Authors' pre-print version.

Manuscript published in Molecular Biology and Evolution MBE-13-0254.R2

DOI of published manuscript: 10.1093/molbev/mst164

Article
Discoveries Section

Distribution and Evolution of the Mobile *vma-1*b Intein.

Kristen S. Swithers[1,3], Shannon M. Soucy[1], Erica Lasek-Nesselquist[1,5], Pascal Lapierre[2,4] and Johann Peter Gogarten[1*]

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, United States of America

[2] University of Connecticut Biotechnology Center, University of Connecticut, Storrs, Connecticut, United States of America

[3] Current affiliation: Department of Cell Biology, Yale University Medical School, New Haven, Connecticut, United States of America

[4] Current affiliation: New York State Department of Health, Wadsworth Center, Albany, New York, United States of America

[5] Current affiliation: Department of Biology, University of Scranton, Scranton, PA, United States of America

Communicating author: Johann Peter Gogarten
Communicating author email: jpgogarten@gmail.com, gogarten@uconn.edu

Abstract

Inteins are self-splicing parasitic genetic elements found in all domains of life. These genetic elements are found in highly conserved positions in conserved proteins. One protein family that has been invaded by inteins is the vacuolar and archaeal catalytic ATPase subunits (*vma-1*). There are two intein insertion sites in this protein, "a" and "b". The "b" site was previously thought to be only invaded in archaeal lineages. Here we survey the distribution and evolutionary histories of the "b" site inteins and show the intein is present in more lineages than previously annotated including a bacterial lineage, *Mahalla australienses* 50-1 BON. We present evidence, through ancestral character state reconstruction and substitution ratios between host genes and inteins, for several transfers of this intein between divergent species, including an interdomain transfer between the Archaea and Bacteria. While inteins may persist within a single population or species for long periods of time, transfer of the *vma*-1b intein between divergent species contributed to the distribution of this intein.

Keywords:

intein, parasitic gene, homing endonuclease, gene transfer, co-evolution of genes

Introduction

Inteins are self-splicing, parasitic, mobile genetic elements that splice out of their flanking proteins sequence (extein) after translation via an autocatalytic mechanism (Cooper and Stevens 1995; Perler et al. 1994). Inteins can be divided into two different groups based on their size - large and mini inteins - which are differentiated based on the presence or absence of a homing endonuclease (HE) domain (Elleuche and Poggeler 2010; Liu 2000). The large inteins contain a HE domain, an enzyme with endonuclease activity that has a 14 – 40 bp recognition site (Chevalier and Stoddard 2001), and N- and C-terminal splicing domains (Pietrokovski 1998). This HE domain provides mobility to the intein and the ability to move into new intein-less target sites through a process called homing (Chevalier and Stoddard 2001; Gogarten and Hilario 2006). Transfer of the intein to hosts that contain an intein-less target site is necessary for the HE to remain under purifying selection (Goddard and Burt 1999). The transfers of the HE containing inteins can occur either within or between species (Barzel et al. 2011; Clerissi et al. 2013; Gogarten and Hilario 2006; Koufopanou et al. 2002; Macgregor et al. 2013; Okuda et al. 2003; Yahara et al. 2009). The mini intein is simply composed of the N and C-terminal splicing domains with a linker region between them.

Inteins are found in all three domains of life and in viruses, and tend to be present in conserved regions of proteins (exteins) with high sequence similarity (Swithers et al. 2009). Targeting conserved sites in conserved proteins is advantageous for two reasons. First, it provides selective pressure on the splicing domains, which ensures the intein is spliced out exactly to maintain a functional extein. Mutations to the splicing domains of the intein will result in improper splicing, resulting in a nonfunctional extein that could be detrimental to the cell. Second, targeting the most conserved regions of the most conserved genes will facilitate transfer to inteinless alleles across species or even domain boundaries.

One such conserved protein family invaded by inteins is the vacuolar and archaeal catalytic ATPase subunits (*vma-1*) (Hirata et al. 1990; Kane et al. 1990; Senejani et al. 2001). The vacuolar, archaeal, and bacterial ATPases (or ATP synthases) are a family of multi-subunit proteins that are present in all three domains of life and share a common ancestor (Gogarten et al. 1989; Zimniak et al. 1988). The eukaryotic version, called the vacuolar ATPase (V-type), is involved in the intravesicular acidification of the endomembrane system (vacuoles, lysosome, endosomes and trans Golgi network) and it is also found in the plasma membrane of specialized cells of transport epithelia (Harvey and Nelson 1992). The archaeal (A-type) and bacterial (F-type) counterparts are used for ATP generation and/or ion transport. Several cases of horizontal gene transfer of the A-type ATPase to Bacteria have been documented (Hilario and Gogarten 1993; Lapierre 2007; Lapierre et al. 2006).

Two distinct sites in this family of ATPases have been invaded by inteins; the “a” site in the vacuolar ATPases in yeasts and the “b” site in the Archaea. These two insertion sites have been shown to be in the most conserved parts of the protein (Swithers et al. 2009). The intein in the "a" site was frequently transferred between different yeast species (Koufopanou et al. 2002; Okuda et al. 2003). The intein in the "b" site has a wide phylogenetic distribution among Archaea. To date the intein database (InBase) annotates seven *vma-1*b inteins in the *Thermoplasma* and *Pyrococcus* (Perler 2002) genera. Here we report on additional *vma-1*b inteins, show that they are found in more diverse species, and discuss their evolutionary histories.

Table 1. Distribution of *vma-1*b inteins.

Organisms	Lineage	AA Length of Intein	Isolated From	Reference
Pyrococcus furiosus DSM 3638	Archaea; Euryarchaeota; Thermococci	426	Shallow marine solfatara at Vulcano Island off southern Italy	(Fiala and Stetter 1986)
Pyrococcus abyssi GE5	Archaea; Euryarchaeota; Thermococci	428	Hydrothermal Vent in the North Fiji Basin in the Pacific Ocean	(Erauso et al. 1993)
Pyrococcus sp. NA2	Archaea; Euryarchaeota; Thermococci	428	Hydrothermal vent in Papua New Guinea-Australia-Canada-Manus (PACMANUS) field	(Lee et al. 2011)
Pyrococcus horikoshii OT3	Archaea; Euryarchaeota; Thermococci	375	Hydrothermal vent in the Okinawa Trough in the Pacific Ocean	(Gonzalez et al. 1998)
Thermococcus litoralis DSM 5473	Archaea; Euryarchaeota; Thermococci	428	Shallow submarine hot spring at Lucrino Beach near Naples	(Neuner et al. 1990)
Thermoplasma acidophilum 122-1B2 ATCC 25905	Archaea; Euryarchaeota; Thermoplasmata	174	Self-heating coal refuse pile from the Friar Tuck mine in Indiana, USA	(Searcy 1975; Senejani et al. 2001)
Thermoplasma acidophilum DSM 1728	Archaea; Euryarchaeota; Thermoplasmata	173	Self-heating coal refuse pile from the Friar Tuck mine in Indiana, USA	(Darland et al. 1970)
Thermoplasma volcanium GSS1	Archaea; Euryarchaeota; Thermoplasmata	185	Solfataric field	(Segerer et al. 1988)
Thermoplasmatales archaeon I-plasma	Archaea; Euryarchaeota; Thermoplasmata	179	Ultraback A location (UBA site) in the Richmond Mine, Iron Mountain, CA	(Dick et al. 2009)
Ferroplasma sp. Type II	Archaea; Euryarchaeota; Thermoplasmata	535	5-way CG acid mine drainage site in the Richmond Mine, Iron Mountain, CA	(Tyson et al. 2004)
Picrophilus torridus DSM 9790	Archaea; Euryarchaeota; Thermoplasmata	332	Dry solfataric field in northern Japan	(Schleper et al. 1995)
Methanothermus fervidus DSM 2088	Archaea; Euryarchaeota; Methanobacteria	357	Hot solfataric spring from Iceland	(Anderson et al. 2010)
Candidatus Nanosalinarum sp. J07AB56	Archaea; Euryarchaeota; Nanohaloarchaea	522	Hypersaline lake NW Victoria, Australia	(Narasingarao et al. 2012)
Thetis Sea contig00086	Metagenome gi number 3548555569	521	Deep-sea hypersaline anoxic lake Thetis	(Ferrer et al. 2012)
Mahella australiensis 50-1 BON	Bacteria; Firmicutes; Clostridia; Thermoanaerobacterales	522	Riverslea oil field in the Bowen-Surat basin, Australia	(Salinas et al. 2004)

Results

Distribution of *vma-1*b inteins. Databases at the NCBI were surveyed for *vma-1* exteins and inteins that reside in the b insertion position (*vma-1*b). To date we identified fifteen *vma-1*b inteins within complete flanking extein sequences present in these databases (Table 1). Thirteen are found within the Euryarchaeota, one is present in Clostridia and another is found in a deep sea hypersaline metagenomic sequencing project. Additionally, we identified two inteins in contigs from a marine metagenome (gi:129952466) and a hot springs metagenome (gi:290482983) that only encode partial extein sequences. These partial sequences did not contain sufficient information for phylogenetic placement of the host organisms. The multiple sequence alignment shows the *vma-1*b intein family is composed of both mini inteins and larger inteins with homing endonuclease domains (see supplementary materials, alignment of intein sequences). The larger inteins all have the LAGLIDADG homing endonuclease domains, suggesting these may be active homing endonucleases. However, additional experimental evidence is required to verify their activity. Compared to the currently annotated inteins in InBase, we report an additional ten inteins at this insertion site in three new classes of prokaryotes.

Breakpoints within the extein/intein alignment. The genetic algorithm for recombination detection (GARD) determined breakpoints in the extein/intein alignment. This algorithm analyzes fragments of the overall alignment and compares the goodness of fit of phylogenies from these smaller alignments under a maximum likelihood framework using the corrected Akaike Information Criterion. Significant breakpoints were identified three amino acids before the insertion site (position 245 $p<0.01$) and at three amino acids before the end of the intein (position 891 $p<0.01$) (Figure S1). The fact that the GARD algorithm detects breakpoints three amino acids upstream of the insertion site might be due to the conserved, phylogenetically uninformative nature of the splice site. The last three amino acids at the carboxy terminal end of the intein are conserved in all the inteins included in this study. This level of conservation is more typical for the extein than for the intein sequences (see histograms for conservation in Figure S8). Regardless of the slight misplacement of the breakpoint, the GARD analysis does suggest that extein and intein have different evolutionary histories, which is also corroborated by the intein and extein trees (cf. Figure 2).

Archaeal-type ATPase distribution. A survey of the bacterial domain reveals that all major clades of Bacteria have representatives containing an archaeal-type ATPase (uninvaded extein sequence). A phylogenetic analysis of the extein protein using representatives from all clades shows three bacterial clusters interspersed within the archaea (Supplementary Figure S2). The phylogeny of the archaeal exteins is in good agreement with the ribosomal protein phylogeny (see below). Thus, the finding that the Bacteria are interspersed within the Archaea suggests that the three distinct bacterial groups (Supplementary Figure S2) represent independent gene transfer events from Archaea to Bacteria.

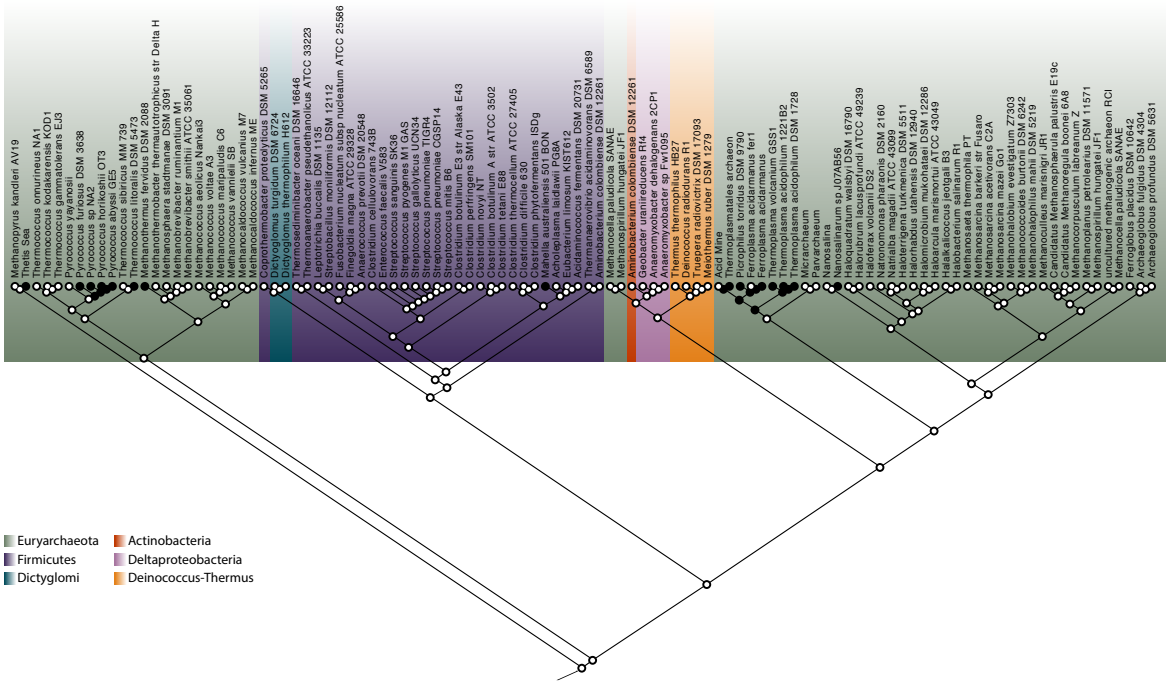


Figure 1. Maximum Parsimony Ancestral Character State Reconstruction. The depicted cladogram represents the part of the VMA-1 tree where inteins are present (see Supplemental Figure S2 for full tree). Presence and absences of the intein are represented at the tips of the tree; black circles correspond to presence of the intein and white circles correspond to absences of the intein. The single most parsimonious reconstruction with nine steps indicated seven independent gains and two losses of the intein. The gains are once for the Thetis sea sequence, once for the *Pyrococcus* spp., once for the *Thermococcus litoralis*, once for *Methanothermus fervidus*, once for *Mahella australiensis*, once for the Thermoplasmatales and once for *Candidatus* Nanosalinarum sp. J07AB56. The two losses are in the Acid Mine sequence and *Ferroplasma acidarium* fer1. A maximum likelihood reconstruction using the MK1 model (Lewis 2001) produced nearly identical results (see supplemental figure S3).

Archaeal extein and ribosome trees. Overall, the extein and ribosomal protein reference phylogeny display a high degree of topological similarity, particularly at the ordinal level. Many clades show identical branching orders, such as the Methanosarcinales, Methanococcales, and Methanobacteriales (Supplementary Figure S4). Deeper phylogenetic relationships were also similar, including the sister relationship between the Archaeoglobi and a Halobacteria-Methanomicrobia clade. Less resolved positions on the tree tended to stem from taxa on long branches, prone to artifact, such as *Micrarchaeum* and *Nitrosopumilis* in the ribosomal phylogeny.

The reduced extein and ribosomal protein trees were congruent (Supplementary Figure S5) with the exception of the "*misplacement*" of *Candidatus* Nanosalinarum sp. in the ribosome phylogeny due to LBA. Both SH and AU tests indicated that the extein and ribosomal topologies were not significantly different from each other (*p*-values = 0.28 and 0.73 for AU and SH tests, respectively). This suggests that the archaeal exteins follow ribosomal evolution (considered to be vertical evolution) and have not been transferred between divergent archaeal lineages. This finding also implies that transfers of the intein occurred independently of any extein transfers.

Ancestral Character State Reconstruction. Maximum parsimony and maximum likelihood ancestral state reconstructions (ASR) were performed using the phyletic patterning of the intein relative to its host gene across the prokaryotes (Figure 1). There was only one most parsimonious reconstruction with nine steps given the dataset, which is validated by the maximum likelihood (ml) ASR analysis. The reconstructions suggest seven independent gains and two losses of the intein. The reconstructions suggest seven independent gains and two losses of the intein. According to the parsimony and the ml reconstruction the Thetis sea sequence (probability from mlASR P=100%), the ancestor of the four *Pyrococcus* spp. (P=99%), *T. litoralis* (P=100%), *M. australienses* (P=100%), *Candidatus* Nanosalarium (P=100%), *M. fervidus* (P=100%) and the ancestor of the *Thermoplasma*/*Ferroplasma*/*Picrophilus*/Thermoplasmatales clade (P=70%) all independently gained the intein. See the discussion section for consideration of the overestimation of intein absence in the species represented by the leaves of the cladogram in Figure 1.

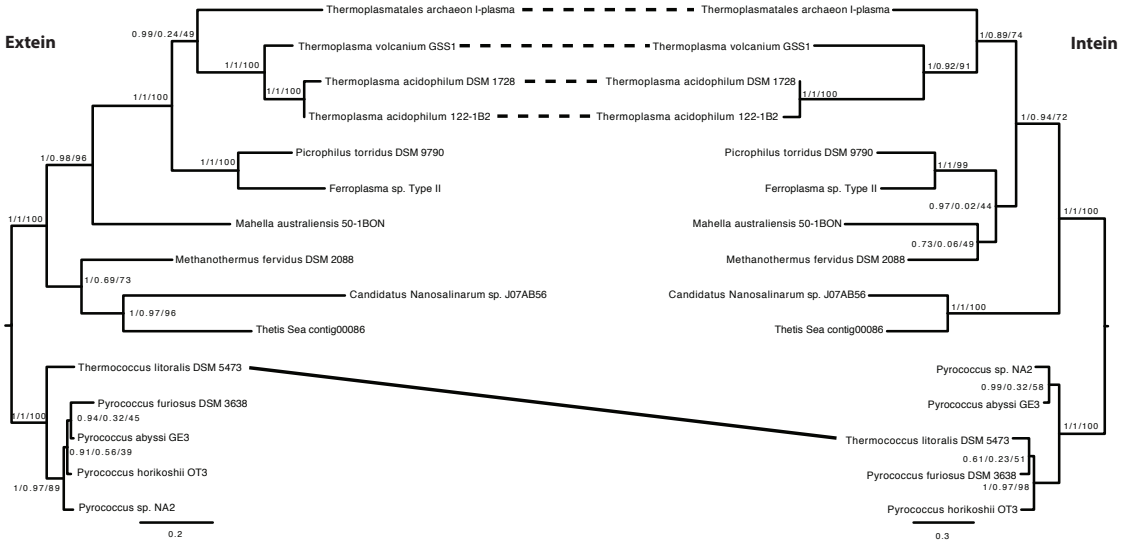


Figure 2. Extein and Intein Phylogenetic Trees. Phylogenetic trees depicting the extein (left) and intein (right). Support values (from left to right) were determined by posterior probability, approximate likelihood ratio test, and bootstrap replicates. Supported incongruencies are found between the extein and intein tree in the Pyrococcus/Thermococcus clade.

Multiple horizontal gene transfers of the *vma-1*b intein. Phylogenetic reconstructions were preformed for the extein and intein splicing domains. These trees are incongruent with each other by both the AU and SH tests ($p \leq 0.01$). This significant incongruence is due to a well-supported transfer of the intein from within the *Pyrococcus* spp. to *T. litorialis* (Figure 2). Unfortunately, none of the other discrepancies between the extein and intein trees show high bootstrap support. Including the HE domain in reconstruction of the intein's phylogeny did not improve the placement of the bacterial intein sequence from *Mahella australiensis* relative to the *Methanothermus* and *Picrophilus*/*Ferroplasma* inteins (Supplementary Figure S6). The ratio of extein to intein substitution rates provides additional insight on intein transfers relative to exteins.

Differences in divergence rates (Novichkov et al. 2004) and higher than expected sequence similarity (e.g., Podell and Gaasterland 2007) are frequently used to infer HGT events. Immediately after an intein

has been transferred between two species, the intein sequences are identical between donor and recipient and more conserved than the extein. This unexpected high sequence similarity of two inteins is an indcation of transfer of an intein (Liu et al. 1997). Moreover, the extein and intein evolve under very different selection pressures. While the ATPase catalytic subunit belongs to one of the most conserved protein families (Gogarten 1994; Swithers et al. 2009), the inteins evolve much faster making sequence based phylogenetic reconstruction difficult (Gogarten et al. 2002; Perler et al. 1997). This is reflected in the conservation profiles of the extein and intein (Supplemental Figure S7). The intein has a significantly higher average sequence variation score than the extein, 3.42 and 2.85 respectively ($p<0.01$, T-test). The higher score also indicates a higher substitution rate, because it represents the number of different amino acids in a sliding window. Among site rate variation for the extein sequences is more extreme than for the intein sequences, i.e. the exteins contain many sites under strong purifying selection (Supplemental Figure S8). To assess the divergence rate of the inteins relative to extein sequences, we used the well-aligned splicing domains of the intein, omitting the less conserved linker and homing endonuclease domains (see Alignments, Supplementary Material). The phylogenetic reconstructions of the extein and intein datasets are the same for the *Thermoplasma* genus and the Thermoplasmatales I-plasma archaeon, suggesting that in this group (in the following designated as *Thermoplasma* group) the extein and intein were inherited together. The inteins in these sequences (Table 1) also have lost their HE domain, which suggests they are no longer mobile and further strengthens the notion that the intein was inherited together with the extein for this clade. Therefore, the pairwise ratios of the substitution rates of extein to intein in this clade should represent ratios expected under co-inheritance of extein and intein. Ratios significantly greater than these (i.e. the intein is less divergent than expected), likely indicates horizontal transfer of the intein relative to the extein (see Table 2 for ratios). The chosen cut-off level of 0.7 corresponds to a significance level of $p<0.002$ given the rate ratios within the *Thermoplasma* group. To assess false positive rates obtained using this ratio approach we simulated sequence evolution along the extein tree using the parameters estimated for the intein and extein sequences. Using the rate ratio of 0.31, corresponding to the ratio of the lengths of intein and extein subtree in the *Thermoplasma* group, the rate of false positives was smaller than 0.000005. Incorporating uncertainty of the extein/intein ratio into the simulation, the false positive rate increases to 0.0025, corresponding to an expectation of 0.5 false identifications of transfer in Table 2.

The most parsimonious explanation for the ratios is mapped on the extein phylogeny in Figure 3. Although these values cannot resolve direction of transfer they do show the *vma-1*b inteins are mobile genetic elements that have undergone several transfers throughout the evolution of the gene family encoding the archaeal ATPase catalytic subunit, including several transfers between divergent organisms.

Table 2. Pairwise Extein to Intein Substitution Ratios.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Thermoplasmatales archaeon I-plasma														
2 *Thermoplasma volcanium* GSS1	0.39													
3 *T. acidophilim* DSM 1728	0.35	0.14												
4 *T. acidophilum* DSM 122-1B2	0.39	0.17	0.53											
5 *Feroplasma acidarmanus* Type II	0.49	0.40	0.37	0.41										
6 *Picrophilus torridus* DSM 9790	0.56	0.33	0.44	0.48	0.63									
7 *Mahella australiensis* 50-1 BON	0.55	0.44	0.42	0.46	**0.75**	**0.73**								
8 *Candidatus* Nanosalinarum so. J07AB56	0.59	0.57	0.59	0.72	**0.71**	**0.75**	0.64							
9 Thetis_Sea_contig_00086	0.60	0.51	0.49	0.58	0.60	0.59	0.56	**1.22**						
10 *Methanothermus fervidus* DSM 2088	**0.71**	0.50	0.48	0.50	**0.86**	**0.84**	**0.74**	0.48	0.37					
11 *Thermococcus litoralis* DSM 5473	0.45	0.37	0.31	0.34	0.53	0.44	0.34	0.48	0.42	0.32				
12 *Pyrococcus abyssi* GE5	0.48	0.38	0.32	0.35	0.54	0.42	0.36	0.53	0.47	0.38	0.41			
13 *Pyrococcus* sp. NA2	0.47	0.38	0.33	0.36	0.51	0.42	0.35	0.53	0.46	0.36	0.41	0.50		
14 *P. furiosus* DSM 3638	0.45	0.38	0.32	0.35	0.53	0.44	0.37	0.53	0.45	0.34	**1.29**	0.27	0.30	
15 *P. horikoshii* OT3	0.43	0.38	0.31	0.35	0.55	0.42	0.35	0.50	0.43	0.32	**0.76**	0.09	0.13	0.59

Number of amino acid substitutions were calculated for extein and intein sequences and corrected for multiple substitutions. Ratios greater than inside the *Thermoplasma* group (the top four entries in the table) are taken to be deviations from a ratio of extein to intein that represents co-inheritance of intein and extein. Bold font indicates ratios significantly greater than the ones within the *Thermoplasma* group.

Figure 3. Pairwise substitution rate ratios mapped onto the extein tree. Lines connect taxa with extein to intein substitution ratios greater than 0.7. Although direction of intein movement cannot be determined using these ratios, the ratios reveal several putative transfers of the *vma-1*b intein.

Discussion

Compared to the currently annotated sequences in InBase we have identified ten additional inteins belonging to the *vma-1*b family of inteins. Three of which are in new taxonomic classes that were not previously described, the Methanobacteria, Nanohaloarchaea, and Clostridia. This is the first time the *vma-1*b intein is found in a bacterial lineage.

Similar to the PRP8 inteins (Bokor et al. 2012) the *vma-1*b family of inteins is composed of mini inteins and inteins with homing endonuclease domains. These likely represent inteins in different states of the homing or life-cycle. In the homing cycle an exogenous intein with a homing endonuclease comes in contact with a population that does not contain an intein and spreads through the population through super-Mendelian rates of inheritance (Goddard and Burt 1999; Gogarten and Hilario 2006; Gogarten et al. 2002). Once all target sites in a population are occupied by inteins containing HEs there is reduced selection on the HE and it starts to degrade over time, followed by loss of the intein. The mini inteins of the Thermoplasmatales order may be representatives of the later part of the homing cycle, indicating a more ancient invasion of the *vma-1b* site in this clade. Even in cases where the homing cycle does not go to complete fixation in the intein containing allele, the individual intein insertion sites progress through the same life history from empty target site, occupation by an intein with functioning HE, and then eventual decay of the HE activity (Barzel et al. 2011; Yahara et al. 2009).

The evolutionary history of the vacuolar, archaeal and bacterial ATPase catalytic subunit is characterized by several gene transfer events, including transfers between the archaeal and the bacterial domain. Due to the sequence conservation of the catalytic ATPase subunit, and because the ATPase phylogeny can be rooted by an ancient gene duplication event (Gogarten et al. 1989), these transfer events have been well established (Hilario and Gogarten 1993; Lapierre 2007; Lapierre et al. 2006; Olendzenski et al. 2000).

Although there were multiple transfers of the host gene, all of our analyses were preformed relative to the host gene in order to distinguish the intein transfers from the transfers of the extein. Breakpoint analyses, consideration of pairwise substitution rates, and ancestral character state reconstructions all suggest that the inteins repeatedly invaded the *vma-1* host gene in independent transfer events. In particular these phylogenetic methods suggest that the intein in *Mahella australiensis* was gained independently from the between domain transfer of the host A-ATPase operon. The alternate explanation that the intein was gained only once in the ancestor of the *vma-1* protein family and that the extant presence/absence pattern is a result of vertical descent and loss of the intein is incompatible with the above listed evidence. In addition, under the loss-only scenario there would have to be a minimum of 28 steps to explain the patterning, one gain and 27 losses, opposed to only nine steps to explain the intein distribution through independent gains and losses of the intein. The scenario with the minimum number of events is depicted in Figure 1. It involves seven intein gains and two losses. Study of inteins and homing endonucleases in natural populations as well as theoretical considerations suggest that alleles with and without inteins can coexist in populations for long periods of time (Barzel et al. 2011; Butler et al. 2006; Gogarten and Hilario 2006; Yahara et al. 2009). Given that the sequences used in this study only represent one member of a larger population, assigning a species branch a character state of intein absence ignores the possibility that other members of the species might contain the intein. Thus, only taking into consideration the ASR, one needs to entertain the possibility that the loss events

inferred for branches represent only absence in the particular sampled genome, and not absence in the population or species. Likewise, the inference that the ancestor of the *Pyrococcus* and *Thermococcus* genera did not harbor the intein is not strongly supported through the parsimony reconstruction. The most parsimonious scenario depicted in figure 1 corresponds to two gains; the presence of the intein in the *Pyrococcus - Thermococcus* ancestor corresponds to three loss events within the *Pyrococcus* and *Thermococcus* group. Given that intein absence possibly is overestimated, three apparent loss events *versus* two gains do not provide a strong argument for intein gain in *T. litoralis*. However, the gain of the intein in *T. litoralis* is also supported by the phylogeny of the intein-splicing domain, which groups *T. litoralis* inside the cluster of homologs from the *Pyrococcus* species, supporting the hypothesis that the intein in *T. litoralis* was recently acquired from a *Pyrococcus* species. The general loss-only scenario is incompatible with a comparison of intein and extein phylogenies. If the phyletic pattern could be explained with one gain and subsequent losses, the datasets of the host extein phylogenetic tree and the intein tree should be compatible. However, the AU and SH tests, and the GARD analysis all confirm these trees are incompatible.

The GARD algorithm identifies breakpoints three amino acids upstream of the extein intein junctures. This slight deviation from the true intein extein boundary likely reflects a limitation of the approach due to the insertion site and the last three amino acids of the intein being highly conserved. Consequently, little phylogenetic information is contained in the amino acids directly neighboring the intein/extein boundary. However, alternative or additional explanations are possible: the positions in the extein upstream of the intein participate in catalyzing the splicing reaction (Paulus 2000; Saleh and Perler 2006), thus it is not be surprising that they have a signature similar to the intein's splicing domain; during homing the sequence surrounding the target site also is copied from the invading template containing the intein (Chevalier and Stoddard 2001).

The incongruence between the extein and intein trees in conjunction with the extein to intein divergence ratio data and the ancestral state reconstruction data all suggest that the inteins in the *vma-1*b position were horizontally acquired several times. The phylogenetic incongruence between the extein and intein tree provides strong support for the horizontal transfer of the *T. litoralis* intein from within the *Pyrococcus* spp.. The extein to intein divergence ratios suggest several additional transfers between *P. horikoshii* and *P. furious*; *T. litorialis* and *P. horikoshii*; *T. litorialis* and *P. furious*. The most likely scenario to explain these findings would be a within group transfer between the two *Pyrococcus* spp. followed by a transfer to *T. litorialis*. The phylogenetic incompatibility and the extein and intein substitution rate ratios taken together are strong support for a transfer from a *Pyrococcus* sp. to *T. litorialis*.

This is the first report of the *vma-1*b intein found in a bacterial lineage. Although the region where *M. australienses* falls on the intein tree is not well resolved, the extein to intein ratios suggest several transfers of the intein between the *Candidatus* Nanosalinarium and Thetis sea lineages, the *Methanothermus* and *Mahella* lineages, the *Methanothermus* and *Ferroplasma*/*Picrophilus* clade, and the *Mahella* and *Ferroplasma*/*Picrophilus* clade. The phylogenetic distribution relative to its host gene and the ratio data of the *vma-1*b intein taken together suggest an interdomain transfer between *M. australienses* and the Archaea. However, we cannot conclude that the intein was directly transferred between specific lineages; it is possible that both lineages recently received the intein from a third, currently unsampled lineage.

Conclusions

Here we surveyed the distribution of the *vma-1*b intein and showed that this family of inteins is found in more diverse species than previously reported. These inteins were transferred between divergent organisms in several independent horizontal transfer events, including an HGT from within the *Pyrococci* to *Thermococcus litorialis,* a likely transfer to *Candidatus* Nanosalinarum and a transfer across the domain boundaries between the Archaea and Bacteria. While an intein may persist within a lineage for a long time without transfer between species (Barzel et al. 2011; Butler et al. 2006; Gogarten and Hilario 2006; Yahara et al. 2009), our data convincingly show that transfer of the intein between divergent species did occur and contributed to the observed modern day distribution of this intein.

Materials and Methods

Sequences. For the large archaeal type ATPase tree sequences were gathered from the NCBI's non-redundant database. Sequences for the smaller fifteen sequence datasets were gathered from either the NCBI non-redundant, whole genome shotgun, or metagenome databases. The MG-RAST database was also surveyed for additional inteins but did not reveal any. All databases were queried in February 2012. The A-ATPases sequence from the Thetis sea metagenome (Ferrer et al. 2012) was translated from contig 00086 (gil3548555569) after correction of frame shifts.

Phylogenetic Trees. For most datasets (except the ribosomal reference) sequences were aligned with SATe 2.03 (Liu et al. 2009) using the following options: MAFFT for the initial alignment, MUSCLE for the merger, RAxML with a gamma plus invariant sites LG substitution model (PROTGAMMAILGF) for tree estimation. Intein splicing domain and extein sequences were also aligned using muscle (Edgar 2004) and PRANK (Loytynoja and Goldman 2010). Maximum likelihood phylogenies calculated from these alignments (WAG Gamma + F + I) for the exteins were identical to the one given in Figure 2. For both the muscle and the PRANK alignments, the *T. litoralis* intein sequence grouped as sister to *P. horikoshii* within the clade formed by the *Pyrococcus* homologs. For both alignments the grouping of the *T. litoralis* intein with *P. furiosus* and *P. horikoshii* inside the *Pyrococcus* clade was supported by a bootstrap support value of 96%. In the intein maximum likelihood phylogeny calculated from the PRANK alignment, the *Ferroplasma* and *Picrophilus* inteins group with the *Thermoplasma* homologs, but without significant support (52%). The phylogenetic trees were reconstructed using PhyML v3.0, with the best parameters determined by ProtTest3 Maximum likelihood phylogenies, approximate Likelihood Ratio Test (aLRT) and bootstrap support values were determined using PhyML v3.0. under the WAG+G+F+I model. Posterior probabilities were calculated in MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) using the WAG+G+F+I substitution model (Whelan and Goldman 2001) and two runs. After 70,000 generations the standard deviation of split frequencies were 0.004317 and 0.008746 for the extein and intein datasets, respectively (small standard deviation of split frequencies indicate adequate convergence of the two runs, the MrBayes tutorial recommends smaller than 0.01 as stop criterion). After inspection of the likelihood trace, the first 700 and 450 generations were discarded as burnins (beginning phase of the tree search that hasn't reach stationarity) for the extein and intein datasets, respectively.

Archaeal ribosomal phylogeny. We generated an archaeal ribosomal reference tree from a concatenated alignment of 62 ribosomal proteins (supplementary multiple sequence alignments) extracted from genomes downloaded from the NCBI ftp site. The ribosome tree included all archaeal taxa from the extein dataset except those that did not have genome representation, such as Thetis Sea contig from a metagenome project (Ferrer et al. 2012). This led to a total of 81 Archaea. The number of taxa in each alignment ranged from 19-81 (see Supplementary Table 1 for the complete taxon set). BLASTP (Altschul et al. 1997) searches with reference sets of bacterial and archaeal ribosomal proteins against each archaeal genome and an e-value cut-off of 1E-10 identified homologs. In most cases, each sequence from a reference set returned the same best BLAST hit from a genome; interpreted as strong evidence of homology. We also accepted instances where the majority of reference sequences returned the same best BLAST hit as evidence of homology (such as ratios of 80:1, where 80 references sequences returned the same best BLAST hit and only one differed). Muscle (Edgar 2004) aligned each protein set and Trimal v.1.2 (Capella-Gutierrez et al. 2009) with the "automated1" option trimmed ambiguous regions from all alignments, which were concatenated into a supermatrix with an in-house Python script. Phyml generated a maximum likelihood phylogeny with 100 bootstrap replicates under the same parameters as those used to generate the extein genealogy.

Reduced taxa genealogies We generated an extein genealogy and corresponding ribosomal phylogeny exclusively for organisms that contained inteins involved in HGT to highlight the correspondence between gene and reference tree. Phyml constructed ML trees under the same parameters as those employed for the full archaeal trees with 100 bootstrap replicates. PhyloBayes v.3.3e (Lartillot et al. 2009)performed Bayesian analyses under the CAT mixture model (Lartillot and Philippe 2004) with global exchange rates estimated by WAG and two chains for each analysis. The bpcomp option estimated the convergence of the two chains from the maximum difference of the bipartition frequencies. A total of 72,787 and 14,650 cycles were run for the reduced extein and ribosomal trees, respectively. The maximum bipartition frequency difference for the extein genealogy was 0.02 and 0.06 for the ribosomal phylogeny. We employed the CAT model for its efficacy in reducing or eliminating the effects of long-branch attraction. Three taxa were susceptible to LBA artifacts in the ribosomal tree due to amino acid composition and incomplete ribosomal datasets from incomplete genomes – *Micrarchaeum acidiphilium*, *Nanosalinarum sp*., and *Nanosalina sp*. The bacterial sequence from *Mahella australiensis* (which served as an outgroup) attracted *Nanosalinarum* to the base of the Archaea in the reduced ribosomal phylogeny and even the CAT model failed to overcome this LBA artifact.

Topology tests CONSEL v.0.2 performed both AU and SH tests to determine whether trees and their bootstrap replicates were significantly different from each other. Site-likelihoods were calculated in RAxML v.7.3.5 (Stamatakis et al. 2008).

Breakpoint Analysis. The Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond et al. 2006) as implemented on the Datamonkey web server (Delport et al. 2010) was used to determine potential points of recombination in the fifteen *vma*-1 proteins that harbor inteins. Sequences were aligned using SATé 2.1 (Liu et al. 2012), and sequences for the homing endonucleases were deleted prior to the analysis.

Substitution Rate Ratios. Pairwise maximum likelihood distances were calculated for each extein and each intein protein splicing domain alignment using TREE-PUZZLE 5.2 (Schmidt et al. 2002) using the WAG+F +I model for substitution and a Gamma distribution approximated through four discrete rate categories, such that the model was the same for all other phylogenic reconstructions. The extein and the

intein splicing domain were extracted from the SATe alignment (see above). The estimated alpha shape parameter for the intein splicing domains was 3.01 (S.E. 0.11), whereas the extein sequences showed a stronger ASRV with an estimated alpha of 1.37 (S.E. 0.04). To estimate the substitution rate ratio between extein and intein sequences under vertical inheritance, we used the distances estimated with TreePuzzle between *Thermoplasma acidophilum* DSM 122-1B2, *Thermoplasma volcanium* GSS1, and the Thermoplasmatales archaeon I-plasma (see table 1 and the spreadsheet in supplementary materials). The inteins in these Thermoplasmatales do not contain a homing endonuclease domain indicating they are no longer mobile and were likely inherited together with the extein (see figure 2). To obtain a phylogenetically independent estimate of the rate ratio, trees were calculated from the intein and extein distances within the *Thermoplama* using FITCH from the PHYLIP package (Felsenstein 2011), and rate ratios were calculated from the branch lengths. Ratios that were significantly larger than the within *Thermoplasma* ratios (mean ratio 0.30; SEM 0.13), which represent co-inheritance of intein and extein, suggest that the inteins are more similar to each other than under the assumption of co-inheritance with the exteins, indicating an intein transfer relative to the extein. Although direction of the intein transfer cannot be determined using these ratios recent transfers can be detected. Values greater than 0.70 were considered to be significantly greater than the *Thermoplasma* ratios ($p \leq 0.002$). The high cut-off level was chosen to avoid false positives.

False positive rates for the substitution ratio test. We performed two simulation studies to estimate the rate of false positives associated with the rate ratio cut-off used to identify putative transfer events. An intein tree, under the assumption of vertical inheritance only, was derived from the extein tree by dividing the branch lengths in the extein tree by 0.3145. This value was obtained from the ratio of the tree lengths within the *Thermoplama* group between the actual extein and intein trees. For the first simulation we used the aforementioned trees and simulated 10,000 sequence sets for the intein and extein tree using Evolver (Yang 2007). The sequences were simulated under the WAG model and alpha values were taken from the actual trees, which were 3.01 and 1.37 for the intein and extein, respectively. The substitution ratios were calculated. Extein distances less then 0.5 were excluded, and the number of ratios above 0.7 counted as false positives.

We also estimated the false positive rates incorporating the uncertainty in determining the rate ratio. This was done by randomly sampling 1000 rate ratios from a folded normal distribution with mean 0.3145 and standard deviation of 0.1343 (the SEM for the determined rate ratio). The sampled ratios were used to generate new intein trees. The simulated intein and extein trees were evaluated as described above.

Maximum Parsimony Ancestral State Reconstruction. A parsimony ancestral state reconstruction and a maximum likelihood ansectral state reconstruction were both performed to determine where among the *vma-1* protein family the inteins were gained. The presence/absence patterns of the inteins were converted to a binary form and the maximum parsimony ancestral state reconstruction was calculated using the ordered model and the maximum likelihood ancestral state reconstruction was calculated under the MK1 model (Lewis 2001). Both reconstructions were implemented in Mesquite (Maddison and Maddison 2011).

Supplementary Materials

fasta_files.zip: Multiple sequence alignments in fasta format:

Content:

sate_alignment_of_Intein_sequences.fasta
sate_alignment_of_Extein_and_Intein_sequences_(withoutHE).fasta
sate_alignment_of_alignment_of_extein_sequences.fasta
Concatenated_alignment_of_archaeal_ribosomal_proteins.fasta

Interleaved_alignments.pdf: interleaved alignments with annotations (SEE BELOW)

ratios.xls: Spreadsheet used for calculation of substitution ratios

Supplementary_Table_1.pdf: Ribosomal proteins in supermatrix and number of taxa per alignment. (SEE BELOW)

Supplementary_Figures.pdf (SEE BELOW)

Acknowledgements

This work was supported by the National Aeronautics and Space Administration Astrobiology, Exobiology and Evolutionary Biology (grant numbers NNX08AQ10G and NNX13AI03G) and the NSF Assembling the Tree of Life (ATOL) (DEB0830024) programs.

We thank Matthew Fullmer, David Williams, Tim Harlow, Ofir Cohen, and an anonymous reviewer for providing insightful discussions on the manuscript.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.

Anderson I, Djao OD, Misra M, Chertkov O, Nolan M, Lucas S, Lapidus A, Del Rio TG, Tice H, Cheng JF, Tapia R, Han C, Goodwin L, Pitluck S, Liolios K, Ivanova N, Mavromatis K, Mikhailova N, Pati A, Brambilla E, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Sikorski J, Spring S, Rohde M, Eichinger K, Huber H, Wirth R, Goker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Klenk HP, Kyrpides NC. Complete genome sequence of Methanothermus fervidus type strain (V24S). Stand Genomic Sci 2010;3(3):315-24.

Barzel A, Obolski U, Gogarten JP, Kupiec M, Hadany L. Home and Away- The Evolutionary Dynamics of Homing Endonucleases. BMC Evolutionary Biology 2011;11(1):324.

Bokor AA, Kohn LM, Poulter RT, van Kan JA. PRP8 inteins in species of the genus Botrytis and other ascomycetes. Fungal Genet Biol 2012.

Butler MI, Gray J, Goodwin TJ, Poulter RT. The distribution and evolutionary history of the PRP8 intein. BMC Evolutionary Biology 2006;6:42.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25(15):1972-3.

Chevalier BS, Stoddard BL. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res 2001;29(18):3757-74.

Clerissi C, Grimsley N, Desdevises Y. Genetic exchanges of inteins between prasinoviruses (phycodnaviridae). Evolution; international journal of organic evolution 2013;67(1):18-33.

Cooper AA, Stevens TH. Protein splicing: self-splicing of genetically mobile elements at the protein level. Trends Biochem Sci 1995;20(9):351-6.

Darland G, Brock TD, Samsonoff W, Conti SF. A thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. Science 1970;170(3965):1416-8.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 2010;26(19):2455-7.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. Community-wide analysis of microbial genome sequence signatures. Genome Biol 2009;10(8):R85.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 2004;32(5):1792-1797.

Elleuche S, Poggeler S. Inteins, valuable genetic elements in molecular biology and biotechnology. Appl Microbiol Biotechnol 2010;87(2):479-89.

Erauso G, Reysenbach A-L, Godfroy A, Meunier J-R, Crump B, Partensky F, Baross J, Marteinsson V, Barbier G, Pace N, Prieur D. Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Archives of Microbiology 1993;160(5):338-349.

Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.67. Distributed by the author. . Seattle: Department of Genome Sciences, University of Washington; 2011.

Ferrer M, Werner J, Chernikova TN, Bargiela R, Fernandez L, La Cono V, Waldmann J, Teeling H, Golyshina OV, Glockner FO, Yakimov MM, Golyshin PN. Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study. Environmental microbiology 2012;14(1):268-81.

Fiala G, Stetter KO. Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Archives of Microbiology 1986;145(1):56-61.
Goddard MR, Burt A. Recurrent invasion and extinction of a selfish gene. Proc Natl Acad Sci U S A 1999;96(24):13880-5.
Gogarten J, Kibak H, Dittrich P, Taiz L, Bowman E, Bowman B, Manolson M, Poole R, Date T, Oshima T, Konishi J, Denda K, Yoshida M. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 1989;86(17):6661-6665.
Gogarten JP. Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. J Mol Evol 1994;39(5):541-3.
Gogarten JP, Hilario E. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol Biol 2006;6:94.
Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. Inteins: structure, function, and evolution. Annu Rev Microbiol 2002;56:263-87.
Gonzalez JM, Masuchi Y, Robb FT, Ammerman JW, Maeder DL, Yanagibayashi M, Tamaoka J, Kato C. Pyrococcus horikoshii sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. Extremophiles 1998;2(2):123-30.
Harvey WR, Nelson N. V-ATPases (J Exp Biol). Cambridge, UK: The Company of Biologists; 1992.
Hilario E, Gogarten JP. Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems 1993;31(2-3):111-9.
Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y. Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. J Biol Chem 1990;265(12):6726-33.
Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001;17(8):754-5.
Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science 1990;250(4981):651-7.
Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. GARD: a genetic algorithm for recombination detection. Bioinformatics 2006;22(24):3096-8.
Koufopanou V, Goddard MR, Burt A. Adaptation for Horizontal Transfer in a Homing Endonuclease. Molecular biology and evolution 2002;19(3):239-246.
Lapierre P. PhD Thesis: The impact of horizontal gene transfers on prokaryotic genome evolution. Storrs: University of Connecticut; 2007.
Lapierre P, Shial R, Gogarten JP. Distribution of F- and A/V-type ATPases in Thermus scotoductus and other closely related species. Syst Appl Microbiol 2006;29(1):15-23.
Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 2009;25(17):2286-8.
Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution 2004;21(6):1095-109.
Lee HS, Bae SS, Kim MS, Kwon KK, Kang SG, Lee JH. Complete genome sequence of hyperthermophilic Pyrococcus sp. strain NA2, isolated from a deep-sea hydrothermal vent area. J Bacteriol 2011;193(14):3666-7.
Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol 2001;50(6):913-25.
Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 2009;324(5934):1561-4.

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Systematic Biology 2012;61(1):90-106.

Liu L, Laufer H, Gogarten PJ, Wang M. cDNA cloning of a mandibular organ inhibiting hormone from the spider crab Libinia emarginata. Invertebrate Neuroscience 1997;3(2-3):199-204.

Liu XQ. Protein-splicing intein: Genetic mobility, origin, and evolution. Annu Rev Genet 2000;34:61-76.

Loytynoja A, Goldman N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC bioinformatics 2010;11:579.

Macgregor BJ, Biddle JF, Teske A. Mobile Elements in a Single-Filament Orange Guaymas Basin Beggiatoa ("Candidatus Maribeggiatoa") sp. Draft Genome: Evidence for Genetic Exchange with Cyanobacteria. Applied and environmental microbiology 2013;79(13):3974-85.

Maddison WP, Maddison DR. 2011 Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 2012;6(1):81-93.

Neuner A, Jannasch HW, Belkin S, Stetter KO. <i>Thermococcus litoralis</i> sp. nov.: A new species of extremely thermophilic marine archaebacteria. Archives of Microbiology 1990;153(2):205-207.

Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, Koonin EV. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. Bacteriol. 2004;186(19):6575-6585.

Okuda Y, Sasaki D, Nogami S, Kaneko Y, Ohya Y, Anraku Y. Occurrence, horizontal transfer and degeneration of VDE intein family in Saccharomycete yeasts. Yeast 2003;20(7):563-73.

Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP. Horizontal transfer of archaeal genes into the Deinococcaceae: Detection by molecular and computer-based approaches. Journal of molecular evolution 2000;51(6):587-99.

Paulus H. Protein splicing and related forms of protein autoprocessing. Annu Rev Biochem 2000;69:447-96.

Perler FB. InBase: the Intein Database. Nucleic Acids Res 2002;30(1):383-4.

Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J, Belfort M. Protein splicing elements: inteins and exteins--a definition of terms and recommended nomenclature. Nucleic Acids Res 1994;22(7):1125-7.

Perler FB, Olsen GJ, Adam E. Compilation and analysis of intein sequences. Nucleic Acids Res 1997;25(6):1087-93.

Pietrokovski S. Modular organization of inteins and C-terminal autocatalytic domains. Protein Sci 1998;7(1):64-71.

Podell S, Gaasterland T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol 2007;8(2):R16.

Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003;19(12):1572-4.

Saleh L, Perler FB. Protein splicing in cis and in trans. Chemical record 2006;6(4):183-93.

Salinas MB, Fardeau ML, Thomas P, Cayol JL, Patel BK, Ollivier B. Mahella australiensis gen. nov., sp. nov., a moderately thermophilic anaerobic bacterium isolated from an Australian oil well. Int J Syst Evol Microbiol 2004;54(Pt 6):2169-73.

Schleper C, Piihler G, Kuhlmorgen B, Zillig W. Life at extremely low pH. Nature 1995;375(6534):741-742.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002;18(3):502-4.

Searcy DG. Histone-like protein in the prokaryote Thermoplasma acidophilum. Biochimica et biophysica acta 1975;395(4):535-47.

Segerer A, Langworthy TA, Stetter KO. Thermoplasma acidophilum and Thermoplasma volcanium sp. nov. from Solfatara Fields. Systematic and Applied Microbiology 1988;10(2):161-171.

Senejani AG, Hilario E, Gogarten JP. The intein of the Thermoplasma A-ATPase A subunit: structure, evolution and expression in E. coli. BMC Biochem 2001;2:13.

Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. Systematic biology 2008;57(5):758-71.

Swithers KS, Senejani AG, Fournier GP, Gogarten JP. Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. BMC Evol Biol 2009;9:303.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 2004;428(6978):37-43.

Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution 2001;18(5):691-9.

Yahara K, Fukuyo M, Sasaki A, Kobayashi I. Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer. Proc Natl Acad Sci U S A 2009;106(44):18861-6.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 2007;24(8):1586-91.

Zimniak L, Dittrich P, Gogarten JP, Kibak H, Taiz L. The cDNA sequence of the 69-kDa subunit of the carrot vacuolar H+- ATPase. Homology to the beta-chain of F0F1-ATPases. The Journal of biological chemistry 1988;263(19):9102-12.

Supplementary Table 1. Ribosomal proteins in supermatrix and number of taxa per alignment.

Ribosomal protein	# sequences in alignment
L13e	19
S25e	20
S26e	21
S30e	21
S1P	35
L34e	41
L14e	43
L30e	66
L37e	71
L39e	74
S27Ae	74
L24e	77
L29	78
L31e	78
S27e	78
L44e	79
S17e	79
S24e	79
L10	80
L14	80
L15e	80
L18ae	80
L18e	80
L1	80
L21e	80
L22	80
L23	80
L2	80
L3	80
L40e	80
L4	80
L7Ae	80
S11	80
S13	80
S14	80
S15	80
S17	80
S19e	80
S19	80
S28e	80
S2	80
S3Ae	80
S3p	80
S4	80
S6e	80
S8e	80
S9	80
L15	81
L16P	81
L18	81
L19e	81
L24	81
L30p	81
L32e	81
L5	81
L6	81
S10	81
S12	81
S4e	81
S5	81
S7	81
S8	81

Supplementary Figures

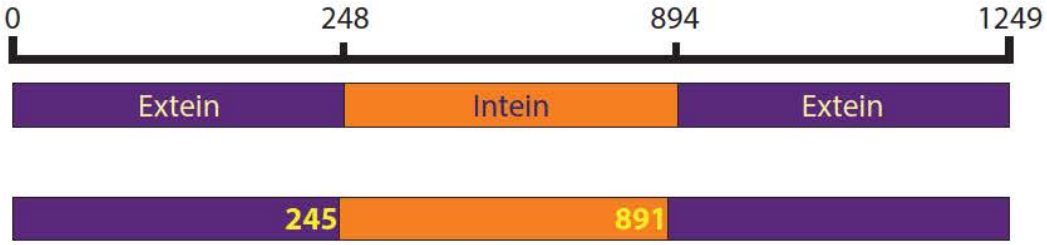


Figure S1. Breakpoint Analysis of the VMA-1b Extein and Intein sequences. A GARD analysis (Kosakovsky Pond et al. 2006) was performed for organisms for which complete extein and intein sequences were available (table 1). The purple is the extein and the orange is the intein. Significant breakpoints (p=<0.01) were determined to be at amino acid positions 245 and 891 on the alignment. These breakpoints are 3 amino acids before the intein/extein boundaries.

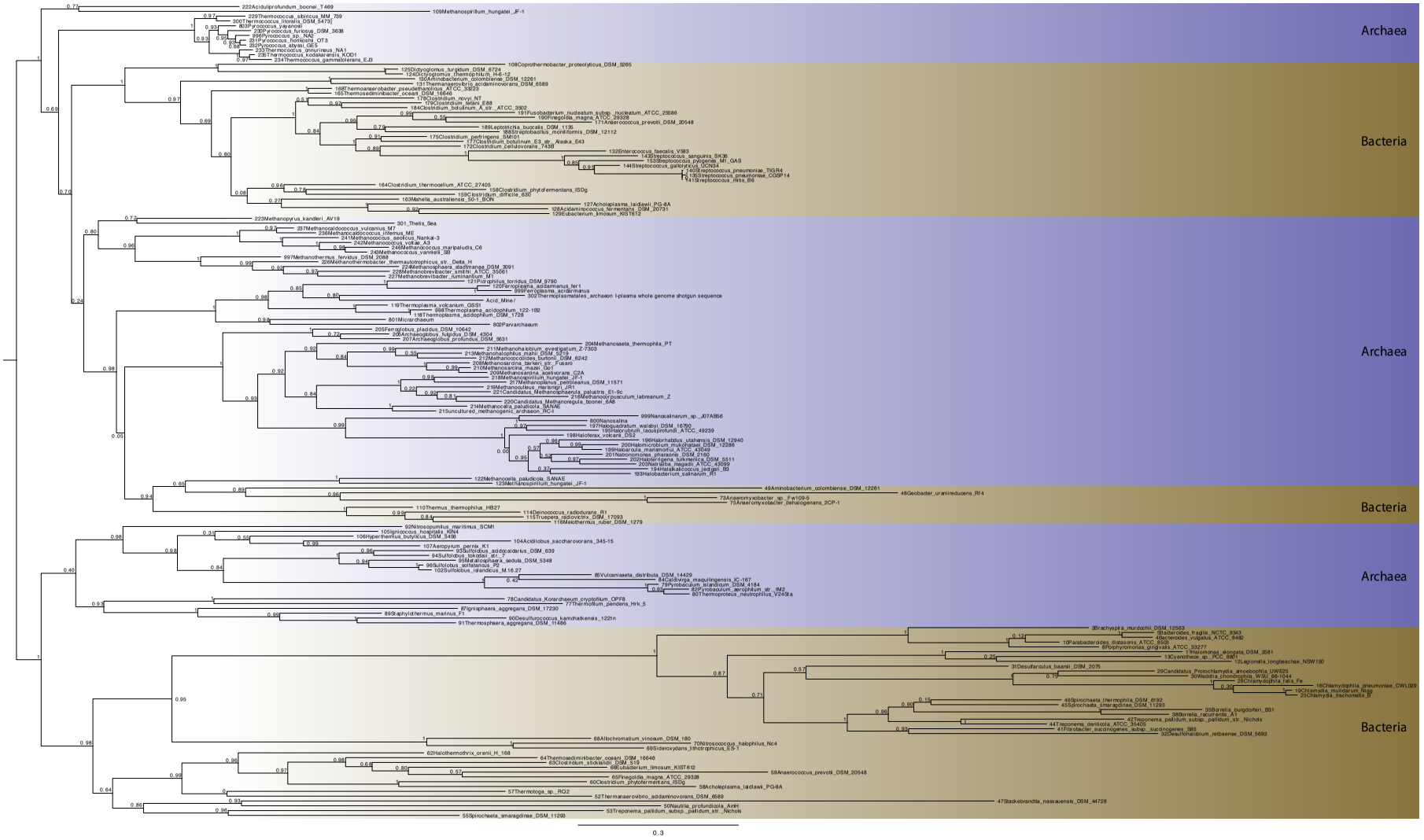


Figure S2. Maximum Likelihood Phylogenetic Tree of the VMA-1 Protein Family Illustrating the Distribution of the Catalytic ATPase Subunit. Support values on the nodes of the tree were determined by the approximate likelihood ratio test. Highlighted in brown are bacterial and in purple are archaeal representatives.

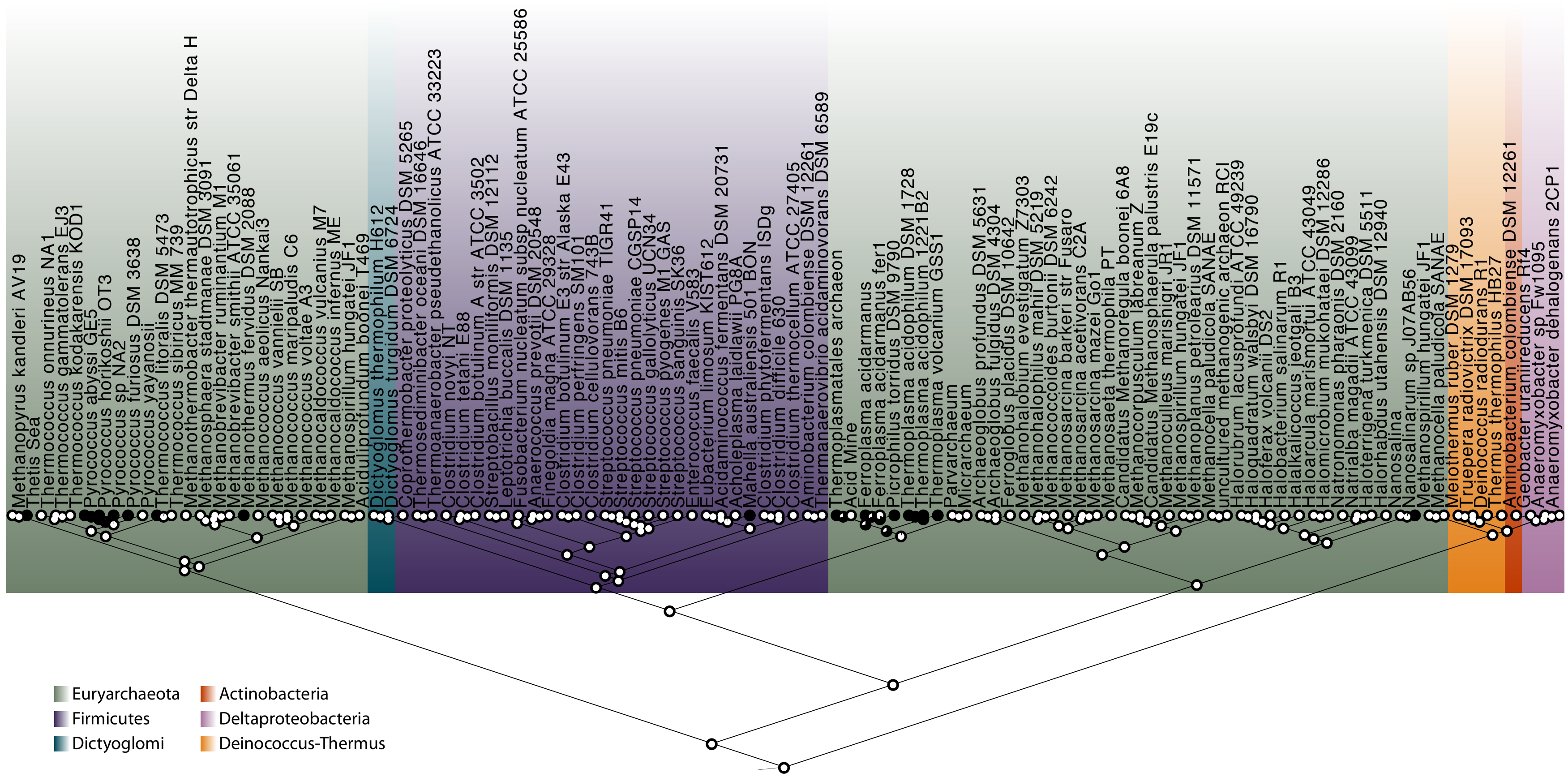


Figure S3. Maximum Likelihood Ancestral Character State Reconstruction of *vma*-1b Intein Presence and Absence. The depicted cladogram represents the part of the VMA-1 tree where inteins are present (see Supplemental Figure S2 for full tree). Presence and absences of the intein are represented at the tips of the tree; black circles correspond to presence of the intein and white circles correspond to absences of the intein. Partially filled circles represent the probability of an intein present at the node. This reconstruction was calculated under the MK1 model (Lewis 2001) as implemented in mesquite (Maddison and Maddison 2011).

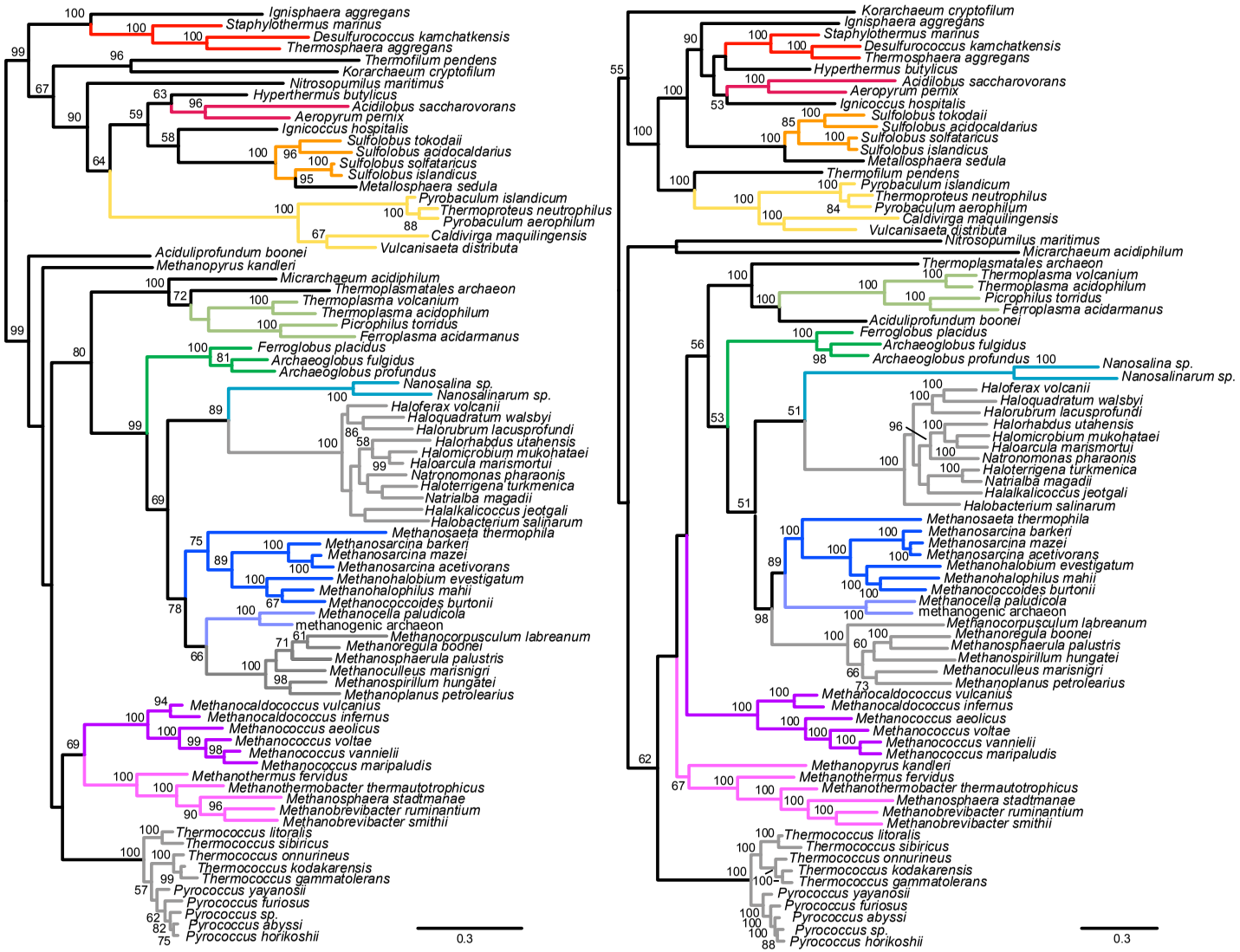


Figure S4. Phylogenetic Trees of the Ribosomal Reference (right) and Corresponding Extein Sequences (left). Maximum likelihood trees were generated by Phyml. Trees are rooted between Euryarchaeota and Crenarchaeota. Colored branches represent identical subtree topologies between the extein and ribosome trees. Gray branches indicate clades comprised of the same members between the two trees but different internal topologies. Bootstrap values < 50% not shown; 0.3, 0.3 amino acid substitutions per site.

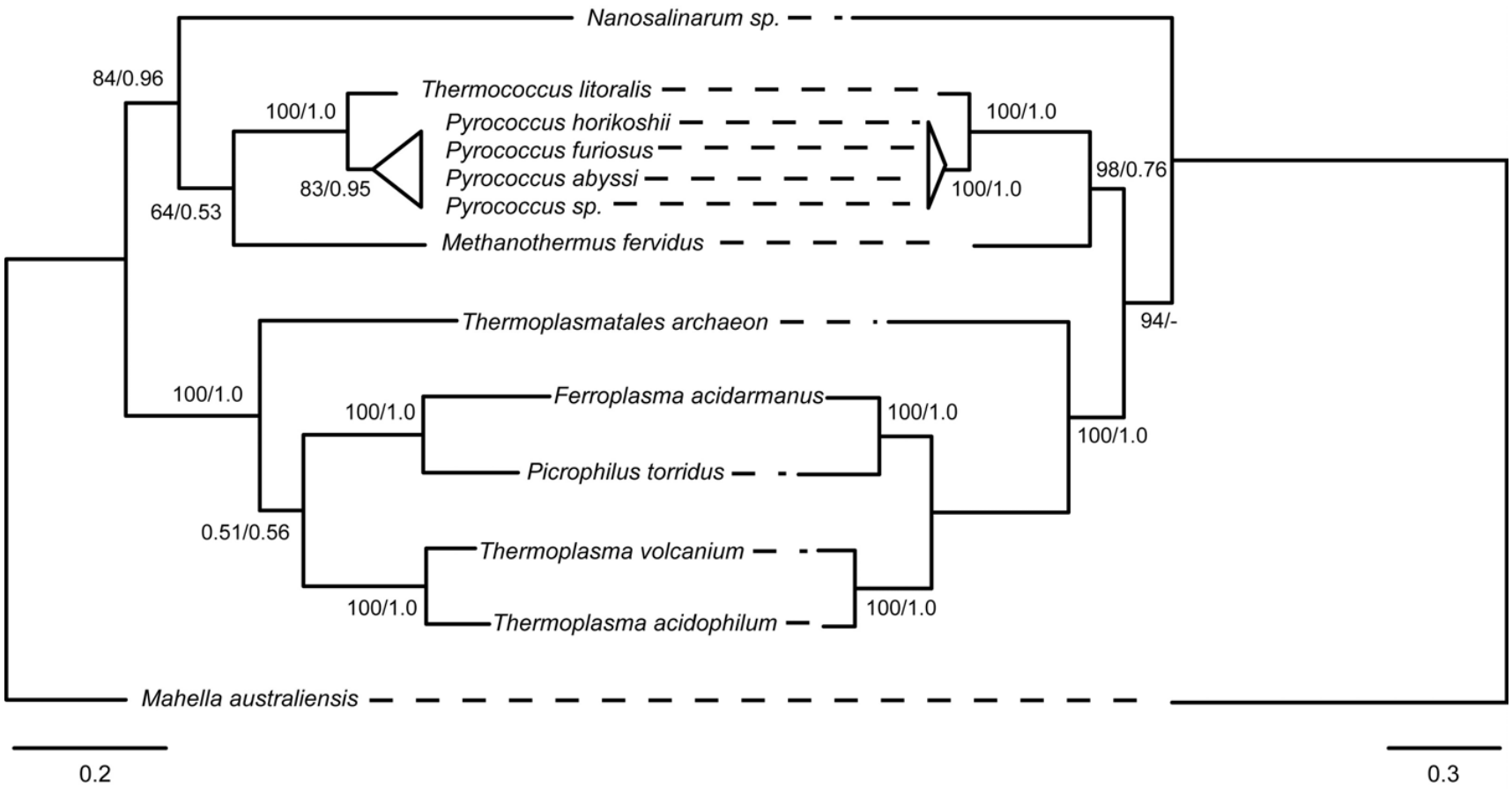


Figure S5. Reduced Archaeal Extein (left) and Ribosome Reference (right) Trees for Taxa Showing Evidence of Intein HGT. Bootstrap support values for the maximum likelihood trees generated in Phyml and posterior probabilities for the Bayesian trees constructed in PhyloBayes are shown to the left and right of the forward slash, respectively. Trees are rooted with the bacterium, *Mahella australiensis*, which also shows evidence of intein HGT. While the intein genealogy displays incongruence with the extein tree - suggestive of HGT - the extein tree is congruent with the ribosome reference tree, which is not prone to HGT. The only topological difference stems from the incorrect placement of *Candidatus* Nanosalinarum sp. in the ribosome tree due to LBA from unusual amino acid composition and an incomplete ribosomal protein dataset. Note, with greater taxonomic sampling in the complete archaeal tree (see supplementary figure 3), *Candidatus* Nanosalinarum sp. groups with the Halobacteria, which is considered the proper placement for this taxon.

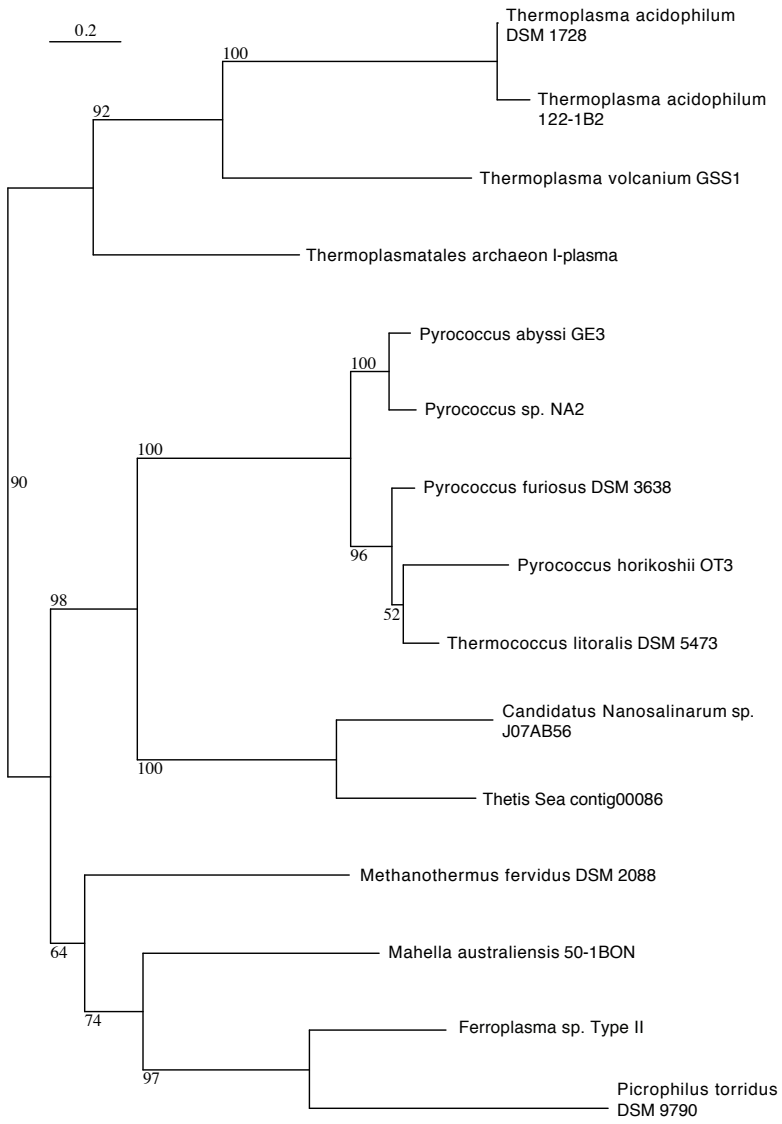


Figure S6. Nonparametric Bootstrap Phylogenetic Analysis of Intein Sequences including the Homing Endonuclease Domain. The analysis was performed using phyml with the WAG model, and Gamma plus I model with estimated parameters for among site rate variation. See material and methods for details.

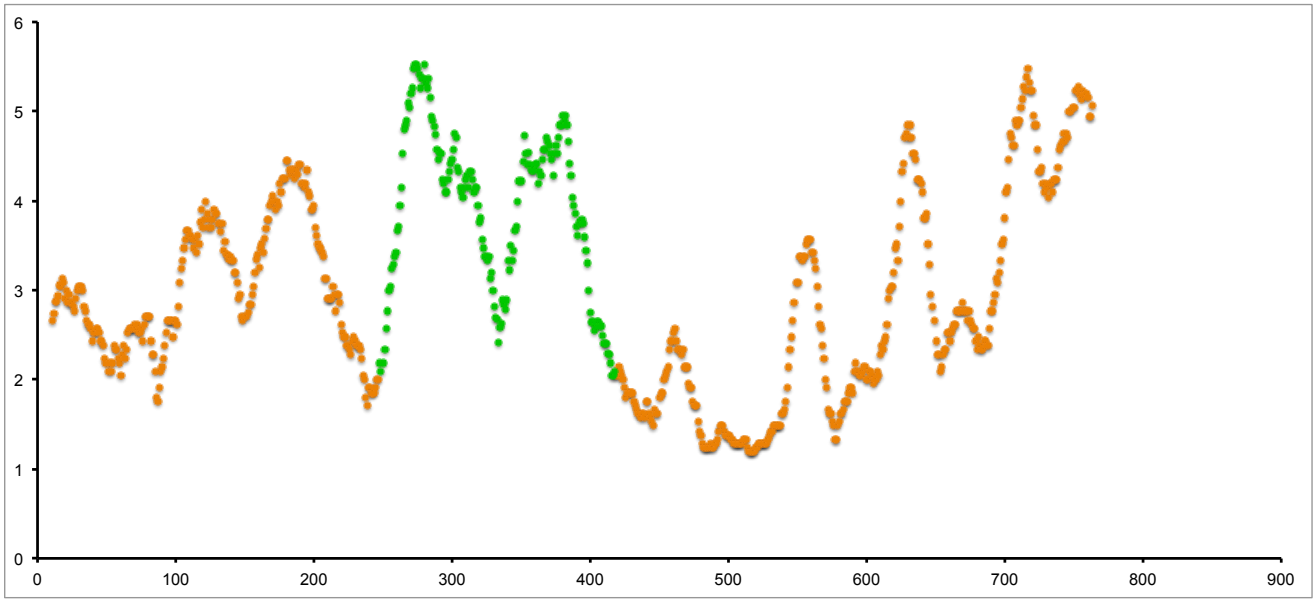


Figure S7. Conservation Profile of the Extein and Splicing Domain. The y-axis gives the average number of amino acid in a ten amino acid sliding window of the alignment and the x-axis is the sequence position in the alignment. In orange is the extein and in green is the intein splicing domains.

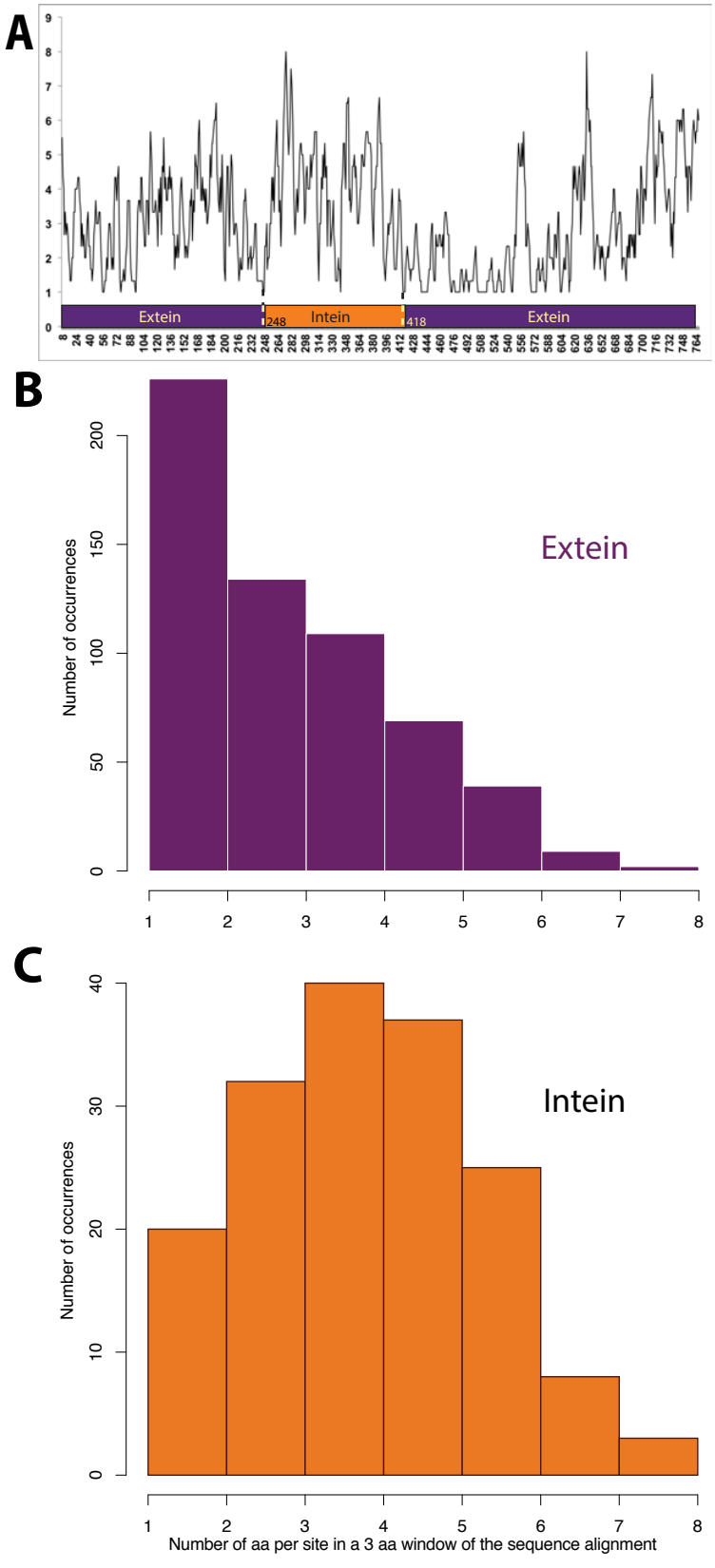
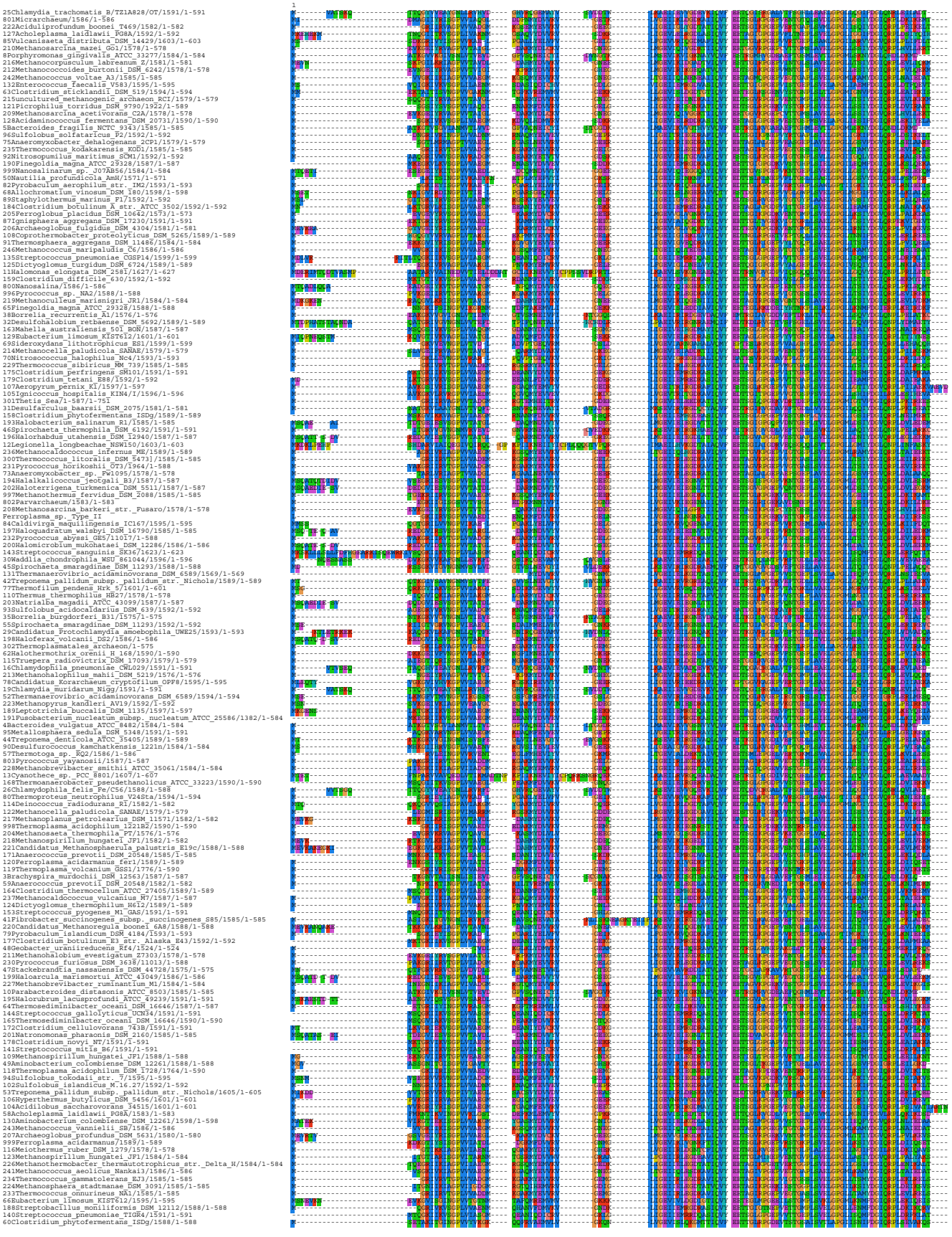

A
Extein
248
Intein
418
Extein
B
Extein
Number of occurrences
C
Intein
Number of occurrences
Number of aa per site in a 3 aa window of the sequence alignment

Figure S8. Site Conservation in VMA-1b Extein and Intein Sequences. Panel A gives the conservation profile for the extein and intein splicing domain in a three amino acid sliding window. The bar gives the location of extein (purple) and intein (orange) sequences. Dotted lines indicate the breakpoints determined in the GARD analysis. Panel B and C give histograms for average conservation per site in three amino acid windows of extein (B) and intein (C) sequences respectively. Note that the three amino acids at the carboxy terminal end of the intein are completely conserved in this data set, and that this level of conservation occurs more frequently in the extein sequences.

Interleaved Alignments with Annotations

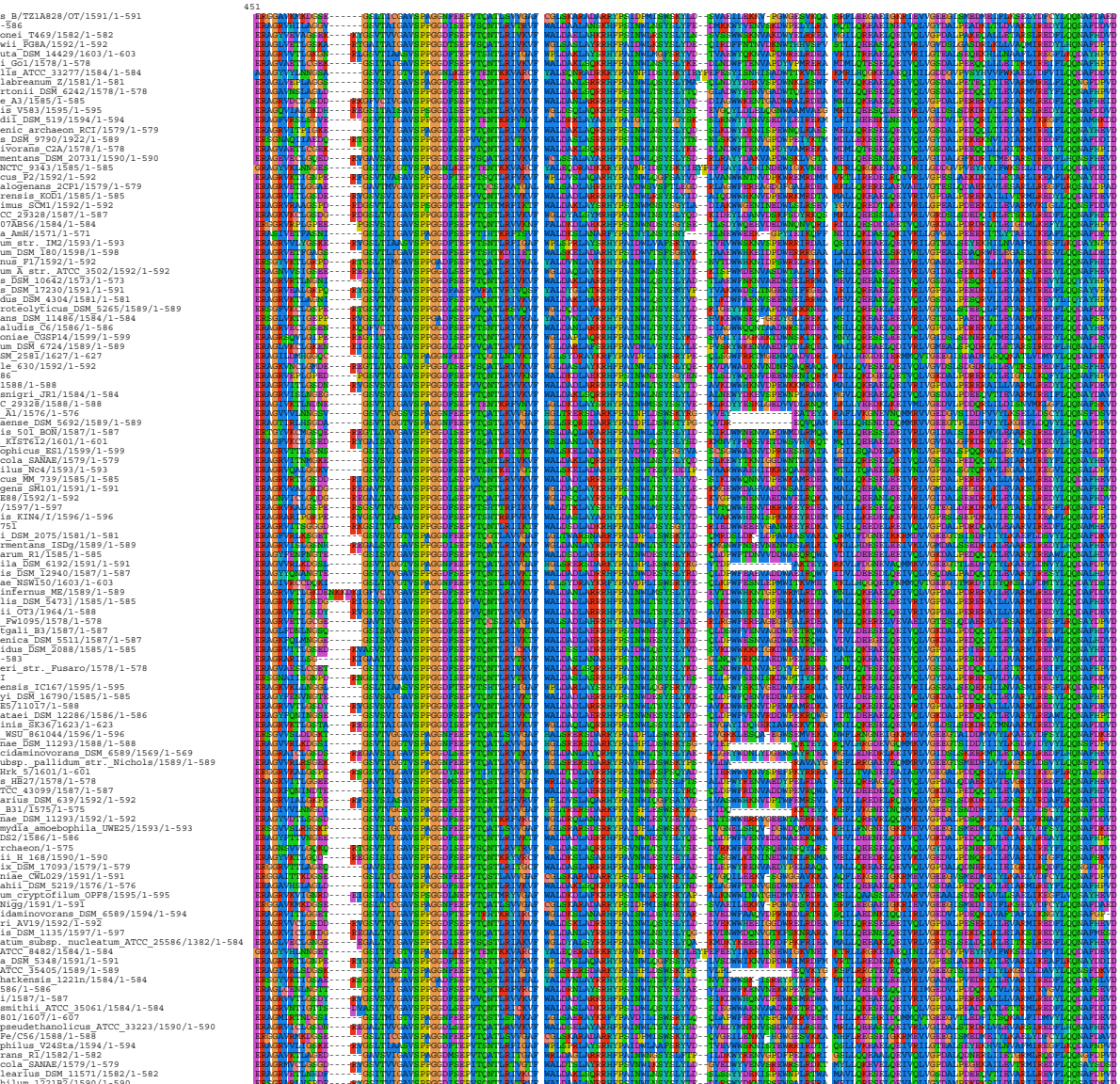
Content:

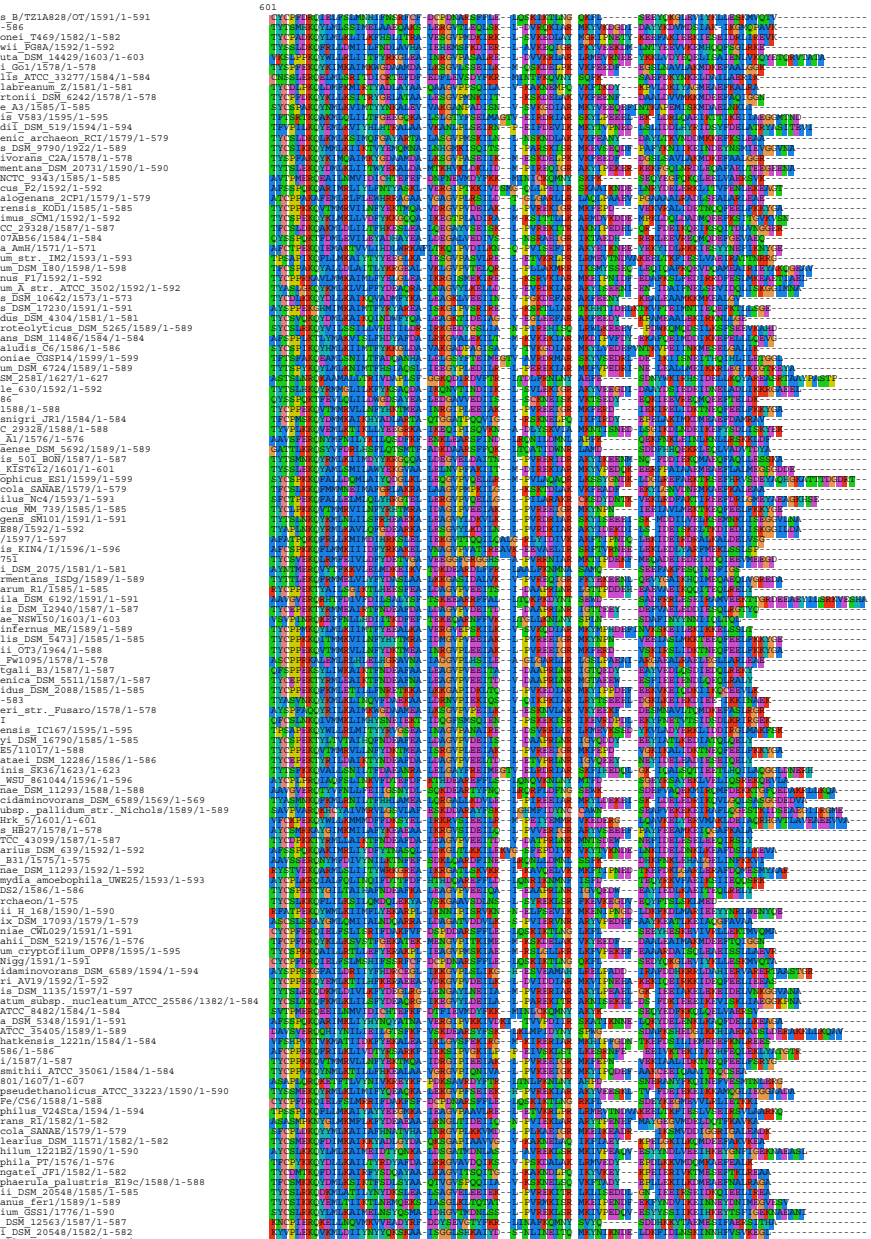
Alignment: /Users/jpgogarten/Dropbox/supplementary material for upload/alignement_of_extein_sequences.fasta
Seaview [blocks=50 fontsize=4 LETTER] on Fri Jul 19 09:36:36 2013



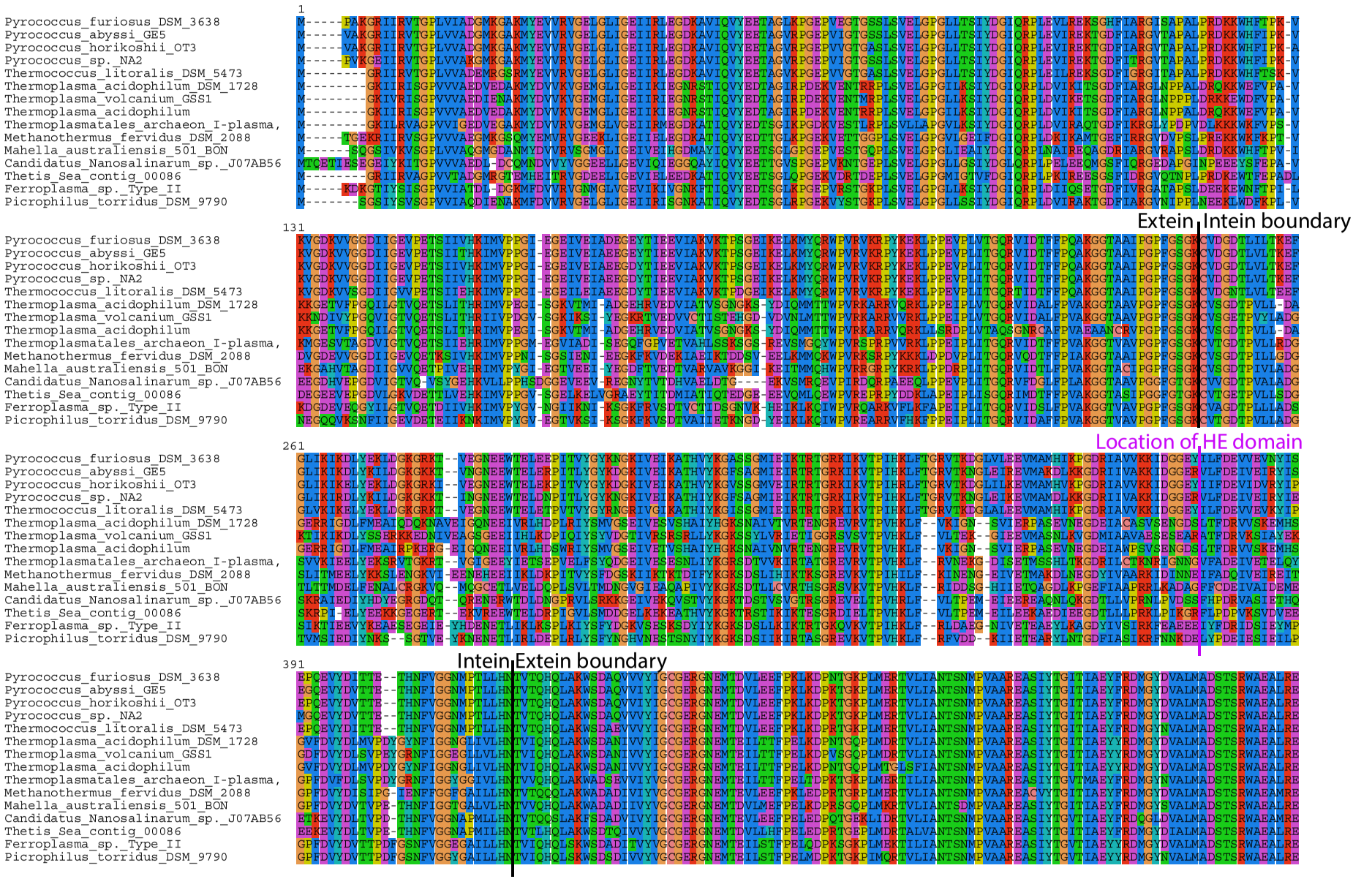

151
25Chlamydia_trachomatis_B/TZ1A828/OT/1591/1-591
801Micrarchaeum/1586/1-586
222Aciduliprofundum_boonei_T469/1582/1-582
127Acholeplasma_laidlawii_PG8A/1592/1-592
85Vulcanisaeta_distributa_DSM_14429/1603/1-603
210Methanosarcina_mazei_Go1/1578/1-578
8Porphyromonas_gingivalis_ATCC_33277/1584/1-584
216Methanocorpusculum_labreanum_Z/1581/1-581
212Methanococcoides_burtonii_DSM_6242/1578/1-578
242Methanococcus_voltae_A3/1585/1-585
132Enterococcus_faecalis_V583/1595/1-595
63Clostridium_sticklandii_DSM_519/1594/1-594
215uncultured_methanogenic_archaeon_RCI/1579/1-579
121Picrophilus_torridus_DSM_9790/1922/1-589
209Methanosarcina_acetivorans_C2A/1578/1-578
128Acidaminococcus_fermentans_DSM_20731/1590/1-590
5Bacteroides_fragilis_NCTC_9343/1585/1-585
96Sulfolobus_solfataricus_P2/1592/1-592
75Anaeromyxobacter_dehalogenans_2CP1/1579/1-579
235Thermococcus_kodakarensis_KOD1/1585/1-585
92Nitrosopumilus_maritimus_SCM1/1592/1-592
190Pinegolidia_magna_ATCC_29328/1587/1-587
999Nanosalinarum_sp._J07AB56/1584/1-584
50Nautilia_profundicola_AmH/1571/1-571
82Pyrobaculum_aerophilum_str._IM2/1593/1-593
68Allochromatium_vinosum_DSM_180/1598/1-598
89Staphylothermus_marinus_F1/1592/1-592
184Clostridium_botulinum_A_str._ATCC_3502/1592/1-592
205Perroglobus_placidus_DSM_10642/1573/1-573
87Ignisphaera_aggregans_DSM_17230/1591/1-591
206Archaeoglobus_fulgidus_DSM_4304/1581/1-581
108Coprothermobacter_proteolyticus_DSM_5265/1589/1-589
91Thermosphaera_aggregans_DSM_11486/1584/1-584
246Methanococcus_maripaludis_C6/1586/1-586
135Streptococcus_pneumoniae_CGSP14/1599/1-599
125Dictyoglomus_turgidum_DSM_6724/1589/1-589
11Halomonas_elongata_DSM_2581/1627/1-627
159Clostridium_difficile_630/1592/1-592
800Nanosalina/1586/1-586
996Pyrococcus_sp._NA2/1588/1-588
219Methanoculleus_marisnigri_JR1/1584/1-584
65Pinegoldia_magna_ATCC_29328/1588/1-588
38Borrelia_recurrentis_A1/1576/1-576
32Desulfohalobium_retbaense_DSM_5692/1589/1-589
163Mahella_australiensis_50I_BON/1587/1-587
129Eubacterium_limosum_KIST612/1601/1-601
69Sideroxydans_lithotrophicus_ES1/1599/1-599
214Methanocella_paludicola_SANAE/1579/1-579
70Nitrosococcus_halophilus_Nc4/1593/1-593
229Thermococcus_sibiricus_MM_739/1585/1-585
175Clostridium_perfringens_SM101/1591/1-591
179Clostridium_tetani_E88/1592/1-592
107Aeropyrum_pernix_K1/1597/1-597
105Ignicoccus_hospitalis_KIN4/I/1596/1-596
301Thetis_Sea/1-597/1-591
31Desulfarculus_baarsii_DSM_2075/1581/1-581
158Clostridium_phytofermentans_ISDg/1589/1-589
193Haloterrigena_turkmenica_DSM_5511/1587/1-587
46Spirochaeta_thermophila_DSM_6192/1591/1-591
196Halorhabdus_utahensis_DSM_12940/1587/1-587
12Legionella_longbeachae_NSW150/1603/1-603
236Methanocaldococcus_infernus_ME/1589/1-589
300Thermococcus_litoralis_DSM_5473]/1585/1-585
231Pyrococcus_horikoshii_OT3/1964/1-588
73Anaeromyxobacter_sp._PW1095/1578/1-578
194Halalkalicoccus_jeotgali_B3/1587/1-587
202Haloterrigena_turkmenica_DSM_5511/1587/1-587
997Methanothermus_fervidus_DSM_2088/1585/1-585
802Parvarchaeum/1583/1-583
208Methanosarcina_barkeri_str._Fusaro/1578/1-578
Ferroplasma_sp._Type_II
84Caldivirga_maquilingensis_IC167/1595/1-595
197Haloquadratum_walsbyi_DSM_16790/1585/1-585
232Pyrococcus_abyssi_GE5/1101/1-588
200Halomicrobium_mukohataei_DSM_12286/1586/1-586
143Streptococcus_sanguinis_SK36/14623/1-623
30Waddlia_chondrophila_WSU_861044/1596/1-596
45Spirochaeta_smaragdinae_DSM_11293/1588/1-588
131Thermanaerovibrio_acidaminovorans_DSM_6589/1569/1-569
42Treponema_pallidum_subsp._pallidum_str._Nichols/1589/1-589
77Thermofilum_pendens_Hrk_5/1601/1-601
110Thermus_thermophilus_HB27/1578/1-578
203Natrialba_magadii_ATCC_43099/1587/1-587
93Sulfolobus_acidocaldarius_DSM_639/1592/1-592
35Borrelia_burgdorferi_B31/1575/1-575
55Spirochaeta_smaragdinae_DSM_11293/1592/1-592
29Candidatus_Protochlamydia_amoebophila_UWE25/1593/1-593
198Haloferax_volcanii_DS2/1586/1-586
302Thermoplasmatales_archaeon/1-575
62Halothermothrix_orenii_H_168/1590/1-590
115Truepera_radiovictrix_DSM_17093/1579/1-579
16Chlamydophila_pneumoniae_CWL029/1591/1-591
213Methanohalophilus_mahii_DSM_5219/1576/1-576
78Candidatus_Korarchaeum_cryptofilum_OPF8/1595/1-595
19Chlamydia_muridarum_Nigg/1591/1-591
52Thermanaerovibrio_acidaminovorans_DSM_6589/1594/1-594
221Methanopyrus_kandleri_AV19/1592/1-592
189Leptotrichia_buccalis_DSM_1135/1597/1-597
191Pusobacterium_nucleatum_subsp._nucleatum_ATCC_25586/1382/1-584
4Bacteroides_vulgatus_ATCC_8482/1584/1-584
95Metallosphaera_sedula_DSM_5348/1591/1-591
44Treponema_denticola_ATCC_35405/1589/1-589
90Desulfurococcus_kamchatkensis_1221n/1584/1-584
57Thermotoga_sp._RQ2/1586/1-586
803Pyrococcus_yayanosii/1587/1-587
228Methanobrevibacter_smithii_ATCC_35061/1584/1-584
13Cyanothece_sp._PCC_8801/1607/1-607
168Thermoanaerobacter_pseudethanolicus_ATCC_33223/1590/1-590
26Chlamydophila_felis_Fe/C56/1588/1-588
80Thermoproteus_neutrophilus_V24Sta/1594/1-594
114Deinococcus_radiodurans_R1/1582/1-582
122Methanocella_paludicola_SANAE/1579/1-579
217Methanoplanus_petrolearius_DSM_11571/1582/1-582
998Thermoplasma_acidophilum_1221B2/1590/1-590
204Methanosaeta_thermophila_PT/1576/1-576
218Methanospirillum_hungatei_JF1/1582/1-582
221Candidatus_Methanosphaerula_palustris_E1-9c/1588/1-588
171Anaerococcus_prevotii_DSM_20548/1585/1-585
120Ferroplasma_acidarmanus_fer1/1589/1-589
119Thermoplasma_volcanium_GSS1/1776/1-590
3Brachyspira_murdochii_DSM_12563/1587/1-587
59Anaerococcus_prevotii_DSM_20548/1582/1-582
164Clostridium_thermocellum_ATCC_27405/1589/1-589
237Methanocaldococcus_vulcanius_M7/1587/1-587
124Dictyoglomus_thermophilum_H6I2/1589/1-589
153Streptococcus_pyogenes_M1_GAS/1591/1-591
41Pirobacter_succinogenes_subsp._succinogenes_S85/1585/1-585
220Candidatus_Methanoregula_boonei_6A8/1588/1-588
79Pyrococcus_islandicum_DSM_4184/1593/1-593
177Clostridium_botulinum_E3_str._Alaska_E43/1592/1-592
48Geobacter_uraniireducens_Rf4/1524/1-524
211Methanohalobium_evestigatum_Z7303/1578/1-578
230Pyrococcus_furiosus_DSM_3638/11013/1-588
47Stackebrandtia_nassauensis_DSM_44728/1575/1-575
199Haloarcula_marismortui_ATCC_43049/1586/1-586
227Methanobrevibacter_ruminantium_M1/1584/1-584
10Parabacteroides_distasonis_ATCC_8503/1585/1-585
195Halorubrum_lacusprofundi_ATCC_49239/1591/1-591
64Thermosediminibacter_oceani_DSM_16646/1587/1-587
144Streptococcus_gallolyticus_UCN34/1591/1-591
165Thermosediminibacter_oceani_DSM_16646/1590/1-590
172Clostridium_cellulovorans_743B/1591/1-591
201Natronomonas_pharaonis_DSM_2160/1585/1-585
178Clostridium_novyi_NT/1591/1-591
141Streptococcus_mitis_B6/1591/1-591
109Methanospirillum_hungatei_JF1/1588/1-588
49Aminobacterium_colombiense_DSM_12261/1588/1-588
118Thermoplasma_acidophilum_DSM_1728/1764/1-590
94Sulfolobus_tokodaii_str._7/1595/1-595
102Sulfolobus_islandicus_M.14.25/1592/1-592
53Treponema_pallidum_subsp._pallidum_str._Nichols/1605/1-605
106Hyperthermus_butylicus_DSM_5456/1601/1-601
104Acidilobus_saccharovorans_34Y15/1601/1-601
58Acholeplasma_laidlawii_PG8A/1583/1-583
130AminobacterIum_colombiense_DSM_12261/1598/1-598
243Methanococcus_vannielii_SB/1586/1-586
207Archaeoglobus_profundus_DSM_5631/1580/1-580
999Ferroplasma_acidarmanus/1589/1-589
116Meiothermus_ruber_DSM_1279/1578/1-578
123Methanospirillum_hungatei_JF1/1584/1-584
226Methanothermobacter_thermautotrophicus_str._Delta_H/1584/1-584
241Methanococcus_aeolicus_Nankai3/1586/1-586
234Thermococcus_gammatolerans_EJ3/1585/1-585
224Methanosphaera_stadtmanae_DSM_3091/1585/1-585
233Thermococcus_onnurineus_NA1/1585/1-585
66Eubacterium_limosum_KIST612/1595/1-595
188Streptobacillus_moniliformis_DSM_12112/1588/1-588
140Streptococcus_pneumoniae_TIGR4/1591/1-591
60Clostridium_phytofermentans_ISDg/1588/1-588

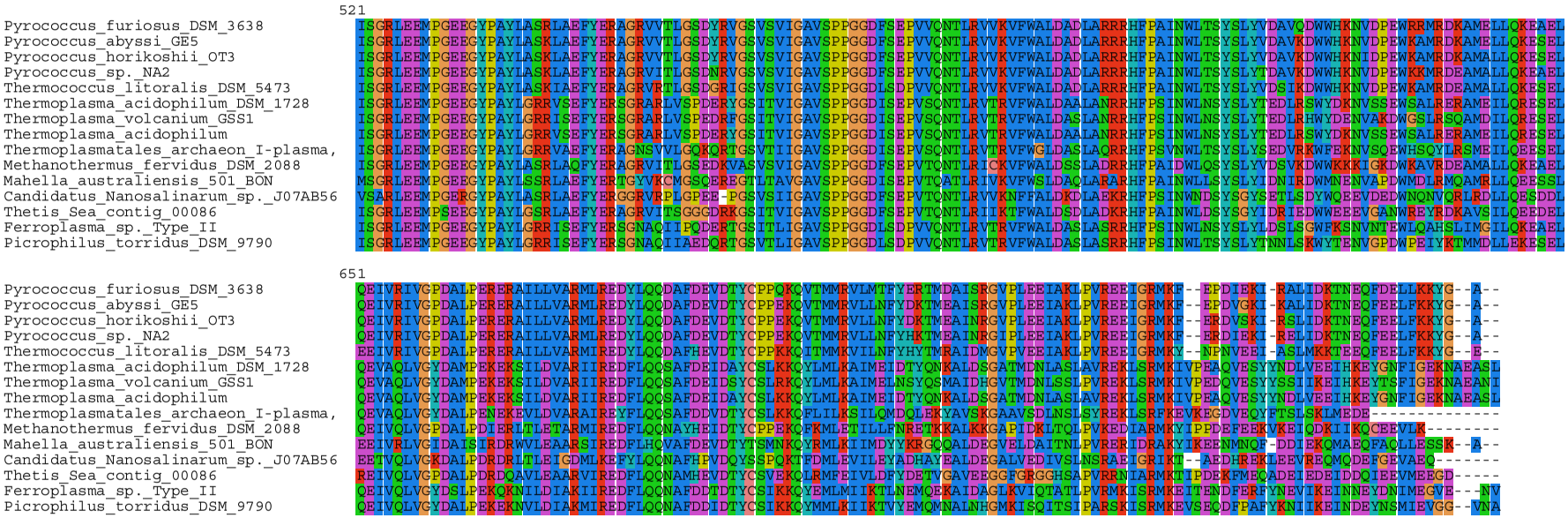
Position of *vma-1*b intein


451
25Chlamydia_trachomatis_B/TZ1A828/OT/1591/1-591
801Micrarchaeum/1586/1-586
222Aciduliprofundum_boonei_T469/1582/1-582
127Acholeplasma_laidlawii_PG8A/1592/1-592
85Vulcanisaeta_distributa_DSM_14429/1603/1-603
210Methanosarcina_mazei_Go1/1578/1-578
8Porphyromonas_gingivalis_ATCC_33277/1584/1-584
216Methanocorpusculum_labreanum_Z/1581/1-581
212Methanococcoides_burtonii_DSM_6242/1578/1-578
242Methanococcus_voltae_A3/1585/1-585
132Enterococcus_faecalis_V583/1595/1-595
63Clostridium_sticklandii_DSM_519/1594/1-594
215uncultured_methanogenic_archaeon_RCI/1579/1-579
121Picrophilus_torridus_DSM_9790/1922/1-589
209Methanosarcina_acetivorans_C2A/1578/1-578
128Acidaminococcus_fermentans_DSM_20731/1590/1-590
5Bacteroides_fragilis_NCTC_9343/1585/1-585
96Sulfolobus_solfataricus_P2/1592/1-592
75Anaeromyxobacter_dehalogenans_2CP1/1579/1-579
235Thermococcus_kodakarensis_KOD1/1585/1-585
92Nitrosopumilus_maritimus_SCM1/1592/1-592
190Pinegoldia_magna_ATCC_29328/1587/1-587
999Nanosalinarum_sp._J07AB56/1584/1-584
50Nautilia_profundicola_AmH/1571/1-571
82Pyrobaculum_aerophilum_str._IM2/1593/1-593
68Allochromatium_vinosum_DSM_180/1598/1-598
89Staphylothermus_marinus_F1/1592/1-592
184Clostridium_botulinum_A_str._ATCC_3502/1592/1-592
205Perroglobus_placidus_DSM_10642/1573/1-573
87Ignisphaera_aggregans_DSM_17230/1591/1-591
206Archaeoglobus_fulgidus_DSM_4304/1581/1-581
108Coprothermobacter_proteolyticus_DSM_5265/1589/1-589
917Thermosphaera_aggregans_DSM_11486/1584/1-584
246Methanococcus_maripaludis_C6/1586/1-586
135Streptococcus_pneumoniae_CGSP14/1599/1-599
125Dictyoglomus_turgidum_DSM_6724/1589/1-589
11Halomonas_elongata_DSM_2581/1627/1-627
159Clostridium_difficile_630/1592/1-592
800Nanosalina/1586/1-586
996Pyrococcus_sp._NA2/1588/1-588
219Methanoculleus_marisnigri_JR1/1584/1-584
65Finegoldia_magna_ATCC_29328/1588/1-588
38Borrelia_recurrentis_A1/1576/1-576
32Desulfohalobium_retbaense_DSM_5692/1589/1-589
163Mahella_australiensis_50I_BON/1587/1-587
129Bubacterium_limosum_KIST612/1601/1-601
69Sideroxydans_lithotrophicus_ES1/1599/1-599
214Methanocella_paludicola_SANAE/1579/1-579
70Nitrosococcus_halophilus_Nc4/1593/1-593
229Thermococcus_sibiricus_MM_739/1585/1-585
175Clostridium_perfringens_SM101/1591/1-591
179Clostridium_tetani_E88/1592/1-592
107Aeropyrum_pernix_K1/1597/1-597
105Ignicoccus_hospitalis_KIN4/I/1596/1-596
301Thetis_Sea/1-587/1-751
31Desulfarculus_baarsii_DSM_2075/1581/1-581
158Clostridium_phytofermentans_ISDg/1589/1-589
193Halobacterium_salinarum_R1/1585/1-585
46Spirochaeta_thermophila_DSM_6192/1591/1-591
196Halorhabdus_utahensis_DSM_12940/1587/1-587
12Legionella_longbeachae_NSW150/1603/1-603
236Methanocaldococcus_infernum_ME/1589/1-589
300Thermococcus_litoralis_DSM_5473]/1585/1-585
231Pyrococcus_horikoshii_OT3/1964/1-588
73Anaeromyxobacter_sp._PW1095/1578/1-578
194Halalkalicoccus_jeotgali_B3/1587/1-587
202Haloterrigena_turkmenica_DSM_5511/1587/1-587
997Methanothermus_fervidus_DSM_2088/1585/1-585
802Parvarchaeum/1583/1-583
208Methanosarcina_barkeri_str._Fusaro/1578/1-578
Ferroplasma_sp._Type_II
84Caldivirga_maquilingensis_IC167/1595/1-595
197Haloquadratum_walsbyi_DSM_16790/1585/1-585
232Pyrococcus_abyssi_GE5/1101/1-588
200Halomicrobium_mukohataei_DSM_12286/1586/1-586
143Streptococcus_sanguinis_SK36/1623/1-623
30Waddlia_chondrophila_WSU_861044/1596/1-596
45Spirochaeta_smaragdinae_DSM_11293/1588/1-588
131Thermanaerovibrio_acidaminovorans_DSM_6589/1569/1-569
42Treponema_pallidum_subsp._pallidum_str._Nichols/1589/1-589
77Thermofilum_pendens_Hrk_5/1601/1-601
110Thermus_thermophilus_HB27/1578/1-578
203Natrialba_magadii_ATCC_43099/1587/1-587
93Sulfolobus_acidocaldarius_DSM_639/1592/1-592
35Borrelia_burgdorferi_B31/1575/1-575
55Spirochaeta_smaragdinae_DSM_11293/1592/1-592
29Candidatus_Protochlamydia_amoebophila_UWE25/1593/1-593
198Haloferax_volcanii_DS2/1586/1-586
302Thermoplasmatales_archaeon/1-575
62Halothermothrix_orenii_H_168/1590/1-590
115Truepera_radiovictrix_DSM_17093/1579/1-579
16Chlamydophila_pneumoniae_CWL029/1591/1-591
213Methanohalophilus_mahii_DSM_5219/1576/1-576
78Candidatus_Korarchaeum_cryptofilum_OPF8/1595/1-595
19Chlamydia_muridarum_Nigg/1591/1-591
52Thermanaerovibrio_acidaminovorans_DSM_6589/1594/1-594
221Methanopyrus_kandleri_AV19/1592/1-592
189Leptotrichia_buccalis_DSM_1135/1597/1-597
191Pusobacterium_nucleatum_subsp._nucleatum_ATCC_25586/1382/1-584
4Bacteroides_vulgatus_ATCC_8482/1584/1-584
95Metallosphaera_sedula_DSM_5348/1591/1-591
44Treponema_denticola_ATCC_35405/1589/1-589
90Desulfurococcus_kamchatkensis_1221n/1584/1-584
57Thermotoga_sp._RQ2/1586/1-586
803Pyrococcus_yayanosii/1587/1-587
228Methanobrevibacter_smithii_ATCC_35061/1584/1-584
13Cyanothece_sp._PCC_8801/1607/1-607
168Thermoanaerobacter_pseudethanolicus_ATCC_33223/1590/1-590
26Chlamydophila_felis_Fe/C56/1588/1-588
80Thermoproteus_neutrophilus_V24Sta/1594/1-594
114Deinococcus_radiodurans_R1/1582/1-582
122Methanocella_paludicola_SANAE/1579/1-579
217Methanoplanus_petrolearius_DSM_11571/1582/1-582
998Thermoplasma_acidophilum_1221B2/1590/1-590
204Methanosaeta_thermophila_PT/1576/1-576
218Methanospirillum_hungateI_JF1/1582/1-582
221Candidatus_Methanosphaerula_palustris_E1-9c/1588/1-588
171Anaerococcus_prevotii_DSM_20548/1585/1-585
120Ferroplasma_acidarmanus_fer1/1589/1-589
119Thermoplasma_volcanium_GSS1/1776/1-590
3Brachyspira_murdochii_DSM_12563/1587/1-587
59Anaerococcus_prevotii_DSM_20548/1582/1-582
164Clostridium_thermocellum_ATCC_27405/1589/1-589
237Methanocaldococcus_vulcanius_M7/1587/1-587
124Dictyoglomus_thermophilum_H-6-12/1589/1-589
153Streptococcus_pyogenes_M1_GAS/1591/1-591
41Pirobacter_succinogenes_subsp._succinogenes_S85/1585/1-585
220Candidatus_Methanoregula_boonei_6A8/1588/1-588
79Pyrobaculum_islandicum_DSM_4184/1593/1-593
177Clostridium_botulinum_E3_str._Alaska_E43/1592/1-592
48Geobacter_uraniireducens_Rf4/1524/1-524
211Methanohalobium_evestigatum_Z7303/1578/1-578
230Pyrococcus_furiosus_DSM_3638/11013/1-588
47Stackebrandtia_nassauensis_DSM_44728/1575/1-575
199Haloarcula_marismortui_ATCC_43049/1586/1-586
227Methanobrevibacter_ruminantium_M1/1584/1-584
10Parabacteroides_distasonis_ATCC_8503/1585/1-585
195Halorubrum_lacusprofundi_ATCC_49239/1591/1-591
64Thermosediminibacter_oceani_DSM_16646/1587/1-587
144Streptococcus_gallolyticus_UCN34/1591/1-591
165Thermosediminibacter_oceani_DSM_16646/1590/1-590
172Clostridium_cellulovorans_743B/1591/1-591
201Natronomonas_pharaonis_DSM_2160/1585/1-585
178Clostridium_novyi_NT/1591/1-591
141Streptococcus_mitis_B6/1591/1-591
109Methanospirillum_hungateI_JF1/1588/1-588
49Aminobacterium_colombiense_DSM_12261/1588/1-588
118Thermoplasma_acidophilum_DSM_1728/1764/1-590
94Sulfolobus_tokodaii_str._7/1595/1-595
102Sulfolobus_islandicus_M.14.25/1592/1-592
53Treponema_pallidum_subsp._pallidum_str._Nichols/1605/1-605
106Hyperthermus_butylicus_DSM_5456/1601/1-601
104Acidilobus_saccharovorans_34Y15/1601/1-601
58Acholeplasma_laidlawii_PG8A/1583/1-583
130AminobacterIum_colombiense_DSM_12261/1598/1-598
243Methanococcus_vannielii_SB/1586/1-586
207Archaeoglobus_profundus_DSM_5631/1580/1-580
999Ferroplasma_acidarmanus/1589/1-589
116Meiothermus_ruber_DSM_1279/1578/1-578
123Methanospirillum_hungateI_JF1/1584/1-584
226Methanothermobacter_thermautotrophicus_str._Delta_H/1584/1-584
241Methanococcus_aeolicus_Nankai3/1586/1-586
234Thermococcus_gammatolerans_EJ3/1585/1-585
224Methanosphaera_stadtmanae_DSM_3091/1585/1-585
233Thermococcus_onnurineus_NA1/1585/1-585
66Eubacterium_limosum_KIST612/1595/1-595
188Streptobacillus_moniliformis_DSM_12112/1588/1-588
140Streptococcus_pneumoniae_TIGR4/1591/1-591
60Clostridium_phytofermentans_ISDg/1588/1-588


601
25Chlamydia_trachomatis_B/TZ1A828/OT/1591/1-591
801Micrarchaeum/1586/1-586
222Aciduliprofundum_boonei_T469/1582/1-582
127Acholeplasma_laidlawii_PG8A/1592/1-592
85Vulcaniaseta_distributa_DSM_14429/1603/1-603
210MethanosarcIna_mazei_GO1/1578/1-578
8Porphyromonas_gingivalis_ATCC_33277/1584/1-584
216Methanocorpusculum_labreanum_Z/1581/1-581
212Methanococcoides_burtonii_DSM_6242/1578/1-578
242Methanococcus_voltae_A3/1585/1-585
132Enterococcus_faecalis_V583/1595/1-595
63Clostridium_sticklandii_DSM_519/1594/1-594
215uncultured_methanogenic_archaeon_RCI/1579/1-579
121Picrophilus_torridus_DSM_9790/1922/1-589
209Methanosarcina_acetivorans_C2A/1578/1-578
128Acidaminococcus_fermentans_DSM_20731/1590/1-590
5Bacteroides_fragilis_NCTC_9343/1585/1-585
96Sulfolobus_solfataricus_P2/1592/1-592
75Anaeromyxobacter_dehalogenans_2CP1/1579/1-579
235Thermococcus_kodakarensis_KOD1/1585/1-585
92Nitrosopumilus_maritimus_SCM1/1592/1-592
190Pinegoldia_magna_ATCC_29328/1587/1-587
999Nanosalinarum_sp._J07AB56/1584/1-584
50Nautilia_profundicola_AmH/1571/1-571
82Pyrobaculum_aerophilum_str._IM2/1593/1-593
68Allochromatium_vinosum_DSM_180/1598/1-598
89Staphylothermus_marinus_F1/1592/1-592
184Clostridium_botulinum_A_str._ATCC_3502/1592/1-592
205Perroglobus_placidus_DSM_10642/1573/1-573
87Ignisphaera_aggregans_DSM_17230/1591/1-591
206Archaeoglobus_fulgidus_DSM_4304/1581/1-581
108Coprothermobacter_proteolyticus_DSM_5265/1589/1-589
917Thermosphaera_aggregans_DSM_11486/1584/1-584
246Methanococcus_maripaludis_C6/1586/1-586
135Streptococcus_pneumoniae_CGSP14/1599/1-599
125Dictyoglomus_turgidum_DSM_6724/1589/1-589
11Halomonas_elongata_DSM_2581/1627/1-627
159Clostridium_difficile_630/1592/1-592
800Nanosalina/1586/1-586
996Pyrococcus_sp._NA2/1588/1-588
219Methanoculleus_marisnigri_JR1/1584/1-584
65Finegoldia_magna_ATCC_29328/1588/1-588
38Borrelia_recurrentis_A1/1576/1-576
32Desulfohalobium_retbaense_DSM_5692/1589/1-589
163Mahella_australiensis_50I_BON/1587/1-587
129Eubacterium_limosum_KIST612/1601/1-601
69Sideroxydans_lithotrophicus_ES1/1599/1-599
214Methanocella_paludicola_SANAE/1579/1-579
70Nitrosococcus_halophilus_Nc4/1593/1-593
229Thermococcus_sibiricus_MM_739/1585/1-585
175Clostridium_perfringens_SM101/1591/1-591
179Clostridium_tetani_E88/1592/1-592
107Aeropyrum_pernix_K1/1597/1-597
105Ignicoccus_hospitalis_KIN4/I/1596/1-596
301Thetis_Sea/1-587/1-751
31Desulfarculus_baarsii_DSM_2075/1581/1-581
158Clostridium_phytofermentans_ISDg/1589/1-589
193Halobacterium_salinarum_R1/1585/1-585
46Spirochaeta_thermophila_DSM_6192/1591/1-591
196Halorhabdus_utahensis_DSM_12940/1587/1-587
12Legionella_longbeachae_NSW150/1603/1-603
236Methanocaldococcus_infernus_ME/1589/1-589
300Thermococcus_litoralis_DSM_5473]/1585/1-585
231Pyrococcus_horikoshii_OT3/1964/1-588
73Anaeromyxobacter_sp._PW1095/1578/1-578
194Halalkalicoccus_jeotgali_B3/1587/1-587
202Haloterrigena_turkmenica_DSM_5511/1587/1-587
997Methanothermus_fervidus_DSM_2088/1585/1-585
802Parvarchaeum/1583/1-583
208Methanosarcina_barkeri_str._Fusaro/1578/1-578
Ferroplasma_sp._Type_II
84Caldivirga_maquilingensis_IC167/1595/1-595
197Haloquadratum_walsbyi_DSM_16790/1585/1-585
232Pyrococcus_abyssi_GE5/1101/1-588
200Halomicrobium_mukohataei_DSM_12286/1586/1-586
143Streptococcus_sanguinis_SK36/1623/1-623
30Waddlia_chondrophila_WSU_861044/1596/1-596
45Spirochaeta_smaragdinae_DSM_11293/1588/1-588
131Thermanaerovibrio_acidaminovorans_DSM_6589/1569/1-569
42Treponema_pallidum_subsp._pallidum_str._Nichols/1589/1-589
77Thermofilum_pendens_Hrk_5/1601/1-601
110Thermus_thermophilus_HB27/1578/1-578
203NatrialbA_magadii_ATCC_43099/1587/1-587
93Sulfolobus_acidocaldarius_DSM_639/1592/1-592
35Borrelia_burgdorferi_B31/1575/1-575
55Spirochaeta_smaragdinae_DSM_11293/1592/1-592
29Candidatus_Protochlamydia_amoebophila_UWE25/1593/1-593
198Haloferax_volcanii_DS2/1586/1-586
302Thermoplasmatales_archaeon/1-575
62Halothermothrix_orenii_H_168/1590/1-590
115Truepera_radiovictrix_DSM_17093/1579/1-579
16Chlamydophila_pneumoniae_CWL029/1591/1-591
213Methanohalophilus_mahii_DSM_5219/1576/1-576
78Candidatus_Korarchaeum_cryptofilum_OPF8/1595/1-595
19Chlamydia_muridarum_Nigg/1591/1-591
52Thermanaerovibrio_acidaminovorans_DSM_6589/1594/1-594
223Methanopyrus_kandleri_AV19/1592/1-592
189Leptotrichia_buccalis_DSM_1135/1597/1-597
191Pusobacterium_nucleatum_subsp._nucleatum_ATCC_25586/1382/1-584
4Bacteroides_vulgatus_ATCC_8482/1584/1-584
95Metallosphaera_sedula_DSM_5348/1591/1-591
44Treponema_denticola_ATCC_35405/1589/1-589
90Desulfurococcus_kamchatkensis_1221n/1584/1-584
57Thermotoga_sp._RQ2/1586/1-586
803Pyrococcus_yayanosii/1587/1-587
228Methanobrevibacter_smithii_ATCC_35061/1584/1-584
13Cyanothece_sp._PCC_8801/1607/1-607
168Thermoanaerobacter_pseudethanolicus_ATCC_33223/1590/1-590
26Chlamydophila_felis_Fe/C56/1588/1-588
80Thermoproteus_neutrophilus_V24Sta/1594/1-594
114Deinococcus_radiodurans_R1/1582/1-582
122Methanocella_paludicola_SANAE/1579/1-579
217Methanoplanus_petrolearius_DSM_11571/1582/1-582
998Thermoplasma_acidophilum_1221B2/1590/1-590
204Methanosaeta_thermophila_PT/1576/1-576
218Methanospirillum_hungateI_JF1/1582/1-582
221Candidatus_Methanosphaerula_palustris_E1-9c/1588/1-588
171Anaerococcus_prevotii_DSM_20548/1585/1-585
120Ferroplasma_acidarmanus_fer1/1589/1-589
119Thermoplasma_volcanium_GSS1/1776/1-590
3Brachyspira_murdochii_DSM_12563/1587/1-587
59Anaerococcus_prevotii_DSM_20548/1582/1-582
164Clostridium_thermocellum_ATCC_27405/1589/1-589
237Methanocaldococcus_vulcanius_M7/1587/1-587
124Dictyoglomus_thermophilum_H6I2/1589/1-589
153Streptococcus_pyogenes_M1_GAS/1591/1-591
41Pirobacter_succinogenes_subsp._succinogenes_S85/1585/1-585
220Candidatus_Methanoregula_boonei_6A8/1588/1-588
79Pyrococcus_islandicum_DSM_4184/1593/1-593
177Clostridium_botulinum_E3_str._Alaska_E43/1592/1-592
48Geobacter_uraniireducens_Rf4/1524/1-524
211Methanohalobium_evestigatum_Z7303/1578/1-578
230Pyrococcus_furiosus_DSM_3638/11013/1-588
47Stackebrandtia_nassauensis_DSM_44728/1575/1-575
199Haloarcula_marismortui_ATCC_43049/1586/1-586
227Methanobrevibacter_ruminantium_M1/1584/1-584
10Parabacteroides_distasonis_ATCC_8503/1585/1-585
195Halorubrum_lacusprofundi_ATCC_49239/1591/1-591
147Thermosediminibacter_oceani_DSM_16646/1587/1-587
144Streptococcus_gallolyticus_UCN34/1591/1-591
165Thermosediminibacter_oceani_DSM_16646/1590/1-590
172Clostridium_cellulovorans_743B/1591/1-591
201Natronomonas_pharaonis_DSM_2160/1585/1-585
178Clostridium_novyi_NT/1591/1-591
141Streptococcus_mitis_B6/1591/1-591
109Methanospirillum_hungatei_JF1/1588/1-588
49Aminobacterium_colombiense_DSM_12261/1588/1-588
118Thermoplasma_acidophilum_DSM_1728/1764/1-590
94Sulfolobus_tokodaii_str._7/1595/1-595
102Sulfolobus_islandicus_M.16.27/1592/1-592
53Treponema_pallidum_subsp._pallidum_str._Nichols/1605/1-605
106Hyperthermus_butylicus_DSM_5456/1601/1-601
104Acidilobus_saccharovorans_34Y15/1601/1-601
58Acholeplasma_laidlawii_PG8A/1583/1-583
130AminobacterIum_colombiense_DSM_12261/1598/1-598
243Methanococcus_vannielii_SB/1586/1-586
207Archaeoglobus_profundus_DSM_5631/1580/1-580
999Ferroplasma_acidarmanus/1589/1-589
116Meiothermus_ruber_DSM_1279/1578/1-578
123Methanospirillum_hungatei_JF1/1584/1-584
226Methanothermobacter_thermautotrophicus_str._Delta_H/1584/1-584
241Methanococcus_aeolicus_Nankai3/1586/1-586
234Thermococcus_gammatolerans_EJ3/1585/1-585
224Methanosphaera_stadtmanae_DSM_3091/1585/1-585
233Thermococcus_onnurineus_NA1/1585/1-585
66Eubacterium_limosum_KIST612/1595/1-595
188Streptobacillus_moniliformis_DSM_12112/1588/1-588
140Streptococcus_pneumoniae_TIGR4/1591/1-591
60Clostridium_phytofermentans_ISDg/1588/1-588

Alignment: /Users/jpgogarten/Dropbox/supplementary material for upload/sate_alignment_of_Extein_and_Intein_sequences(withoutHE).fasta
Seaview [blocks=130 fontsize=7 LETTER-landscape] on Fri Jul 19 09:49:28 2013




521
Pyrococcus_furiosus_DSM_3638
Pyrococcus_abyssi_GE5
Pyrococcus_horikoshii_OT3
Pyrococcus_sp._NA2
Thermococcus_litoralis_DSM_5473
Thermoplasma_acidophilum_DSM_1728
Thermoplasma_volcanium_GSS1
Thermoplasma_acidophilum
Thermoplasmatales_archaeon_I-plasma,
Methanothermus_fervidus_DSM_2088
Mahella_australiensis_501_BON
Candidatus_Nanosalinarum_sp._J07AB56
Thetis_Sea_contig_00086
Ferroplasma_sp._Type_II
Picrophilus_torridus_DSM_9790
651
Pyrococcus_furiosus_DSM_3638
Pyrococcus_abyssi_GE5
Pyrococcus_horikoshii_OT3
Pyrococcus_sp._NA2
Thermococcus_litoralis_DSM_5473
Thermoplasma_acidophilum_DSM_1728
Thermoplasma_volcanium_GSS1
Thermoplasma_acidophilum
Thermoplasmatales_archaeon_I-plasma,
Methanothermus_fervidus_DSM_2088
Mahella_australiensis_501_BON
Candidatus_Nanosalinarum_sp._J07AB56
Thetis_Sea_contig_00086
Ferroplasma_sp._Type_II
Picrophilus_torridus_DSM_9790

Alignment of Intein Sequences