Temperature and pH universally govern protein diversity in hydrothermal spring communities, but they do so differently

Juan Rivas-Santisteban*[1], Jaime Alcorta[2,3], Beatriz Díez[3,4,5], Javier Tamames[1], and Carlos Pedrós-Alió*[1]

[1]*Systems Biology Department, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain*
[2]*Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain*
[3]*Millennium Institute Center for Genome Regulation (MI-CRG), Santiago, Chile*
[4]*GEMA Center for Genomics, Ecology & Environment, Universidad Mayor, Santiago, Chile*
[5]*Center for Climate and Resilience Research (CR2), University of Chile, Santiago, Chile*

September 28, 2025

Abstract

It has been suggested that temperature and pH are strong explanatory causes of protein evolutionary constraint. However, a general understanding of how changes in temperature and pH may universally impact protein evolution is lacking. While pH is locally regulated in microbial compartments, temperature is not. These variables may act at different rates in a given genome, since there are genes encoding for proteins allocated in different compartments. To test whether this is true, we can use a proxy for the number of evolutionary events. One way is to calculate the diversity of unique sequences assigned to an ortholog across species. The expectation for any ortholog is to be represented by a similar number of sequences. We examine sequence diversity among 17 metagenomes from El Tatio geothermal field (Chile), spanning temperatures of 45-62ºC, and pH values of 7.2-9.3. When controlling for metagenomic abundance, higher temperature increases, while basic pH decreases the expected by-function diversity. We thus provide evidence against temperature-induced random gene loss hypotheses. The inclusion of these variables improved the prediction of sequence diversity from metagenomic abundance. Importantly, between-compartment diversity was more affected by pH than by temperature, partially supporting our initial hypothesis

Introduction

(1) How does the overall diversity of orthologous proteins scale with the metagenomic abundance of a KEGG annotation? (2) Which factors influence this prediction over evolutionary timescales? While question (1) is deceptively straightforward (the more abundant a KEGG is across taxa, the more unique sequences will contribute to its abundance), the resolution of question (2) may further elucidate the genetic basis behind the observed functional diversity. The number of different sequences contributing to a KEGG across species is a decent proxy of the number of evolutionary events suffered by an ortholog in a metastable community [1]; events like mutation [2], duplication [3], indels [4], gene length variation [5], or horizontal transfer [6, 7]; and the relation of these to genome architecture: gene loss [8], genome reduction [9], tandem and antiparallel gene overlapping [10], or GC content [11]. Thus, knowing which factors make the by-KEGG sequence diversity grow or diminish should render a clearer image of the causes for protein evolution.

Unfortunately, a clear description of these universal factors is lacking, although their existence is granted. This is because a long-standing assumption in the nearly-neutral model of molecular evolution is that species with large effective population sizes (like most microbial organisms) are under strong selection [12–15]. Then selection, not drift, is expected to be the primary responsible of protein evolution and genome size in many prokaryotes [16]. But, as said, the mechanistic causes for selection are less clear [14, 17]. The intuition that temperature and other environmental variables are most responsible for the evolutionary constraint of microbial proteins has a varied body of evidence, but of unproven magnitude [18–21]. In agreement with this anticipation are recent observations of correlation between physicochemical variables and rates of bacterial protein evolution [22]. Another kind of evidence are the signature adaptive genomic responses to different temperatures during experimental evolution in *E. coli* [23]. Moreover, convergent mutations in independent populations grown under thermal stress have been observed [24].

But not all environmental variables are of the same nature, and they might elicit quite different responses from microbial protein evolution. Additionally, redox potential and pH are regulated by specific genetic machinery that transforms and compartmentalizes the local chemistry [25].

Indeed, many oxidative functions are confined within the gram-negative periplasm [26], whereas a reduced cytoplasm is a common feature of all cellular life [27–29]. In contrast, temperature changes irrespectively of the cell compartment. In other words, any single-celled organism must, by definition, encode all its proteins to function optimally at a single temperature range, but different segments of its genome may be still adapted to a variety of pH optima [30]. Thus, our hypothesis posits that environmental pH and temperature (i) constrain protein evolution in distinct ways, (ii) will produce distinguishable rates of evolution (and number of evolutionary events) among proteins located in different compartments.

To the best of our knowledge, a simple but intriguing question is yet to be addressed: how do temperature and pH affect the total number of ortholog variants? Here we have the opportunity to ask this to a metagenomic collection from El Tatio geothermal field, at over 4,200 meters above sea level in the Atacama Desert, Chile. The collection comprises 17 metagenomes from 13 different thermal locations, with temperature ranging from 45º to 62ºC, and pH ranging from 7.2 to 9.3 (Fig. 1). El Tatio geographic isolation, combined with extreme – yet stable – conditions [31,32] make it a natural laboratory for investigating microbial evolutionary processes. If temperature or pH had a genuine influence on the evolution of these protein repertoires, this relationship should emerge clearly regardless of particular taxonomic trends. Since our dataset encompasses only a photograph in the evolution of the organisms inhabiting these peculiar environments, we cannot derive conclusions for detailed mechanisms or evolutionary rates among their orthologs. In spite this limitation, we can safely assume that much of the variation, if any, in actual sequence diversity per ortholog will have been caused by adaptation to such particular environments.

Overall, we found that pH reduces, while temperature increases, the number of unique sequences when controlling for abundance: this supports our primary hypothesis. Contrary to our expectations, we found that pH has a consistently higher explanatory power than temperature. We also found evidence supporting our secondary hypothesis, as coefficients of pH and temperature on sequence diversity are distinctive among compartments. Finally, when interrogating protein length distributions, we find that warmer metagenomes are enriched in shorter proteins, and depleted in very long proteins (> 1000aa). Our results broadly argue against the

hypotheses relating genome reduction via random gene loss to thermophilic adaptation [33–36], as by-ortholog sequence diversity is found to be enriched in the warmer springs.

Sequence diversity is predicted by temperature categories

In this section, we analyze the most represented taxonomic domain in our samples (Bacteria: 60.41% of the ORFs, Fig. S1). Analyses for other domains are shown in the Figure S2. To facilitate the interpretation of results, we collapsed ORFs from different samples corresponding to three temperature bins with the same taxonomical diversity (see Methods). We checked whether the distributions of sequence diversity (average number of sequences contributing to a KEGG) were different (Fig. 2a). All pairwise comparisons were significant ($p < 0.0001$; 45-50°C: median = 9.0, mean = 30.35; 54-55°C: median = 13.0, mean = 38.82; 56-62°C: median = 7.5, mean = 25.41). Thus, the lowest diversity was found in the 56-62°C range. To better capture the relationship between abundance and sequence diversity, we perfomed bivariate regressions by temperature category. All three linear regressions were significantly different from that of the general dataset (Fig. 2b). However, the slopes might not be accurate as different temperature categories had differences in sequence diversity driven by the lack of more abundant KEGGs; e.g. the latter are non-randomly distributed across categories. Thus, we checked if the residual distributions (on the overall abundance-diversity regression) were still different. We found that every pairwise t-test to be highly significant ($p < 0.0001$, Fig. 2c). Residuals averaged -2.75 at the lowest temperature range, indicating overprediction (i.e. the overestimation of higher diversity was due to a lack of control for abundance), peaked 1.45 at the mid range (indicating underprediction), and remained positive (0.35) in the warmest. Medians showed the same trends.

Sequence diversity is predicted by pH categories

As pH is not correlated with temperature (Fig. S3), we performed the latter analysis for three distinct pH categories (Fig. 3). The distributions of sequence diversity were likewise different ($p < 0.0001$; 7.2–7.5: median = 7.0, mean = 19.88; 7.6–8.0: median = 10.0, mean = 31.32;

8.1-9.3: median = 13.0, mean = 40.96) (Fig. 3a). Without controlling for abundance, the highest diversity is found within the more basic pH category. Every regression slope did depart significantly from the general dataset (Fig. 3b). While checking the residual distributions, we found that every pairwise t-test was significant ($p < 0.0001$; Fig. 3c). The intermediate pH category showed the highest residuals (mean of 1.74). Thus, pH categories also had a genuine effect on sequence diversity when controlling for abundance, although its optimum was reached at the intermediate category (the slope also hints this in Fig. 3b, being lower afterwards).

Effects of pH and temperature are variable among taxonomic domains when controlled for abundance

To answer whether the observed differences between pH and temperature categories provide a universal trend on the expected by-KEGG sequence diversity, we computed the sum of squares to quantify the explained variance in sequence diversity by each environmental factor, but calculating it for every taxonomic domain assigned to our queries (Table 1).

A considerable amount of the explained variance remained when controlling for abundance in bacterial sequence diversity, with the pH being more important. This suggests a genuine effect of temperature and pH on the number of predicted ORFs per function. We wondered if this effect is universal, or alternatively whether it is bound to different compartmentalization strategies in different life domains. In our dataset, we found that the effect was universal for the ORFs linked to cellular life (pH was always explaining more variance), although it was stronger in Archaea and weaker in Eukarya. On the other hand, the diversity of sequences associated with Acytota (viral entities) was found to be barely explained by temperature, despite significance. Whether this result may reflect the lack of control of both environmental variables inside viral particles or not, needs to be clarified.

Table 1: Proportion of the variance explained (%) by pH and temperature across domains (Sum of squares). Significance is indicated; ***: p-value < 0.001; **: p-value < 0.01.

Domain	Factor	Gene counts	Residuals
Bacteria	pH	7.42^{***}	7.37^{***}
	Temperature	3.96^{***}	3.94^{***}
Acytota	pH	1.81^{***}	1.83^{***}
	Temperature	0.08^{**}	0.09^{**}
Archaea	pH	8.61^{***}	8.44^{***}
	Temperature	6.48^{***}	5.76^{***}
Eukarya	pH	4.04^{***}	4.04^{***}
	Temperature	2.38^{***}	2.39^{***}

Higher temperatures increase sequence diversity, while basic pH reduces it

Although illustrative of the effect size, the latter result has two limitations: (i) does not capture its directionality – whether it increases or reduces the number of sequences associated to a given function, and (ii) it does not capture the full breadth of specific conditions either, as we collapsed temperature/pH values into categories to improve visualisation. To confirm or reject our interpretation, we examined the residuals from an ordinary least squares (OLS) regression of sequence diversity on metagenomic abundance without considering categories (i.e.: temperature and pH as continuous variables).

The results suggest a universal effect of pH and temperature on prokaryote sequence diversity: basic pH reduces the by-function diversity, while higher temperatures increase it (Fig. 4). While ORFs bounded to particular domains exhibit different trends (Fig. 4), the polarity of the effects is mostly maintained in bacterial ORFs – which as said represent 60.41% of the counts, but comprise 97.17% of unique KEGG identifiers × samples in which they are found.

Remarkably, pH was often not as significant as temperature, but had a consistently larger effect size (Fig. 4), also predicted before (Table 1). This larger variance suggests a nonlinearity regarding the relationship between sequence diversity and the tested range of pH values.

Variability in sequence diversity among different subcellular locations increases more with environmental pH

We have shown good evidence for pH and temperature acting in opposite directions when predicting sequence diversity. Our primary hypothesis (outlined in the Introduction) predicted a different behavior caused by the different nature of the variables. These would act fostering or preventing events of protein divergence within an ortholog regardless of the taxonomic trends. As shown, our results are already supporting this interpretation. However, this says little about the validity of our secondary hypothesis – i.e.: is the differential control that cells exert among their compartments over external factors responsible for different degrees of protein constraint? This hypothesis predicts (i) distinctive dN/dS (i.e. abundance-controlled evolutionary rates) among proteins allocated in different subcellular compartments (as shown in [37]), (ii) a distinctive by-ortholog sequence diversity in particular compartments, but that is found regardless of taxonomic trends, (iii) that the latter phenomena is expected only as a response of the differential exposure to some environmental variables like pH or conductivity.

To check whether our data are consistent with predictions (ii) and (iii), we downloaded information of subcellular location from UniProt using the KEGG orthology identifiers. Then we calculated the abundance-controlled sequence diversities from an overall regression – as we did in the previous section, but now segregating by compartment. The results are shown in Figure 5. As before, all compartments have a negative slope with pH, and the reverse is true for temperature. However, we now do grasp more information: (1) the between-compartment coefficients are widely more variable with pH than with temperature increases (Fig. 5a), (2) the secreted proteins are only significantly affected by pH (Fig. 5a), (3) the virus/spore compartments are less affected by both (Fig. 5a), (4) none of the latter is due to the characteristic sequence diversity of a compartment (Fig. S4), (5) there is heterogeneity among compartments in both pH and temperature coefficients (Fig. 5b), and (6) we observe significant pairwise differences among compartments only in pH-derived coefficients (Fig. 5b). We thus find these results mostly compatible with the confined evolution hypothesis: points 1–4 support (ii), while (iii) is supported by points 2, 3 and 6.

These results validate, for example, that periplasmic sequence diversity is more affected by the external pH oscillations (it has more variability than cytoplasm), but that the negative influence of pH on sequence diversity appears diminished when comparing it to that of the cytoplasm. This is in strong agreement with the hypothesis. However, we found heterogeneity in residual diversity with temperature (5). At first sight, this point is contradictory with prediction (iii) and will be revisited in more detail in the discussion.

Many abundant functions are absent in warmer environments, but overrepresented in habitats with basic pH

We showed strong evidence for temperature increasing abundance-controlled sequence diversity and pH diminishing it. But, why is this so? We expected a reduction in sequence diversity with temperature, but we observed the opposite. We wondered if this could be explained by a differential loss of whole functions, instead of random gene loss. This is to say, temperature would increase sequence diversity per function, increasing gene redundancy, but it would also prevent biochemical functions from being exceedingly abundant. We analysed both halves of the abundance distribution: the top 50% and the bottom 50% abundant KEGGs (Fig. 6).

In regard to temperature, we found that the more abundant KEGG identifiers were significantly depleted with temperature (p = 0.0003, -0.081 Spearman's ρ). This analysis is performed with temperature as a continous variable (Fig. 6b). We also performed a Monte Carlo test (10,000 extractions) to visualize how enriched in negative correlations with temperature are the more abundant KEGGs (observed mean ρ (Top 50%): -0.0814; mean of random draws: -0.0645 ± 0.0034; empirical p-value (2-sided): 0.0002; Fig. S5a). But does this mean that the more abundant KEGGs are completely lost in warmer temperatures? To give answer to this question, we performed a regression between the summed abundance of a KEGG and the average temperature of the samples where its abundance was found to be 0 (KEGG identifier is lost in those samples). We found that the slope was positive and significant (slope = 0.45; p = 0.0125; Fig. S5b). As total abundance of a KEGG increases, it makes more probable for that KEGG to become "extinct" at higher temperatures. In absolute values, sequence diver-

sity is of lesser magnitude in hotter springs (Fig. 2a), but this is only caused by the fact that low-abundance KEGGs are preferred (Figs. 6a, S5) – corresponding high diversities are not achieved, as anticipated in Figure 2b.

When interrogating the data for increasing pH values, we find the opposite behavior: the more abundant functions are best represented in basic springs ($p < 0.0001$; 0.03 Spearman's ρ Fig. 6b). Again, Monte Carlo tests are in agreement with the latter (observed mean ρ (Top 50%): 0.0217; mean of random draws: 0.0072 ± 0.0034; empirical p-value (2-sided): <0.000; Fig. S6a). We performed the regression to ascertain in which pH a KEGG is lost as its total abundance increases. This time we found that the slope was slightly negative (slope = -0.096; p = 0.02; Fig. S6b). As abundance increases, the KEGG "extinction" probability for higher pH values is decreasing. This suggest that thriving at more basic pH values may require certain kinds of functions to be overrepresented at a metagenomic level, despite the fact that these functions are losing sequence diversity (Figs. 3 & 4). Surprisingly, this represents an opposite response to that induced by temperature. This is especially meaningful given that pH is not correlated with temperature in our dataset (Fig. S3).

To check whether correlation data differ among variables, we directly compared correlations with pH and temperature of the more abundant KEGGs (top 10%). We confirmed that they were significantly different (mean ρ (Temperature): -0.08 ± 0.15; mean ρ (pH): 0.02 ± 0.14; t = –28.14, p = 0.000; Fig. S7).

Proteins are shorter in warmer and basic environments

We contend that the above analyses disclosed the effects of pH and temperature on total sequence diversity well enough. But, if these were to be true, should we not expect differences compatible with these conclusions, such as discrepancies in protein length? Indeed, the protein length distribution is considered mostly uniform across the tree of life [38]. As our metagenomes include proteins from astranged branches of the tree of life and of similar taxnomic diversity, our protein length distributions should remain fairly uniform across pH and temperature categories, if they do not have a true effect.

When interrogating our data, the results showed clear trends (Fig. 7). Significantly shorter protein lengths corresponded only to the warmest temperature category (average residues: 45-50°C = 164.61; 54-55°C = 164.79; 56-62°C = 148.66; pairwise t-test p-values indicated in Fig. 7a). Likewise, the fraction of Long Proteins (LP; $> 1,000$ aa) was significantly depleted as temperature increased (Fig. 7b). Interestingly, while a more basic pH constrained protein length (average residues: 7.2-7.5 = 177.91; 7.6-8.0 = 160.37 8.1-9.3 = 157.16; Fig. 7c), LP enrichment was found in both the more acidic (neutral) and basic ranges.

Individual taxa partially exhibit the general pattern

The observed effects of temperature and pH on sequence diversity held up to all our falsification efforts. Thus, we were curious to see whether the general trend would also occur in the main taxa inhabiting these springs. Specifically, we analysed the most abundant genera in Chloroflexota (*Roseiflexus* spp., total unique KEGGs: 1460) and Cyanobacteria (*Fischerella* spp., total unique KEGGs: 1472). They are common inhabitants of hot springs, but with opposite abundance trends: while *Roseiflexus* spp. increases in abundance with temperature, *Fischerella* spp. abundance is expected to diminish in hotter springs due to photosynthetic inhibition [39–41]. Considering the number of ORFs per temperature and pH categories, however, this trend was masked – temperature drove the fraction of total ORFs from 3.0% to 1.3% in *Fischerella* spp., as expected, while *Roseiflexus* spp. was less constrained (from 3.0% to 2.1%; Fig. S8). In contrast, pH seemed to have a random association with the fraction of total ORFs (Fig. S8).

But how do these results fit with the fact that the by-function sequence diversity should increase with temperature? The straightfoward explanation may be the same as before: the more abundant KEGGs are lost preferentially as temperature increases. Indeed, the regressions retained the same form – slopes were steeper with temperature category in both genera, regardless of their differential representation across temperatures (Fig. S9). In both genera we observed that the more abundant KEGGs (those at the extreme of the X-axis) were in the colder range, followed by the intermediate range. The hottest category only barely occupied the abundance axis, not reaching even a third of it. Abundant KEGGs were indeed absent with temperature

in both genera, however, we found major differences among the two taxa with respect of the abundance (Fig. S10). For example, *Roseiflexus* spp. was not significantly depleted of abundant KEGGs in the 54-55ºC category, while *Fischerella* sp. was strongly depleted of them.

With regard to pH, we observed a different trend than in the overall analyses (Fig. S11). Here, basic pH appears to accounted for more sequence diversity, with the exception of that the most abundant KEGGs were not clearly influenced by pH (Fig. S10).

Discussion

Here we examined how the general relationship of sequence diversity and metagenomic abundance is broadly influenced by temperature and pH in organisms from hot springs. Our results were mostly consistent with the interpretation that pH and temperature are major actors in determining the pace of functional evolution, but acknowledging that they act in a different direction. Within the range of values inspected, pH diminishes, while temperature fosters sequence diversity over metastable populations. These effects are found when normalizing by the characteristic metagenomic abundance of each pH/temperature category; i.e. by interrogating the residual diversities not explained by abundance. On the other hand, the more extreme temperatures consistently underpredicted the number of sequences found, but this happens to be explained by a preferential loss of abundant KEGGs in the warmer categories. In addition, here we have shown that the effect of pH on by-ortholog sequence diversity depends more than temperature on the subcellular localization of the protein, and that pairwise comparisons of the residual sequence diversities among compartments are only significant with increasing pH values. Since microorganisms cannot control temperature among their subcellular compartments, but they do control their pH, this supports the confined evolution hypothesis. However, we found between-compartment heterogeneity among temperature coefficients (Fig. 5b), but our model was not anticipating that result. It may simply imply that that this is a naive model, not accurate enough to fit all the observations. For example, we assumed that temperature is not different among locations because it is not controlled, but an alternate model could be that cell's size is an important factor: the compartments from species with strong differential cell volume

might experience external temperature oscillations distinctly.

As mentioned, we observed higher sequence diversity in warmer springs. This appears to contradict the interpretation that thermophilic organisms are more prone to genomic reduction via gene loss [33–35]. However, they do not argue against a reduced genome size as a consequence of thermal adaptation. In fact, this was indirectly predicted by the biosynthetic model, where higher body temperatures are the cause for an –admittedly– adaptive reduction in genome size, as depicted: "*In other words, if there is a selective advantage maximizing the rate of biosynthesis, then genome size is expected to decrease with increasing temperature*" [42]. To reconcile this with our results, we propose that genomic reduction would then be achieved either by a loss of highly abundant biochemical functions –and not of random genes–, or by a temperature-driven gene shortening effect, or both (we summarize this other hypothesis in a scheme in Figure S12). This interpretation is supported by the observation of *Fischerella thermalis* ecotypes with an increase of divergence among orthologous genes as temperature increases, where exclusive genes appeared at 66ºC [39]. Furthermore, the observation of increased sequence diversity cannot be explained by increases of taxonomic diversity with temperature, as no significant differences were found (Fig. S13).

But why would temperature induce such a generalised response? Our observation of consistently shorter proteins (Fig. 7a), and depletion of very long proteins (Fig. 7b) with increasing temperatures could be simply explained with the fitness cost of thermophilic adaptation in evolutionary timescales [42, 43]. However, our results partially argue against the simplistic fitness explanation: (i) when considering proteins in the 54-55ºC category, they were not significantly shorter than those from 45-50ºC (Fig. 7a), and (ii) when looking to the characteristic sequence diversity found in each temperature category, residuals were higher with increases in temperature (Figs. 2 & 4). If the fitness explanation were sufficient, an unequivocal positive relationship between temperature and loss of sequence diversity would have been found, because the optimal number of genes or variants per biochemical function should be minimized with temperature. But the highest median and mean residuals were found within the 54-55°C samples, followed by those in the 56-62°C category. This effect is better revealed by the OLS, where we analyse temperature as a continous (uncategorized) variable affecting residuals (Fig. 4). In addition, the

sequence diversity is thus growing faster with abundance in the 56-62°C category, as the slope is greater, but there is a lack of abundant KEGGs: X-axis (abundance) is not as occupied by other categories, preventing corresponding levels in the Y-axis (sequence diversity).

Interestingly, while the effects of the more extreme temperatures on sequence diversity are fairly positive and constant across representatives of the tree of life, pH showed an overall broad relationship, although a domain-dependent variability on polarity and significance. However, it remains less clear how our results should be interpreted with regard to the generality of protein adaptation to different pH ranges. What can be likely said about the effects of environmental pH is that they were (i) important (they had more effect size on sequence diversity than temperature, even when normalized by the narrower range of predictor values; Fig. 4), (ii) opposite to those of temperature regarding sequence diversity (Fig. 4), (iii) contrary to those of temperature regarding preferential depletion of abundant KEGGs (Figs. 6, S5, & S6), (iv) dependent on the domain (Fig. 4). The latter preliminarly supports the intuition that the compartmentalisation strategy could indeed be a cause for the differential responses to these variables. Indeed, we did observe variable responses of sequence diversity from functions allocated in the different compartments, and these were more acute for pH increases (Fig. 5a–b). However, if the confined evolution hypothesis were to be true, evolutionary rates, not only the by-function gene diversity, should be different among compartments. Evidence consistent with this prediction is provided for *E. coli* and *S. cerevisiae* in [37]. Contrary to the extended complexity hypothesis [44], distinctive evolutionary rates and/or sequence diversities among localisations would be not explained by the number of functional connections of a protein, but would respond to the differential ability of cells to buffer environmental stresses from one compartment to another.

Further discussion regarding particularities of taxonomical domains is available in the Supplementary Information.

In this study, we used by-function sequence diversity as a proxy for many different evolutionary events (eg. gene duplication, mutation, etc.). However, a limitation of this work is that we use KEGG orthology as an indicator of true unique molecular functions. MAGs have been shown to yield functionally coherent reconstructions, comparable to single-amplified genomes, thereby justifying their use in KEGG-based evolutionary inference [45]. However,

these metagenomes mostly encompassed bacterial and viral ORFs (Fig. S1), and possible biases in this regard could arise from the use of Prodigal for protein prediction. So our conclusions on Archaea and Eukarya should be handled as preliminary. Another kind of considerations is that the sequencing depth of samples might be different, or that our approach considered only annotated ORFs (avoiding broader measures of sequence diversity). Nevertheless, (i) we used TPM averages for each condition, (ii) the non-pareil analysis exhibited sequencing coverage (C values) above the 95% threshold (range: 96.34–99.52%) for all but one sample (93.21%), indicating that the observed patterns likely reflect the true community composition. Finally, as our results only encompassed hydrothermal organisms, the more general conclusions are yet to be addressed.

Methods

Sampling & sequencing

Samples were obtained from 13 hot spring manifestations of the El Tatio geothermal field in January 2020, which their metadata and sequences were reported in previous studies [46, 47] and are available in Table S1. Briefly, triplicate samples of approximately 2 ml were collected with a punch from each microbial mat, kept in cryogenic vials containing RNAlater (Thermo Fisher Scientific, Vilnius, Lithuania), and then stored at -80ºC until DNA extraction. DNA extractions were performed according an optimized protocol for microbial mat samples from hot springs [48]. An equimolar amount of DNA (400 ng) from each replicate was pooled and sequenced at the Roy J. Carver Biotechnology Center (University of Illinois at Urbana-Champaign, IL, USA with the Illumina NovaSeq 6000 platform (S1 flowcell, 2×150 bp).

Metagenomic analysis

The resulting fastq files for metagenomes were filtered for quality using Cutadapt [49] with the parameters: paired-end mode, a perfect match of at least 10 bp (-O 10) against the standard Illumina adapters, hard trimming of the first 10 to 15 bases of the 5' (-u 10/15) and 3' (-U 10/15) ends depending on quality and base distributions, removal of N-containing reads (–trim-n 1)

and retention of only sequences longer than 50 bp (-m 50). De novo assemblies of trimmed reads were generated using SPAdes v3.10.1 [50] with the –meta setup. Sequencing effort was assessed with the Nonpareil v3.401 software [51] with the -T alignment and -f fasta options.

The CDS were predicted with the Prodigal V2.6.3 software [52] with the -p meta option. Therefore, at the wider level, all the contigs were classified with geNomad v1.8.1 (using the end-to-end option; [53]) in viral and plasmid contigs, and then the unclassified >1000 pb (–minsize 1000) contigs were assigned to prokaryotes and eukaryotes with whokaryote software [54], leaving all other contigs as unclassified. Then, the taxonomic affiliation of prokaryotic and plasmid contigs and proteins was obtained with the refineM tool [55], using the entire assemblies as input for the modules scaffold_stats and the assignment with the module taxon_profile using also the previously predicted CDS and a custom protein/taxonomic diamond v2.1.9.163 [56] database with all the proteins from the representative genomes from the GTDB R220 (available at https://data.ace.uq.edu.au/public/gtdb/data/releases/release220/220.0/genomic_files_reps/) and formatted as required with a custom python script.

To functionally annotate the predicted ORFs, DRAM v1.5.0 software [57] was used with the annotate_genes and distill options and retrieving the results from the KOfam [58] database. Then, reads were aligned to the nucleotide sequences of the predicted ORFs with the coverM v0.6.1 (https://github.com/wwood/CoverM; [59]) with the following parameters: coverm contig -p bwa-mem –methods covered_bases mean rpkm tpm length count covered_bases –min-read-percent-identity 95 –min-read-aligned-percent 50 and –min-covered-fraction 0 against their respective metagenomes. The functional, taxonomic, protein length, and abundance information was parsed to each contig and each ORF with custom bash and python scripts.

Modelling of sequence diversity

For each sample we computed the total transcript abundance (TPM) and the number of open reading frames (ORFs). In total, the dataset comprised 17 samples with a cumulative TPM of 17,000,193 and 1,381,486 ORFs.

To assess how environmental variables influence sequence diversity beyond what is expected by metagenomic abundance, we implemented a two-stage linear model. This approach attempts

to separate the influence of abundance on sequence diversity from that of the two environmental variables tested, allowing us to estimate whether conditions (pH and temperature) are associated with consistent deviations in the number of sequences per KEGG identifier. We thus normalised TPM within categories per sample (thus, we display TPM averages). In the first stage, we fitted a linear model to predict the number of observed sequences per KEGG identifier from the summed abundance associated with that ortholog across samples. We used the base-10 logarithm of the total (`sum_TPM`) as a continuous predictor:

$$\texttt{sequence diversity} = \beta_0 + \beta_1 \log_{10}(\texttt{sum_TPM}) + \varepsilon$$

The residuals from this model represent variation in by-function sequence diversity not explained by abundance (and thus to be explained by the other variables). Positive residuals indicate more frequent ORF retrieval than expected, while negative residuals indicate under-representation. In the second stage, we modelled these residuals as a function of environmental conditions:

$$\texttt{residual } d = \beta_0 + \beta_1 \cdot \texttt{pH} + \beta_2 \cdot \texttt{Temperature} + \varepsilon$$

Both pH and temperature were treated as continuous variables. This model estimates the direction and magnitude of environmental effects on residual sequence diversity, effectively identifying whether gene functions tend to be over- or underrepresented in different environments (by the expected abundance).

This approach represents a specific application of the methodology developed in [60]. However, we restricted the analysis to genes annotated with KEGG orthology identifiers. Then, pH and temperature values were extracted from sample metadata. Observations missing pH, temperature, TPM, or KEGG annotation were excluded. Each KEGG ortholog was considered a single observation in the model, with summarised values for TPM and environmental metadata derived from its full set of occurrences across samples.

Model fitting was performed using ordinary least squares (OLS) implemented in the `statsmodels` library in Python. Coefficient estimates represent the average change in residual sequence diversity per unit change in the environmental variable, holding abundance constant.

Standard errors and t-statistics were used to assess the statistical significance of each term. We visualised the regression coefficients and their 95% confidence intervals to interpret both the direction and strength of environmental effects.

Temperature, pH, and subcellular categorization

As described, TPM values were normalized by the number of unique samples within each environmental category (temperature or pH). Environmental variables were discretized into arbitrary intervals: 7.2–7.5, 7.6–8.0, and 8.1–9.3 for pH; and 45–50°C, 54–55°C, and 56–62°C for temperature. For display, ranges were rounded to the nearest tenth (e.g. 9.3 corresponds to 9.27). These bins were selected to reflect with clarity the trends in sequence diversity variation while ensuring an approximately even distribution of metagenomes across categories. Subcellular localization was assigned by querying UniProt with cleaned protein descriptions from KEGG identifiers. For each entry, the first UniProt match was retrieved and its annotated subcellular compartment extracted from the JSON API. We collapsed localisations for clarity and statistical resolution. "Membrane" corresponds to any membrane that is not "outer". We removed ambiguous fetches.

Further methods available in the Supplementary Information.

Data availability

The custom code to generate the figures and analyses (with specific methods) are available at 10.5281/zenodo.15701078. The metagenomic raw reads can be found under the NCBI BioProject accession PRJNA858297 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA858297/).

Acknowledgements

We want to thank Alvar Lavin, Pablo Yubero, Juan F. Poyatos, Sofia Radrizzani, and Laurence D. Hurst for a variety of inputs helpful to this work. We also thank the organizing comitee and members of the National Network of Extremophiles (RedEx) for discussing the hypotheses presented in this manuscript.

Work supported by Ph.D. grant PRE2020-096130 from the Spanish Ministerio de Ciencia e Innovación and the European Social Fund, and ANID-Fondecyt 1230217, Millennium Institute Center for Genome Regulation (ANID – Millennium Science Initiative Program – ICN2021_044).

Authors' Contributions

J.R.S.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Validation, Resources, Data Curation, Visualization, Writing - Original Draft, Writing – Review & Editing, Supervision; J.A.: Methodology, Validation, Resources, Data Curation, Writing – Review & Editing, Visualization; B.D.: Validation, Resources, Writing – Review & Editing; J.T.: Resources, Project administration, Funding acquisition, Writing – Review & Editing; C.P.A.: Conceptualization, Investigation, Validation, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

References

[1] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2014.

[2] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, 2017.

[3] Eva Nỳvltová, Courtney W Stairs, Ivan Hrdỳ, Jakub Rídl, Jan Mach, Jan Pačes, Andrew J Roger, and Jan Tachezy. Lateral gene transfer and gene duplication played a key role in the evolution of mastigamoeba balamuthi hydrogenosomes. *Molecular biology and evolution*, 32(4):1039–1055, 2015.

[4] Simon K Chan, Michael Hsing, Fereydoun Hormozdiari, and Artem Cherkasov. Relationship between insertion/deletion (indel) frequency of proteins and essentiality. *BMC bioinformatics*, 8:1–13, 2007.

[5] Vladislav Grishkevich and Itai Yanai. Gene length and expression level shape genomic novelties. *Genome research*, 24(9):1497–1503, 2014.

[6] Filip Husnik and John P McCutcheon. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2):67–79, 2018.

[7] Eugene V Koonin. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5:F1000–Faculty, 2016.

[8] Victor Kunin and Christos A Ouzounis. The balance of driving forces during genome evolution in prokaryotes. *Genome Research*, 13(7):1589–1594, 2003.

[9] James O McInerney, Alan McNally, and Mary J O'connell. Why prokaryotes have pangenomes. *Nature microbiology*, 2(4):1–5, 2017.

[10] Zackary I Johnson and Sallie W Chisholm. Properties of overlapping genes are conserved across microbial genomes. *Genome research*, 14(11):2268–2272, 2004.

[11] En-Ze Hu, Xin-Ran Lan, Zhi-Ling Liu, Jie Gao, and Deng-Ke Niu. A positive correlation between gc content and growth temperature in prokaryotes. *BMC genomics*, 23(1):110, 2022.

[12] Motoo Kimura. The neutral theory of molecular evolution: a review of recent evidence. *The Japanese Journal of Genetics*, 66(4):367–386, 1991.

[13] Tomoko Ohta. The neutral theory is dead. the current significance and standing of neutral and nearly neutral theories. *BioEssays*, 18(8):673–677, 1996.

[14] Martin Kreitman. The neutral theory is dead. long live the neutral theory. *Bioessays*, 18(8):678–683, 1996.

[15] Chris P Ponting. Biological function in the twilight zone of sequence conservation. *BMC biology*, 15(1):71, 2017.

[16] Louis-Marie Bobay and Howard Ochman. Factors driving effective population size and pan-genome evolution in bacteria. *BMC evolutionary biology*, 18:1–12, 2018.

[17] Douglas J Futuyma. Evolutionary constraint and ecological consequences. *Evolution*, 64(7):1865–1884, 2010.

[18] Konstantin B Zeldovich, Peiqiu Chen, and Eugene I Shakhnovich. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences*, 104(41):16152–16157, 2007.

[19] Sergey V Venev and Konstantin B Zeldovich. Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. *Molecular Biology and Evolution*, 35(1):211–224, 2018.

[20] Anindya S Panja, Smarajit Maiti, and Bidyut Bandyopadhyay. Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges. *Scientific reports*, 10(1):1–9, 2020.

[21] Qinyi Zhao. Molecular and thermodynamic mechanisms for protein adaptation. *European Biophysics Journal*, 51(7-8):519–534, 2022.

[22] Jeffrey M Dick and Delong Meng. Community-and genome-based evidence for a shaping influence of redox potential on bacterial protein evolution. *Msystems*, 8(3):e00014–23, 2023.

[23] Daniel E Deatherage, Jamie L Kepner, Albert F Bennett, Richard E Lenski, and Jeffrey E Barrick. Specificity of genome evolution in experimental populations of escherichia coli evolved at different temperatures. *Proceedings of the National Academy of Sciences*, 114(10):E1904–E1912, 2017.

[24] Andrea González-González, Tiffany N Batarseh, Alejandra Rodríguez-Verdugo, and Brandon S Gaut. Patterns of fitness and gene expression epistasis generated by beneficial mu-

tations in the rho and rpob genes of escherichia coli during high-temperature adaptation. *Molecular Biology and Evolution*, 41(9):msae187, 2024.

[25] Ian R Booth. Regulation of cytoplasmic ph in bacteria. *Microbiological reviews*, 49(4):359–378, 1985.

[26] Dominique Missiakas and Satish Raina. Protein folding in the bacterial periplasm. *Journal of bacteriology*, 179(8):2465–2471, 1997.

[27] Hiroshi Kobayashi. A proton-translocating atpase regulates ph of the bacterial cytoplasm. *Journal of biological chemistry*, 260(1):72–76, 1985.

[28] Young-Mi Go and Dean P Jones. Redox compartmentalization in eukaryotic cells. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1780(11):1273–1290, 2008.

[29] Joseph R Casey, Sergio Grinstein, and John Orlowski. Sensors and regulators of intracellular ph. *Nature reviews Molecular cell biology*, 11(1):50–61, 2010.

[30] Joan L Slonczewski, Makoto Fujisawa, Mark Dopson, and Terry A Krulwich. Cytoplasmic ph measurement and homeostasis in bacteria and archaea. *Advances in microbial physiology*, 55:1–317, 2009.

[31] Vernon R Phoenix, Philip C Bennett, AS Engel, SW Tyler, and FG Ferris. Chilean high-altitude hot-spring sinters: a model system for uv screening mechanisms by early precambrian cyanobacteria. *Geobiology*, 4(1):15–28, 2006.

[32] Constanza Nicolau, Martin Reich, and Bridget Lynne. Physico-chemical and environmental controls on siliceous sinter formation at the high-altitude el tatio geothermal field, chile. *Journal of Volcanology and Geothermal Research*, 282:60–76, 2014.

[33] Hannah B Rappaport and Angela M Oliverio. Extreme environments offer an unprecedented opportunity to understand microbial eukaryotic ecology, evolution, and genome biology. *Nature Communications*, 14(1):4959, 2023.

[34] Andrei S Steindorff, Maria Victoria Aguilar-Pontes, Aaron J Robinson, Bill Andreopoulos, Kurt LaButti, Alan Kuo, Stephen Mondo, Robert Riley, Robert Otillar, Sajeet Haridas,

et al. Comparative genomic analysis of thermophilic fungi reveals convergent evolutionary adaptations and gene losses. *Communications Biology*, 7(1):1124, 2024.

[35] HB Rappaport and Angela M Oliverio. Lessons from extremophiles: functional adaptations and genomic innovations across the eukaryotic tree of life. *Genome biology and evolution*, 16(8):evae160, 2024.

[36] Pongpan Laksanalamai, Timothy A Whitehead, and Frank T Robb. Minimal protein-folding systems in hyperthermophilic archaea. *Nature Reviews Microbiology*, 2(4):315–324, 2004.

[37] Juan Rivas-Santisteban, Pablo Yubero, and Laurence D Hurst. Cell compartment is a predictor of protein rate of evolution, but not in the manner expected: evidence against the extended complexity hypothesis. *Genome Biology and Evolution*, 17(7):evaf126, 2025.

[38] Yannis Nevers, Natasha M Glover, Christophe Dessimoz, and Odile Lecompte. Protein length distribution is remarkably uniform across the tree of life. *Genome Biology*, 24(1):135, 2023.

[39] Jaime Alcorta, Sebastián Espinoza, Tomeu Viver, María E Alcamán-Arias, Nicole Trefault, Ramon Rosselló-Móra, and Beatriz Díez. Temperature modulates fischerella thermalis ecotypes in porcelana hot spring. *Systematic and applied microbiology*, 41(6):531–543, 2018.

[40] María E Alcamán-Arias, Carlos Pedrós-Alió, Javier Tamames, Camila Fernández, Danilo Pérez-Pantoja, Mónica Vásquez, and Beatriz Díez. Diurnal changes in active carbon and nitrogen pathways along the temperature gradient in porcelana hot spring microbial mat. *Frontiers in microbiology*, 9:2353, 2018.

[41] Annastacia C Bennett, Senthil K Murugapiran, and Trinity L Hamilton. Temperature impacts community structure and function of phototrophic chloroflexi and cyanobacteria in two alkaline hot springs in yellowstone national park. *Environmental microbiology reports*, 12(5):503–513, 2020.

[42] Xuhua Xia. Body temperature, rate of biosynthesis, and evolution of genome size. *Molecular Biology and Evolution*, 12(5):834–842, 1995.

[43] Andrew Clarke. Costs and consequences of evolutionary temperature adaptation. *Trends in Ecology & Evolution*, 18(11):573–581, 2003.

[44] Stéphane Aris-Brosou. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Molecular biology and evolution*, 22(2):200–209, 2005.

[45] Johannes Alneberg, Christofer MG Karlsson, Anna-Maria Divne, Claudia Bergin, Felix Homa, Markus V Lindh, Luisa W Hugerth, Thijs JG Ettema, Stefan Bertilsson, Anders F Andersson, et al. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome*, 6:1–14, 2018.

[46] Carla Barbosa, Javier Tamayo-Leiva, Jaime Alcorta, Oscar Salgado, Linda Daniele, Diego Morata, and Beatríz Díez. Effects of hydrogeochemistry on the microbial ecology of terrestrial hot springs. *Microbiology Spectrum*, 11(5):e00249–23, 2023.

[47] Oscar Salgado, Sergio Guajardo-Leiva, Ana Moya-Beltrán, Carla Barbosa, Christina Ridley, Javier Tamayo-Leiva, Raquel Quatrini, Francisco JM Mojica, and Beatriz Díez. Global phylogenomic novelty of the cas1 gene from hot spring microbial communities. *Frontiers in Microbiology*, 13:1069452, 2022.

[48] Jaime Alcorta, Sebastián Espinoza, Tomeu Viver, María E. Alcamán-Arias, Nicole Trefault, Ramon Rosselló-Móra, and Beatriz Díez. Temperature modulates fischerella thermalis ecotypes in porcelana hot spring. *Systematic and Applied Microbiology*, 41(6):531–543, 2018.

[49] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[50] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[51] Luis M Rodriguez-R, Santosh Gunturu, James M Tiedje, James R Cole, and Konstantinos T Konstantinidis. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *MSystems*, 3(3):10–1128, 2018.

[52] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:1–11, 2010.

[53] Antonio Pedro Camargo, Simon Roux, Frederik Schulz, Michal Babinski, Yan Xu, Bin Hu, Patrick SG Chain, Stephen Nayfach, and Nikos C Kyrpides. Identification of mobile genetic elements with genomad. *Nature Biotechnology*, 42(8):1303–1312, 2024.

[54] Lotte JU Pronk and Marnix H Medema. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microbial Genomics*, 8(5):000823, 2022.

[55] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, 2(11):1533–1542, 2017.

[56] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368, 2021.

[57] Michael Shaffer, Mikayla A Borton, Bridget B McGivern, Ahmed A Zayed, Sabina Leanti La Rosa, Lindsey M Solden, Pengfei Liu, Adrienne B Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, et al. Dram for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic acids research*, 48(16):8883–8900, 2020.

[58] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[59] Samuel TN Aroney, Rhys JP Newell, Jakob N Nissen, Antonio Pedro Camargo, Gene W Tyson, and Ben J Woodcroft. Coverm: Read alignment statistics for metagenomics. *arXiv preprint arXiv:2501.11217*, 2025.

[60] Juan Rivas-Santisteban, Pablo Yubero, Semidán Robaina-Estévez, José M González, Javier Tamames, and Carlos Pedrós-Alió. Quantifying microbial guilds. *ISME communications*, 4(1):ycae042, 2024.

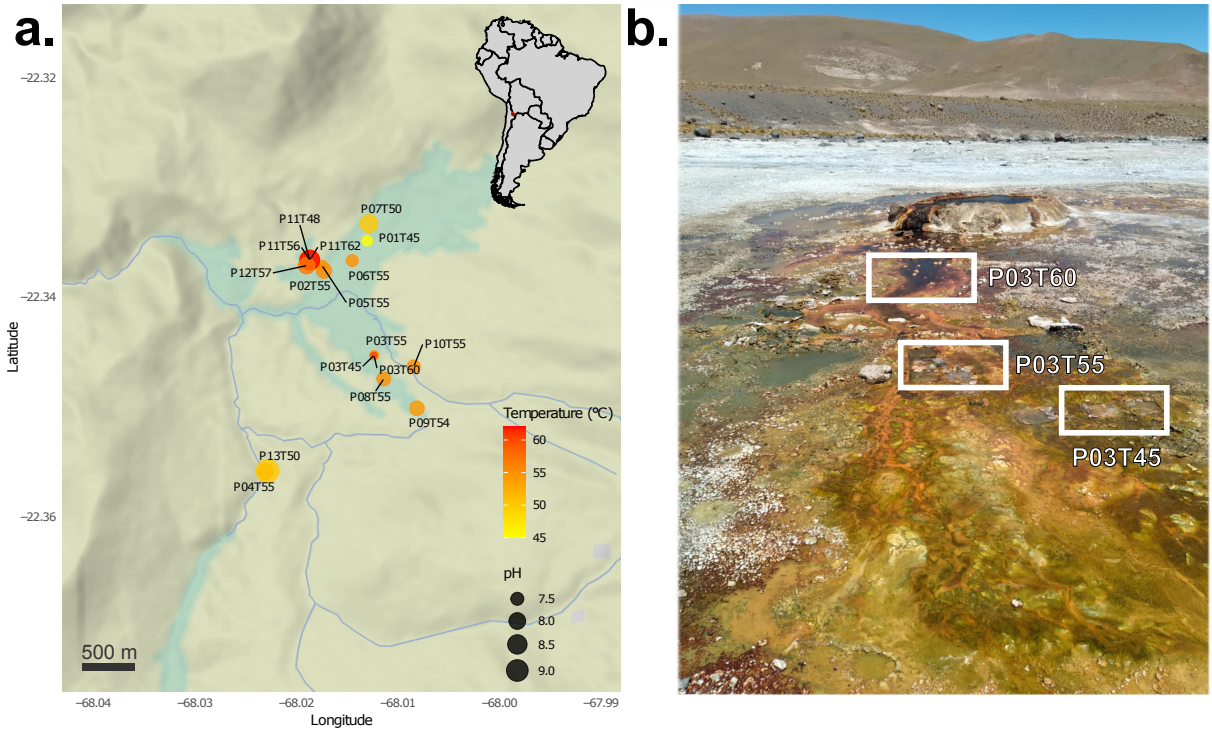

Figure 1: (a). Sampling spots in El Tatio geothermal system. (b). Example of site (P03) in which three different temperatures were sampled (in this case, 45ºC, 55ºC and 60ºC).

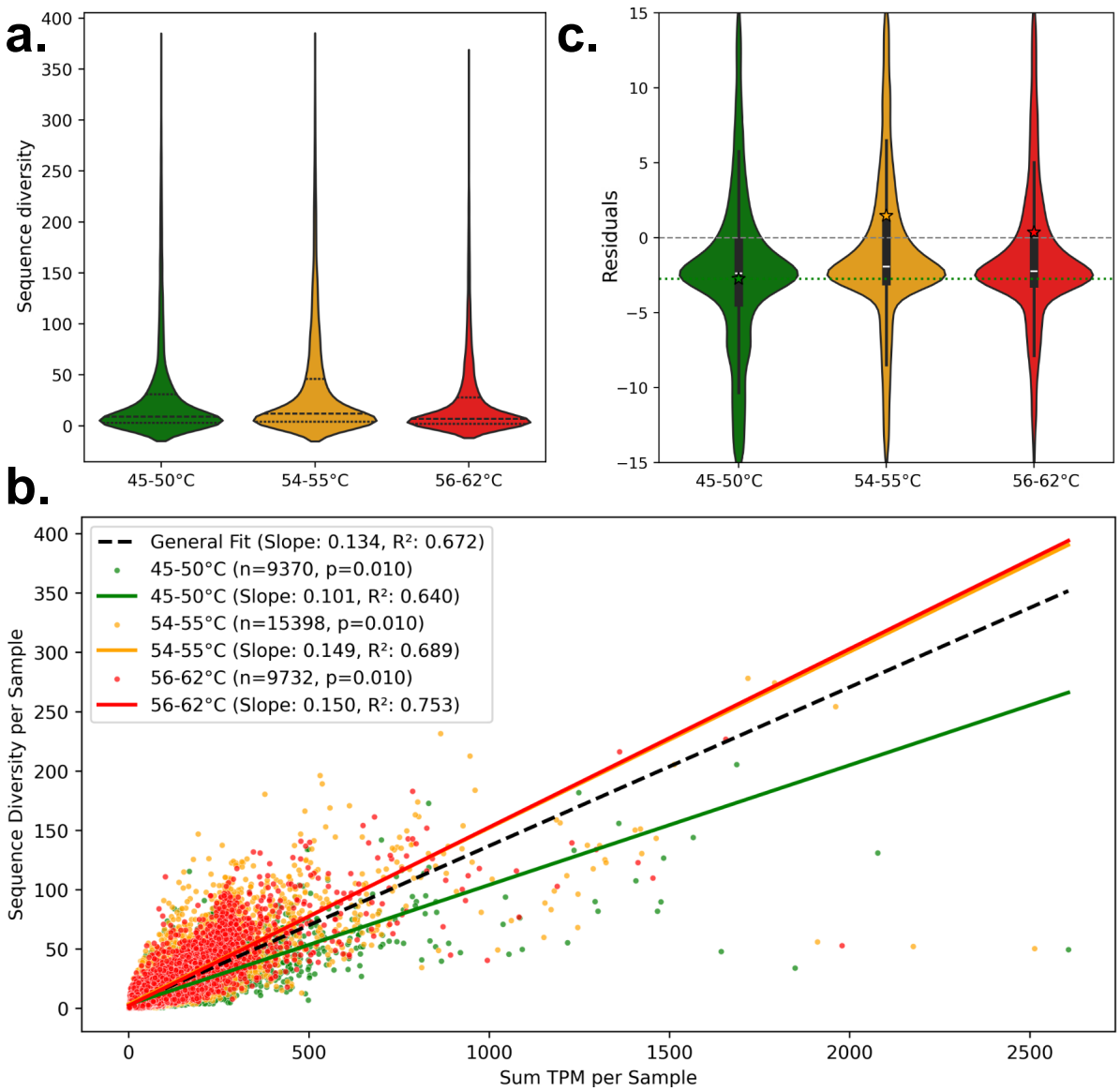

Figure 2: a. Distributions of sequence diversity (predicted bacterial ORFs for the same KEGG) are significantly different in all temperature categories tested (paired t-tests, $p < 0.0001$). *A priori*, the lowest diversity of sequences per function is found in the 56-62ºC category. b. General and specific linear regressions among temperature categories. The explanatory power of genomic abundance consistently increases with temperature. Every slope differs from the general regression (10,000 iterations, p-values indicated). We observe that sequence diversity grows faster with abundance, but warmer temperatures are depleted of abundant functions. c. After allowing for abundance, residual distributions are found to be significantly different across temperatures ($p < 0.0001$), with the 54-55ºC category containing the higher by-function diversity. Stars indicate means. Range limited to ±15 for visualization. The variance is gradually reduced with temperature (full range in Fig. S2), which explains the better linear fit within warmer metagenomes.

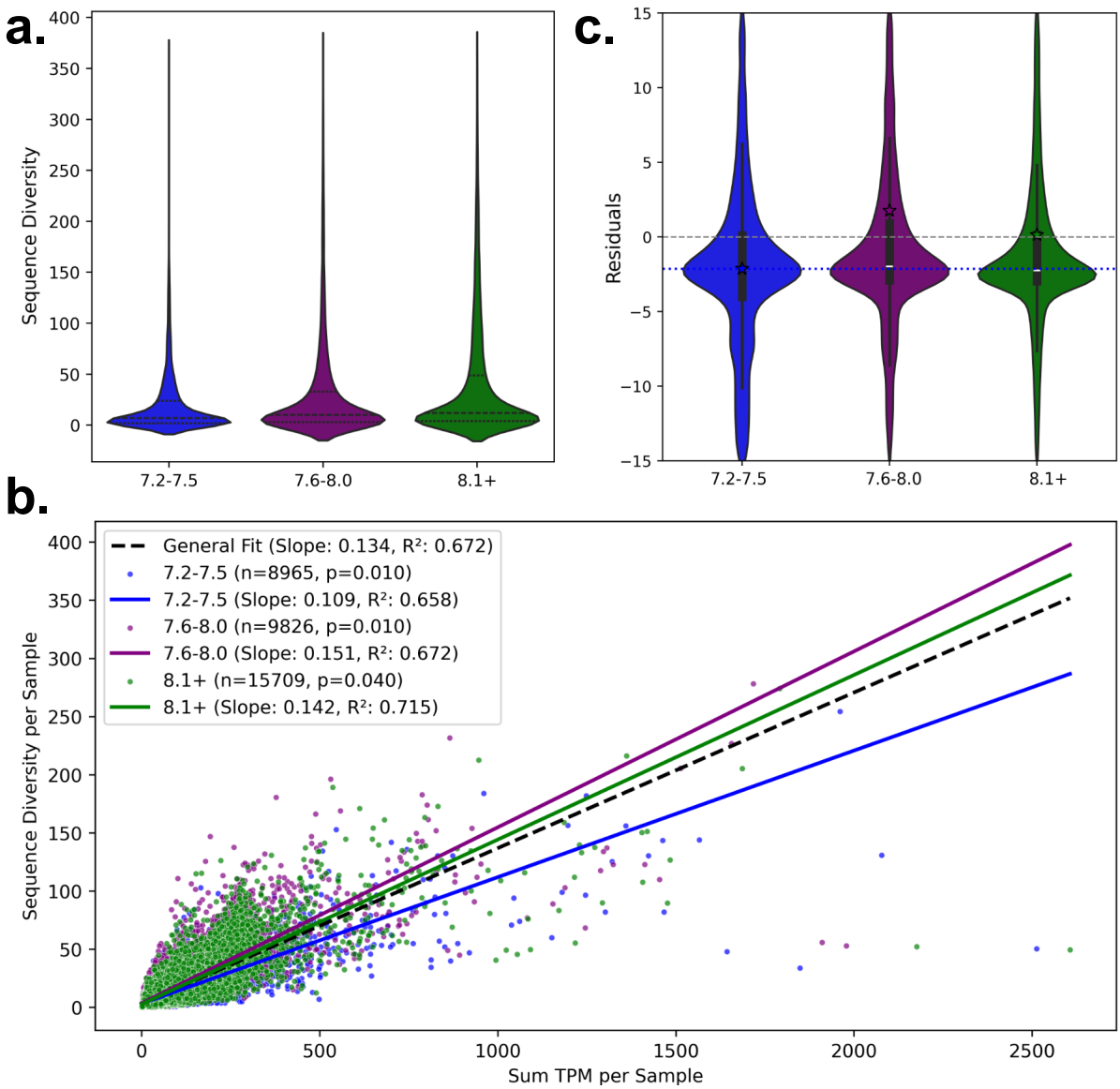

Figure 3: a. Distributions of gene counts are also different among pH categories (paired t-tests, $p < 0.0001$). b. General and specific linear regressions among temperature classes. In this case, the explanatory power of genomic abundance is barely increasing with pH. The 8.1-9.3 pH range slope is not differing from the general regression. c. Higher pH residual distribution is likewise significantly lower ($p < 0.0001$), although the higher pH range has a lower significance ($p = 0.04$). Stars indicate means. Range limited to ±15 for visualization. Unlike temperature, residuals show more variance as pH increases (full range in Fig. S2).

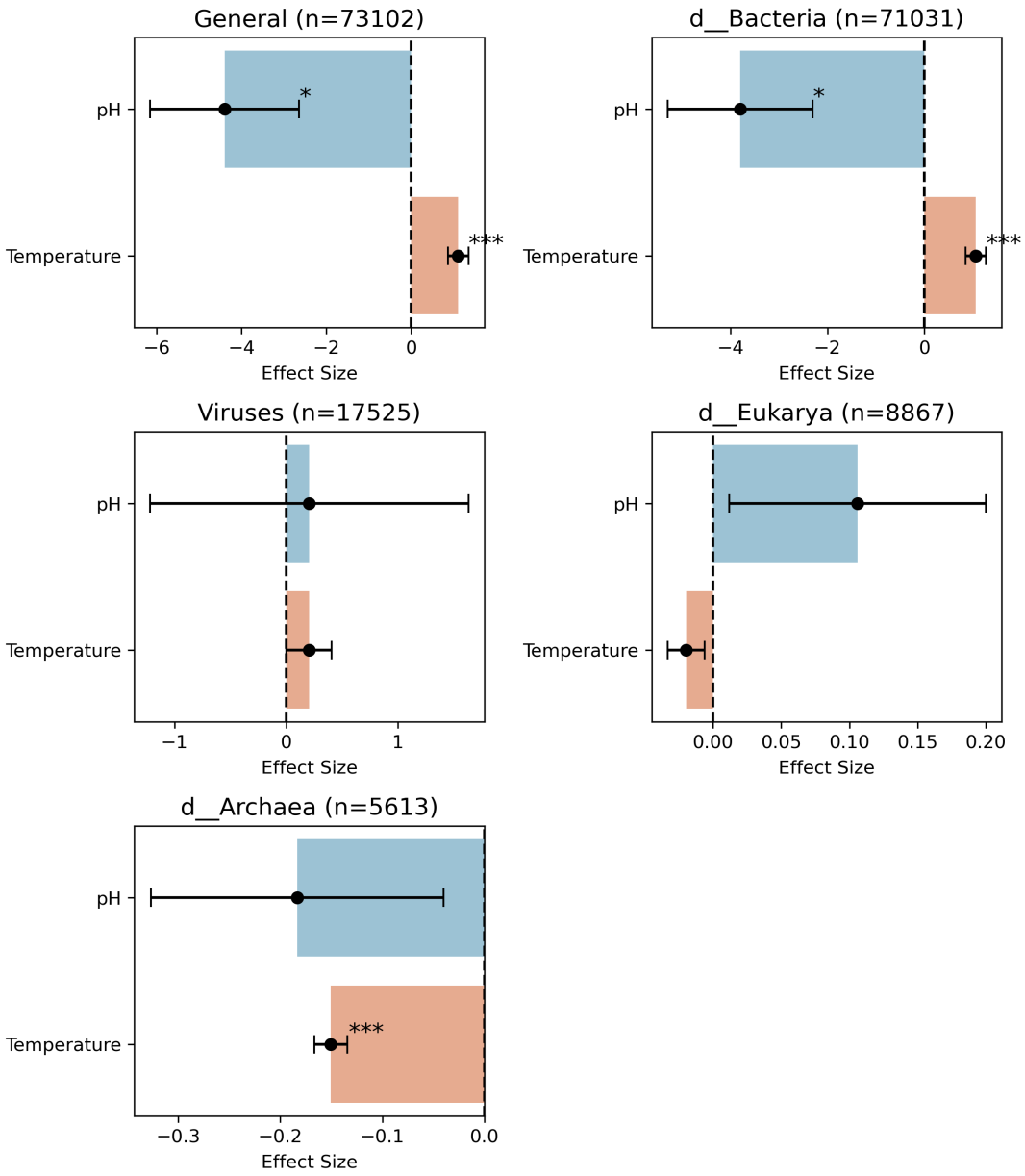

Figure 4: Environmental determinants of residual sequence diversity across taxonomic domains. Each panel depicts the estimated coefficients from linear regressions evaluating the effects of pH and temperature on the residuals obtained from a prior linear model: `sequence_diversity` ~ log(`sum_TPM`). This residual-based approach quantifies the extent to which environmental variables predict gene presence beyond what is attributable to metagenomic abundance alone. Positive coefficients indicate a greater-than-expected sequence diversity at higher values of the predictor variable. The number of unique by-condition KEGG identifiers analysed in each domain is shown in the corresponding panel title (e.g. bacterial ORFs are associated with almost every by-condition KEGG found in the dataset). Statistical significance is disclosed: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

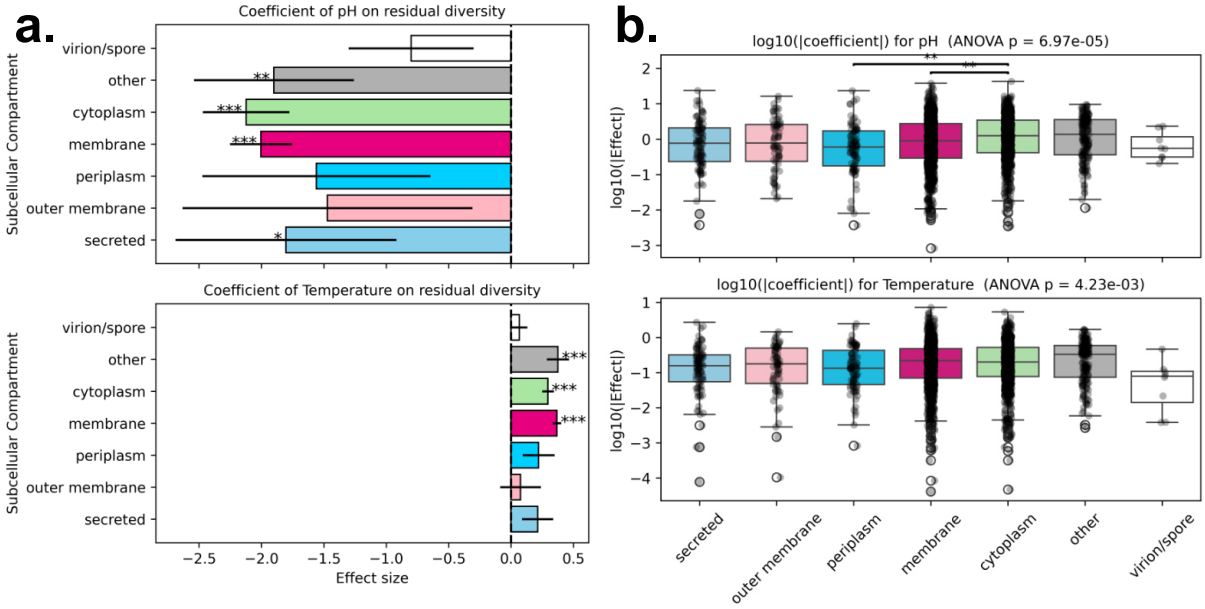

Figure 5: Environmental determinants have a variable influence among the between-compartment sequence diversity (abundance-controlled). a. Above, effects of pH. Below, of temperature. Secreted proteins are more significantly influenced by pH compared to temperature (functions there have less unique sequences). Likewise, variability on the coefficients is found to be of much higher magnitude with pH than with temperature. We still observe the universal trends: higher pH constrains while higher temperatures fosters sequence diversity; the latter with a more similar size effect among compartments than the former. b. Absolute log(coefficient) value distribution per compartment. Although significant heterogeneity is found with both variables, paired t-tests (bonferroni corrected) were only significant with increasing pH. The periplasmic average by-ortholog diversity is more positive and variable than the cytoplasmic. These results broadly support the confined evolution hypothesis.

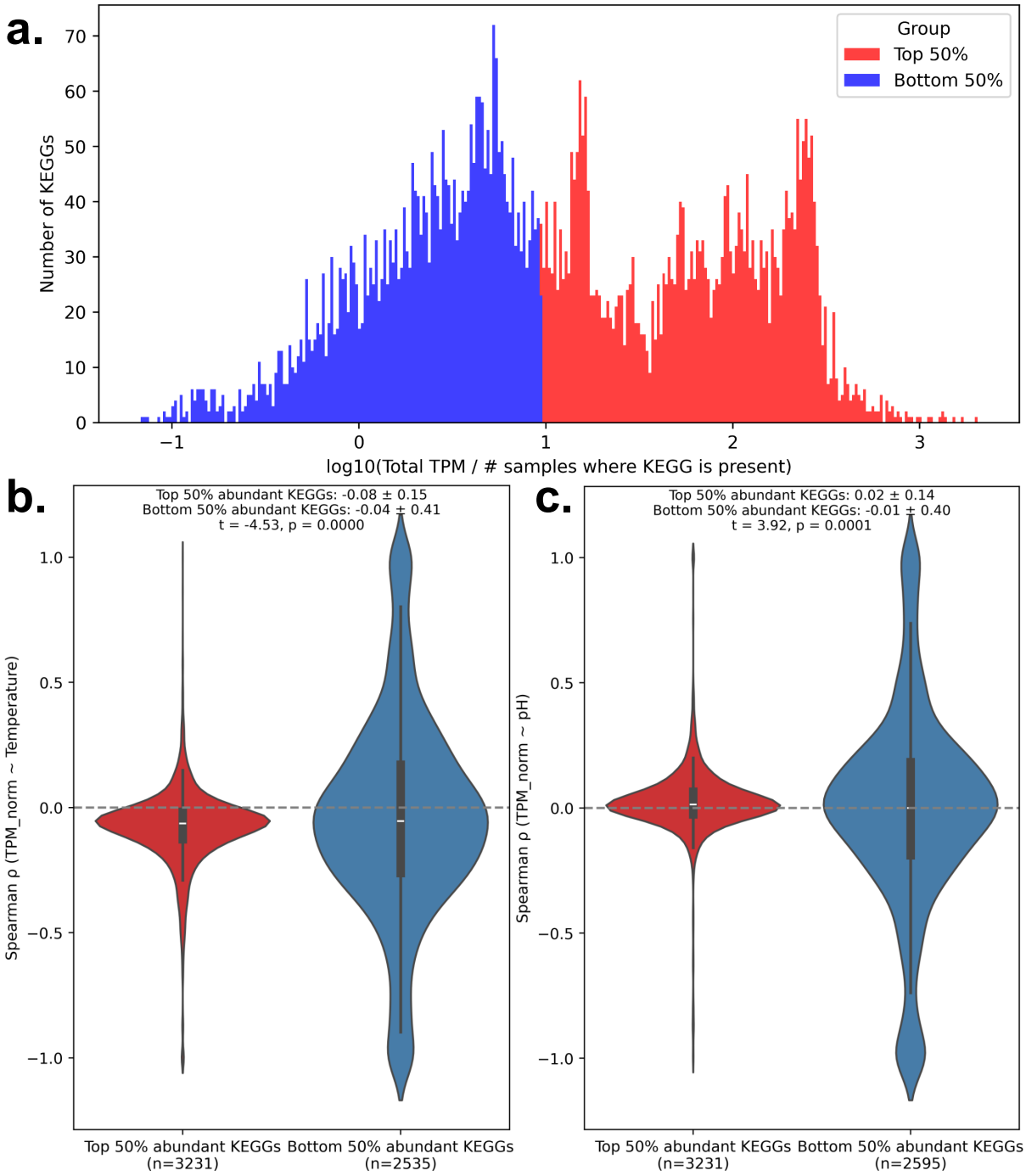

Figure 6: (a). Distribution of KEGG abundance. To make a conservative test, we split the distribution in two, to search if there is a characteristic behavior of the most abundant fraction with the variables. (b). Top 50% abundant KEGGs are significantly depleted (decrease in abundance) with increasing temperature compared to those with less abundance. Top vs bottom: $p<0.0000$; Top vs All: $p = 0.003$ (KEGGs are less abundant in the top with temperature increases); Bottom vs All: $p = 0.016$ (KEGGs are more abundant in the bottom with temperature increases) (c). In contrast, the more basic environments are significantly enriched by Top 50% abundant KEGGs. Top vs bottom: 0.0001; Top vs All: $p = 0.0014$ (KEGGs are more abundant in the top as pH increases) Bottom vs All: $p = 0.04$ (KEGGs are less abundant in the bottom as pH increases). These results are supported by 10% top comparison (Fig. S7) and regressions of KEGG loses, in Figures S5 and S6.

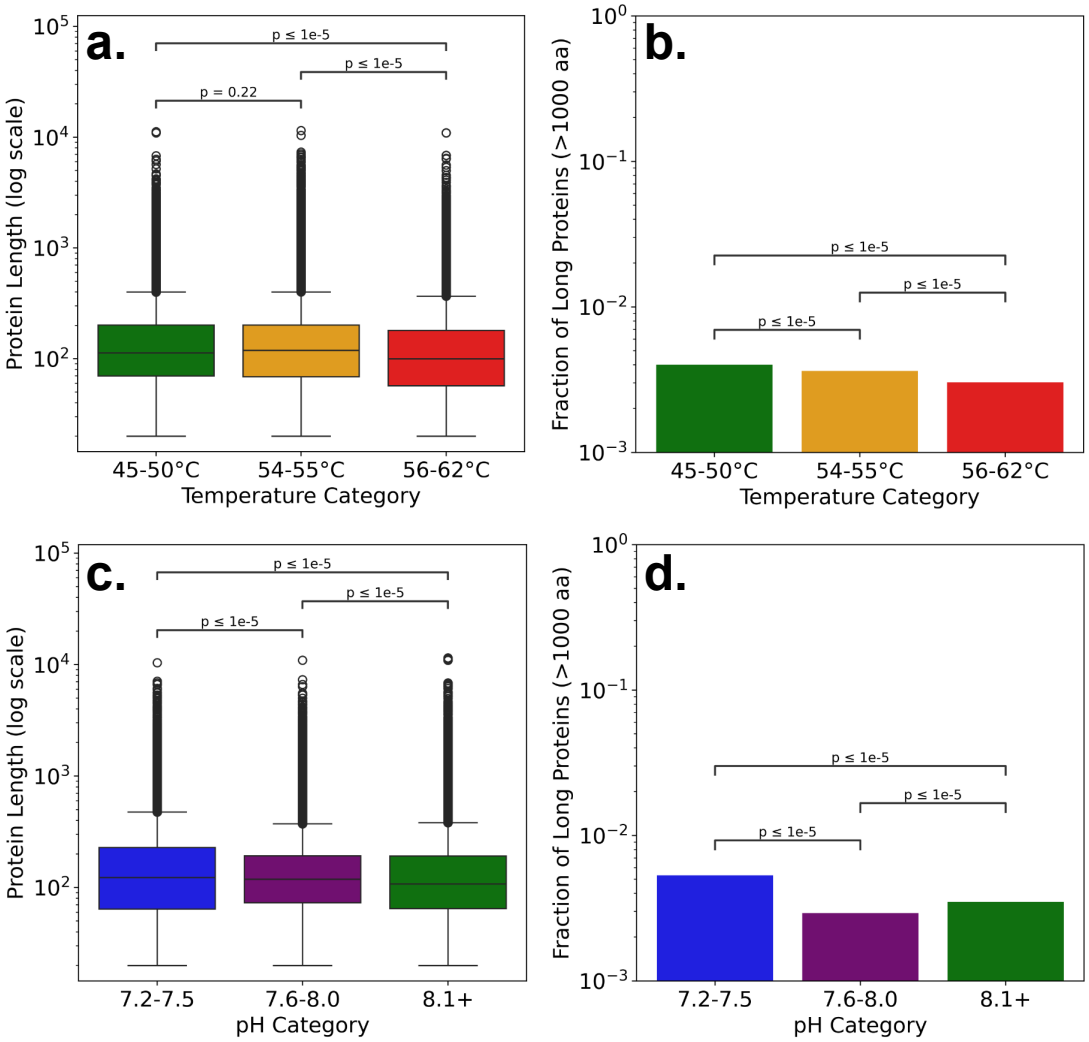

Figure 7: a. Distributions of protein length by temperature categories (all domains). Only higher temperature bin (red) was shown significantly shorter proteins (paired t-tests). b. Log fraction of proteins spanning more than 1,000 aa classified by temperature. The distributions showed a significant depletion of long proteins that correspond to higher temperatures (confidence intervals are negligible). c. Distributions of protein length by pH categories (all domains). In this case, every category showed significant differences. d. Log fraction of proteins spanning more than 1,000 aa classified by pH category. This result suggest that enrichment of long proteins is anticipated in metagenomes retrieved from an environment outside the 7.6–8.0 pH range (purple), and that this also applies to more basic environments.