Article

Explorative Meta-Analysis of 417 Extant Archaeal Genomes to Predict Their Contribution to the Total Microbiome Functionality

Robert Starke [1,*], Maysa Lima Parente Fernandes [1], Daniel Kumazawa Morais [1], Iñaki Odriozola [1], Petr Baldrian [1] and Nico Jehmlich [2]

1   Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, 142 00 Praha, Czech Republic; maysaparente1@gmail.com (M.L.P.F.); daniel.morais@brmicrobiome.org (D.K.M.); inaki.odriozola@biomed.cas.cz (I.O.); baldrian@biomed.cas.cz (P.B.)
2   Molecular Systems Biology, Helmholtz-Center for Environmental Research, UFZ, 04318 Leipzig, Germany; nico.jehmlich@ufz.de
*   Correspondence: Robert.starke@biomed.cas.cz

Abstract: Revealing the relationship between taxonomy and function in microbiomes is critical to discover their contribution to ecosystem functioning. However, while the relationship between taxonomic and functional diversity in bacteria and fungi is known, this is not the case for archaea. Here, we used a meta-analysis of 417 completely annotated extant and taxonomically unique archaeal genomes to predict the extent of microbiome functionality on Earth contained within archaeal genomes using accumulation curves of all known level 3 functions of KEGG Orthology. We found that intergenome redundancy as functions present in multiple genomes was inversely related to intragenome redundancy as multiple copies of a gene in one genome, implying the tradeoff between additional copies of functionally important genes or a higher number of different genes. A logarithmic model described the relationship between functional diversity and species richness better than both the unsaturated and the saturated model, which suggests a limited total number of archaeal functions in contrast to the sheer unlimited potential of bacteria and fungi. Using the global archaeal species richness estimate of 13,159, the logarithmic model predicted 4164.1 ± 2.9 KEGG level 3 functions. The non-parametric bootstrap estimate yielded a lower bound of 2994 ± 57 KEGG level 3 functions. Our approach not only highlighted similarities in functional redundancy but also the difference in functional potential of archaea compared to other domains of life.

Keywords: archaea; functional diversity; microbiome functionality

1. Introduction

The biochemical transformations conducted by a community of microbes from all domains of life mediate ecosystem functioning [1]. Even though ecological studies tend to focus on bacteria and fungi, archaea as a major part of global ecosystems [2] are ubiquitous in both terrestrial and aquatic environments [3,4]. Particularly, archaea make up between 20% and 30% of the total prokaryotes in marine environments [5] and between 0% and 10% in soil environments [4,6]. Increases in the proportion of archaea were found in extreme habitats such as acidity and low temperature [7]. Functionally, archaea play key roles in global carbon (e.g., methanogenesis or CO_2 fixation) and nitrogen (e.g., N_2 fixation or oxidation of ammonia) cycles [8], but they also have complex relationships with both bacteria and fungi [9]. In a microbial community, multiple organisms with different taxonomy may have similar if not identical roles in ecosystem functionality, the so-called functional redundancy [10]. Indeed, interspecies redundancy was reported to be very high with several hundreds to thousands of different taxa to express the same function in a habitat [11]. These functions can be statistically inferred based on the homology to experimentally characterized genes and proteins in specific organisms to find orthologs present in a microbiome [12–14]. This ortholog annotation is used by KEGG Orthology [15,16], which

covers a wide range of functional classes (level 1 of KEGG) comprising cellular processes, environmental information processing, genetic information processing, human diseases, metabolism, organismal system, brite hierarchies, and functions not included in the annotation of the two databases pathway or brite. KEGG level 2 functions provide more detail, i.e., differentiating glycoside hydrolases and glycosyltransferases within carbohydrate active enzymes (level 1), whereas KEGG level 3 is the enzyme itself, i.e., the glycogen phosphorylase (K00688, EC: 2.4.1.1.) that belongs to the glycosyltransferases (more information can be obtained under https://www.genome.jp/kegg/kegg3.html, accessed on 17 July 2020). However, the bottleneck of reporting microbiome functions is the low number of fully sequenced and annotated genomes as only organisms are captured that have undergone isolation and extensive characterization [12–14] with respect to the expected total diversity. Hence, the lower the share of known species or the higher the predicted total diversity, the weaker the prediction itself. Problematically, the vast majority of organisms have not yet been studied [17,18] which is why the annotation is based on the similarity to the genomes of the very few studied model organisms [12–14]. Consequentially, microbiome functionality is inferred based on the taxonomic composition of the community and its relation to functional parameters [19], indicated by the frequent use of the 16S rRNA gene metabarcoding to describe the prokaryotic community. Even though the description of microbial communities is important to assess the drivers of the occurrence of individual taxa and the composition of their communities [12], the mere taxonomic composition itself did not provide detailed answers about its functional diversity [20]. The functional diversity for both bacteria [13] and fungi [12–14] were recently predicted to comprise millions of different functions using meta-analyses of proteins [13] and genomes [12,14], most of which are unknown today. However, our understanding of functional redundancy in archaea and their contribution to the total microbiome functionality is still scarce.

Here, we used both parametric and non-parametric estimators of functional richness with the aim to predict the total archaeal functionality on Earth and to unveil the relationship between taxonomy and function in the archaeal domain. To do so, we obtained all completely annotated genomes of taxonomically unique archaeal species (n = 417) from the integrated microbial genomes and microbiomes (IMG) of the Joint Genome Institute (JGI) (https://img.jgi.doe.gov/, accessed on 17 July 2020) with taxonomic annotation on the species level and functional annotation of KEGG on level 3 (referred to as KEGG function). We used a parametric estimation based on accumulation curves (AC) [21] that are characterized by the increasing number functions with increasing species. The AC was fitted to saturated, unsaturated, and logarithmic models, and the best fitting model was chosen based on its fitness in comparison to the other models. As a non-parametric estimator, Chao-1 was used for every 50 randomly picked species of all 417 in the database each with 20 replicates. The precision of both the parametric and the non-parametric approach generally depends on the proximity to the asymptote of the model, with greater extrapolation to the total count resulting in greater error [22]. We therefore hypothesized more precise estimates of the contribution of archaea to the total microbiome functionality than previously proposed for both bacteria [13] and fungi [12–14] due to the higher coverage of the predicted taxonomic diversity in archaea.

2. Materials and Methods

2.1. Metadata Collection of the Total Known Archaeal Microbiome Functions

To predict the contribution to microbiome functions and to compare the genome content across archaeal species, habitats, and temperature ranges, available genomes from archaea were downloaded from the integrated microbial genomes and microbiomes (IMG) of the Joint Genome Institute (JGI) on 17 July 2020. A genome was randomly selected in the case of multiple sequenced genomes from the same species to obtain taxonomically distinct archaeal species. For each genome, the gene counts for each KEGG function [15,16] were retrieved. Our database comprised 417 completely annotated archaeal genomes with, in total, 2835 KEGG functions (Supplementary Table S1). Noteworthy,

761 archaeal metagenome-assembled genomes (MAG) were available in the database (as of 31 August 2020) but only seven with high quality. Even though many archaeal genomes and functions are derived from non-cultivable species, we wanted to use complete information for precise modeling. For one genome, the sequencing status was denoted as "Draft", for 217 as "Permanent Draft", and for 199 as "Finished". Only three genomes were available to describe psychrophilic archaea and the taxa were therefore excluded from further analysis. Intergenome redundancy was calculated as the number of KEGG functions covered by one randomly chosen species in comparison to the total number of functions in all species [12]. Intragenome redundancy or gene redundancy was estimated as the average of genes per individual KEGG function in one species [12]. The gene counts and KEGG functions per archaeal phylum, habitat, and temperature range were retrieved as the average with standard deviation from the database. To estimate the specific differences, both intergenome and intragenome redundancy were calculated for every phylum, habitat, and temperature range as described for the total database above.

2.2. Accumulation Curves (AC)

Archaeal species were added in intervals of one to 417 species using 1000 random permutations per step via the function *specaccum* from the R package *vegan* [23]. A saturated (Equation (1)) and an unsaturated model (Equation (2)) with the critical point estimated by the term $3A_f$ [24] was then fitted to the AC of the database permutation. Due to the plateauing shape of the AC, a logarithmic model was used as well. The fit of all models was validated by the analysis of the Akaike Information Criterion (AIC) [25] with a penalty per parameter set to k equals two. The total number of KEGG functions in archaea on Earth was predicted using a global species richness estimate of 13,159 archaeal species [26] to calculate the potential maximum of KEGG functions via uncertainty propagation and Monte Carlo simulation of the function *predictNLS* in the R package *propagate* [27]. The non-parametric estimation of functional richness was calculated by Chao-1 [28,29]. This method was developed to estimate the asymptotic species richness in a set of samples. Since our objective was to estimate the asymptotic functional richness, genomes took the role of samples and KEGG functions took the role of species in our analysis. Resampling and repeating computations for lower levels of sample accumulation generated a smoother curve of the estimations. A reliable estimator would reach its own asymptote before the species accumulation curve does [21]. To test whether this occurred in our dataset, Chao-1 was estimated using a random subset of every 50 picked archaeal genomes in the database starting with two species (Equation (3)). Additionally, asymptotic functional richness was estimated using a first order jackknife (jack-1) and the bootstrap "boot" methods with the function *specpool* in the R package *vegan* [23] to check if these two alternative methods yielded estimations comparable to the parametric and Chao-1 estimates.

$$\text{Functional richness} = f_{max} \times \text{Species richness}/(A_f + \text{Species Richness}), \tag{1}$$

$$\text{Functional richness} = f_{max} \times \text{Species richness}/(A_f + \text{Species Richness}) + (k \times \text{Species richness}), \tag{2}$$

$$\text{Chao-1} = \text{Functional richness} + (a_1{}^2/2a_2), \tag{3}$$

Here, f_{max} is the maximum functional richness, A_f the accretion rate of functions with an increasing number of species, and k the constant of the additive term. Functions found only once or twice are indicated by a_1 as singletons and a_2 as doubletons, respectively.

2.3. Statistical Analysis

The differences between gene counts, KEGG functions, and their functional redundancy were estimated by Tukey's honestly significant difference (HSD) test [30] using the package *agricolae* [31].

3. Results

3.1. Gene Counts and Number of KEGG Level 3 Functions

The gene count per species was significantly higher (HSD-test) in *Euryarchaeota* as compared to *Crenarchaeota* and *Thaumarchaeota* (Figure 1a). On the level of habitats, archaea isolated from fresh water, sediments, or soils had on average significantly more genes than archaea enriched from the deep sea or hot springs. A comparable number of archaeal genomes were sampled from each habitat, ranging from 8 in sludge to 37 in hot springs. On the level of temperature preferences, mesophilic archaea comprised significantly (HSD-test) more genes than thermophilic and hyperthermophilic archaea. Similar significant differences were apparent in the number of KEGG level 3 functions on all three prior investigated levels (Figure 1b).
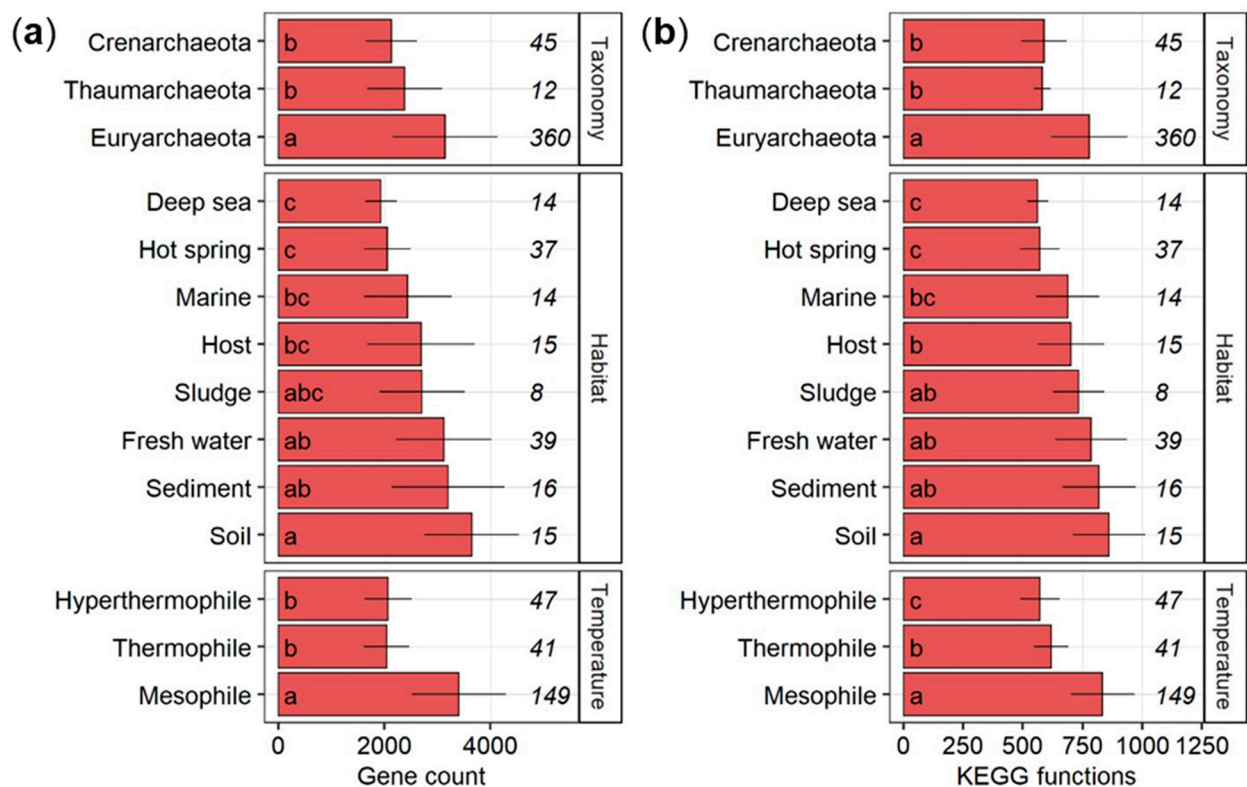

Figure 1. Counts (**a**) and the number of different KEGG functions (**b**) per genome across archaeal phyla, habitats, and temperature ranges shown as average with standard deviation. The number of archaeal genomes is given in italics. Groups followed by a different letter are significantly different according to the HSD-test ($p < 0.05$).

3.2. Inter- and Intragenome Functional Redundancy

Intergenome functional redundancy is a proxy for the performance of one metabolic function by multiple taxonomically distinct organisms, while intragenome functional redundancy describes the number of replicated functions within one genome [12,14]. Across all 417 archaeal genomes, the median of intergenome functional redundancy was found to be 0.06 (Figure 2a). Most functions were found with low redundancy as 1650 KEGG functions were present in less than 10% of the species. In comparison, only 172 KEGG functions were present in more than 90% of the archaeal genomes. Together, 65.3% of all functions showed either high or low redundancy while the rest appeared intermediate with an intergenome functional redundancy between 0.1 and 0.9 with a particularly high abundance at around 0.24. The median of intragenome functional redundancy across all 417 archaeal genomes was found to be 1.02 gene copies per KEGG function with a maximum of 72 gene copies (Figure 2b).
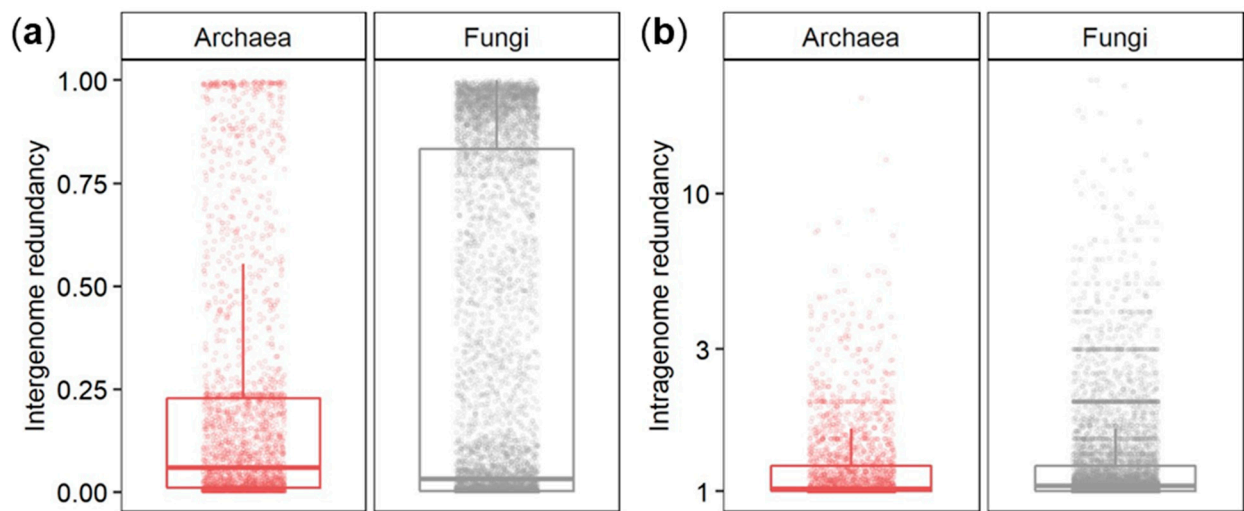
Figure 2. The distribution of intergenome functional redundancy as the total share of functions within archaea relative to the total number of archaeal species in the database (**a**) and intragenome functional redundancy as the number of replicated KEGG functions within one archaeal species in the database (**b**) compared to the previously published distributions in fungi [12].

Among archaeal phyla, *Thaumarchaeota* showed a significantly higher (HSD-test) intergenome functional redundancy compared to *Crenarchaeota* and *Euryarchaeota* (Figure 3a). Within habitats, the intergenome redundancy in the deep sea, hot springs, sediments, and sludges was significantly higher (HSD-test) than in fresh water, host, marine, and soil habitats. On the level of temperature preferences, intergenome redundancy was highest in hyperthermophilic archaea, followed by thermophilic and mesophilic ones. The inverse pattern was found for intragenome functional redundancy for all three investigated levels (Figure 3b). Significantly higher intergenome redundancy was accompanied by significantly lower intragenome redundancy and vice versa regardless the taxonomy, habitat, and temperature preference of archaea.

3.3. Parametric and Non-Parametric Estimation of the Archaeal Contribution to the Total Microbiome Functionality

The logarithmic model comprised a significantly better fit of the dependence of functions on archaeal species richness than both the saturated and the unsaturated model, estimated by lower akaike information criterion (AIC) (Figure 4a) to imply a plateau of functional richness with higher species richness. Considering the estimate of 13,159 archaeal species on Earth [26] and assuming that the relationship between species richness and functional richness will be logarithmic with the addition of new species, we propagated the logarithmic model with the result of a total archaeal functionality of 4164.2 ± 2.9 KEGG functions (with 4158.6 and 4169.9 as 95% confidence intervals). Similarly, the non-parametric estimator of functional richness that assumes the existence of a maximum functional richness indeed plateaued for the 417 archaeal genomes (Figure 4b). Estimations obtained with more than 200 archaeal genomes generated broadly overlapping confidence intervals indicative of the reliability of the estimation of the asymptotic functional richness. The three non-parametric estimators yielded comparable estimations of asymptotic functional richness: 3128 ± 42 KEGG functions using the Chao-1 index, 3169 ± 78 using the first order jackknife, and 2994 ± 57 using the bootstrap method (Figure 4b).
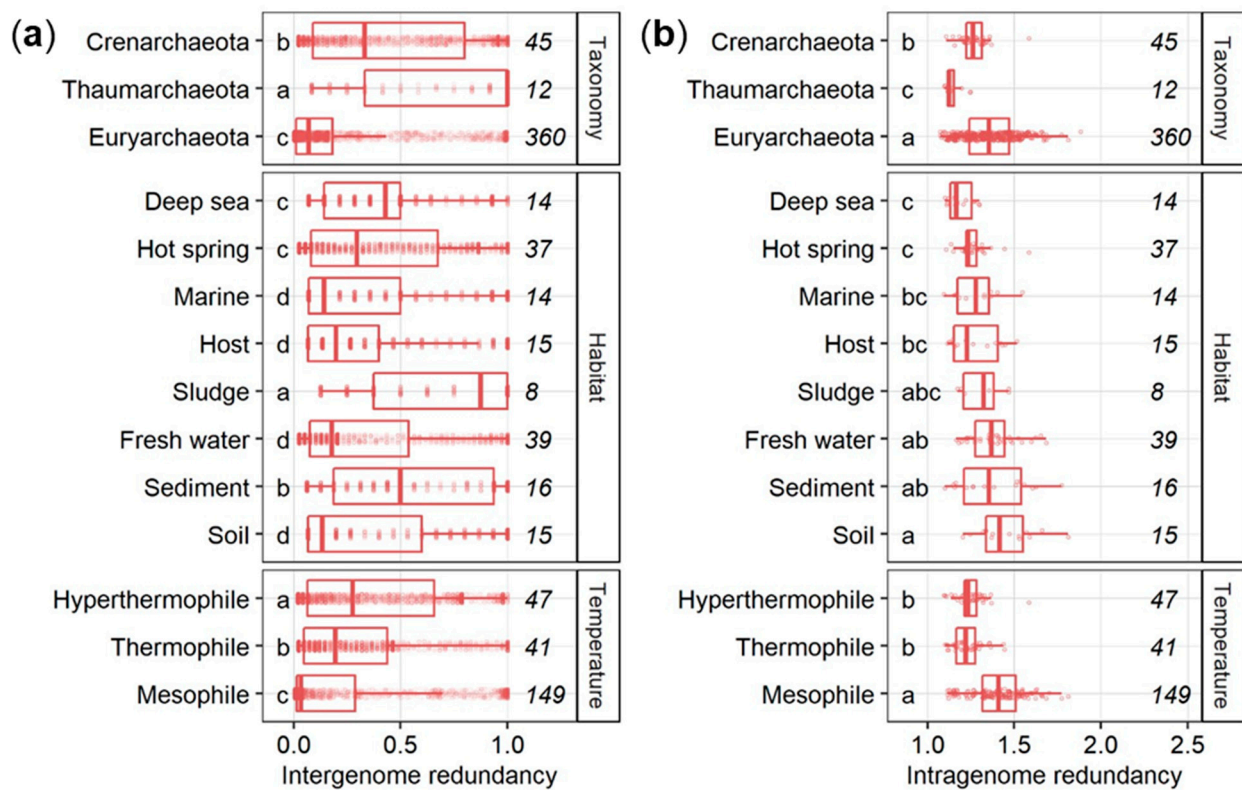
Figure 3. Intergenome (**a**) and intragenome functional redundancy (**b**) in archaeal phyla, habitats, and temperature ranges. The number of archaeal genomes is given in italic. Groups followed by a different letter are significantly different according to the HSD-test ($p < 0.05$).
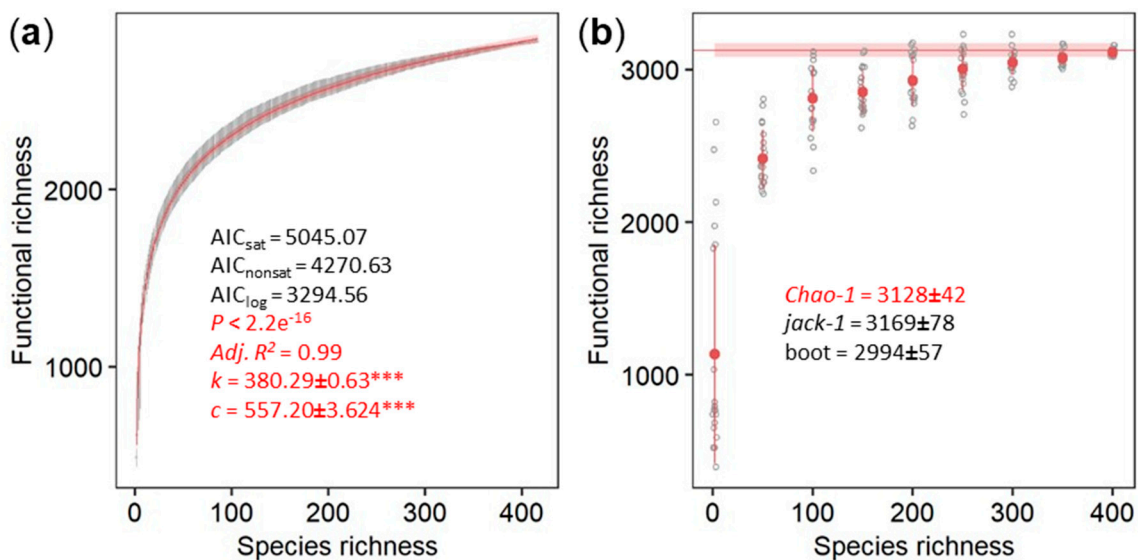

Figure 4. Parametric (**a**) and non-parametric (**b**) estimation of the total functional richness. The logarithmic model of the accumulation curves as gray points with error bars as 95% confidence intervals for the total known archaeal microbiome functions derived from the KEGG database by 1000 random permutations for every one species richness with 95% confidence intervals. Significance of the parameter estimates are indicated by asterisks (*** equals $p < 0.001$). The Chao-1 index as lower bound and non-parametric estimate was calculated using 20 replicates shown in gray for every 50 randomly picked archaeal genomes in the database starting with two species. The Chao-1 index of all 417 archaeal genomes with standard errors is shown as red line.

4. Discussion

4.1. Genome Content

The genome size of an organism generally reflects its developmental and ecological needs [32]. Larger genomes are directly related to increases in both cellular and nuclear volumes [33] that help to cushion fluctuations in concentrations of regulatory proteins or to protect coding DNA from spontaneous mutation [34]. Variation in genome size is therefore a result of the adaptive needs or of natural selection in different organisms [32]; the so-called adaptive theory of genome evolution. The smaller genomes of archaea could be directly related to a higher evolutionary rate [35]. Indeed, the 417 archaeal species generally comprised smaller genomes compared to bacteria [36], but with statistically significant differences among phyla, habitats, and temperature ranges that were mirrored by the number of KEGG level 3 functions in each genome. Particularly, archaea inhabiting extreme habitats such as deep sea or hot springs characteristic with high local temperatures not only had significantly fewer total genes but also fewer KEGG functions. Otherwise, environments of higher complexity and diversity such as soils or sediments contained archaea with a larger functional potential that may have allowed them more options for the competition for or the utilization of a wider range of nutrients.

4.2. Functional Redundancy

A limited set of metabolic pathways found in a variety of taxonomic groups drive most biogeochemical reactions [37] which is why the diversity in the community is correlated strongly with its functional diversity [38]. Functions are classified into two groups [12–14]: (i) Highly redundant across different species present in more than 90% of all species or (ii) unique to only a few species present in less than 10% of all species. Here, intergenome redundancy was either high or low for roughly two thirds of all the KEGG functions; fewer than the 77.3% were found in fungal genomes [12,14]. However, the presence of a higher share of functions of intermediate redundancy that are present between the two thresholds suggested the presence of more than two groups [12–14] that could be particularly important for organisms with smaller genomes such as archaea and bacteria. A set of functions present in a quarter of all archaea indicated that the presence of a driving phylotype or environment may drive intergenome redundancy. Indeed, most functions (151/194) with an intergenome redundancy between 22% and 26% belonged to the phylum *Crenarchaeota*, the habitat hot springs, and the temperature range of hyperthermophilic archaea, mainly affiliated with amino acid utilization, fermentation, methanogenesis, and nucleic acid metabolism. The median intergenome redundancy was twice as high as found for fungi before [12,14], implying a higher share of functions shared among archaea on average. However, only half the gene copies (1.02 in archaea compared to 2.0 in fungi) were present, highlighting the close relationship between intergenome and intragenome redundancy. Indeed, the archaeal genomes revealed that low intergenome redundancy is generally related to high intragenome redundancy and vice versa. Presumably, every organism must choose between additional copies of functionally important genes or a higher number of different genes, especially in reduced genomes. Similarly to the pattern found in fungi before [12,14], functions belonging to the maintenance apparatus such as S-adenosylmethionine synthetase (EC 2.5.1.6, K00789) involved in the biosynthesis of amino acids were with both high intergenome and high intragenome redundancy, allowing for more complex regulation of the gene, i.e., when more transcripts are needed. Otherwise, functions with low intergenome and low intragenome redundancy are highly specialized processes only performed by a few archaea such as the drug transporter MFS transporter, DHA1 family, multidrug resistance protein (K19578) found in the crenarchaeote *Thermofilum adornatus*.

4.3. Archaeal Contribution to the Total Microbiome Functionality

The parametric approach estimated the archaeal contribution to the total microbiome functionality to roughly 4200 KEGG functions; a magnitude less than predicted for both bacteria [13] and fungi [12–14]. The lower bound estimate of functional richness derived from the non-parametric approaches yielded roughly 3000 KEGG functions. A plateau of functional richness with higher species richness made the predictions for archaea more reliable as the errors decreased with proximity to the asymptote [22]. Theoretically, a higher number of species must be sequenced until no additional functions are unveiled and the accumulation curve reaches the actual asymptote [39]. However, practically, this is nearly impossible as a prohibitively large number of species are needed to be sampled in order to reach an asymptote [40]. In our meta-analysis, admittedly, the 417 genomes of distinct archaeal species only spanned three archaeal phyla from all 21 proposed phyla [41,42] and covered only a small part of the predicted taxonomic diversity in archaea; with databases containing up to 13,159 archaeal species [26], the prediction of 5000 archaeal genera [43], and the finding of 669 distinct archaeal species among 10,575 prokaryotic genomes [44]. Hence, the addition of genomes from novel archaeal species with potentially new KEGG functions could change both the parametric and the non-parametric estimates of functional richness. However, the differences in the estimates are likely not as tremendous as the potential differences in the estimates for both bacteria [13] and fungi [12–14] as the accumulation curve already plateaued with 417 taxonomically distinct archaeal species. Noteworthy, it is unclear how well new functions are recovered in archaea. As there is notably less interest in archaea compared to bacteria, functional annotations might generally miss archaea-specific functions to a larger extent than bacteria-specific functions missed in bacteria. As of today, our understanding of the contribution of archaea to the total microbiome functionality covers the majority of the KEGG functions, but many as-yet unknown archaea-specific functions could exist.

5. Conclusions

Our results suggest a limited contribution of archaea to the total functional potential of the microbiome, with most archaeal functions already identified as of today. However, the existence of archaea-specific functions must be validated by novel and more sophisticated methods. The accumulation curve describing the increase of functional categories with the number of sequenced genomes in archaea was closer to the asymptote than in bacteria [13] and fungi [12–14]. This made the estimate of archaeal contribution to the total microbiome functionality more precise, although it is still uncertain if the functional diversities of different domains can easily be compared. Noteworthy and similar to fungi, only one quarter of all genes in archaeal genomes on average were affiliated with a KEGG function, which demonstrates the limitations of the annotation because the prediction of microbiome functionality technically excluded three quarters of the entire functional potential in archaea. Different ortholog databases such as COG or Pfam could further improve our understanding of functional diversity, especially in archaea, as those covered three times more genes than KEGG did. Still, different approaches and definitions of functions are necessary to estimate the actual functional diversity of the microbiome.

References

1. Woese, C.R.; Kandler, O.; Wheelis, M.L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 4576–4579. [CrossRef] [PubMed]
2. Delong, E.F. Everything in moderation: Archaea as "non-extremophiles.". *Curr. Opin. Genet. Dev.* **1998**, *8*, 649–654. [CrossRef]
3. Timonen, S.; Bomberg, M. Archaea in dry soil environments. *Phytochem. Rev.* **2009**, *8*, 505–518. [CrossRef]
4. DeLong, E.F.; Pace, N.R. Environmental diversity of bacteria and archaea. *Syst. Biol.* **2001**, *50*, 470–478. [CrossRef] [PubMed]
5. Stoica, E.; Herndl, G.J. Contribution of Crenarchaeota and Euryarchaeota to the prokaryotic plankton in the coastal northwestern Black Sea. *J. Plankton Res.* **2007**, *29*, 699–706. [CrossRef]
6. Bates, S.T.; Berg-Lyons, D.; Caporaso, J.G.; Walters, W.A.; Knight, R.; Fierer, N. Examining the global distribution of dominant archaeal populations in soil. *Isme J.* **2011**, *5*, 907–917. [CrossRef] [PubMed]
7. Korzhenkov, A.A.; Toshchakov, S.V.; Bargiela, R.; Gibbard, H.; Ferrer, M.; Teplyuk, A.V.; Jones, D.L.; Kublanov, I.V.; Golyshin, P.N.; Golyshina, O.V. Archaea dominate the microbial community in an ecosystem with low-to-moderate temperature and extreme acidity. *Microbiome* **2019**, *7*, 11–14. [CrossRef] [PubMed]
8. Leininger, S.; Urich, T.; Schloter, M.; Schwark, L.; Qi, J.; Nicol, G.W.; Prosser, J.I.; Schuster, S.C.; Schleper, C. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **2006**, *442*, 806–809. [CrossRef] [PubMed]
9. Bengtson, P.; Sterngren, A.E.; Rousk, J. Archaeal abundance across a pH gradient in an arable soil and its relationship to bacterial and fungal growth rates. *Appl. Env. Microbiol.* **2012**, *78*, 5906–5911. [CrossRef]
10. Hubbell, S.P. Neutral theory in community ecology and the hypothesis of functional equivalence. *Funct. Ecol.* **2005**, *19*, 166–172. [CrossRef]
11. Žifčáková, L.; Větrovský, T.; Lombard, V.; Henrissat, B.; Howe, A.; Baldrian, P. Feed in summer, rest in winter: Microbial carbon utilization in forest topsoil. *Microbiome* **2017**, *5*, 1–12. [CrossRef]
12. Starke, R.; Capek, P.; Morais, D.; Jehmlich, N.; Baldrian, P. Explorative Meta-Analysis of 377 Extant Fungal Genomes Predicted a Total Mycobiome Functionality of 42.4 Million KEGG Functions. *Front. Microbiol.* **2020**, *11*, 143. [CrossRef] [PubMed]
13. Starke, R.; Capek, P.; Morais, D.; Callister, S.J.; Jehmlich, N. The total microbiome functions in bacteria and fungi. *J. Proteom.* **2020**, *2013*, 103623. [CrossRef]
14. Starke, R.; Capek, P.; Morais, D.K.; Jehmlich, N.; Baldrian, P. The Total Fungal Microbiome Functionality. 2019. Available online: https://www.biorxiv.org/content/biorxiv/early/2020/08/04/2020.08.04.236075.full.pdf (accessed on 17 July 2020).
15. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [CrossRef] [PubMed]
16. Kanehisa, M.; Sato, Y.; Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **2016**, *428*, 726–731. [CrossRef]
17. Pham, V.H.T.; Kim, J. Cultivation of unculturable soil bacteria. *Trends Biotechnol.* **2012**, *30*, 475–484. [CrossRef]
18. Martiny, A.C. High proportions of bacteria are culturable across major biomes. *ISME J.* **2019**, *13*, 2125–2128. [CrossRef]
19. Starke, R.; Jehmlich, N.; Bastida, F. Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *J. Proteom.* **2018**, *198*, 50–58. [CrossRef]
20. Větrovský, T.; Kohout, P.; Kopecký, M.; Macháč, A.; Man, M.; Bahnmann, B.D.; Brabcová, V.; Choi, J.; Meszárošová, L.; Human, Z.R.; et al. A meta-analysis of global fungal distribution reveals climate-driven patterns. *Nat. Commun.* **2019**, *10*, 1–9. [CrossRef] [PubMed]
21. Gotelli, N.J.; Colwell, R.K. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **2001**, *4*, 379–391. [CrossRef]
22. Thompson, G.G.; Withers, P.C.; Pianka, E.R.; Thompson, S.A. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. *Austral. Ecol.* **2003**, *28*, 361–383. [CrossRef]
23. Oksanen, J.; Blanchet, F.G.; Friendly, M.; Kindt, R.; Legendre, P.; McGlinn, D.; Minchin, P.R.; O'Hara, R.B.; Simpson, G.L.; Solymos, P.; et al. Vegan: Community Ecology Package. 2015. Available online: https://cran.r-project.org/web/packages/vegan/index.html (accessed on 17 July 2020).
24. Čapek, P.; Kotas, P.; Manzoni, S.; Šantrůčková, H. Drivers of phosphorus limitation across soil microbial communities. *Funct. Ecol.* **2016**, *30*, 1705–1713. [CrossRef]
25. Bertrand, P.V.; Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. Akaike Information Criterion Statistics. *J. R. Stat. Soc. Ser. A* **2006**, *151*, 567–568. [CrossRef]

26. Yarza, P.; Yilmaz, P.; Pruesse, E.; Glöckner, F.O.; Ludwig, W.; Schleifer, K.H.; Whitman, W.B.; Euzéby, J.; Amann, R.; Rosselló-Móra, R. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **2014**, *12*, 635–645. [CrossRef] [PubMed]

27. Spiess, A.N. Propagate: Propagation of Uncertainty. 2018. Available online: https://cran.r-project.org/web/packages/vegan/index.html (accessed on 17 July 2020).

28. Chao, A. Non-parametric estimation of the classes in a population. *Scand. J. Stat.* **1984**, *11*, 265–270. [CrossRef]

29. Chao, A. Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics* **1989**, *45*, 427–438. [CrossRef]

30. Tukey, J.W. Comparing Individual Means in the Analysis of Variance. *Biometrics* **1949**, *5*, 99–114. [CrossRef]

31. De Mendiburu, F. Agricolae: Statistical Procedures for Agricultural Research. 2014. Available online: https://cran.r-project.org/web/packages/agricolae/index.html (accessed on 17 July 2020).

32. Petrov, D.A. Evolution of genome size: New approaches to an old problem. *Trends Genet.* **2001**, *17*, 23–28. [CrossRef]

33. Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* **1978**, *24*, 247–278.

34. Vinogradov, A.E. Buffering: A possible passive-homeostasis role for redundant DNA. *J. Biol.* **1998**, *193*, 197–199. [CrossRef]

35. Wang, H.Y.; Guo, S.Y.; Huang, M.R.; Thorsten, L.H.; Wei, J.C. Ascomycota has a faster evolutionary rate and higher species diversity than Basidiomycota. *Sci. China Life Sci.* **2010**, *53*, 1163–1169. [CrossRef]

36. Větrovský, T.; Baldrian, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE* **2013**, *8*, e57923. [CrossRef] [PubMed]

37. Falkowski, P.G.; Fenchel, T.; Delong, E.F. The microbial engines that drive earth's biogeochemical cycles. *Science* **2008**, *320*, 1034–1039. [CrossRef]

38. Rineau, F.; Courty, P.E. Secreted enzymatic activities of ectomycorrhizal fungi as a case study of functional diversity and functional redundancy. *Anna. For. Sci.* **2011**, *68*, 69–80. [CrossRef]

39. Gotelli, N.; Colwell, R. Estimating species richness. *Biol. Divers. Front. Meas. Assess.* **2011**, *12*, 39–54. [CrossRef]

40. Chao, A.; Colwell, R.K.; Lin, C.W.; Gotelli, N.J. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **2009**, *90*, 1125–1133. [CrossRef] [PubMed]

41. Williams, T.A.; Szöllosi, G.J.; Spang, A.; Foster, P.G.; Heaps, S.E.; Boussau, B.; Ettema, T.J.G.; Martin Embley, T. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E4602–E4611. [CrossRef]

42. Castelle, C.J.; Banfield, J.F. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **2018**, *172*, 1181–1197. [CrossRef]

43. Amann, R.; Rosselló-Móra, R. After all, only millions? *MBio* **2016**, *7*, e00201-16. [CrossRef] [PubMed]

44. Zhu, Q.; Mai, U.; Pfeiffer, W.; Janssen, S.; Asnicar, F.; Sanders, J.G.; Belda-Ferre, P.; Al-Ghalith, G.A.; Kopylova, E.; McDonald, D.; et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef] [PubMed]

45. Starke, R.; Fernandes, M.L.P.; Morais, D.K.; Odriozola, I.; Jehmlich, N.; Baldrian, P. Explorative Meta-Analysis of 417 Extant Archaeal Genomes to Predict Their Contribution to the Total Microbiome Functionality. 2020. Available online: https://www.biorxiv.org/content/10.1101/2020.08.04.236075v1 (accessed on 17 July 2020).