g] The mean size of confidence interval generated by each reconstruction method (Mean CI size) was compared to that generated by ancestral sequence reconstruction (0.0246) using a T-test.

[h] The final columns show the T-test statistic and p-value (a low p-value indicates a significant difference in confidence interval size).

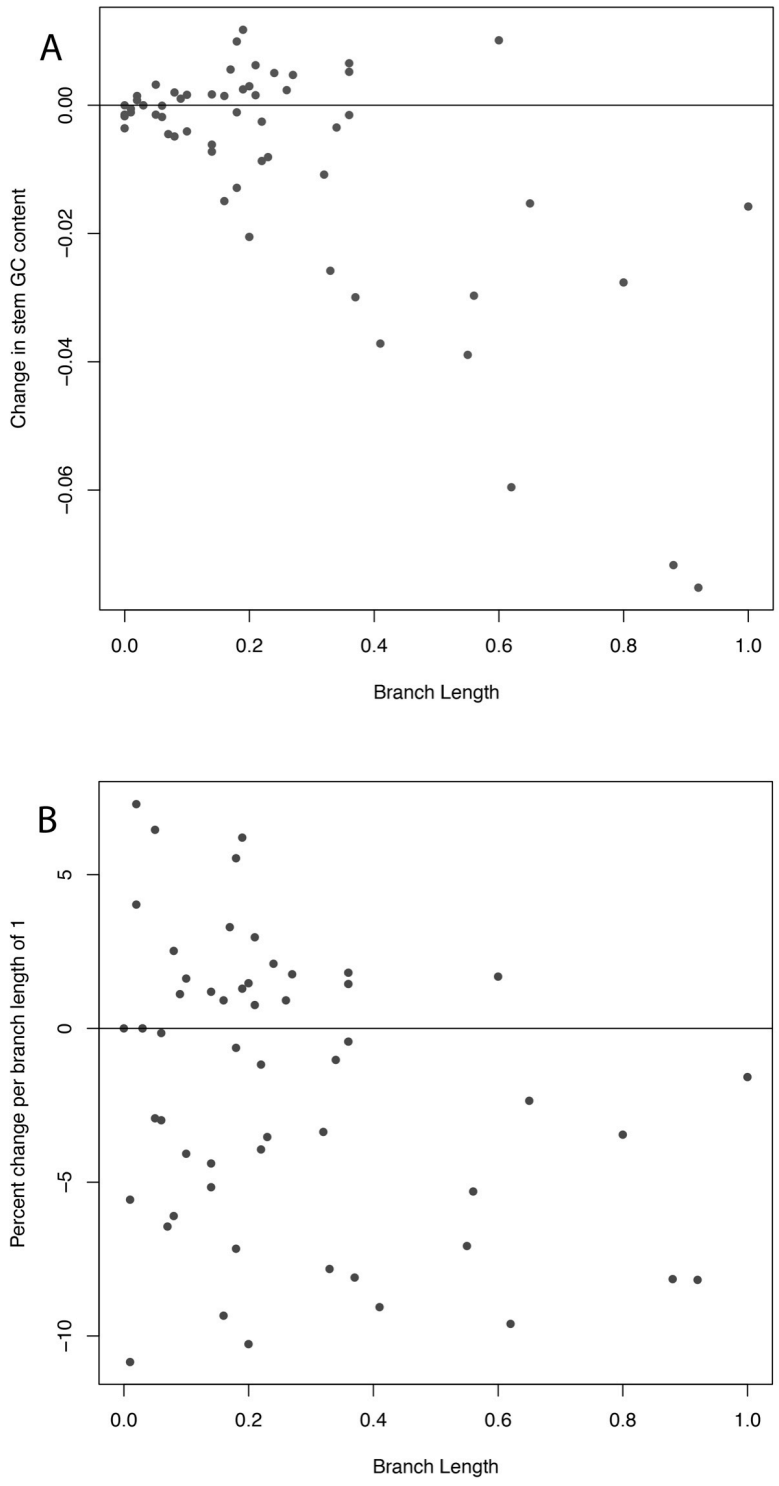

Figure 5: Changes in stem GC content as a function of the length of the branch on which they occur **(a)**, and percent change in stem GC content for a branch length of one as a function of the branch length on which each change occurs **(b)**. Under a random walk simulation, we expect the rates of change to converge to 0 as the branch lengths increase.

Biased Random Walk Simulation

To attempt to explain the tendency of the GC content to decrease, the distribution of phenotypic effects produced by point mutations to the *T. maritima* MSB8 ribosome was used to simulate evolution of the Thermotogales. The results are summarized in (tab. 4). The lineages with low stem GC content do no longer fall outside the 95% confidence interval for random walk evolution, unlike in the random walk simulation that did not use the estimated substitution probabilities. However, many of the lineages with high stem GC content now fall outside the 95% confidence interval, including *Thermotoga maritima* MSB8, *Thermotoga* sp. RQ2, *Thermotoga neapolitana, Thermotoga petrophila, Thermotoga neapolitana, Thermotoga petrophila, Thermotoga naphthophila,* and *Thermotoga thermarum*.

Comparison of ancestral character reconstruction methods

The results of each ancestral reconstruction method, excluding Bayestraits, were compared to the ancestral stem GC content values obtained from sequence reconstruction. The G-test was used to test the congruence of the confidence intervals. The T-test was used to test for difference in mean size of confidence interval generated. The results are summarized in (tab. 5). No reconstruction method produced both values congruent with those estimated by ancestral sequence reconstruction (ASR) and confidence intervals of a similar size, as indicated by the G-test and T-test results.

Confidence intervals for reconstructed stem GC content were calculated by 10,000 replicate sequence reconstructions in Bppancestor, sampling from the probability distribution for each site, rather than using the most probable nucleotide (tab. 4). The 95% confidence interval for total GC content was used, because calculation of the secondary structures for 10,000 replicates was computationally unfeasible. A comparison of confidence interval generated by each method is provided in (tab. 5).

Using the constant-variance random walk model implemented in Bayestraits, the ancestral node was reconstructed at a stem GC content of 0.727, which falls outside of the 95% confidence interval generated by the GC content of the sequence reconstruction (0.751-0.777). The 95 credibility interval for the constant-variance random walk model was 0.699 - 0.755, which overlaps with the 95% confidence interval generated by the G-C content of the sequence reconstruction, although this is a large confidence interval range. The directional random walk for Bayestraits estimated an ancestral stem GC content of 0.791, which was higher than the confidence interval produced by sequence reconstruction. The 95 credibility interval for the directional random walk model was 0.752 - 0.829, which overlaps with the credibility interval produced by sequence reconstruction, though the confidence interval is again large.

Of the models of ancestral reconstruction tested, only REML performs adequately. PIC, GLSB, and GLSG fail the t-test for size of confidence intervals, meaning that the confidence intervals produced by those methods are significantly larger than those produced by ASR. SCP fails the G-test for overlap of confidence intervals, meaning that SCP fails to produce estimates similar to those produced by ASR. The only method that does not fail both of these tests is REML, indicating it produces similar estimates and confidence intervals to those made by ASR.

Both the random and directional models implemented in BayesTraits produced ancestral

values with confidence intervals that overlapped the confidence interval produced by ASR, though these confidence intervals were quite large. The directional model produced an estimate of ancestral stem GC content much higher than ASR did, and the random model produced a much lower estimate than ASR.

Discussion

In this paper we used ancestral sequence reconstruction to reconstruct ancestral character states, and systematically compared this method to traditional methods of ancestral character state reconstruction. The advantage of this approach is that it calculates the property at the ancestral node based on sequence reconstruction, a procedure that relies on well-established models of sequence evolution, and does not rely on averaging of extant values of character states. In addition, this method does not reduce the value of the trait to a single numeric estimate, and instead considers the nucleotide sequence underlying that estimation. The accuracy of the method based on sequence reconstruction has been demonstrated in previous studies, which were able to successfully reconstruct sequences from deep nodes in the tree of life, and use those sequences to make evolutionary inferences (Boussau et al. 2008; Groussin and Gouy 2011). In addition, our method does not rely on pairwise distance between the extant species to test for random walk evolution, as pairwise distances are not independent data points (Felsenstein 1985). Indeed, when our method of ancestral character estimation based on sequence reconstruction is compared to conventional methods, the conventional methods perform poorly, either failing to reconstruct values within the confidence intervals of our sequence based method, or producing very large confidence intervals on their reconstructed points, rendering the estimates practically useless.

Our analyses indicate that the ancestor to the Thermotogae grew at a relatively high optimal growth temperature, 76 ± 3° C. The majority of lineages in this clade have undergone a decrease in optimal growth temperature over time (for example, *Mesotoga prima* has undergone a decrease of 39° C), while a few have maintained or increased their optimal growth temperature, but only by a small amount (e.g., the lineage leading to *Thermotoga maritima* MSB8 has undergone an increase of 4° C in its optimal growth temperature since its divergence from the ancestral node). While this differs from earlier analyses that suggested an ancestral growth temperature of over 80° C (Zhaxybayeva et al. 2009), our results are based on a dataset with many more taxa and a more sophisticated method of sequence reconstruction (*i. e.* non-homogeneous implementation of substitution models).

Initial evolutionary simulations seemed to indicate that the lower temperatures members of the Thermotogae are not evolving according to an unbiased random walk, and are in fact under directional selection. We have shown that the extant values of stem GC content in the majority of lineages falls within the 95% confidence interval produced by the simulations. However, a few organisms, *Petrotoga mexicana, Petrotoga halophila, Petrotoga olearia,* and *Petrotoga sibirica* fall below the 95% confidence interval of the values obtained from the simulation. These results would indicate that these lineages have been undergoing directional evolution toward a lower stem GC content, and thus a lower optimal growth temperature, if not for the mutation bias that further simulations revealed.

We provide evidence that a trait may appear to be under directional selection when in fact

there is a mutation bias affecting that trait. This was done by a simulation of point mutations to the *T. maritima* MSB8 16S rRNA, which demonstrated that the number of possible mutations that would decrease the stem GC content is almost twice as large as the number of mutations that would increase it. There is also a significantly greater variance in the number of possible decreases, indicating that it may be possible to make greater changes over a shorter period of time. However, when this distribution of possible mutations is used to simulate evolution along the tree, the high stem GC content lineages are found to be outside the 95% confidence interval for evolution according to a random walk. The high stem GC content lineages, and not the low ones, are the lineages under directional selection, which is only detected when the proper distribution of possible mutations is used to simulate evolution. This demonstrates that a one-dimensional random walk does not adequately reflect the possible mutations that underlie evolution of OGT, and may fail to detect evolutionary patterns such as directional selection.

The lineages that have undergone a decrease in stem GC content, and therefore a decrease in OGT, tend to be found on longer branches, even when branch lengths are calculated independently of stem GC content (fig. 5). One may assume that this sharp decrease in stem GC content, and therefore OGT, is due to directional selection, mutation bias, or a combination of the two. However, this fails to provide an explanation for why those lineages tend to be found at a greater distance from the root, because the branch lengths are calculated independently of stem GC content, which is presumably the trait under selection. This may be due to increased directional selection on certain organisms to adapt to lower optimal growth temperature, so more mutations are allowed to reach fixation, or due to an increased number of mutations that become non-detrimental to the organism as it gains the ability to survive at lower temperatures. A previous study provides evidence for the second explanation, demonstrating by biophysical simulation that the maximum permissible mutation rate, i.e. the rate above which populations go extinct, in thermophiles is less than one third of the maximum permissible rate in mesophiles (Zeldovich et al. 2007b), in support of the earlier observation that high temperature lineages are more slowly evolving (Woese 1987).

We acknowledge the possibility that one substitution model is not adequate to reflect the evolution of both the stem and loop regions of the 16S rRNA molecule. However, attempts, to create a phylogeny using only the stem regions of the alignment resulted in poorly supported tree topologies that were incongruent with the well-supported tree produced by using the entire sequences. It is possible that the alignment of only stem regions does not contain adequate information to recover the correct tree topology (see alignment 3, supplementary information), or that the substitution models available do not adequately reflect the evolution of the stem regions. Advancement of models of site-dependent evolution will provide better tools to address this question (Williams et al. 2011)

Conclusions

We have shown that many lineages in the Thermotogae have adapted to lower OGT over time, whereas few have maintained or increased their OGT by a small amount. Because there is a higher GC content in thermophiles our observations can be explained by the larger number of possible mutations in the 16S rRNA that cause a decrease in stem GC content, and therefore a decrease in OGT. Presumably, there are also more possible mutations that would cause a decrease in the fitness of finely adaptive protein and DNA structures in thermophiles than there

are those that would increase it. Therefore, adaptation to lower temperatures is likely easier for a thermophile than further adaptation to even higher temperatures. This suggests that it is relatively more difficult to adapt from a lower to higher temperature, and relatively easy to adapt from a higher to lower temperature.

It is important to address the order of events in this proposed process of adaptation to lower OGT. Organisms that have undergone a steep decrease in stem GC content, and therefore OGT, must have already been able to survive at a lower temperature. After moving into a lower temperature niche, where the temperature is survivable but not optimal, their proteins and rRNA would have begun to adapt to function optimally at the new temperature. This would be facilitated by the greater number of possible mutations available to decrease the stem GC content. In addition, a greater number of mutations would be permissible at these lower temperatures (Zeldovich et al. 2007b).

Methods

Sequence alignment and tree reconstruction

The 16S rRNA sequences of 30 members of the Thermotogales were downloaded from the GenBank database at NCBI. See (tab. 6) for accession numbers and optimal growth temperature information obtained from characterization papers. Phenotype data on optimal growth temperatures were obtained from characterization papers, see (tab. 6). Organisms that have not been characterized were not included in the study. For the purpose of tree and ancestral sequence reconstruction, our dataset also included bacterial and archaeal outgroups: *Thermoanaerobacter pseudethanolis* (CP000924.1), *Thermoanaerobacter tengcongensis* (NR_074701.1), *Carboxydothermus hydrogenoformans* (NR_074395.1), *Hydrogenobaculum* sp. Y04AAS1(NR_074960.1), *Aquifex aeolicus (AJ309733.1), Sulfurihydrogenibium* sp. YO3AOP1 (NR_074557.1), *Pyrococcus furiosus (NR_074375.1),* and *Thermococcus kodakerensis (NR_028216.1)*, for a total of 38 taxa. These sequences were downloaded from the NCBI GenBank database.

The 16S rRNA sequences were initially aligned using muscle v. 3.8.31 (Edgar 2004) with the default settings. The alignment was then refined to include structural information using RNASalsa (Stocsits et al. 2009), using stringency settings s1, s2, and s3 = 0.9, and a constraint structure file for *Thermotoga maritima* MSB8, obtained from the Comparative RNA Website and Project (Cannone et al. 2002). See Supplementary Materials for the full structural alignment of 16S sequences.

To determine the best substitution model to use for constructing a tree from the alignment, we used the phangorn package for R (Schliep 2011). This software tests the substitution models JC, F81, K80, HKY, SYM, and GTR, with a proportion of invariant sites, gamma rate categories, both, and neither to find the model with the best fit to the data. This model was GTR+G+I. We used PhyML v.3.0 (Guindon and Gascuel 2003) to create a tree from the alignment (fig. 1), using the parameters specified by model test, and allowing estimation of the rate categories and proportion of invariant sites from the dataset.

Table 6: Thermotogae species used in this study, 16S sequence accession numbers, and optimal growth temperatures (OGT).

Organism	Accession #	OGT	Source
Thermotoga maritima MSB8	NR_102775.1	80	(Huber et al. 1986)
Thermotoga sp. RQ2	AJ872273.1	80	(Swithers et al. 2011)
Thermotoga neapolitana	NR_074959.1	80	(Jannasch et al. 1988)
Thermotoga petrophila	CP000702.1	80	(Takahata et al. 2001)
Thermotoga naphthophila	NR_074952.1	80	(Takahata et al. 2001)
Thermotoga lettingae	NR_074951.1	65	(Balk et al. 2002)
Thermotoga elfii	NR_026201.1	66	(Ravot et al. 1995)
Thermotoga subterranea	NR_025969.1	70	(Jeanthon et al. 1995)
Thermotoga hypogea	NR_029205.1	70	(Fardeau et al. 1997)
Thermotoga thermarum	CP002351.1	70	(Windberger et al. 1989)
Thermosipho atlanticus	NR_029020.1	65	(Urios et al. 2004)
Thermosipho geolei	NR_025389.1	70	(Haridon et al. 2001)
Thermosipho japonicus	NR_024726.1	72	(Takai and Horikoshi 2000)
Thermosipho africanus	NR_102773.1	75	(Huber et al. 1989).
Thermosipho melanesiensis	CP000716.1	70	(Antoine et al. 1997)
Fervidobacterium islandicum	NR_044730.1	70	(Nam et al. 2002)
Fervidobacterium changbaicum	NR_043248.1	77.5	(Cai et al. 2007)
Fervidobacterium nodosum	NR_074093.1	70	(Patel et al. 1985)
Fervidobacterium gondowanense	NR_036997.1	66.5	(Andrews and Patel 1996)
Kosmotoga olearia	NR_044583.1	65	(Dipippo et al. 2009)
Mesotoga prima	CP003532.1	37	(Nesbø et al. 2012)
Marinitoga hydrogenitolerans	NR_042320.1	60	(Postec et al. 2005)
Marinitoga piezophila	NR_027541.1	65	(Alain et al. 2002)
Marinitoga okinawensis	NR_041466.1	57.5	(Nunoura et al. 2007)
Marinitoga camini	NR_028907.1	55	(Wery et al. 2001)
Petrotoga mexicana	NR_029058.1	55	(Miranda-Tello et al. 2004)
Petrotoga halophila	NR_043201.1	60	(Miranda-Tello et al. 2007)
Petrotoga mobilis	NR_074401.1	60	(Lien et al. 1998)
Petrotoga olearia	NR_028947.1	55	(L'Haridon et al. 2002)
Petrotoga sibirica	NR_025466.1	55	(L'Haridon et al. 2002)

Note. - In cases where an optimal growth range was given in the characterization paper, the midpoint was used as the OGT.

Ancestral Sequence Reconstruction

We used the BppML program in the Bio++ package (Dutheil and Boussau 2008) to refine the branch lengths of the tree produced in PhyML and optimize the model parameters for ancestral sequence reconstruction. To refine this tree, we defined a non-homogeneous substitution model using GTR+G+I, with a different set of parameters on each of the four major clades on the tree: the more thermophilic Thermotogales group (genera *Thermotoga*, *Thermosipho*, and *Fervidobacterium*), the less thermophilic Thermotogales group, the bacterial outgroup, and the archaeal outgroup. The parameters of the model were optimized in BppML (Dutheil and Boussau 2008), and used to reconstruct the ancestral sequences at each node of the tree in Bppancestor (Dutheil and Boussau 2008).

Gap inferences in ancestral sequences

We inferred the position of gaps in the ancestral sequences. The reconstructed ancestral sequences were the length of the original alignment, but contained no gaps, because gaps are not treated as a character in the substitution model we used. To calculate the position of gaps in the ancestral sequences, we changed all of the gaps in existing sequences to C's, and all of the nucleotides to A's. We then used the F84 substitution model in Bppancestor to determine ancestral 'sequences', which represent the position of gaps as nucleotides, for all of the extant nodes. The substitution model is appropriate for our application because no guanine or thymine, which do not represent anything in our model, will be introduced to the sequences. This model has three parameters: the GC content is *theta*, the G/(G+C) ratio is *theta1*, and the A/(A+T) ratio is *theta2* (PERL scripts are available in Supplementary Materials). More importantly, the GC content (or in this case C content, *i.e.* the number of gaps), will remain constant throughout the tree, because the extant sequences all have a similar number of gaps, and presumably the 16S sequences have not shortened or lengthened significantly over time. We used Bppancestor to calculate these three parameters based on the input sequences, and then reconstructed ancestral sequences from those parameters. We then used the position of C's in the output ancestral sequences to infer the position of gaps in the actual reconstructed ancestral sequences using in-house PERL scripts (available in Supplementary Materials). A comparison of this method to another available method of gap reconstruction is provided in the supplementary material.

Determination of stem GC content of ancestral sequences

The structure of the ancestral sequences was inferred using RNASalsa (Stocsits et al. 2009) with stringency settings s1, s2, s3 = 0.9 and a constraint structure file for *Thermotoga maritima* MSB8. A full alignment of the extant and reconstructed ribosomal sequences can be found in the supplementary information. After the structures of the ancestral sequences were inferred, we determined the stem GC content of all of the sequences, ancestral and extant, using in-house PERL scripts (available in Supplementary Materials).

Confirming correlation between optimal growth temperature and stem GC content

We calculated a regression equation between the optimal growth temperature and stem GC content of extant species. As comparisons between traits of related species are prone to correlation due to shared ancestry, i.e., the traits of related species are not independent data points, we used independent contrasts (Felsenstein 1985), as implemented in the ape package for R (Paradis 2004) to confirm that these two traits are correlated in the Thermotogae.

Calculating time intervals and step size for random walk simulation

After we inferred the optimal growth temperature at every node in the tree, we then calculated the change in stem GC content along each branch of the tree. In our tree, the branch lengths are determined by the number of differences between sequences, and the trait of interest is a bias in the sequence - therefore the two are not independent. To determine a time variable independent of the trait being measured, we wrote PERL scripts (available in Supplementary Materials) to determine the number of changes along each branch that did not affect stem GC content. These sums were then scaled so that the lowest number of changes that occurred on a branch (n=0) corresponded to a branch length of 0, and the largest number within the Thermotogae group (n=100) corresponded to a branch length of 1. These calculated branch lengths were all increased by 10^{-9}, to avoid dividing by zero in cases when the branch length was 0. These calculated branch lengths were used for the following random walk analysis.

Testing the data against a random walk simulation

Ancestral reconstruction data was used to test the hypothesis that evolution of optimal growth temperature in the Thermotogae proceeds by a random walk, using stem GC content as a proxy for optimal growth temperature. This was done using the function *rTraitCont*, available in the ape package for R (Paradis et al. 2003), which simulates evolution of a continuous character from the root of a given phylogeny to the tips. We simulated the evolution of the stem GC content 10,000 times, along the tree with branch lengths calculated as explained above. One outlier was excluded when calculating the standard deviation of the change along the branches for use in the simulation. The simulation resulted in 10,000 values for the possible phenotype at each node, obtained according to a random walk model of trait evolution.

We simulated all possible point mutations to the *T. maritima* MSB8 16S rRNA sequence using an in-house PERL script to make all the possible point mutations (available in the supplementary material). The resulting sequences each had one point mutation. The secondary structure of these sequences was determined using RNAsalsa with constraint values of 1, appropriate for the highly similar sequences. In-house PERL scripts were used to calculate the number of mutations that increased, decreased, or did not affect the stem GC content.

A second random walk simulation was performed using the probability of the possible point mutations in the *T. maritima* ribosome. Evolution was again simulated from the root to the tips, sampling from the distribution of possible point mutations. The number of samples per branch was calculated by multiplying the branch length (which gives the substitutions per site) by the length of the entire sequence, 1462. Evolution was simulated in this fashion 10,000 times.

Ancestral character state reconstruction

Using the phylogeny generated above, we compared various methods of ancestral character state reconstruction. To compare different reconstruction methods we implemented a maximum likelihood reconstruction for continuous characters using a Brownian motion model (Schluter et al. 1997), as well as reconstructions based on least squares (i. e., phylogenetically independent contrasts (PIC) (Felsenstein 1985)), generalized least squares with a Brownian (GLSB) and Grafen (GLSG) model (Cunningham et al. 1998; Martins and Hansen 1997), and squared change parsimony (SCP) (Maddison 1991); all implemented in the APE package for R (Paradis et al. 2004; R-Development-Core-Team 2012). Restricted maximum likelihood (REML)

results in unbiased estimates of the variance of the Brownian motion process while maximum likelihood gives a downward bias so we used the restricted maximum likelihood approach for the Brownian motion model (Heyde 1997).

To test the congruence of our reconstructed estimates of ancestral stem GC content with those estimated from the sequence data, we used G-tests (Sokal and Rohlf 1995). While reconstruction estimates were often given as point estimates, it was also possible to generate confidence intervals for these estimates. As a much more conservative test of the congruence between methods, we also used the number of nodes on a tree for which the confidence intervals for the nodes had any overlap. By definition our expected frequency of overlap in confidence intervals should be >0.05, while point estimates should fall in the confidence intervals 95% of the time. We also tested whether confidence intervals were of significantly different sizes between the reconstruction methods by using t-tests.

In addition, we tested two models of trait evolution in BayesTraits (Pagel et al. 2004); a constant-variance random walk model and a directional random-walk implemented in a Bayesian framework. These two models of trait evolution were compared using likelihood ratio test (Pagel 1999). BayesTraits allows for the reconstruction of the most recent common ancestor of all taxa in the tree and estimates from the two models of character evolution were compared with the results estimated from the G-C content of the sequence data to see if point fell within the 95% confidence interval and whether confidence intervals and credibility intervals overlapped. Default settings were used in BayesTrait with the exception of the rate deviance parameter, which was set at 0.15 for the random walk and 0.1 for the directional walk to get an acceptance ratio between 20 and 40 % as suggested in the BayesTraits manual.

Supplementary Materials

scripts.docx: Collection of PERL and R scripts used to perform the described analyses.
http://gogarten.uconn.edu/articles/OGT_Evolution/scripts.docx

alignments.zip This folder contains the alignments in FASTA sequential, and clustalw, philip and pdf interleaved formats.

- Extant seqs structural alignment: The alignment of extant 16S sequences used in the study, initially aligned using Muscle v 3.8.31 and then refined based on secondary structure using RNASalsa.
- All seqs structural alignment: The alignment of extant 16S sequences and reconstructed ancestral 16S sequences, refined based on secondary structure using RNASalsa.
- All seqs structural alignment stemloop encoded: Same as All seqs structural alignment, except the loop regions are all represented as T's and the stem regions are all represented as A's. This allows for the visualization of secondary structures inferred by RNA Salsa.

http://gogarten.uconn.edu/articles/OGT_Evolution/alignments.zip

Gap_reconstruction_method.pdf: Comparison of methods to reconstruct gaps in ancestral sequences.
http://gogarten.uconn.edu/articles/OGT_Evolution/Gap_reconstruction_method.pdf

Funding:
This work was supported by a US National Science Foundation Grant (DEB 0830024 to JPG and DGE-1142336 to JFG) and the Canadian Institutes of Health Research's Strategic Training Initiative in Health Research's Systems Biology Training Program (to JFG).

Acknowledgements:
The authors thank Kenneth Noll, Pascal Lapierre, David Williams, Timothy Harlow, Olga Zhaxybayeva for discussions and the Biotechnology Bioservices Center of the University of Connecticut, Storrs, USA, for technical support.

Works Cited

Alain K, Marteinsson VT, Miroshnichenko ML, Bonch-Osmolovskaya EA, Prieur D, Birrien JL. Marinitoga piezophila sp. nov., a rod-shaped, thermo-piezophilic bacterium isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. International Journal of Systematic and Evolutionary Microbiology 2002;52(4):1331-1339.

Andrews KT, Patel BK. Fervidobacterium gondwanense sp. nov., a new thermophilic anaerobic bacterium isolated from nonvolcanically heated geothermal waters of the Great Artesian Basin of Australia. International Journal of Systematic Bacteriology 1996;46(1):265.

Antoine E, Cilia V, Meunier JR, Guezennec J, Lesongeur F, Barbier G. Thermosipho melanesiensis sp. nov., a new thermophilic anaerobic bacterium belonging to the order Thermotogales, isolated from deep-sea hydrothermal vents in the southwestern Pacific Ocean. International Journal of Systematic Bacteriology 1997;47(4):1118.

Balk M, Weijma J, Stams AJ. Thermotoga lettingae sp. nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. International Journal of Systematic and Evolutionary Microbiology 2002;52(Pt 4):1361.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. Parallel adaptations to high temperature in the Archaean eon. Nature 2008;456:942-5.

Cai J, Wang Y, Liu D, Zeng Y, Xue Y, Ma Y, Feng Y. Fervidobacterium changbaicum sp. nov., a novel thermophilic anaerobic bacterium isolated from a hot spring of the Changbai Mountains, China. International journal of systematic and evolutionary microbiology 2007;57(Pt 10):2333.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 2002;3(2).

Cunningham CW, Omland KE, Oakley TH. Reconstructing ancestral character states: a critical reappraisal. . Trends in Ecology & Evolution 1998;13:361-366.

Dahle H, Hannisdal B, Steinsbu BO, Ommedal H, Einen J, Jensen S, Larsen O, Ovreås L, Norland S. Evolution of temperature optimum in Thermotogaceae and the prediction of trait values of uncultured organisms. Extremophiles 2011;15(4):509-16.

Dipippo JL, Nesbo CL, Dahle H, Doolittle WF, Birkland NK, Noll KM. Kosmotoga olearia gen.

nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. 2009, 59(12):2991-3000

Dutheil J, Boussau B. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evolutionary Biology 2008;8(255).

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004; 5(113).

Fardeau ML, Ollivier B, Patel BK, Magot M, Thomas P, Rimbault A, Rocchiccioli F, Garcia JL. Thermotoga hypogea sp. nov., a xylanolytic, thermophilic bacterium from an oil-producing well. International Journal of Systematic and Evolutionary Microbiology 1997;47(4):1013.

Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist 1985;125(1):1-15.

Galtier N, Lobry JR. Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. Journal of Molecular Evolution 1997;44(6):632-636.

Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science 1999;283:220-221.

Gingerich PD. Quantification and Comparison of Evolutionary Rates. American Journal of Science 1993;293-A:453-478.

Grosjean H, Oshima T. How nucleic acids cope with high temperatures. In: Gerday C, Glansdorff N, editors. Physiology and biochemistry of extremophiles. Volume 1. Washington, DC: ASM Press; 2007. p 39-56.

Groussin M, Gouy M. Adaptation to Environmental Temperature is a Major Determinant of Molecular Evolutionary Rates in Archaea. Mol. Biol. Evol. 2011;28:2661-2674.

Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 2003;52(5):696-704.

Haridon SL, Miroshnichenko ML, Hippe H, Fardeau ML, Bonch-Osmolovskaya E, Stackebrandt E, Jeanthon C. Thermosipho geolei sp. nov., a thermophilic bacterium isolated from a continental petroleum reservoir in Western Siberia. International Journal of Systematic and Evolutionary Microbiology 2001;51(Pt 4):1327.

Heyde CC. Quasi-likelihood and its application: a general approach to optimal parameter estimation. Springer Verlag; 1997.

Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, MOnk CR, Daniel RM, Arcus VL. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of Bacillus. Mol. Biol. Evol 2012;29(825-835).

Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, Sletyr UB, Stetter KO. Thermotoga maritima, sp. nov. represents a new genus of unique extremely thermophilic bacteria growing up to 90oC. Archives of Microbiology 1986;144(4):324-333.

Huber R, Woese CR, Langworthy TA, H. F, Stetter KO. Thermosipho africanus, gen. nov., represents a new genus of thermophilic eubacteria within the Thermotogales. Systematic and Applied Microbiology 1989;12:32-37.

Jannasch HW, Huber R, Belkin S, Stetter KO. Thermotoga neapolitana sp. nov. of extremely thermophilic eubacterial genus Thermotoga. Archives of Microbiology 1988;150:103-104.

Jeanthon C, Reysenbach AL, L'Haridon S, Gambacorta A, Pace NR, Glenat P, Prieur D. Thermotoga subterranea sp. nov., a new thermophilic bacterium isolated from a

continental oil reservoir. Archives of Microbiology 1995;164(2):91.
L'Haridon S, Miroshnichenko ML, Hippe H, Fardeau ML, Bonch-Osmolovskaya EA, Stackebrandt E, Jeanthon C. Petrotoga olearia sp. nov. and Petrotoga sibirica sp. nov., two thermophilic bacteria isolated from a continental petroleum reservoir in Western Siberia. International Journal of Systematic and Evolutionary Microbiology 2002;52(Pt. 5):1715-1722.
Lien T, Madsen M, Rainey FA, Birkeland NK. Petrotoga mobilis sp. nov., from a North Sea oil-production well. International Journal of Systematic Microbiology 1998;48(Pt. 3):1007-1013.
Maddison WP. Squared-Change Parsimony Reconstructions of Ancestral States for Continuous-Valued Characters on a Phylogenetic Tree. Systematic Zoology 1991;40(3):304-314.
Martins EP, Hansen TF. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist 1997:646-667.
Massant J. How thermophiles cope with thermolabile metabolites. In: Gerday C, Glansdorf N, editors. Physiology and biochemistry of extremophiles. Volume 1. Washington, DC: ASM Press; 2007. p 57-74.
Miranda-Tello E, Fardeau ML, Joulian C, Magot M, Thomas P, Tholozan JL, Ollivier B. Petrotoga halophila sp. nov., a thermophilic, moderately halophilic, fermentative bacterium isolated from an offshore oil well in Congo. International Journal of Systematic and Evolutionary Microbiology 2007;57(Pt. 1):40-44.
Miranda-Tello E, Fardeau ML, Thomas P, Ramirez F, Casalot L, Cayol JL, Garcia JL, Ollivier B. Petrotoga mexicana sp. nov., a novel thermophilic, anaerobic and xylanolytic bacterium isolated from an oil-producing well in the Gulf of Mexico. International Journal of Systematic and Evolutionary Microbiology 2004;54(Pt 1):169-174.
Nam GW, Lee DW, Lee HS, Lee NJ, Kim BC, Choe EA, Hwang JK, Suhartono MT, Pyun YR. Native-feather degradation by Fervidobacterium islandicum AW-1, a newly isolated keratinase-producing thermophilic anaerobe. Archives of Microbiology 2002;178(6):538-547.
Nesbø CL, Bradnan DM, Adebusuyi A, Dlutek M, K. PA, J. F, F. DW, M. NK. Mesotoga prima gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. Extremophiles 2012;16:387-393.
Nunoura T, Oida H, Miyazaki M, Suzuki Y, Takai K, Horikoshi K. Marinitoga okinawensis sp. nov., a novel thermophilic and anaerobic heterotroph isolated from a deep-sea hydrothermal field, Southern Okinawa Trough. International Journal of Systematic and Evolutionary Microbiology 2007;57(Pt 3):467-471.
Pagel M. Inferring the historical patterns of biological evolution. Nature 1999;401(6756):877-884.
Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. Systematic Biology 2004;53(5):673-684.
Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 2003;20(2):289-90.
Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004;20(2):289-290.
Patel BKC, Morgan HW, Daniel RM. Fervidobacterium nodosum gen nov. and spec. nov., a new chemoorganotrophic, caldoactive, anaerobic bacterium. Archives of Microbiology

1985;141:63-69.
Postec A, Le Breton C, Fardeau ML, Lesongeur F, Pignet P, Querellou J, Ollivier B, Godfroy A. Marinitoga hydrogenitolerans sp. nov., a novel member of the order Thermotogales isolated from a black smoker chimney on the Mid-Atlantic Ridge. International Journal of systematic and Evolutionary Microbiology 2005;55(Pt 3):1217-1221.
R-Development-Core-Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
Ravot G, Magot M, Fardeau ML, Patel BK, Prensier G, Egan A, Garcia JL, Ollivier B. Thermotoga elfii sp. nov., a novel thermophilic bacterium from an African oil-producing well. International Journal of Systematic Bacteriology 1995;45(2):308-314.
Reysenbach A-L, Banta A, Civello S, Daly J, MItchel K, Lalonde S, Konhauser K, Rodman A, Rusterholtz K, Takacs-Vesbach C. The Aquificales of Yellowstone National Park. . In: Inskeep WP, McDermott TR, editors. Geothermal Biology and Geochemistry in Yellowstone National Park. Volume 1. Bozeman, Montana: Montana State University; 2005. p 129-142.
Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics 2011;27(4):592-593.
Schluter D, Price T, Mooers AØ, Ludwig D. Likelihood of ancestor states in adaptive radiation. Evolution 1997:1699-1711.
Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. New York: W. H. Freeman and Company; 1995. 887 p.
Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF. Accurate and efficient reconstruction of deep phylogenies from structured RNAs. Nucleic Acids Research 2009;37(18):6184–6193.
Suhre K, Claverie J-M. Genomic correlates of hyperthermostability, an update. Journal of Biological Chemistry 2003;278(19):17198-17202.
Swithers KS, DiPippo JL, Bruce DC, Detter C, Tapia R, Han S, Saunders E, Goodwin LA, Han J, Woyke T, Pitluck S, Pennacchio L, Nolan M, Mikhailova N, Lykidis A, Land ML, Brettin T, Stetter KO, Nelson KE, Gogarten JP, Noll KM. Genome sequence of Thermotoga sp. strain RQ2, a hyperthermophilic bacterium isolated from a geothermally heated region of the seafloor near Ribeira Quente, the Azores. Journal of Bacteriology 2011;193(20):5869-5870.
Takahata Y, Nishijima M, Hoaki T, Maruyama T. Thermotoga petrophila sp. nov. and Thermotoga naphthophila sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan. International Journal of Systematic and Applied Microbiology 2001;51(Pt. 5):1901-1909.
Takai K, Horikoshi K. Thermosipho japonicus sp. nov., an extremely thermophilic bacterium isolated from a deep-sea hydrothermal vent in Japan. Extremophiles: life under extreme conditions 2000;4(1):9-17.
Urios L, Cueff-Gauchard V, Pignet P, Postec A, Fardeau ML, Ollivier B, Barbier G. Thermosipho atlanticus sp. nov., a novel member of the Thermotogales isolated from a Mid-Atlantic Ridge hydrothermal vent. International Journal of Systematic and Evolutionary Microbiology 2004;54(Pt. 6):1953-1957.
Wery N, Lesongeur F, Pignet P, Derennes V, Cambon-Bonavita MA, Godfroy A, Barbier G. Marinitoga camini gen. nov., sp. nov., a rod-shaped bacterium belonging to the order Thermotogales, isolated from a deep-sea hydrothermal vent. International Journal of Systematic and Evolutionary Microbiology 2001;51(Pt. 2):495-504.

Williams D, Fournier GP, Lapierre P, Swithers KS, Green AG, Andam CP, Gogarten JP. A Rooted Net of Life. Biology Direct 2011;6(45).

Windberger E, Huber R, Trincone A, Fricke H, Stetter KO. Thermotoga thermarum sp. nov. and Thermotoga neapolitana occuring in African continental solfataric springs. Archives of Microbiology 1989;151:506-512.

Woese CR. Bacterial evolution. Microbiology Reviews 1987;51(2):221-271.

Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA Sequence Determinants of Thermophilic Adaptation. PLoS Comput. Biol 2007a;3:e5.

Zeldovich KB, Chen P, Shakhnovich EI. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proceedings of the National Academy of Sciences 2007b;104(41):16152-16157.

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbø CL, Doolittle WF, Gogarten JP, Noll KM. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proceedings of the National Academy of Sciences 2009;106(14):5865-5870.