Noname manuscript No.
(will be inserted by the editor)

Codon usage bias in prokaryotic genomes and environmental adaptation

Davide Arella · Maddalena Dilucca · Andrea Giansanti

Received: date / Accepted: date

Abstract In each genome, some codons are favored over others by selection likely because they are translated more efficiently and accurately. The selectively favored codons tend to correspond to the most highly expressed tRNAs. It has been recognized that this codon usage bias can influence the cellular fitness and that might be associated with the lifestyle of the organism. To test the impact of environments on genome evolution we studied the codon usage bias of 615 prokaryotes. We found that the extent of codon usage corresponds to the environment in witch the prokaryotes live. In particular, measuring the degree of codon usage bias by the tRNA adaptation index (tAI), we obtained that organisms living in a specialized habitat have high extents of codon usage bias, consistent with their need to adapt efficiently to specific environmental constraints. Differently, organisms able to live in multiple habitats exhibit low codon usage bias as they need to adapt to various physical and chemical conditions. Our results suggest the importance of co-evolution between tRNA availability and codon usage of an organism, in relation with the environmental adaptation.

Keywords codon bias · environment · adaptation · tRNA

D. Arella
Dipartimento di Fisica, Sapienza University of Rome, Rome, Italy
E-mail: davidearella@gmail.com

M. Dilucca
Dipartimento di Fisica, Sapienza University of Rome, Rome, Italy

A. Giansanti
Dipartimento di Fisica, Sapienza University of Rome, Rome, Italy
INFN, Roma1 unit, Rome, Italy

INTRODUCTION

The genetic code is degenerate, *i.e.*, some amino acids are encoded by more than one codon. Although coding for the same amino acid, synonymous codons are not equally used, a phenomenon known as codon usage bias (or shorter, codon bias) [12]. Codon usage can differ widely not only between organisms, but also within a genome and within a single gene [16,25]. A lot of factors might cause different codon usage bias and the selective forces influencing it, such as selection for optimized translation, expression, location within genes, rate of evolution, secondary structure, nucleotide composition, protein length and environment [30]. It was demonstrated that many bacteria and yeast undergo translational selection, with highly expressed genes preferentially using codons assumed to be translated faster and/or more accurately by the ribosome [11,2]. Thus, the codon usage bias within a genome usually reflects the selection pressure for translational optimization of highly expressed genes. The choice of preferred codons in a single genome is most closely correlated with abundance of the cognate tRNA molecules [2,17,18,7] and further influenced by the genome's GC content [5,14].

It was suggested by Ardersson and Kurland [1] and then substantiated by Kudla *et al.* [21] that selection towards highly adapted codons in highly expressed proteins has a global effect on the cell, resulting in an increase in cellular fitness. This suggest that the global extent of codon usage bias of an organism might be associated with its phenotypic traits. Following this idea, Botzman *et al.* determined an association between the lifestyles of several prokaryotic organisms and variations in their codon usage [4]. Their results indicated that species living in a wide range of habitats have low codon usage bias, which is consistent with the need to adapt to different environments. In addition, results also suggest that species may more readily adjust to metabolic variability by maintaining low codon bias. Furthermore, by analyzing 11 diverse microbial community sequencing samples, Roller *et al.* demonstrated that microbes living in the same ecological niche share a common preference for codon usage, regardless of their phylogenetic diversity [28]. Complementing these studies, the analysis of acidophilic bacteria revealed that they preferentially have low codon bias, consistent with both their capacity to live in a wide range of habitats and their slow growth rate [13].

The physical requirements that are optimal for prokaryotic growth vary dramatically for different types of bacteria and archea. As a group, bacteria display the widest variation of all organisms in their ability to inhabit different environments. One of the most prominent differences between prokaryotes is their requirement for, and response to, atmospheric oxygen (O_2). On the basis of oxygen requirements, prokaryotic organisms can be divided into obligate aerobes (they have absolute requirement for oxygen in order to grow), obligate anaerobes (they grow only in the absence of oxygen), facultative anaerobes (they thrive in the presence of oxygen but also grow in its absence), aerotolerant anaerobes (they do not use oxygen but are indifferent to the presence of oxygen) and microaerophiles (they require a minimum level of oxygen for

growth, about 1%–10%). Prokaryotes have adapted to a wide range of temperatures. The National Center for Biotechnology Information (NCBI) Microbial Genome Project Database uses five terms to categorize the temperature range an organism grows at, where cryophilic refers to -30° to -2°C, psychrophilic refers to -1° to +10°C, mesophilic refers to +11° to +45°C, thermophilic refers to +46° to +75°C, hyperthermophilic refers to above +75°C, and organisms that live at ranges that overlap with more than one category are labeled as the one corresponding to the largest overlap [43]. Water is a fundamental requirement for life. Some organisms prefer salty environments and are thus called halophiles. There is a wide range of halophilic microorganisms belonging to the domains Bacteria and Archaea. Halophiles are categorized as slight, moderate, or extreme, by the extent of their halotolerance. Slight halophiles prefer 0.3 to 0.8 M (1.7 to 4.8% seawater is 0.6 M or 3.5%), moderate halophiles 0.8 to 3.4 M (4.8 to 20%), and extreme halophiles 3.4 to 5.1 M (20 to 30%) salt content [23].

To better understand the role of codon usage bias in the adaptation of prokaryotes to their environments, we studied the codon usage for more than 600 prokariotic species.

MATERIALS AND METHODS

We analyzed the extent of codon usage bias in 615 organisms (544 bacteria and 71 archaea) reported in Supplementary Material (see Table 1 of file Excel). Classification of environmental properties and pathogenicity were downloaded from [4]. Nucleotide sequences were downloaded from the FTP server of the National Center for Biotechnology Information [3]. The tRNA gene copy number for each organism was retrieved from the genomic tRNA database (GtRNAdb) [22] (available at the site `http://gtrnadb.ucsc.edu`). To detect different patterns of codon usage among the genes of a species, heat-maps were drawn with the CIMminer software [39], which uses Euclidean distances and the average linkage algorithm (`http://discover.nci.nih.gov/cimminer`).

RSCU values

There are a lot of methods and indices to estimate the degree of codon usage bias in a gene. For an overview of current methods, their classification and rationale see [29]. We used here a basic statistical indicator, the relative synonymous codon usage (RSCU) [33]. The RSCU is the observed frequency of a codon divided by the expected frequency if all the synonymous codons for the amino acid were used equally. The RSCU is computed for each codon of each amino acid and it is formally defined as follows. Let n_i denote the number of synonymous codons encoding for the amino acid i (codon degeneracy) and let X_{ij} denote the number of occurrences of the codon j for amino acid i. The

RSCU for codon j encoding the amino acid i is defined as

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \,. \tag{1}$$

RSCU is a real value comprised between 0 and the number of synonymous codons for that amino acid, *i.e.* n_i . For average synonymous codon usage (no codon bias) the RSCU is 1. For codon usage more infrequent than the average codon usage, the RSCU is less than one, and for more frequent usage than the average for the amino acid, the RSCU is greater than 1.

We calculated these values with a Python script homemade. The RSCU values of the various codons in a gene can be grouped together as the 61 components (excluding the stop codons TAA, TAG and TGA – which are differently used by different species) of a vector which measures the codon usage bias for that given gene.

For each genome we calculated the average vector of RSCU, $\mathrm{RSCU}_{\mathrm{avg}}$, and the distance between the RSCU vector of a gene and the average RSCU vector of the genome using the cosine similarity. The cosine similarity considered as a normalized distance $d \in [0, 1]$ for each gene is estimated as:

$$d = \frac{\mathrm{RSCU} \bullet \mathrm{RSCU}_{\mathrm{avg}}}{\parallel \mathrm{RSCU} \parallel \parallel \mathrm{RSCU}_{\mathrm{avg}} \parallel} \,, \tag{2}$$

where $\bullet$ denotes the scalar product and $\parallel$ RSCU $\parallel$ is the magnitude of the RSCU vector. When d is close to 1, RSCU vector of a single gene is similar to RSCU average vector of the genome.

Principal component analysis

Principal component analysis (PCA) [19] is a multivariate statistical method to transform a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (called principal components) spanning a space of lower dimensionality. The transformation is defined so that the first principal component accounts for the largest possible variance of the data, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (*i.e.*, uncorrelated with) the preceding components.

We used this technique on the space of $\mathrm{RSCU}_{\mathrm{avg}}$ values, so that each organism of dataset is represented as a 61-dimensional vector with coordinates the codons. The eigenvectors of the associated correlation matrix, ordered according to the magnitude of the corresponding eigenvalues, are the principal components of the original data.

We projected in the plane of the first two principal components all genomes of the dataset. Centroids were calculated as average value with relative error bars as standard deviation.

We then performed another PCA adding the tRNA gene copy number (tGCN) values to the $\mathrm{RSCU}_{\mathrm{avg}}$ values to consider the availability of tRNA for each prokaryote.

tAI calculation

The speed of protein synthesis is bound to the waiting time for the correct tRNA to enter the ribosomal A site [37], and thus depends on tRNA concentrations [35]. The consequent adaptation of codon usage to tRNA availability [17, 18] is at the basis of tRNA adaptation index (tAI) [26,8]. It takes advantage of the fact that the tRNA gene copy number across many genomes has a high and positive correlation with tRNA abundance within the cell [17,24,20,9]. The tAI follows the same mathematical model of CAI [32] – defining for each codon i its absolute (W_i):

$$W_i = \sum_{j=1}^{m_i} (1 - s_{ij})\, \mathrm{tGCN}_{ij}\,, \tag{3}$$

where m_i is the number of tRNA isoacceptors that recognize the ith codon (*i.e.*, tRNAs that carry the same amino acid that is encoded by i and that make either WC or wobble pairing with it), tGCN_{ij} is the gene copy number of the j-th tRNA that recognizes the i-th codon and s_{ij} is a selective constraint on the efficiency of the codon-anticodon coupling. From the W_i values the relative adaptiveness value w_i of a codon is obtained as

$$w_i = \begin{cases} W_i/W_{\max} & \text{if} \quad W_i \neq 0 \\ w_{\mathrm{mean}} & \text{else} \end{cases}\,, \tag{4}$$

where $W_{\max}$ is the maximum W_i value and w_{mean} is the geometric mean of all w_i with $W_i \neq 0$. Finally, the tRNA adaptation index tAI_g of a gene g is computed as the geometric mean of the relative adaptiveness values of its codons

$$\mathrm{tAI}_g = \left(\prod_{k=1}^{l_g} w_k \right)^{1/l_g}\,, \tag{5}$$

where k is the codon defined by the k-th triplet in gene g and l_g is the length of the gene in codons (except the stop codon). The critical issue for tAI is the selection of a meaningful set of s_{ij} values, *i.e.*, weights that represent the efficiency of the interactions between codons and tRNAs. Assuming that tRNA usage is maximal for highly expressed genes, these values are chosen in order to optimize the correlation of tAI values with expression levels.

We calculated tAI values using the tAI package provided by Mario dos Reis on GitHub (`https://github.com/mariodosreis/tai`). This is an R package that implements the tAI as described in dos Reis *et al.* [26].

We divided the prokaryotes into groups according to their environmental characteristics and pathogenicity and then we compared the distributions of the average tAI values belonging to these groups. Mann-Whitney U-test was used to verify if distributions were well separated with p-value < 0.05.

GC content

In genetics, the GC content is the percentage of the nitrogenous bases on a DNA molecule that are either guanine or cytosine. The overall GC content was computed using a Perl script from the inspiring study on microsporidia by Xiang *et al.* [40] (`https://github.com/hxiang1019/calcGCcontent.git`). We divided the prokaryotes into groups according to their environmental characteristics and pathogenicity and then we compared the distributions of the GC content values belonging to these groups. Mann-Whitney U-test was used to verify if distributions were well separated with p-value < 0.05.

RESULTS AND DISCUSSION

RSCU values

Previous observations (see [25]) point to the fact that each bacterial species has a specific pattern and level of codon bias, which is strongly shared by all its genes; codon bias in specialized categories of genes appears to be just a modulation of the distinctive codon bias of the species [6]. To check this statement, we computed RSCU values of each codon for our set of prokaryotic genomes. We plotted in Figure 1 an example of heatmap of RSCU values for each gene (in the example, of *Escherichia coli* K12 substrain MG1655) that shows the existence of a finger print of codon bias for this organism. So, we calculated the distances (as cosine similarity) between the RSCU vector of each gene and the $\mathrm{RSCU}_{\mathrm{avg}}$ vector belonging to the species (see Figure 2). The distance distribution shows an average distance near to 1 meaning that the gene vectors are quite similar to the average vector.

Overall, this exploration suggests that there should be a strong correlation between codon bias patterns of each species and his evolutionary history. In our opinion, there is an ecological determinant behind this rough classification based on basic codon bias.

Principal component analysis

We performed PCA over the space of the $\mathrm{RSCU}_{\mathrm{avg}}$ vectors measured for each species in dataset (see Figure 3). The two first principal components (PC_1 and PC_2) turned out to represent as much as 71% of the total variance of codon bias over the genomes. Interestingly, the prokaryotes related to different environmental characteristics are well localized and separated in this reduced

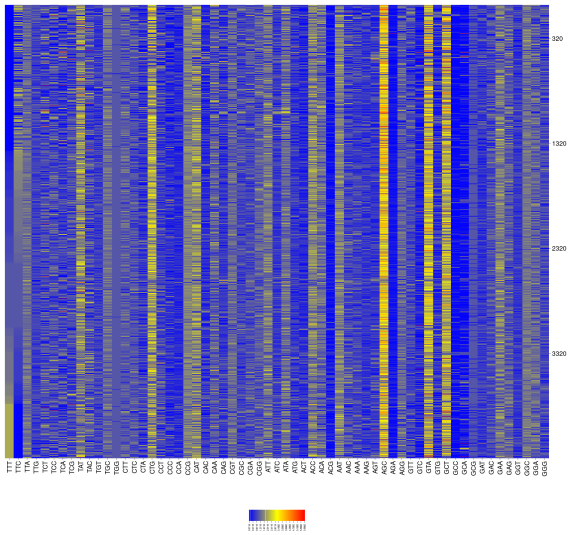


Fig. 1 Heat-map of RSCU values for each gene of *Escherichia coli* strain K12 substrain MG1655. Genes are in rows and codon are in columns. We note that RSCU values of each gene cluster in the axis of codons and RSCU vectors are very close each others. So an average RSCU vector can be considered a finger print for a species.

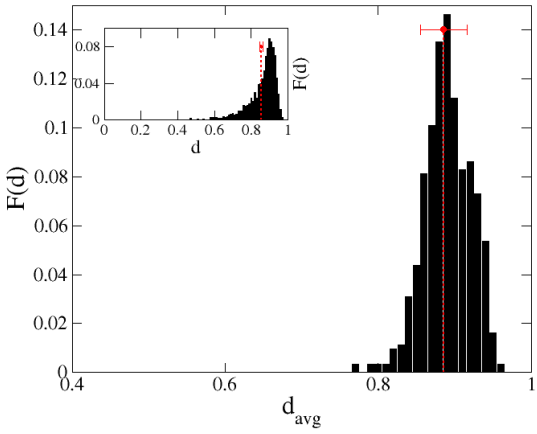


Fig. 2 Relative frequency F(d) of the average distances, calculated as cosine similarity, between RSCU vector of a gene and RSCU average vector of a genome for all the prokaryotes in dataset. Red dotted line denotes mean of the distribution and the error bar its standard deviation. Inset: distance distribution F(d) for all the genes of *Escherichia coli* K12 substrain MG1655. Note that the average distance (0.82 ± 0.02) is near to 1 to denote that the average vector of RSCU is a good descriptor for the codon usage of the whole genome.

space (four panels of Figure 3). The same result is shown when we characterized prokaryotes for different habitats (see Figure 4). In particular, we note that terrestrial organisms are located in the left part of the graph, isolated from

the others. This represents an important evidence: proteins that belong to the a particular environment are well-localized in the space of the two principal components. In other words, if a set of genomes are physically and functionally connected in an environment, their corresponding genes should share common codon bias features.

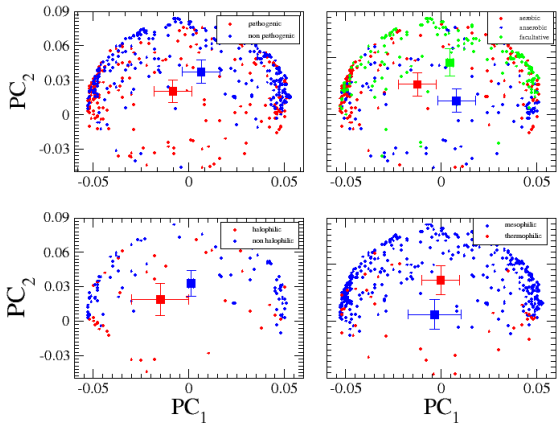


Fig. 3 Prokaryotes show different codon usage bias measured with $RSCU_{avg}$ values, in relation with different environmental characteristics. We project in PC_1-PC_2 planes the prokaryotes of the dataset exhibiting different environmental properties. Centroids are calculated as average value with relative error bars as standard deviation.

To consider the tRNA availability of these genomes we carried out another PCA combining together $RSCU_{avg}$ values and tGCN values measured for each species in the dataset (see Figure 5). The two first principal components (PC_1 and PC_2) turned out to represent as much as 52% of the total variance of codon bias over the genomes (worst than the first PCA with only $RSCU_{avg}$ values). In this case, the prokaryotes related to different environmental characteristics are not well localized and separated in this reduced space (four panels of Figure 5). These plots show that centroids of the groups are not well separated with higher error bars as standard deviation. However, when we observe the pattern of codon usage for different habitats (see Figure 6) organisms continue to be well-localized in the space of the two principal components.

tAI distribution

Pathogenic prokaryotes demonstrated lower codon usage bias than non-pathogenic prokaryotes, in accord with the multiple environments that many pathogens occupy (see Figure 7 and Table 1). In relation to oxygen requirement, facultative prokaryotes showed the lower codon usage bias and aerobic prokaryotes

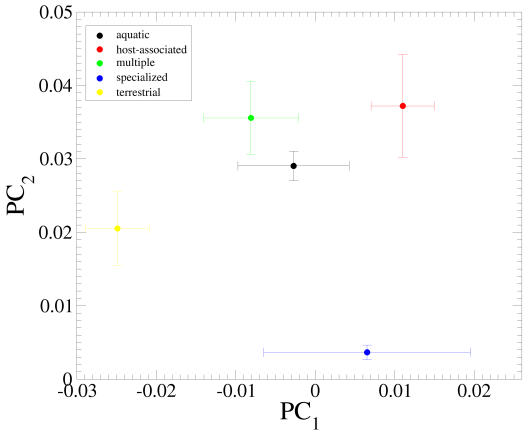


Fig. 4 Prokaryotes show different codon usage bias measured with $RSCU_{avg}$ values, in relation with different habitat conditions. We project in PC_1-PC_2 planes all the prokaryotes living in different habitats. Centroids are calculated as average value with relative error bars as standard deviation.

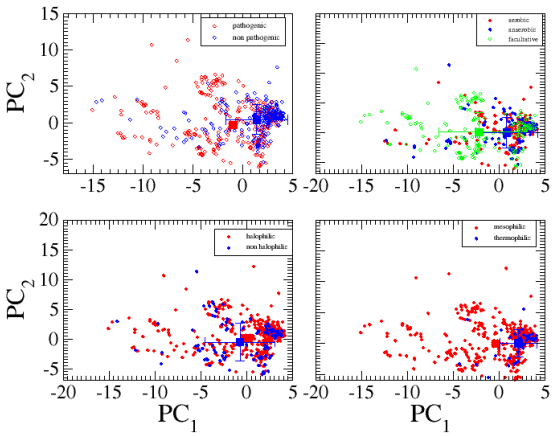


Fig. 5 Results of the second PCA considering as coordinates the $RSCU_{avg}$ values togheter with the tGCN values. We project in PC_1-PC_2 planes the prokaryotes of the dataset exhibiting different environmental properties. Centroids are calculated as average value with relative error bars as standard deviation.

showed the highest extent. Prokaryotes that live in different salinity environments (halophilic or non halophilic) showed a small difference in codon usage, with halophilic organisms exhibiting larger values of tAI. Thermophilic prokaryotes, in agreement with previous studies, had a higher extent of codon usage bias than mesophilic prokaryotes, demonstrating the correlation of codon bias with temperature [34, 15, 42]. Regarding the habitats, organisms living in a specialized habitat had the highest extents of codon usage bias, measured

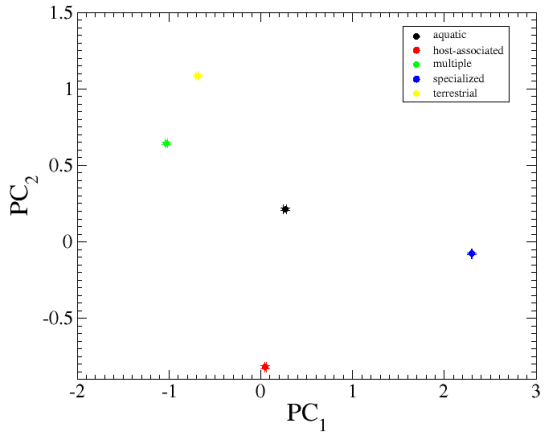


Fig. 6 Results of the second PCA considering as coordinates the $RSCU_{avg}$ values togheter with the tGCN values. We project in PC_1-PC_2 planes all the prokaryotes living in different habitats. Centroids are calculated as average value with relative error bars as standard deviation.

by tAI index (0.434 ± 0.01) consistent with their need to adapt efficiently to specific environmental conditions. They use efficiently always the same codons with a corresponding abundant tRNA availability, whereas organisms able to live in multiple habitats had low codon usage bias measured by tAI index (0.348 ± 0.006), consistent with their need to adapt to various physical and chemical conditions (see Figure 8).

We note that some of our results are anticorrelated with the results in [4], because in our analysis we consider the tRNA availability. Communities of microbes have been shown to share similar tRNA pools to facilitate horizontal gene transfer [36], which also implies a limited choice of preferred codons that are cognate to the shared community tRNA pool. It has also been shown that fast growth rates introduce stronger bias in synonymous codon usage at the level of whole metagenomes [38], much like the effect observed in single microbial species [27,31]. This shows the importance of tRNA co-evolution with codon bias and that the conditions under which a gene is replicated also appear to affect codon preferences.

GC content

Microbes in the same environment live within the same physical and chemical constraints, such as temperature, pH or ion concentration, and it was demonstrated that GC content is metagenome-specific [10]. GC content distributions for groups of prokaryotes living in different environments were calculated. Mann-Whitney test was effectuated to demonstrate if the distributions between two groups were significantly different with p-value < 0.05. The test was not passed only for halophilic and non halophilic organisms. The group of

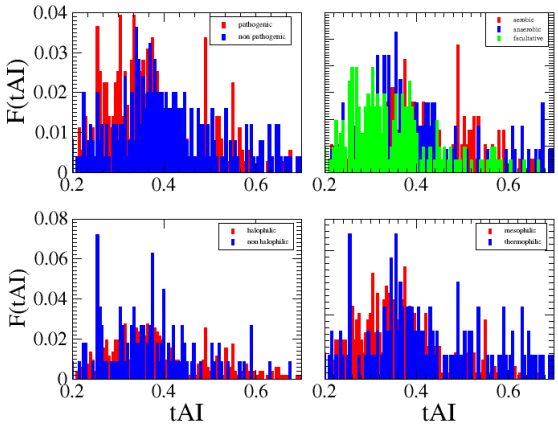


Fig. 7 Relative frequency F(tAI) of average tAI values for different gruops of prokaryoets in dataset classified according to various properties. Mann-Whitney tests were effectuated to demonstrate if two distributions of each panel were well separated with p-value < 0.05. All the differences between the prokaryotic groups are statistically significant. Values of means and standard deviations are reported in Table 1.

Characteristic	**tAI**	σ
pathogenic	0.366	0.005
non pathogenic	0.414	0.001
aerobic	0.406	0.116
anaerobic	0.399	0.126
facultative	0.349	0.100
halophilic	0.390	0.119
non halophilic	0.370	0.104
mesophilic	0.373	0.106
thermophilic	0435	0.139

Table 1 For each prokaryotic characteristic, we calculate average value and standard deviation of tAI. We note that group of thermophiles has higher values of tAI.

host-associated species had higher values of GC content, whereas specialized prokaryotes had lower values.

If we consider all the prokaryotes in dataset, we show in Figure 9 that correlation between tAI and GC content is not present. In Figure 10 we show heatmaps of RSCU average values (upper panel) and tRNA gene copy numbers (bottom panel) for each prokaryote in a given habitat. We sorted codons subdivided in AT-endings and GC-endings. We note that prokaryotes belonging to the same habitat do not cluster in abundance of GC- or AT-endings. It is interesting to show how tRNA availability is equal for half codons in a lot of species.

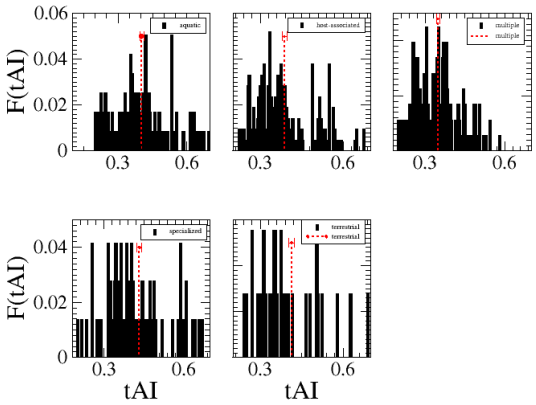


Fig. 8 Relative frequency F(tAI) of average tAI values for different gruops of prokaryoets in dataset classified according to various habitats. Mean values ± standard deviation for each habitat are: aquatic (0.403 ±0.009), host-associated (0.388 ± 0.008), multiple (0.348 ± 0.006), specialized (0.434 ± 0.010) and terrestrial (0.414 ± 0.010).

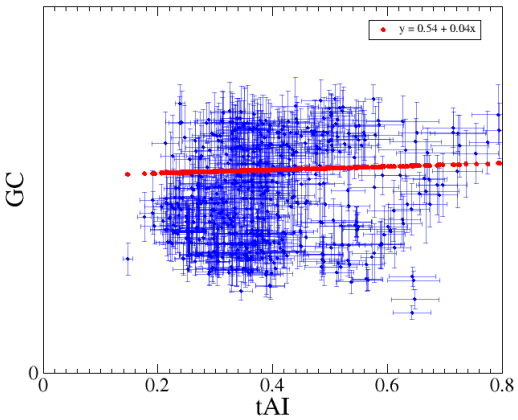


Fig. 9 Correlation between average tAI values and GC content. Linear regression is represented with $R^2 = 0.42$. Very low slope of the line show that GC content of prokaryotes in dataset is not strictly correlated with codon bias. Error bars are calculated as standard deviations.

CONCLUSIONS

The analysis of codon usage bias is very important to characterize similar or different communities of prokaryotes. Our analysis revealed a large variability in codon bias: there are organisms showing very high degrees of codon usage bias and organisms exhibiting very low differences in the use of codons among genes. The aim of this work was to study the codon usage bias of prokaryotes

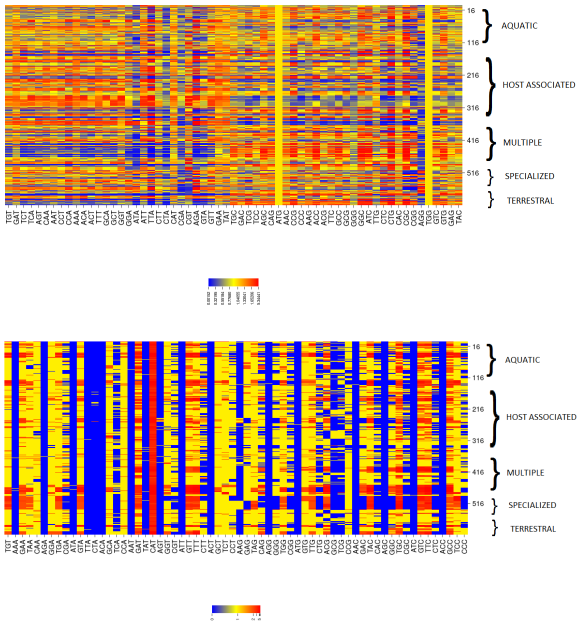


Fig. 10 Heat-maps of $RSCU_{avg}$ and tGCN for prokaryotes in different habitats. We show in upper panel $RSCU_{avg}$ values and in bottom panel tGCN values. We sorted codons subdivided in AT-endings and GC-endings. We note that the prokaryotes do not cluster in abundance of GC- or AT-endings.

in different environments and, remarkably, we found that the codon usage bias correlate to environmental conditions.

Codon usage bias was calculated by RSCU values, that characterize different choices of codons, but only in relation to the statistics. Subsequently, we used the tAI adaptive index, built on availability of tRNA.

The adaptation of codon usage to the genomic tRNA gene pool is a well-known phenomenon in various organisms where translational selection is known to be present. In fact, some authors have yet discussed how the redundancy in the gene number of certain tRNA isoacceptors matches the frequencies of the preferred set of codons. It can be argued that it is the need for translational optimization and hence codon usage that shapes the tRNA pool of organisms. The abundance of distinct tRNAs, even though transferring the same amino acid may affect the speed of translation and protein folding [41].

In the present study, we have sustained that organisms that live in a specialized habitat have higher extents of codon usage bias, consistent with their need to adapt efficiently to a specific environment. On the contrary, prokaryotes that live in multiple environments have shown lower codon usage bias as they need to be more flexible. Accordingly, phatogens, which usually live in heterogeneous physical conditions and facultative anaerobes, which can grow with or without the the presence of oxygen, showed lower codon usage bias. When

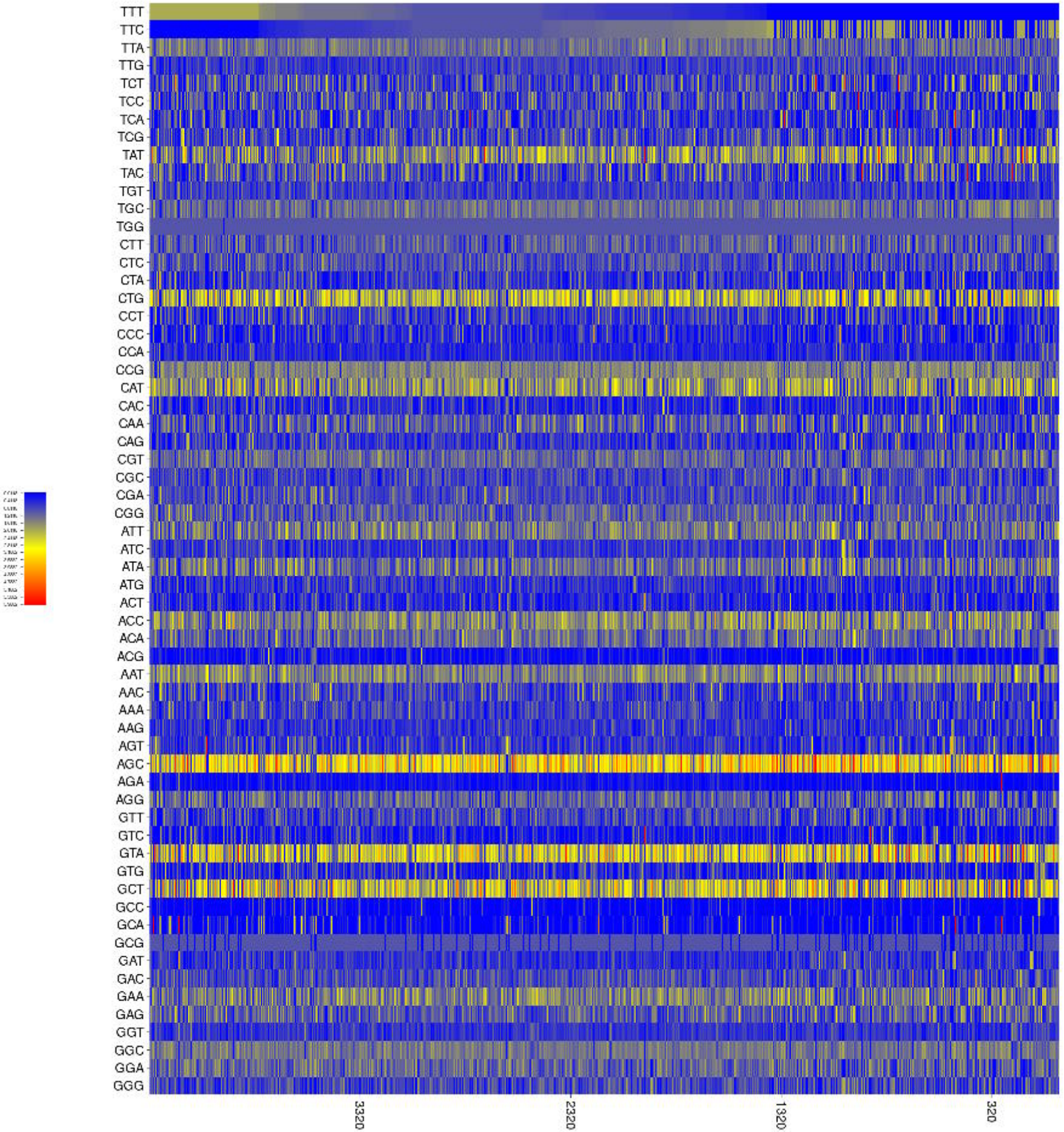
comparing prokaryotes that live in different salinity environments (halophilic or non halophilic) we have found only small differences in codon usage, with halophilic species exhibiting higher values of tAI. As pointed out by previous studies, thermophiles exhibited a higher extent of codon usage bias than mesophiles. We note that some of our results are the opposite of what was found in [4], because in our analysis we take into account the tRNA availability. Our results suggest that the co-evolution of tRNA availability and codon usage bias of an organism plays a role in the adaptation of prokaryotes to their environments.

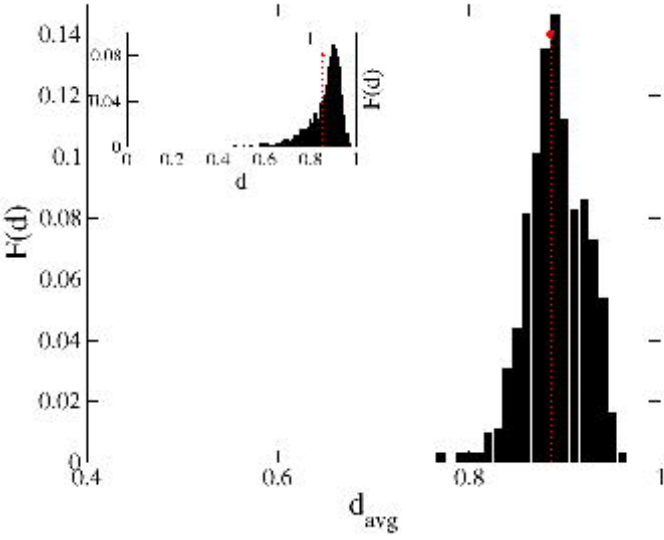
References

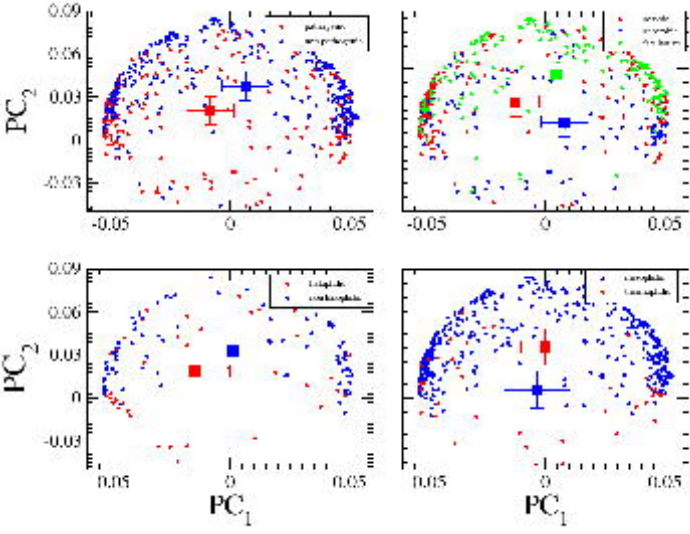
1. Andersson, S., Kurland, C.: Codon preferences in free-living microorganisms. Microbiology and Molecular Biology Reviews **54**(2), 198–210 (1990)
2. Bennetzen, J.L., Hall, B.D.: Codon selection in yeast. Journal of Biological Chemistry **257**(6), 3026–3031 (1982)
3. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. Nucleic acids research **41**(D1), D36–D42 (2012)
4. Botzman, M., Margalit, H.: Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome biology **12**(10), R109 (2011)
5. Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.H.: Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences **101**(10), 3480–3485 (2004)
6. Dilucca, M., Cimini, G., Semmoloni, A., Deiana, A., Giansanti, A.: Codon bias patterns of E. coli's interacting proteins. PLoS ONE **10**(11), e0142127 (2015)
7. Dong, H., Nilsson, L., Kurland, C.G.: Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. Journal of molecular biology **260**(5), 649–663 (1996)
8. Dos Reis, M., Wernisch, L., Savva, R.: Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic acids research **31**(23), 6976–6985 (2003)
9. Duret, L.: tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics **16**(7), 287–289 (2000)
10. Foerstner, K.U., Von Mering, C., Hooper, S.D., Bork, P.: Environments shape the nucleotide composition of genomes. EMBO reports **6**(12), 1208–1213 (2005)
11. Gouy, M., Gautier, C.: Codon usage in bacteria: correlation with gene expressivity. Nucleic acids research **10**(22), 7055–7074 (1982)
12. Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A.: Codon catalog usage and the genome hypothesis. Nucleic acids research **8**(1), 197–197 (1980)
13. Hart, A., Cortés, M.P., Latorre, M., Martinez, S.: Codon usage bias reveals genomic adaptations to environmental conditions in an acidophilic consortium. PloS one **13**(5) (2018)
14. Hershberg, R., Petrov, D.A.: General rules for optimal codon choice. PLoS genetics **5**(7) (2009)
15. Hickey, D.A., Singer, G.A.: Genomic and proteomic adaptations to growth at high temperature. Genome biology **5**(10), 117 (2004)
16. Hooper, S.D., Berg, O.G.: Gradients in nucleotide and codon usage along escherichia coli genes. Nucleic acids research **28**(18), 3517–3523 (2000)
17. Ikemura, T.: Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of molecular biology **151**(3), 389–409 (1981)
18. Ikemura, T.: Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution **2**(1), 13–34 (1985)

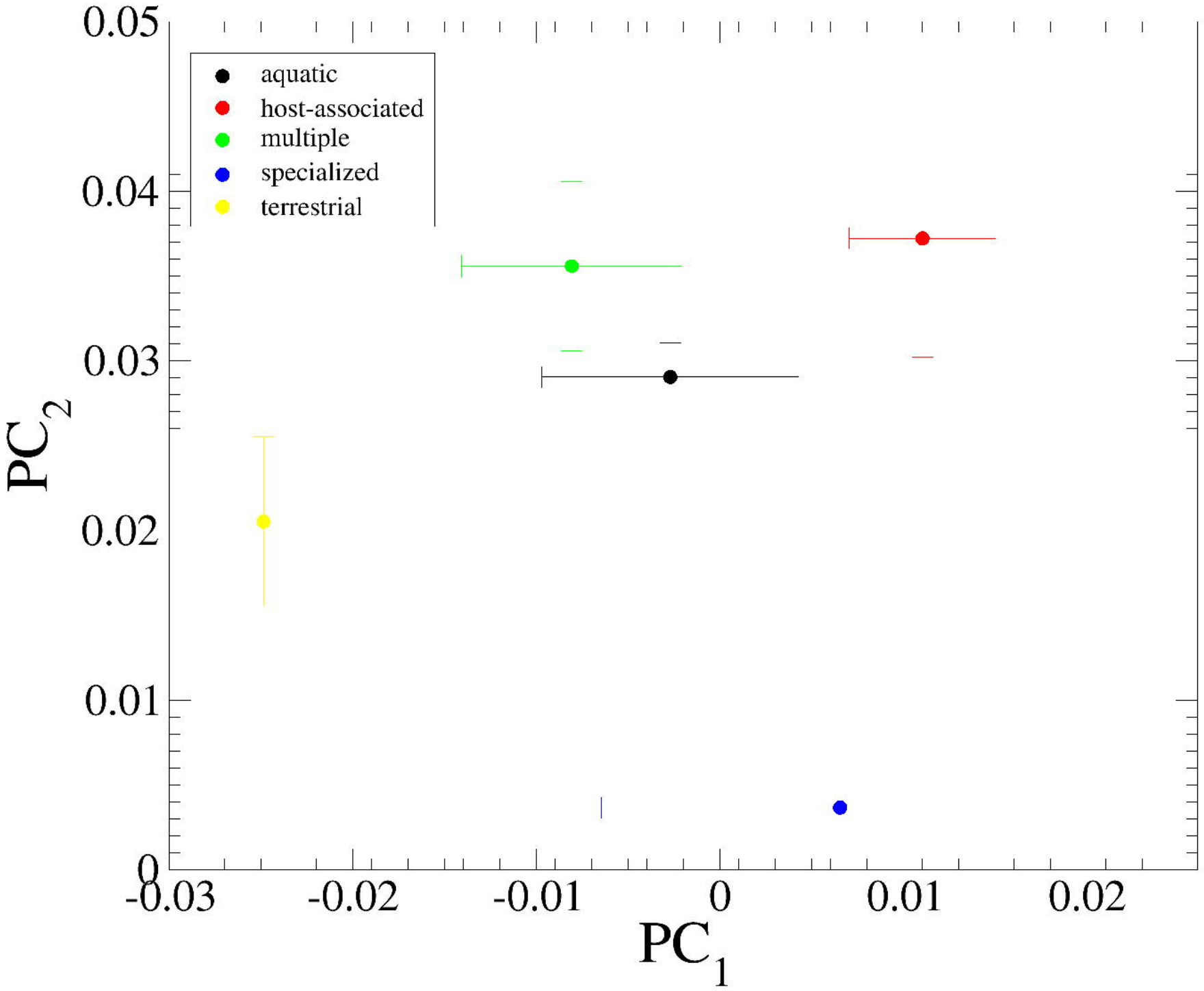
19. Jolliffe, I.: Principal Component Analysis. Springer Series in Statistics. Springer-Verlag (2002). DOI 10.1007/b98835
20. Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T.: Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**(1), 143–155 (1999)
21. Kudla, G., Murray, A.W., Tollervey, D., Plotkin, J.B.: Coding-sequence determinants of gene expression in Escherichia coli. Science **324**(5924), 255–258 (2009)
22. Lowe, T.M., Eddy, S.R.: trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. Nucleic acids research **25**(5), 955–964 (1997)
23. Ollivier, B., Caumette, P., Garcia, J.L., Mah, R.: Anaerobic bacteria from hypersaline environments. Microbiology and Molecular Biology Reviews **58**(1), 27–38 (1994)
24. Percudani, R., Pavesi, A., Ottonello, S.: Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. Journal of molecular biology **268**(2), 322–330 (1997)
25. Plotkin, J.B., Kudla, G.: Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics **12**(1), 32–42 (2011)
26. Reis, M.d., Savva, R., Wernisch, L.: Solving the riddle of codon usage preferences: a test for translational selection. Nucleic acids research **32**(17), 5036–5044 (2004)
27. Rocha, E.P.: Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome research **14**(11), 2279–2286 (2004)
28. Roller, M., Lucić, V., Nagy, I., Perica, T., Vlahoviček, K.: Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic acids research **41**(19), 8842–8852 (2013)
29. Roth, A., Anisimova, M., Cannarozzi, G.M.: Measuring codon usage bias. Codon evolution: mechanisms and models pp. 189–217 (2012)
30. Salim, H.M., Cavalcanti, A.R.: Factors influencing codon usage bias in genomes. Journal of the Brazilian Chemical Society **19**(2), 257–262 (2008)
31. Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., Sockett, R.E.: Variation in the strength of selected codon usage bias among bacteria. Nucleic acids research **33**(4), 1141–1153 (2005)
32. Sharp, P.M., Li, W.H.: The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research **15**(3), 1281–1295 (1987)
33. Sharp, P.M., Tuohy, T.M., Mosurski, K.R.: Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic acids research **14**(13), 5125–5143 (1986)
34. Singer, G.A., Hickey, D.A.: Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene **317**, 39–47 (2003)
35. Sørensen, M.A., Kurland, C., Pedersen, S.: Codon usage determines translation rate in Escherichia coli. Journal of molecular biology **207**(2), 365–377 (1989)
36. Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U., Ruppin, E.: Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic acids research **39**(11), 4743–4755 (2011)
37. Varenne, S., Buc, J., Lloubes, R., Lazdunski, C.: Translation is a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of molecular biology **180**(3), 549–576 (1984)
38. Vieira-Silva, S., Rocha, E.P.: The systemic imprint of growth and its uses in ecological (meta) genomics. PLoS genetics **6**(1) (2010)
39. Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., et al.: An information-intensive approach to the molecular pharmacology of cancer. Science **275**(5298), 343–349 (1997)
40. Xiang, H., Zhang, R., Butler III, R.R., Liu, T., Zhang, L., Pombert, J.F., Zhou, Z.: Comparative analysis of codon usage bias patterns in microsporidian genomes. PloS one **10**(6) (2015)
41. Yoon, J., Chung, Y.J., Lee, M.: STADIUM: Species-Specific tRNA Adaptive Index Compendium. Genomics & informatics **16**(4) (2018)

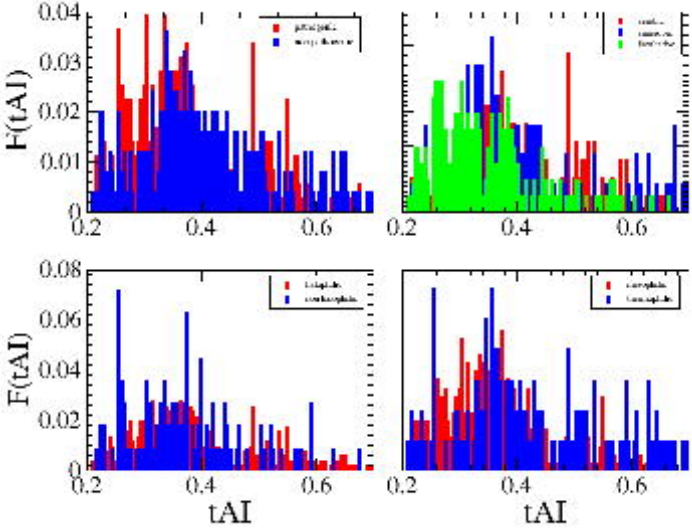
42. Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I.: Protein and DNA sequence determinants of thermophilic adaptation. PLoS computational biology **3**(1) (2007)
43. Zheng, H., Wu, H.: Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species. BMC bioinformatics **11**(11), S7 (2010)

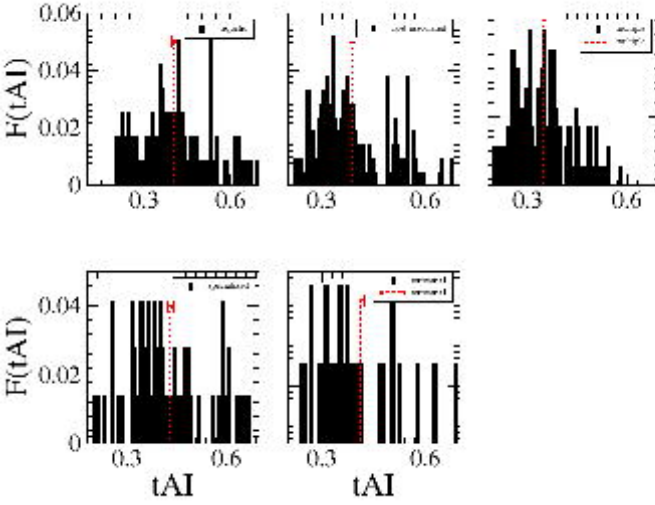

TTT
TTC
TTA
TTG
TCT
TCC
TCA
TCG
TAT
TAC
TGT
TGC
TGG
CTT
CTC
CTA
CTG
CCT
CCC
CCA
CCG
CAT
CAC
CAA
CAG
CGT
CGC
CGA
CGG
ATT
ATC
ATA
ATG
ACT
ACC
ACA
ACG
AAT
AAC
AAA
AAG
AGT
AGC
AGA
AGG
GTT
GTC
GTA
GTG
GCT
GCC
GCA
GCG
GAT
GAC
GAA
GAG
GGT
GGC
GGA
GGG
3320
2320
1320
320

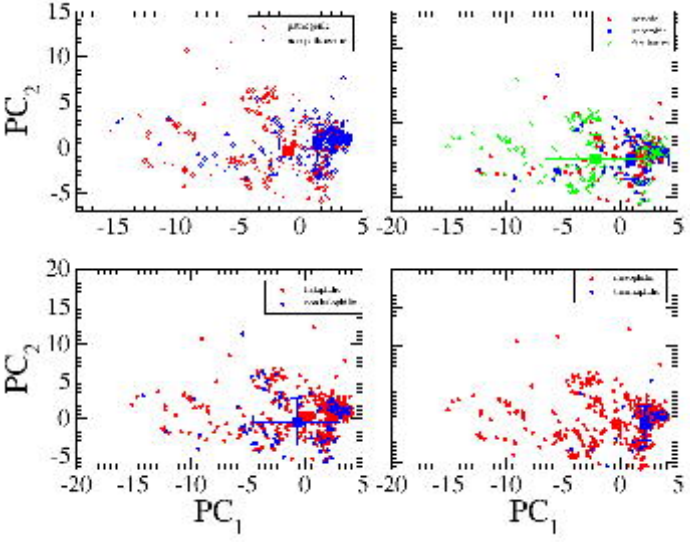

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
F(d)
0.4
0.6
0.8
1
d_{avg}
0.08
0.04
0
0
0.2
0.4
0.6
0.8
1
d
F(d)

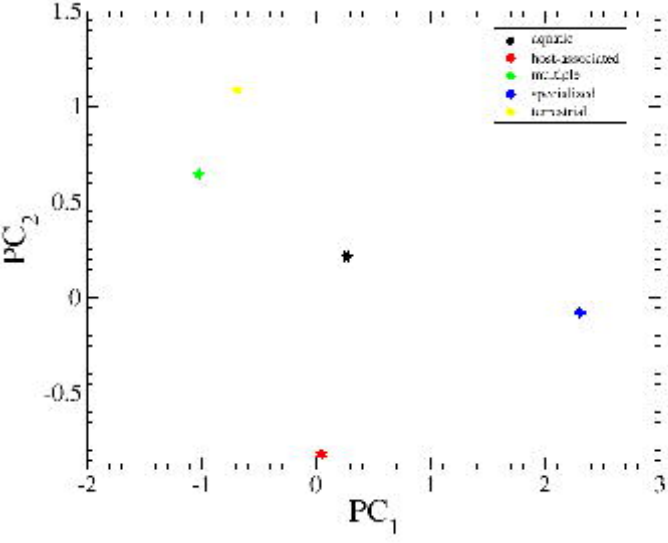

0.09
0.06
0.03
0
-0.03
PC2
-0.05
0
0.05
PC1


aquatic
host-associated
multiple
specialized
terrestrial
0.05
0.04
0.03
0.02
0.01
0
-0.03
-0.02
-0.01
0
0.01
0.02
PC2
PC1

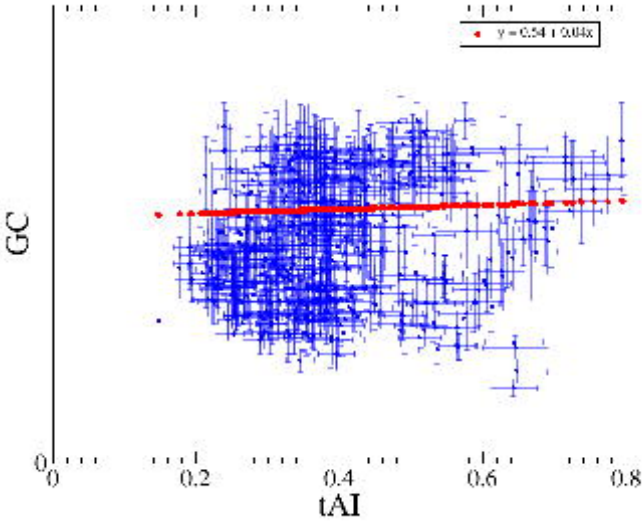

F(tAI)
0.04
0.03
0.02
0.01
0
0.2
0.4
0.6
F(tAI)
0.08
0.06
0.04
0.02
0
0.2
0.4
0.6
tAI
tAI

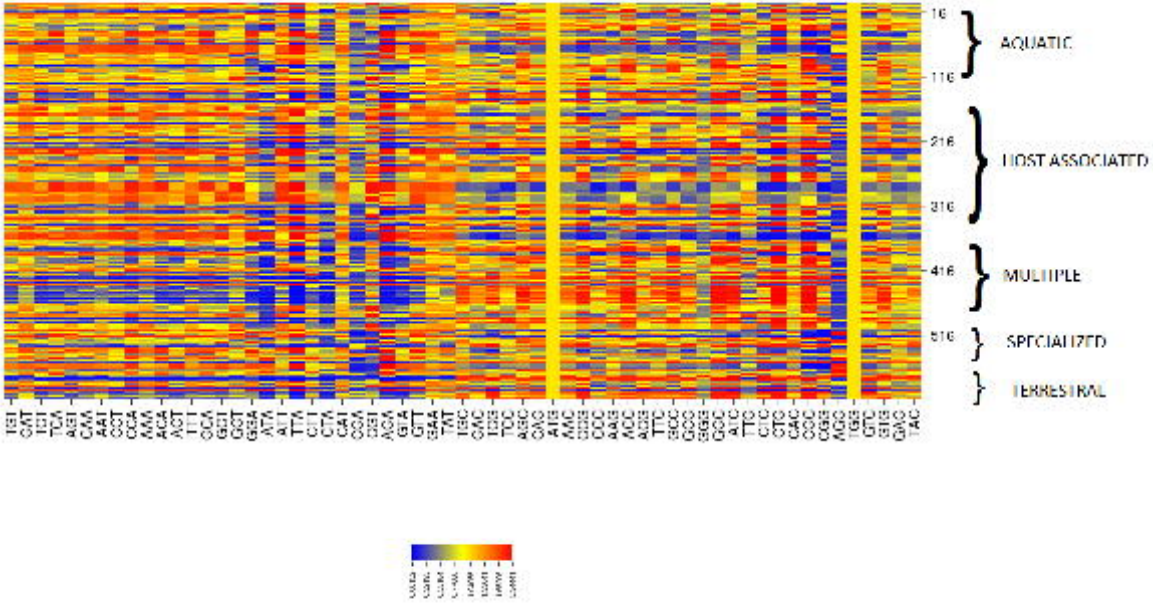

F(tAI)
0.06
0.04
0.02
0
0.3
0.6
tAI

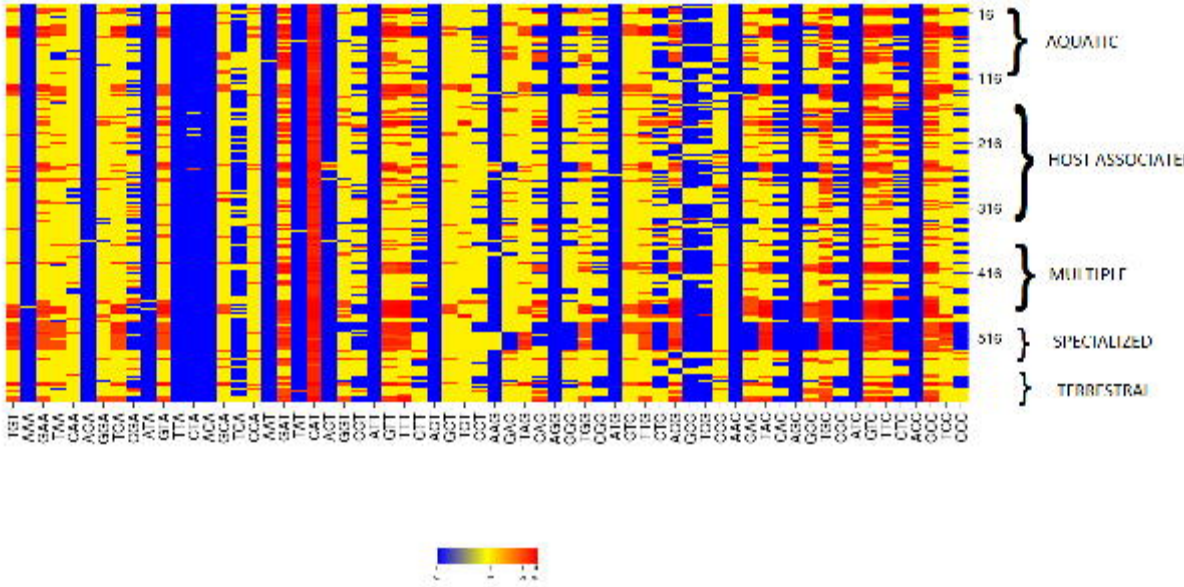

PC2
PC1
15
10
5
0
-5
-15
-10
-5
0
5
-20
-15
-10
-5
0
5
20
15
10
5
0
-5
-20
-15
-10
-5
0
5
-20
-15
-10
-5
0
5



1.5
1
0.5
0
-0.5
-2
-1
0
1
2
3
PC1
PC2
aquatic
host-associated
multiple
specialized
terrestrial


y = 0.54 + 0.04x
GC
0
0
0.2
0.4
0.6
0.8
tAI


1
116
AQUATIC
216
316
HOST ASSOCIATED
416
MULTIPLE
516
SPECIALIZED
TERRESTRIAL


1
115
218
315
418
515
AQUATIC
HOST ASSOCIATED
MULTIPLE
SPECIALIZED
TERRESTRIAL