Kormas KA (in press) In: Sezenna ML (ed) Proteobacteria: Phylogeny, metabolic diversity and ecological effects. Nova Publishers, Hauppauge, NY. ISBN: 978-1-61761-810-9

Interpreting diversity of Proteobacteria based on 16S rRNA gene copy number

Konstantinos Ar. Kormas

Department of Ichthyology & Aquatic Environment, School of Agricultural Sciences, University of Thessaly, 384 46 Nea Ionia, Magnisia, Greece

Tel.: +30-242-109-3082, +30-242-109-3157, E-mail: kkormas@uth.gr

ABSTRACT

The application of the 16S rRNA gene diversity analysis has revealed the immense microbial diversity of our planet. At the same time, and after of more than two decades of using this methodology along with several important improvements and new techniques, there is still no universal golden rule on how to estimate prokaryotic diversity in a natural sample, as there is in macroecology. A general assumption during studies of prokaryotic diversity is that each found 16S rRNA gene found corresponds to one cell. However, in this paper it is shown that recent genomic data reveal that this is not the case for several bacterial phyla. Since the Proteobacteria, along with the Firmicutes, are the most abundant and diverse bacterial phyla, in this paper the average 16S rRNA gene copy number is presented at the sub-phylum (α-, β-, γ-, δ- and ε-Proteobacteria), order and family level of the Proteobacteria phylum. At the sub-phylum level the average 16S rRNA gene copy number varied between 2.1±1.3 and 5.8±2.8. Since the 16S rRNA gene copy number affects the relative abundance of each proteobacterial species/phylotype found in a clone library, and subsequently the estimation of diversity, the corrected relative abundances of the found proteobacterial phylotypes were estimated in 37 clone libraries from six different natural habitats. It is suggested, that at least in the cases where Proteobacteria consist 50-75% of the clone library, the corrected abundances should be used for diversity estimations.

The cultivation of microorganisms from the environment was developed for studying purposes and is a fully accepted, well established and continuously improving methodology. However, cultivation is a human innovation and it does not imply that all microorganisms necessarily will have to conform with our cultivation efforts, since in natural conditions the microorganisms rarely -if ever- encounter cultivation-like conditions. Thus, cultivation-independent approaches (Nocker et al. 2007, Orphan 2009), seem to be more appropriate for the study of the diversity of microorganisms and their communities' structure in the environment.

The analysis of the 16S rRNA gene diversity, amplified directly from environmental samples, is considered a proxy for the occurrence of prokaryotic species. It is the oldest methodology for the investigation of microbial diversity from practically every natural Earth habitat and has been the most widely used approach over the last two decades (Woese et al. 1990). Microbial, namely prokaryotes and microscopic eukaryotes, taxonomy, biology and ecophysiology has benefited enormously by such studies and the race to unravel as much as possible of the hidden microbial diversity remains a fascinating and rewarding scientific field (Cases & de Lorenzo 2002).

In order to assess diversity, it is essential to securely identify the members of the community or population and also their abundances or relative abundances (Bunge 2009). Although in the macroscopic world this is assured by species identification -in most cases based on morphological features- and measuring their abundance, the concept of species in the microbial world is still under debate (Ward 1998, Roselló-Mora & Amann 2001, Konstantinidis et al. 2006). This obstacle has been dealt with the use of phylotypes or operational taxonomic units (OTU) based on 16S rRNA gene sequencing and comparison of similarity (Stackebrandt & Goebel 1994), in most cases with a similarity cut-off limit between 97 and 99% being used.

This approach, assumes either that each of the found 16S rRNA phylotypes corresponds to one prokaryotic cell or that it is unknown to how many cells it corresponds and, thus, only semi-quantitative estimations can be inferred. To a degree this approach is justified by the fact that in most studies there are several unknown phylotypes that are distantly similar to any of the known (cultivated) species. However, for the cases of high similarity with known species, the ample genomic data that is available might be of help. I use the case of the Proteobacteria as is the most diverse and abundant phylum of the domain Bacteria in nature, along with the Firmicutes, and are found almost everywhere on Earth (Kersters et al. 2006).

This paper aims at depicting that the 16S rRNA gene copy numbers for all the available Proteobacteria genomes, varies significantly at different taxonomic levels and shows which taxon provides more safe data for the relative abundance of the 16S rRNA genes in a clone library. The 16S rRNA gene copy number should be taken into account when estimating Proteobacteria diversity from inferred phylogenies as a proxy to corrected relative abundances of the cells that bear them. The only assumption is that for the species with multiple 16S rRNA gene copies, all these copies are identical, although few cases exist which this is not true but recent data suggest that the Proteobacteria are rather on the low range of intra-gene diversity (0.00 – 2.09%, average 0.20%, Pei et al. 2010). These values are very close to the generally accepted threshold of 1.0 – 1.3% difference between two 16S rRNA genes in order to be considered representing two different bacterial species (Stackenbrandt & Ebers 2006). Considering this limitation, this paper aims only to depict that the general rule "one 16S rRNA for every prokaryotic species" which is ubiquitously used and is never stated even as a possible factor of misinterpretation of data in microbial diversity studies, is far from real for the Proteobacteria. It also attempts to reveal trends in the mean number of

Proteobacteria taxa so these can be used for the estimation of more realistic values of Proteobacteria diversity.

In total, 485 genomes of the phylum Proteobacteria were analysed in this study (Table 1). The 16S rRNA gene copy number for each strain was retrieved from the Microbial Genome Resources (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). Only completed genomes as of 01 February 2010 were considered. The ζ-Proteobacteria (Emerson et al. 2007) are not discussed here since there is only one cultivated representative from this sub-phylum (Table 1).

A first glance on the distribution of 16S rRNA gene copy number among the available Proteobacteria genomes, shows that four out the five sub-phyla have multiple gene copy numbers. Only in the α-Proteobacteria, 41.2% of the available genomes have one copy of this gene while for the β-, γ- δ- and ε-Proteobacteria this percentage falls gradually, 9.7%, 8.2%, 5.9% and 3.8%, respectively (Figure 1).

The available genomes of the Proteobacteria, which were used in this study, do not cover equally all the taxa of the phylum and, it is most likely that their representation in the database is irrelevant and accidental with their distribution in nature (Lagesen et al. 2010). For example, there are cases where one family includes genomes only of very few of its genera. In addition, the advent of molecular biology technologies and their applications in microbial ecology the last two decades (Amann et al. 1995) have created a rather large number of "yet-unaffiliated" taxa, i.e. taxa which either do not include a cultured representative or high-level taxa which include only one or very few cultured representatives. Such limitations, render any effort to apply universal rules regarding the 16S rRNA gene copy number per taxon rather impossible at the time being. Instead it depicts the need for incorporating the 16S rRNA gene copy number in studies of microbial diversity, by applying this approach to the ubiquitous phylum of Proteobacteria, since several of its members are far from having only one 16S rRNA gene copy.

The 119 genomes of the α-Proteobacteria represent six orders, 19 families, 49 genera and 33 species of the sub-phylum (Figure 2). The mean copy number of the sub-phylum is 2.1±1.3. Only the Rickettsiales, which includes members with the smallest genomes known to date, and its three families seem to have only one copy of the 16S rRNA gene. The rest of the five orders vary from 1.5±0.7 (Caulobacterales) to 3.6±1.0 (Rhodospirillales). Among these, the Rhizobiales contains the highest number of families, which reflects its wide distribution. In the nine Rhizobiales families, the mean 16S rRNA gene copy number varies significantly from 1.0±0.0 (Aurantimonadaceae) to 5.3±0.9 (Methylobacteraceae). The data from the families Bradyrhizobiaceae (1.6±0.5), Bartonellaceae (2.0±0.0), Brucellaceae (2.5±0.7) and Rhizobiaceae (2.8±0.4) correspond to ≥50% of the known genera of each family and, thus, it is possible to consider their mean 16S rRNA gene copy numbers as representative for each of these families.

The number of available genomes within the β-Proteobacteria sub-phylum is smaller (72), representing seven orders, 12 families, 32 genera and 52 species (Table 1). The mean 16S RNA gene copy number of the sub-phylum is 3.7±1.7, which implies that for this phylum is rather far from realistic to consider it having one copy of the 16S rRNA gene. Six out of the seven sub-phylum's orders include genomes from only one family (Figure 3) and from these families, only for the Nitrosomonadaceae and the Hydrogenophilaceae the available genomes cover ≥50% of the known genera of each family. The Burkholderiales is the dominating order among the β-Proteobacteria available genomes, covering six families, two of which are novel and yet-unclassified. These families show a highly variable mean 16S rRNA gene copy number (1.0±0.0 – 4.5±1.5) but they represent a low number of their known genera, with the exception of the Burkholderiaceae (50%). Thus, at the family level it is not secure yet to

assume specific average 16S rRNA copy numbers, as more genomes from more genera of the Burkholderiaceae families are required.

The γ-Proteobacteria sub-phylum includes the highest number of available genomes (220; Table 1), obviously due to the wide distribution of its members and their importance for the human health and plants/animals of human interest. The available genomes originate from 17 orders, 26 families, 68 genera and 115 species. As a sub-phylum, it is very unlikely to be considered as a single 16S rRNA gene copy taxon due to its high average value is 5.8±2.8 (Figure 4) which is the highest among all the Proteobacteria sub-phyla. Thirteen orders include single family representatives with available genomes. Only the Moraxellaceae family is well represented (66.7% of the family's known genera), whose average number of 16S rRNA gene copies is 5.4±1.0.

Each of the four orders of the δ-Proteobacteria sub-phylum with available genomes (33 in total, Table 1) includes at least two families (Figure 5). The available genomes of this sub-phylum point towards a multiple 16S rRNA gene copies group (3.0±1.5). Although most of the families are not satisfactory represented in terms of available genomes of their known genera, the Desulfovibrionaceae (3.5±1.3), the Myxococcaceae (2.4±0.9) and the Syntrophaceae (1.0+2.0) are fairly well represented (≥50%).

The ε-Proteobacteria contains the smallest number of available genomes (26, Table 1) but the sub-phylum's average 16S rRNA gene copy number is 2.8±0.8. The available genomes refer to the order Campylobacterales and three other orders which are not included in the Bergey's Manual (Staley et al. 2005) and two of them are still unclassified (Figure 6). This is due to the increased scientific interest of the last 10 years which resulted in major breakthroughs for this sub-phylum (Campbell 2006). The two families of the Campylobacterales, namely Campylobacteraceae (3.2±0.6) and Helicobacteraceae (2.2±0.8), have available genomes for more of the 2/3 their known genera, suggesting that the use of a more than one copy of the 16S rRNA gene should be used when phylotypes from any sample are affiliated within these two families. It is indicative for this sub-phylum that only one genome (*Helicobacter hepaticus* ATCC 51449) was found to have single copy of its 16S rRNA gene.

From the above it is evident that the assumption that each found 16S rRNA phylotype in a clone library corresponds to one genome and, subsequently, to one cell, is not valid. Even if one considers the case of a very specialised community consisting practically only from members of the α-Proteobacteria, which have the lowest average 16S rRNA gene copy number, one could ask whether it is safe or not to assume that each found 16S rRNA phylotype corresponds to one cell, in order to calculate some of the most commonly used diversity indices. The answer is probably negative, based on the available data, since just the lower end of the SD for the sub-phylum (0.8) is practically 1.

To further test this issue, the effect of the 16S rRNA gene copy number was investigated in already constructed clone libraries originating from diverse aquatic habitats (Table 2):

- Twelve clone libraries from the sediments of two active marine mud volcanoes, the Kazan MV (Pachiadaki et al. 2010) and the Amsterdam MV (Pachiadaki, de Lange, Kormas in prep).
- Two clone libraries from an urban drinking water distribution network (Kormas et al. 2010).
- Eight clone libraries from a drinking water reservoir (Lymperopoulou, Kormas, Karagouni, in prep.).
- One clone library from the mudshrimp *Pestarella tyrrhena* gut (Demiri et al. 2009) and eight clone libraries from the Norway lobster *Nephrops norvegicus* gut (Meziti et al. in revision).

- Five clone libraries from terrestrial thermal springs (Kormas et al. 2009).
- Two clone libraries from lake water and sediment bacterial communities (Kormas et al. in press).

Overall, the relative abundance of all Proteobacteria in the clone libraries from the above studies varied between 9.9% and 100.0%. For each clone library, the Shannon H index was calculated according to the Shannon–Wiener index H and was calculated as follows:
$H = \Sigma [p_i \ln(p_i)]$
where the summation is over all phylotypes i and p_i is the proportion of phylotypes relative to the sum of all phylotypes (Shannon & Wiever 1949). H was estimated by assuming the "one 16S rRNA gene for one cell" hypothesis (H_{sin}) and then by correcting the relative abundance of each Proteobacteria phylotype according to the mean 16S rRNA gene copy number that is valid for the sub-phylum that it belongs (H_{corr}), as these were mentioned above. The impact of this difference depends upon the overall abundance of the Proteobacteria in each clone library (Figure 7). When a clone library contains <50% Proteobacteria phylotypes, the difference of $H_{corr} - H_{sin}$ (ΔH) varies only slightly, from -0.2 to 0.4, and showed a negative correlation with increasing Proteobacteria presence. The situation changes dramatically when Proteobacteria dominate a clone library between 50% and 75% (ΔH between -1.5 and 0.9). In this case, there is a much stronger positive correlation, which renders the correction of the Proteobacteria relative abundance of the found phylotypes more important. Finally, when Proteobacteria dominate a clone library >75%, no clear trend was observed, since in such libraries what is really important is whether they are dominated from one or more sub-phyla. In the former case, ΔH remains zero, while in the latter depends on how many sub-phyla co-exist and which one of the is dominating.

In conclusion, this paper showed that the widely used assumption that each found 16S rRNA gene corresponds to one cell is not always valid based on the available genomic data for the case of the Proteobacteria. In particular, each sub-phylum has different ranges and average 16S rRNA gene copy numbers and this number varies, also, at the order and family level. By applying the average 16S rRNA gene copy number for each phylum, it was shown that in cases of moderate to high (50-75%) of Proteobacteria abundance, the use of this correction results in different diversity estimations, giving a different picture of the bacterial community. More available genomes from the Proteobacteria, but also from other bacterial phyla as well, are needed in order to will assist us in more accurate and pragmatic estimations of bacterial diversity in natural samples.

REFERENCES

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews 59:143-169

Brenner DJ, Krieg NR, Staley JR, Garriy G (eds) (2005) Bergey's manual of systematic bacteriology: Volume two: The Proteobacteria (Part B): The Gammaproteobacteria. Springer, Berlin

Bunge J (2009) Statistical estimation of uncultivated microbial diversity. In: Epstein S (ed) Uncultivated microorganisms. Springer, Berlin, p 1-18

Campbell B, Summers Engels A, Porter ML, Takai K (2006) The versatile ε-proteobacteria: key players in sulphidic habitats. Nature Reviews Microbiology 4:458-468

Cases I, de Lorenzo V (2002) The grammar of (micro)biological diversity. Environmental Microbiology 4:623-627

Demiri A, Meziti A, Papaspyrou S, Thesslou-Legaki M, Kormas KA (2009) Abdominal setae and midgut Bacteria of the mudshrimp *Pestarella tyrrhena*. Central European Journal of Biology 4:558-566

Emerson D, Rentz JA, Lilburn TG, Davis RE, Aldrich H, Chan C, Moyer CL (2007) A novel lineage of Proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. PloS ONE 2:e667

Garrity G (ed) (2005) Bergey's manual of systematic bacteriology: Volume two: The Proteobacteria (Part A): Introductory essays. Springer, Berlin

Kersters K, de Vos P, Gillis M, Swings J, Vandamme P, Stackenbrandt E (2006) Introduction to the Proteobacteria. Prokaryotes 5:3-37

Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philosophical Transactions Of The Royal Society B-Biological Sciences 361:1929-1940.

Kormas KA, Neofitou C, Pachiadaki M, Koufostathi E (2010) Changes of the bacterial assemblages throughout an urban drinking water distribution system. Environmental Monitoring and Assessment 165:27-38

Kormas KA, Tamaki H, Hanada S, Kamagata Y (2009) Apparent richness and community composition of Bacteria and Archaea in geothermal springs. Aquatic Microbial Ecology 57:113-122

Kormas KA, Vardaka E, Moustaka-Gouni M, Kontoyanni V, Petridou E, Gkelis S, Neofitou C (in press) Molecular detection of potentially toxic Cyanobacteria and their associated Bacteria in lake water column and sediment. World Journal of Microbiology and Biotechnology DOI: 10.1007/s11274-010-0322-x

Lagesen K, Ussery DW, Wassenaar TM (2010) Genome update: the 1000th genome – a cautionary tale. Microbiology. 156:603-608

Nocker A, Burr M, Camper AK (2007) Genotypic microbial community profiling: a critical technical review. Microbial Ecology 54:276-289

Orphan VJ (2009) Methods for unveiling cryptic microbial partnerships in nature. Current Opinion in Microbiology 12:231-237

Pachiadaki MG, Lykousis V, Stephanou EG, Kormas KA (2010) Prokaryotic community structure and diversity in the sediments of an active submarine mud volcano (Kazan MV, East MEditerranean Sea). FEMS Microbiology Ecology 72:429-444

Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, DeSantis T, Brodie E, Nelson K, Pei Z (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. Applied and Environmental Microbiology 76:3886-3897

Roselló-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiology Reviews 25:39-67

Shannon CE & Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana, IL.

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA:DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteria. International Journal of Systematic Bacteriology 44:846-849

Stackebrandt E, Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiology Today 2006:153–155.

Staley JT, Brenner DJ, Garrity G, Boone DR, Krieg NR, De Vos P, Goodfellow M, Rainey FA, Garrity GM, Schleifer K-H (eds) (2005) Bergey's manual of systematic bacteriology: Volume two: The Proteobacteria (Part C): The Alpha-, Beta-, Delta- and Epsilonproteobacteria. Springer, Berlin

Ward D (1998) A natural species concept for prokaryotes. Current Opinion in Microbiology 1:271-277

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proceedings of the National Academy of Sciences USA 87:4576–4579

Table 1. Number of genomes per taxon of the phylum Proteobacteria used in this study. Number of yet-unclassified taxa in parentheses.

Sub-phylum	*Orders*	*Families*	*Genera*	*Species*	**Total genomes**
α-	6	19	49	33	**119**
β-	7 (1)	12 (3)	32	52	**72**
γ-	17 (5)	26 (5)	68	115	**220**
δ-	4	9	18	29	**33**
ε-	4 (2)	5 (2)	9	16	**26**
ζ-	1	1	1	1	**1**
Total	**39**	**72**	**177**	**6**	***485***

Table 2. Percentage of Proteobacteria relative abundance in different clone libraries.

Clone library origin	**% Proteobacteria**	**Reference**
Kazan mud volcano sediments	56.3 – 78.0	Pachiadaki et al. (2010)
Amsterdam mud volcano sediments	15.5 – 66.7	Pachiadaki et al. (in prep.)
Drinking water distribution network	72.4 – 73.6	Kormas et al. (2010)
Drinking water reservoir	19.3 – 44.3	Lymperopou et al. (in prep.)
Mudshrimp *Pestarella tyrrhena* gut	70.2	Demiri et al. (2009)
Norway lobster *Nephrops norvegicus* gut	9.9 – 100.0	Meziti et al. (in revision)
Terrestrial thermal springs	21.1 – 98.2	Kormas et al. (2009)
Lake water and sediment	30.5 – 37.8	Kormas et al. (in press)

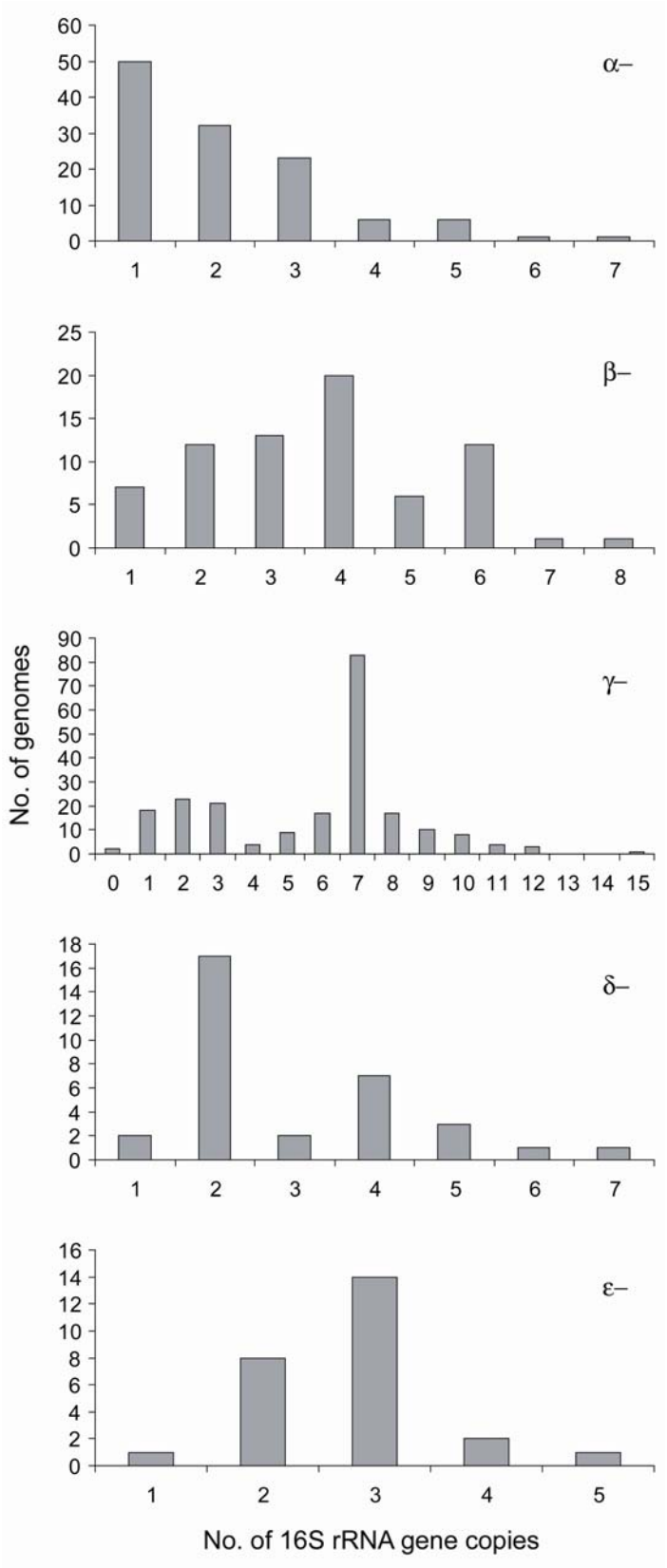

Figure 1. Distribution of available genomes (01 Feb. 2010) of Proteobacteria according to the copy number of the 16S rRNA gene.

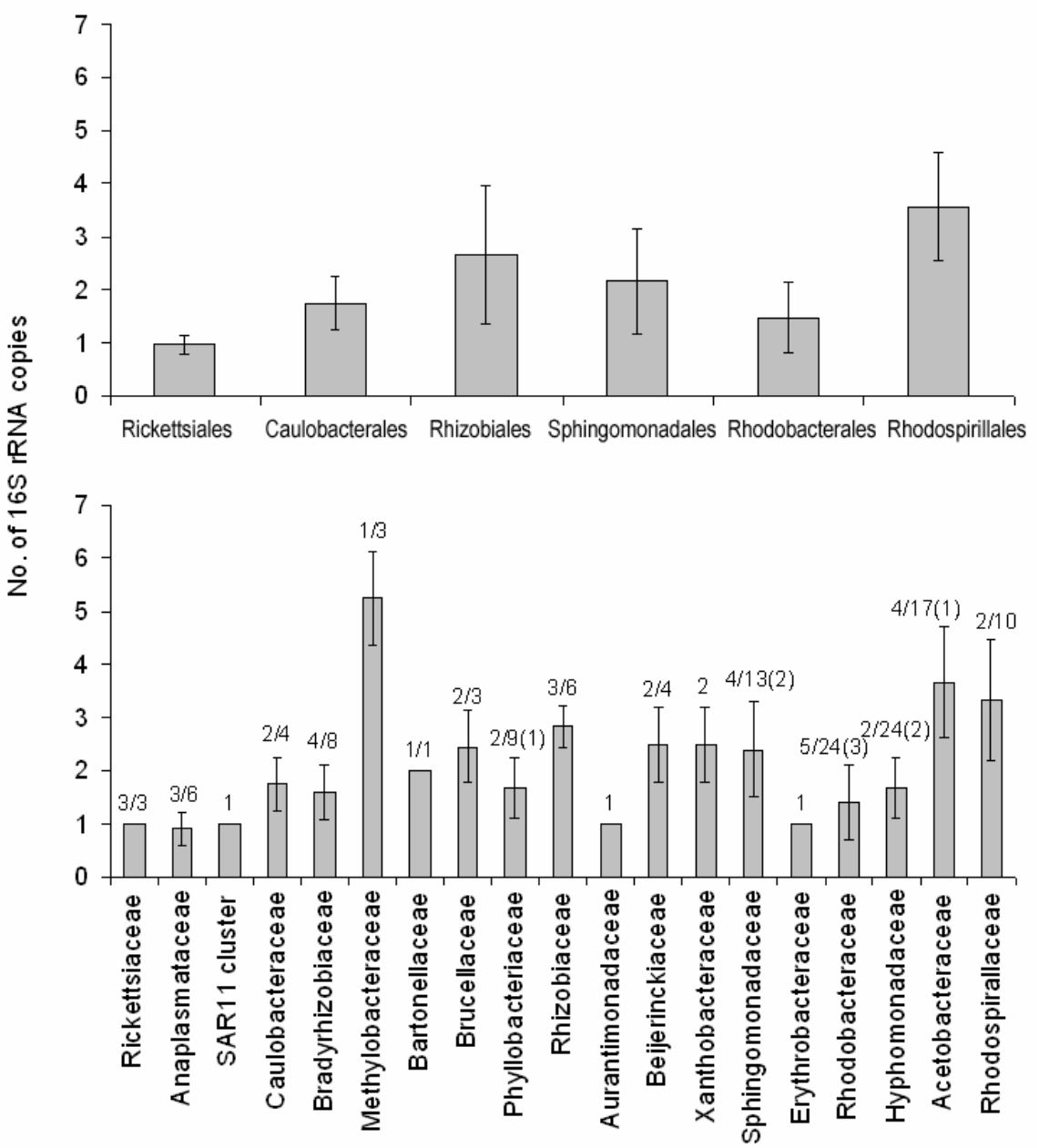

Figure 2. Variance of the 16S rRNA gene copy number among the α-Proteobacteria orders (top) and families (bottom). The ratio of the number of genera with available genomes in GenBank to the number of genera in Bergey's Manual (Staley et al. 2005) is shown above each column; the number of genera not included in Bergey's Manual is in parentheses. Standard deviation is indicated in error bars. No / indicates absence from the Bergey's Manual.

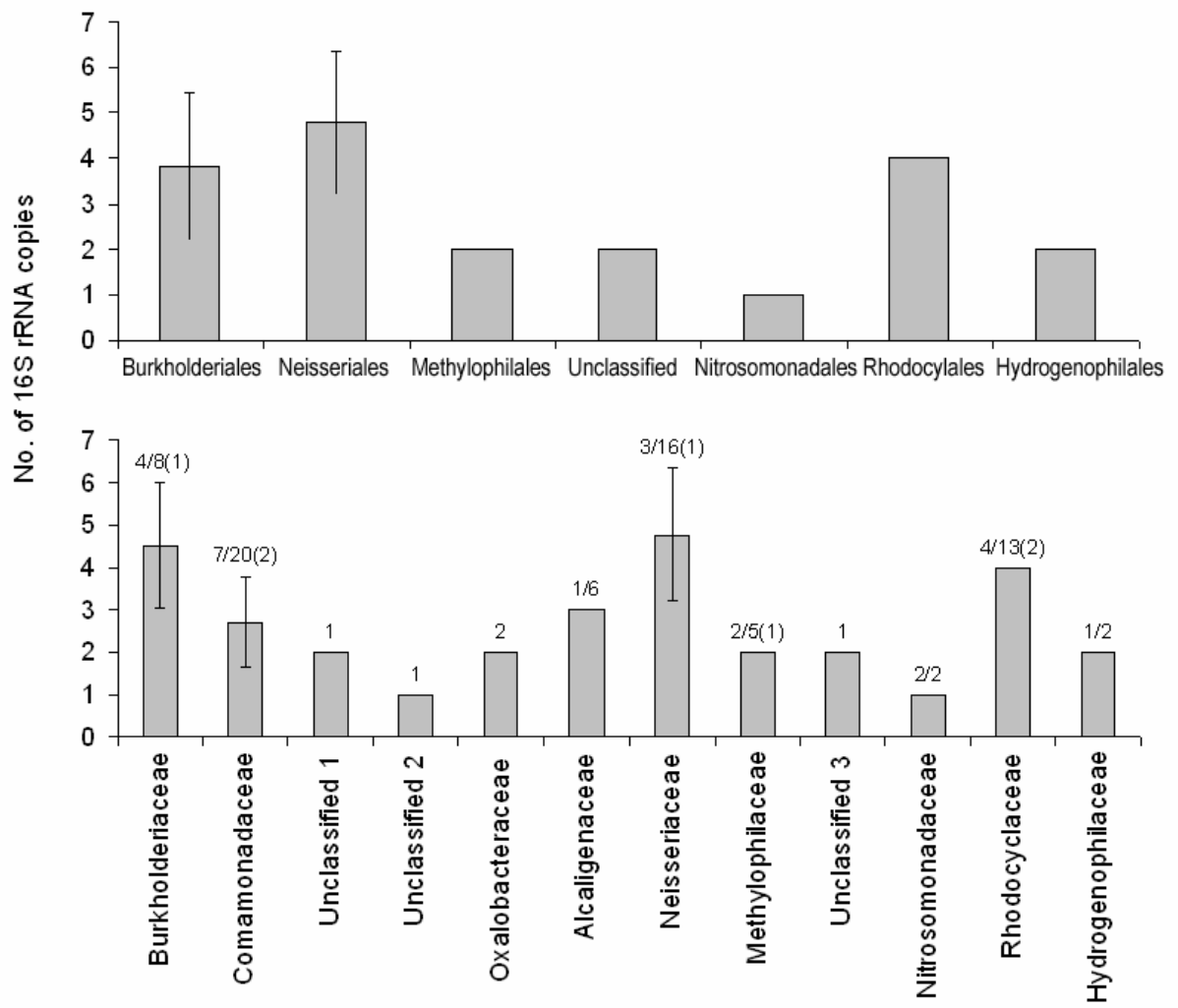

Figure 3. Variance of the 16S rRNA gene copy number among the β-Proteobacteria orders (top) and families (bottom). The ratio of the number of genera with available genomes in GenBank to the number of genera in Bergey's Manual (Staley et al. 2005) is shown above each column; the number of genera not included in Bergey's Manual is in parentheses. Standard deviation is indicated in error bars. No / indicates absence from the Bergey's Manual.

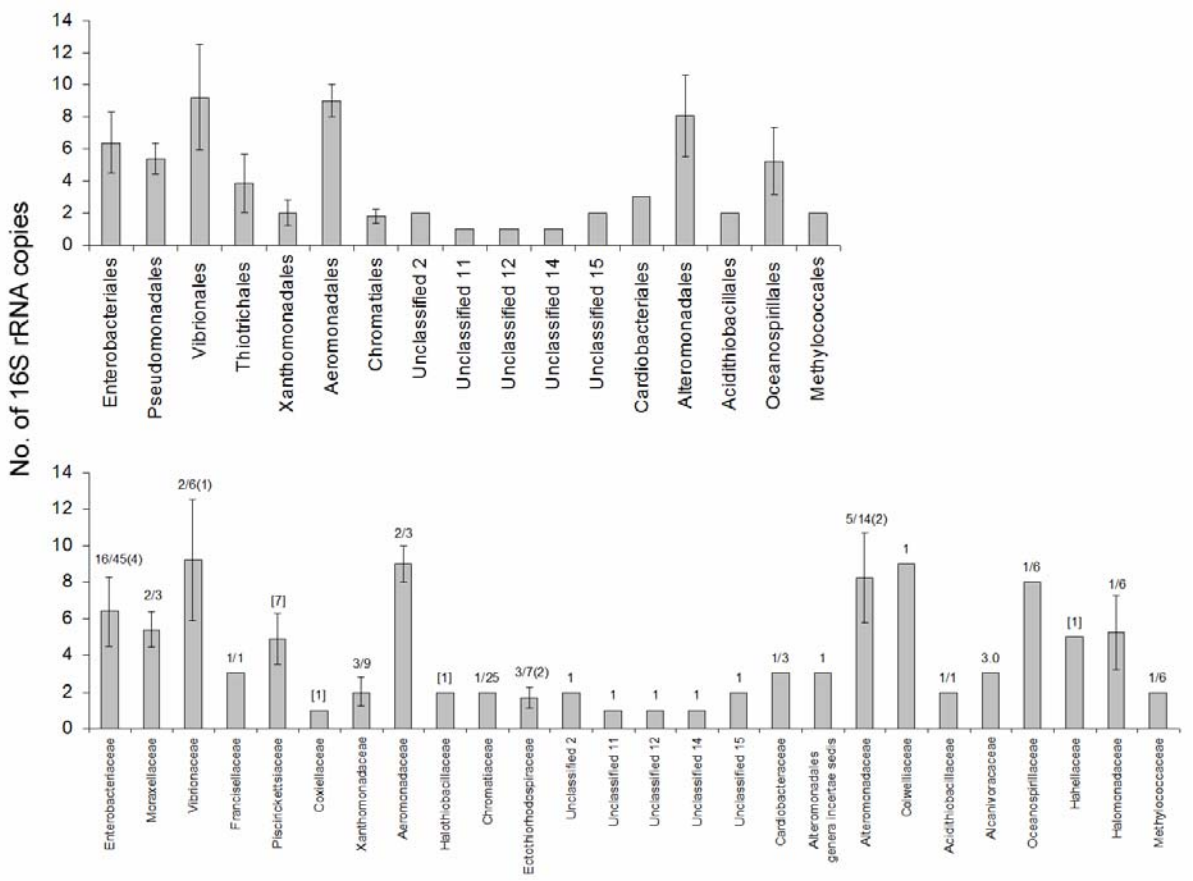

Figure 4. Variance of the 16S rRNA gene copy number among the γ-Proteobacteria orders (top) and families (bottom). The ratio of the number of genera with available genomes in GenBank to the number of genera in Bergey's Manual (Brenner et al. 2005) is shown above each column; the number of genera not included in Bergey's Manual is in parentheses. Standard deviation is indicated in error bars. No / indicates absence from the Bergey's Manual.

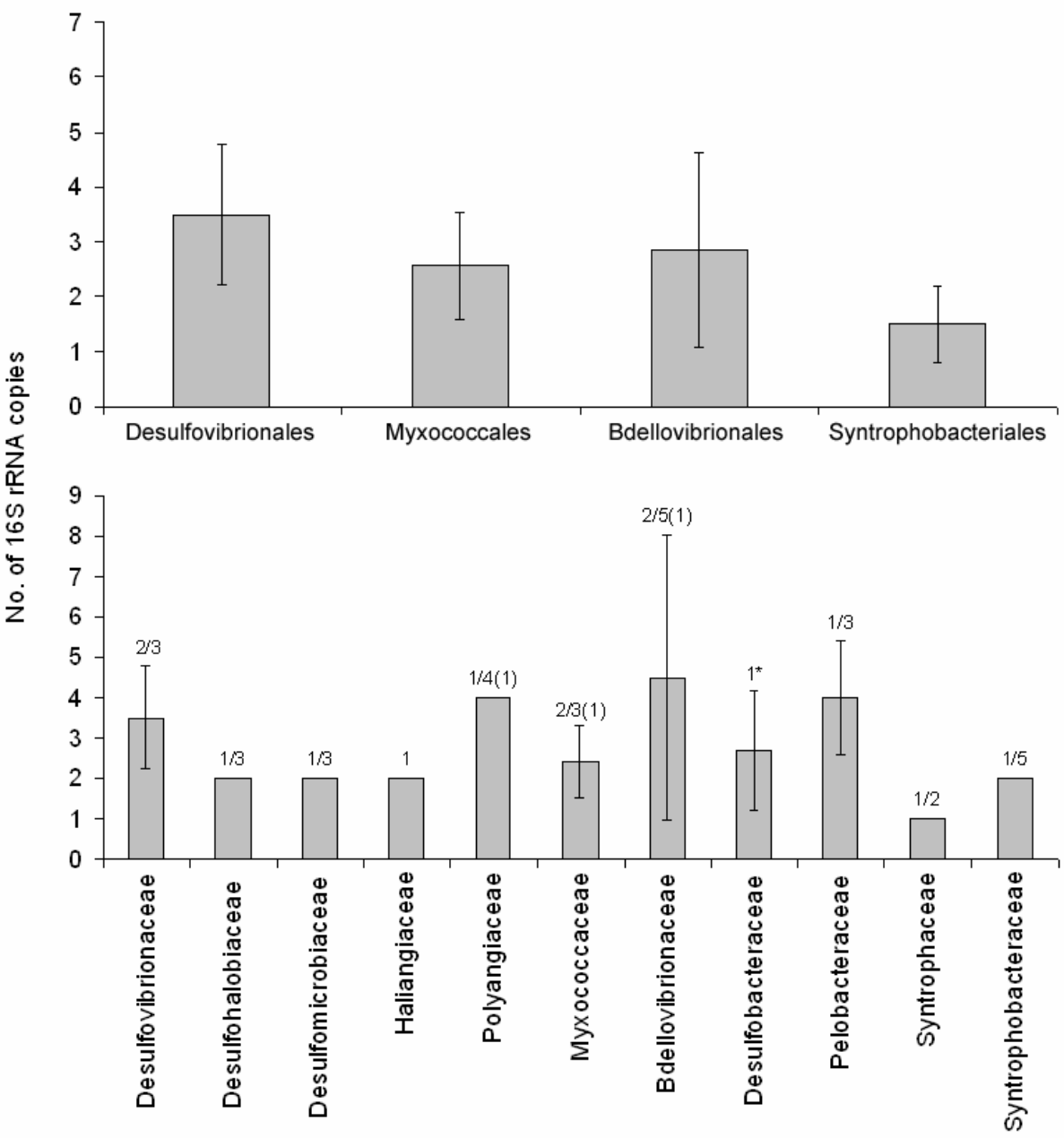

Figure 5. Variance of the 16S rRNA gene copy number among the δ-Proteobacteria orders (top) and families (bottom). The ratio of the number of genera with available genomes in GenBank to the number of genera in Bergey's Manual (Staley et al. 2005) is shown above each column; the number of genera not included in Bergey's Manual is in parentheses. Standard deviation is indicated in error bars. No / indicates absence from the Bergey's Manual.

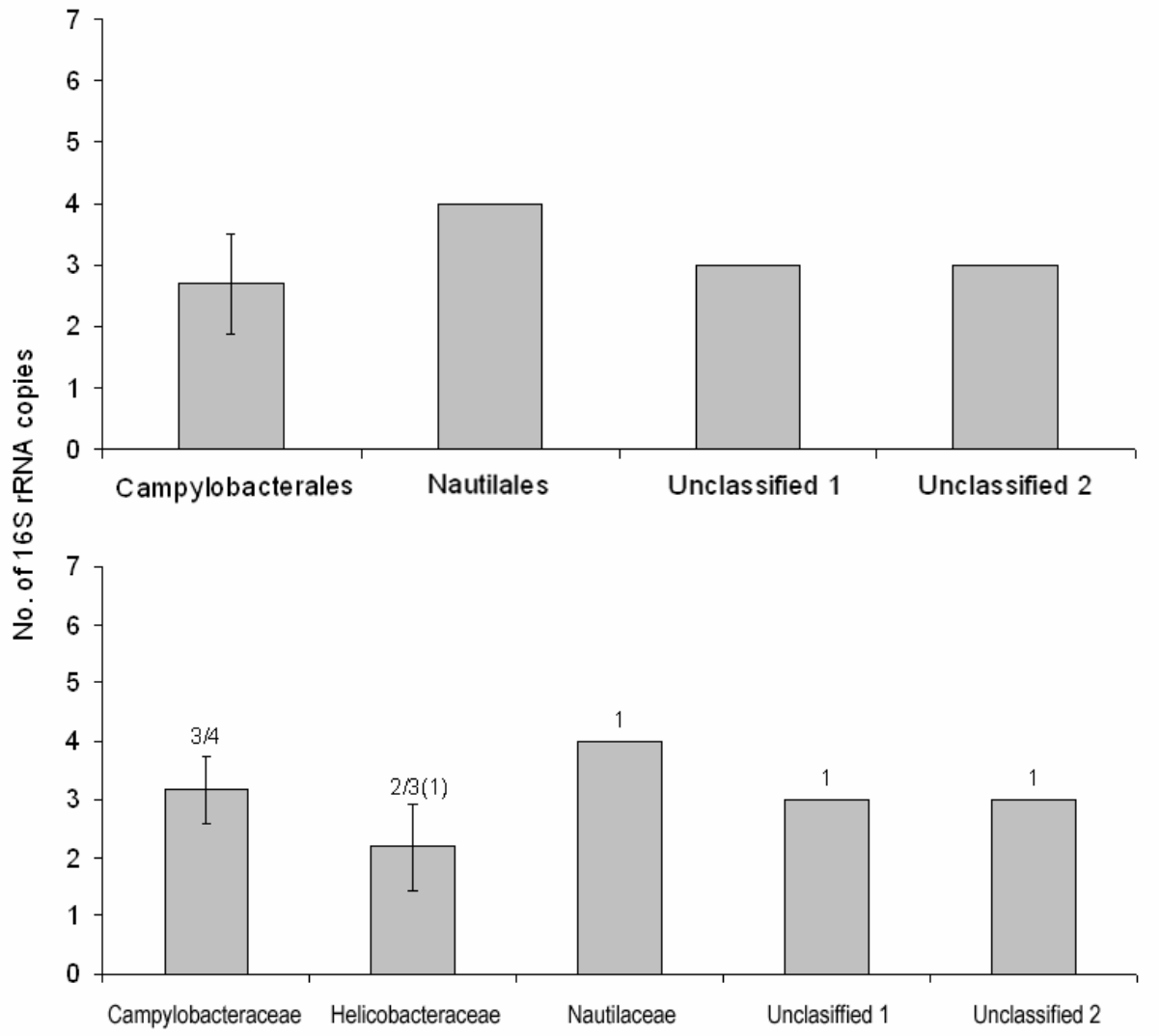

Figure 6. Variance of the 16S rRNA gene copy number among the ε-Proteobacteria orders (top) and families (bottom). The ratio of the number of genera with available genomes in GenBank to the number of genera in Bergey's Manual (Staley et al. 2005) is shown above each column; the number of genera not included in Bergey's Manual is in parentheses. Standard deviation is indicated in error bars. No / indicates absence from the Bergey's Manual.

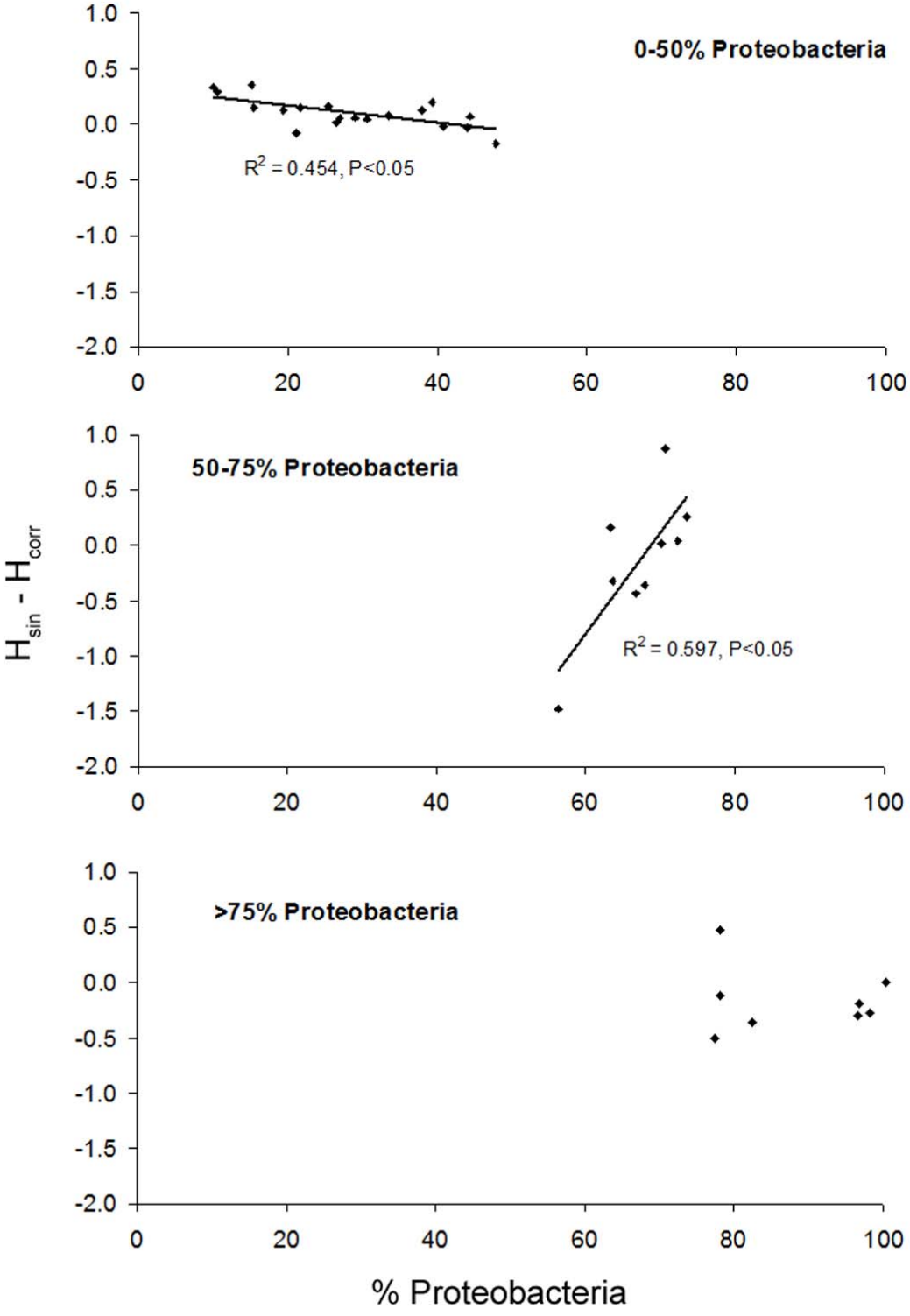

Figure 7. The relationship of the difference between the Shannon-Wiener diversity H calculated with single copy 16S rRNA gene (H_{sin}) and corrected for each of the Proteobacteria sub-phyla (H_{corr}) vs. the relative abundance (%) of Proteobacteria in environmental clone libraries (see details in the text).