

# A Tree of Cellular Life Inferred from a Genomic Census of Molecular Functions

Kyung Mo Kim · Arshan Nasir · Kyuin Hwang · Gustavo Caetano-Anollés

Received: 30 June 2014 / Accepted: 5 August 2014 / Published online: 17 August 2014  
© Springer Science+Business Media New York 2014

**Abstract** Phylogenomics aims to describe evolutionary relatedness between organisms by analyzing genomic data. The common practice is to produce phylogenomic trees from molecular information in the sequence, order, and content of genes in genomes. These phylogenies describe the evolution of life and become valuable tools for taxonomy. The recent availability of structural and functional data for hundreds of genomes now offers the opportunity to study evolution using more deep, conserved, and reliable sets of molecular features. Here, we reconstruct trees of life from the functions of proteins. We start by inferring rooted phylogenomic trees and networks of organisms directly from Gene Ontology annotations. Phylogenies and networks yield novel insights into the emergence and evolution of cellular life. The ancestor of Archaea originated earlier than the ancestors of Bacteria and Eukarya and was thermophilic. In contrast, basal bacterial lineages were non-thermophilic. A close relationship between Plants and Metazoa was also identified that disagrees with the

traditional Fungi-Metazoa grouping. While measures of evolutionary reticulation were minimum in Eukarya and maximum in Bacteria, the massive role of horizontal gene transfer in microbes did not materialize in phylogenomic networks. Phylogenies and networks also showed that the best reconstructions were recovered when problematic taxa (i.e., parasitic/symbiotic organisms) and horizontally transferred characters were excluded from analysis. Our results indicate that functionomic data represent a useful addition to the set of molecular characters used for tree reconstruction and that trees of cellular life carry in deep branches considerable predictive power to explain the evolution of living organisms.

**Keywords** Evolution · Tree of life · Gene ontology · Phylogenomics

## Introduction

Evolutionary genomics embraces the study of phylogenomic relationships between organisms at global scale. Phylogenomic trees are non-reticulated network representations of molecular evolution with branches, nodes, and

Kyung Mo Kim and Arshan Nasir contributed equally to this work.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-014-9637-9) contains supplementary material, which is available to authorized users.

K. M. Kim · K. Hwang  
Microbial Resource Center, Korea Research Institute of  
Bioscience and Biotechnology, Daejeon 305-806, Korea

K. M. Kim · K. Hwang  
Department of Bioinformatics, University of Science and  
Technology, Daejeon 305-350, Korea

A. Nasir · G. Caetano-Anollés (✉)  
Evolutionary Bioinformatics Laboratory, Department of Crop  
Sciences, University of Illinois, Urbana, IL 61801, USA  
e-mail: gca@illinois.edu

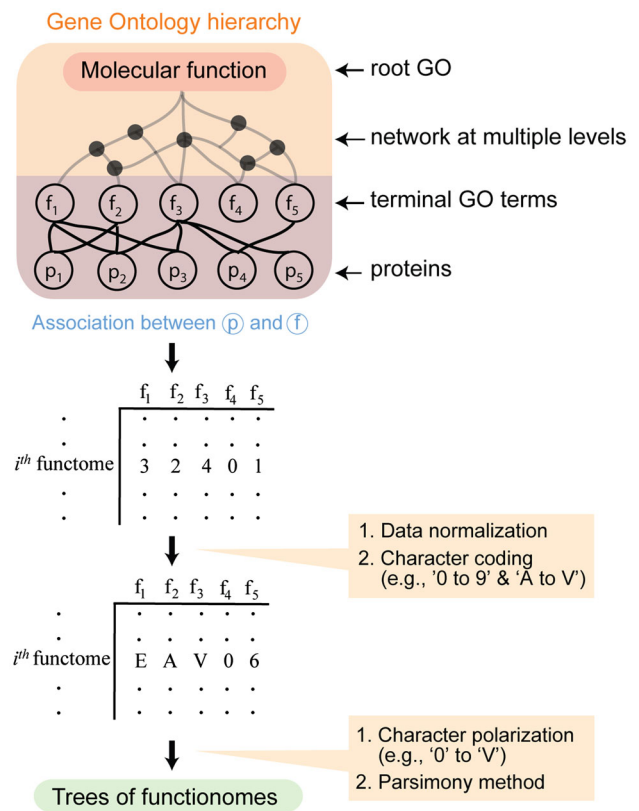
A. Nasir · G. Caetano-Anollés  
Illinois Informatics Institute, University of Illinois, Urbana,  
IL 61801, USA

*Present Address:*  
K. Hwang  
Division of Polar Life Sciences, Korea Polar Research Institute,  
Incheon 406-840, Republic of Korea

leaves (taxa) describing change in features of evolving genomes. Prior to molecular biology, phylogenetics was mostly restricted to the study of morphological, biochemical, and physiological data. This data did not allow systematic comparison across lineages and made impossible the elucidation of the deep evolutionary relationships of organisms belonging to the three superkingdoms of life (reviewed in Delsuc et al. 2005). Advances in molecular biology enabled the use of molecular data for phylogenetic tree reconstruction, including the sequence (Woese et al. 1980), order (Sankoff et al. 1992) and content of genes in genomes (Gu and Zhang 2004), and the atomic structural annotation of gene products (Caetano-Anollés and Caetano-Anollés 2003; Sun and Caetano-Anollés 2008a, b). This led to significant evolutionary discoveries such as recognition of Archaea as the third domain of life (Woese and Fox 1977; Woese 1987), illustration of reductive trends in the genomes and proteomes of cellular organisms (Wang et al. 2007, 2011a, b), and the genetically simple but functionally complex make up of the last universal common ancestor (LUCA) of life (Kim and Caetano-Anollés 2011).

Reconstructing phylogenetic trees from protein and nucleic acid sequences has become common practice. However, the use of sequence information may not be appropriate for studying deep phylogenetic relationships. In fact, mutation, recombination, and gene duplication of molecular sequences occur at relatively fast pace (Forster and Philippe 1999; Penny and Poole 1999; Koonin 2003). This dynamics leads to mutational saturation and paralogy, important processes that limit the validity of phylogenetic analysis to low taxonomy-level snapshots of recent evolutionary history. Although a few highly conserved orthologous genes are still available for reconstructing global phylogenies of living organisms, including the tree of life (ToL), their information cannot fully resolve relationships that are deep (e.g., polytomies in rRNA trees; Pace 2009). A few recent studies have reconstructed ToLs using protein domain repertoires, (Gerstein 1998; Caetano-Anollés and Caetano-Anollés 2003), domain interactomes (Wang and Caetano-Anollés 2006), or metabolic information (Marcet-Houben et al. 2007; Chang et al. 2011). These new kinds of data are regarded as controlled molecular vocabularies that cover the continuous spectrum of evolutionary conservation. While the new phylogenies resemble traditional classifications, they yield novel insights into the emergence and evolution of cellular life. Here, we expand on the idea of reconstructing ToLs from atypical genomic information by producing rooted phylogenies derived directly from the entire repertoires of functions defined at molecular level (functionomes).

The Gene Ontology (GO) database describes the functional annotations and relationships of nearly half a million proteins (Ashburner et al. 2000). This information is



**Fig. 1** Overview of the phylogenomic methodology. The GO hierarchy with multiple levels associates proteins to biological, cellular, and molecular roles. The genomic abundance counts of terminal GO terms (corresponding to molecular functions) were used as phylogenetic characters to describe the evolution of hundreds of functionomes (i.e., repertoire of functions). The abundance matrix was normalized and polarized to resolve compatibility issues with the phylogenetic reconstruction software PAUP\*. Maximum parsimony was used to search for the best possible tree and to reconstruct trees of cellular life built directly from the ontological census of molecular functions

presented in three separate tree-like structures, in which three root GO terms, molecular function (MF), biological process (BP), and cellular component (CC), descend toward a bottom (terminal) level into a multi-level hierarchy of ontological terms. Each of these tree-like structures represents an independent-directed acyclic graph (DAG), where child GO terms can be associated with multiple parents to account for both differing relationships and associations (Fig. 1). In the case of DAG<sub>MF</sub>, GO terms at higher levels represent broader functional categories (e.g., catalytic activity) while those at lower levels indicate more specific functional annotations (e.g., ATPase activity) (Ashburner et al. 2000; Kim and Caetano-Anollés 2010). This hierarchical structure is absent from other existing functional classification schemes such as the Cluster of Orthologous (COG) groups (Tatusov et al. 2003) and the functional classification of the SUPERFAMILY database (Vogel and Chothia 2006). Although the SEED subsystems

provide a hierarchy of multiple functional levels that is similar to the GO, the database specializes in bacterial gene annotation (Overbeek et al. 2005). Consequently, the GO is far more comprehensive than existing databases and has been successfully utilized in the past to describe the evolution of modern MFs (Kim and Caetano-Anollés 2010; Nasir and Caetano-Anollés 2013).

We note that the GO hierarchy can be analogous to an evolutionary hierarchy where higher-level GO terms may be more ancient while lower-level terms seem more modern (Kim and Caetano-Anollés 2010). This notion follows the hypothesis that promiscuous functions can serve as evolutionary starting points (Jensen 1976; Khersonsky and Tawfik 2010), with proteins of ancient origin being promiscuous and serving multiple functions (comparable to higher-level GO terms) and proteins of recent origin carrying more specified functions (comparable to terminal terms). The existence of this link between GO hierarchy and evolution enables sampling GO terms as phylogenetic characters in hundreds of completely sequenced proteomes (which are considered taxa) and studying the evolution of organisms using a new and more biologically controlled vocabulary. One limitation associated with this approach, however, is the possible effect on phylogenetic reconstruction of non-vertical evolutionary processes, such as convergent evolution and horizontal gene transfer (HGT). Because GO terms are structured as DAGs, there are *many-to-many* relationships between child and parent terms. This promiscuity can complicate attempts of ToL reconstruction. In addition, genes whose specific functions are not known can be directly assigned to higher-level GO terms without lower-level GO annotations (Rhee et al. 2008). Consequently, a higher-level GO term is the collection of both evolutionary conserved and functionally unidentified genes.

In this study, we thus restricted the analysis to include only GO terms corresponding to the terminal terms of MF (hereinafter simply referred to as  $GO_{TMF}$  terms), which are highly specialized and represent the majority of MFs of cells (Ashburner et al. 2000). In contrast, BP represents events that are outcomes of molecular activities (e.g., pyrimidine metabolic process) while CC expresses anatomical structures (e.g., ribosome), both of which carry more integrative views and are not as meaningful for evolutionary studies (Kim and Caetano-Anollés 2010). Experimentally, we sampled organisms from the three superkingdoms and counted the number of times each  $GO_{TMF}$  term was present in their functionomes, and transitively, in their associated proteomes. These ‘genomic abundance’ values serve as phylogenetic character states, characterizing the set of functionomes (taxa) that were sampled (Fig. 1). The methodology is similar to the abundance-based approach used previously to study the evolution of

protein domain structures and RNA molecules (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2011a; Harish and Caetano-Anollés 2012; Kim and Caetano-Anollés 2012; Nasir et al. 2012) and is superior to typical sequence-based approaches that are prone to phylogenetic limitations and artifacts, including problems resulting from sequence alignment such as inapplicable characters and indels that make phylogenetic analysis statistically inconsistent (Warnow 2012), mutational saturation, HGT, and violation of assumptions of character independence (Caetano-Anollés and Nasir 2012). Using this new methodology, we show that ToLs reconstructed from the genomic census of  $GO_{TMF}$  terms carry considerable predictive power in their ability to explain the origin and evolution of cellular life.

## Methods

### Data Retrieval and Manipulation

The European Bioinformatics Institute provides Gene Ontology Association (GOA) files for completely sequenced proteomes. We downloaded the GOA files (<http://www.ebi.ac.uk/GOA/proteomes>; November 2009) for a total of 1,595 organisms spanning superkingdoms: Archaea, Bacteria, and Eukarya. We filtered out proteomes that were below the 50 % coverage, with coverage defined as the number of proteins assigned to terminal  $GO_{TMF}$  terms divided by the total number of proteins in a GOA file. We also removed multiple occurrences of the same species, reducing the dataset to 638 non-redundant proteomes. To minimize sampling bias of proteomes between the three superkingdoms, we sampled only one bacterial species per genus, preferentially type strains. In the case of the other two superkingdoms, we retained all sampled proteomes without exclusion. For the remaining 358 proteomes, we studied organism lifestyles using Genomes Online Database (Liolios et al. 2010) and published data (Kim and Caetano-Anollés 2011; Nasir et al. 2011; Kim and Caetano-Anollés 2012). Out of the total 358 organisms, 249 were identified as free-living and 109 either facultative parasitic or obligate parasitic. We generated two datasets: (1) *total* with the complete set of 358 proteomes and (2) *free-living* with only 249 proteomes. We downloaded the OBO flat file from the GO database that gives the standard representation of gene ontologies (<http://www.geneontology.org/GO.downloads.shtml>; November, 2009). Out of the total 8,659 redundant MF terms that were defined in the OBO file, 1,708 were non-redundantly classified as parents and 3,396 as terminal nodes. We scanned for the presence of 3,396 terminal terms in both the *total* and *free-living* datasets. This resulted in 2,050 and 2,039  $GO_{TMF}$  terms identified in the *total* and *free-living* datasets,

respectively. Terms that were not present in the GOA files of our sampled proteomes were excluded from the analysis.

### Phylogenomic Analysis

For both the *total* and *free-living* datasets, we calculated a genomic census by counting the number of times each GO<sub>TMF</sub> term was present in every functionome. We defined this count as the ‘genomic abundance’ value (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2007). This value varies across functionomes as complex organisms encode extremely diverse and specialized functions in comparison to simple organisms. To account for the differences among functionome sizes and unequal variances, and also because most phylogenetic software allow only up to 32 character states, we normalized the genomic abundance values in an alphanumeric format from 0 to 9 and A to V using the following formula (Wang et al. 2007; Kim and Caetano-Anollés 2011).

$$g_{ab\_norm} = \text{Round} [\ln(g_{ab} + 1) / \ln(g_{max} + 1) * 31]$$

Using this formula, the genomic abundance value for each terminal GO<sub>TMF</sub> term in every functionome ( $g_{ab}$ ) is standardized by the maximum value in the matrix ( $g_{max}$ ) and normalized to a scale from 0 to 31. The result is a matrix with rows representing functionome names (taxa) and columns representing GO<sub>TMF</sub> terms (characters) with 32 possible character states (i.e., normalized abundance values) (Fig. 1). These character states are linearly ordered, carry equal weight of one, and are compatible with the phylogenetic reconstruction software PAUP\* ver. 4.0b10 (Swofford 2002). Linear ordering of character states does not violate the assumption of character polarity as changes in both directions, forward (e.g., 18–24) and reverse (e.g., 22–9), are allowed and found to be frequent when traced on the branches of ToL (Nasir et al. 2014a, b). These changes count toward tree length when maximum parsimony (MP) was used as the optimality criterion to search for the best possible tree with the minimum number of character state changes (Fig. 1). MP is the most appropriate optimality criterion for analysis of this kind since we pool the entire set of known genes into a single study. These genes are evolving with different evolutionary rates and in such instances MP is shown to give better performance than any other tree reconstruction method (Kolaczowski and Thornton 2004). Furthermore, convergence is less likely when using large number of multistate characters (Kolaczowski and Thornton 2004; Gough 2005). Trees were polarized using the ANCMETHODS command in PAUP\*, and 0 was specified as the ancestral character state. We assumed that ancient functionomes encoded only a handful of functions and progressively enriched their repertoires

along the evolutionary timeline (Kim and Caetano-Anollés 2010). Trees were rooted using the Lundberg method (Lundberg 1972) that places the root at the most parsimonious location without the need to specify the outgroup taxa (see Kim and Caetano-Anollés 2012 for methodological explanations).

The phylogenetic error (i.e., effect of non-vertical evolutionary processes such as HGT and/or convergent evolution) was estimated by calculating retention indexes ( $r_i$ ) for individual GO<sub>TMF</sub> terms using the ‘DIAG’ option in PAUP\*. The  $r_i$  indicates fit of characters to the phylogeny and is evaluated on a scale from 0 to 1 (Farris 1989). Higher  $r_i$  values indicate better fit of phylogenetic characters and thus lower probability of non-vertical inheritance. The statistical significance of differences between two distributions of  $r_i$  values was evaluated by the Student’s unpaired two-tailed  $t$  test. The reliability of the phylogenetic trees was evaluated by 1,000 non-parametric bootstrap (BS) replicates.

To measure the degree of monophyly of individual taxonomic groups on a phylogenetic tree, we calculated the genealogical sorting index (GSI) using the module *GenealogicalSorting* ver. 0.92 of the R package ver. 2.15.1 with 10,000 permuted replicates (Cummings et al. 2008). The maximum GSI value of 1 signals the complete monophyly of a given taxonomic group and values close to zero indicate increase of dispersal. Trees were visualized using Dendroscope ver. 3 (Huson et al. 2007).

### Reconstruction of rRNA Trees

We downloaded the manually curated aligned sequences of rRNA genes (16S for Archaea and Bacteria, and 18S for Eukarya) for 231 out of 249 genomes of the *free-living* dataset from the SILVA database, release 111, which are reliably curated by considering alignment quality and phylogenetic relationships (Quast et al. 2013). For the remaining 18 genomes, reliable alignments could not be extracted due to differences in naming conventions. All of the 231 rRNA sequences in the alignment were nearly complete in length (longer than 1,200 bp). The alignment file was imported into the ModelTest program (Posada and Crandall 1998) to determine the most appropriate nucleotide substitution model. Based on the results corresponding to the hierarchical likelihood ratio test,  $GTR + I + G$  was identified as the candidate model accounting for both the proportion of invariant sites and gamma-distributed rate variation (Perelman et al. 2011). Sequence alignment and model parameters were then imported into PAUP\* to reconstruct a Neighbor-Joining (NJ) tree (Saitou and Nei 1987). For individual phyla of the NJ tree, GSI values were calculated and compared with the MP trees.

## Reconstruction of Network Trees

Network diagrams are useful indicators of any conflicts that may be present in the phylogenetic model and the reconstructed trees (Bryant and Moulton 2004). These networks are also termed neighbor-nets or network trees. We generated phylogenomic networks using the Neighbor-Net algorithm implemented in the SplitsTree package ver. 4.12.6 (Huson 1998). We transformed the abundance matrices (described above) into occurrence (i.e., presence/absence) matrices for calculation of distance-based phylogenies. To evaluate the amount of ‘vertical’ phylogenetic signal present in our data, we calculated the delta ( $\delta$ ) score, a measure of the reticulation levels of networks on a scale from 0 to 1 (Holland et al. 2002). A  $\delta$ -score of 0 indicates a fully bifurcating tree while a value close to 1 means complete absence of vertical phylogenetic signal or a full network (Holland et al. 2002). Example of modern use of neighbor-nets and  $\delta$ -scores can be found in recent evolutionary studies of language (Wichmann et al. 2011) and culture (Buckley 2012).

## Enrichment Test for HGT

To quantify the degree of HGT affecting phylogenetic trees, we compared 249 *free-living* proteomes to the prokaryotic proteomes listed in the horizontal gene transfer database (HGT-DB; Garcia-Vallve et al. 2003). Only 72 out of 249 proteomes were cross-listed along with GenBank identifiers (GIs) for potential horizontally transferred proteins (HTPs). These proteins were, however, listed with their UniProtKB IDs in the corresponding GOA files. We, therefore, converted the GIs of HTPs to UniProtKB IDs using the online ID MAPPING tool of UniProt ([http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax\\_identifier.cgi](http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi)) and determined GO associations for HTPs present in these 72 proteomes. To evaluate whether a GO<sub>TMF</sub> term should be regarded as a horizontally transferred character or not, we examined how many HTPs corresponded to proteins of a GO<sub>TMF</sub> term. The degree of the association between HTPs and GO<sub>TMF</sub> terms was estimated by conducting a statistical test using the hypergeometric distribution (Forslund et al. 2008; Kim and Caetano-Anollés 2011). We compared the total number of HTPs that were associated with a GO<sub>TMF</sub> term (sample) to the total number of proteins present in the 72 proteomes regardless of HGT (background). The statistical significance of enrichment was evaluated at the 95 % confidence level and using the following equation,

$$P(X = k) = \frac{\binom{M}{k} \binom{N - M}{n - k}}{\binom{N}{n}}, \text{ where } \binom{a}{b} = \frac{a!}{b!(a - b)!},$$

where,  $k$  indicates the multiple occurrences of a GO<sub>TMF</sub> term associated with HTPs in the sample;  $n$  indicates the total redundant numbers of all GO<sub>TMF</sub> terms in the sample;  $M$  indicates the multiple occurrences of a GO<sub>TMF</sub> term associated with HTPs in the background;  $N$  indicates the total redundant numbers of all GO<sub>TMF</sub> terms in the background; and  $P(X = k)$  indicates the probability of enrichment.

## Results

### ToL Reconstructions Describe the Evolution of Functionomes

Since the physiology of an organism is defined by the biological functions of its molecular components, a tree with functionomes as taxa is the closest that is possible to a bonafide tree of organisms and a bonafide ToL. We, therefore, reconstructed rooted ToLs from genomic abundance counts of terminal GO<sub>TMF</sub> terms in the *total* (Supplementary Fig. S1) and *free-living* (Supplementary Fig. S2) datasets, which were used as phylogenetic characters. The *total* dataset included a repertoire of 2,050 GO<sub>TMF</sub> terms from 358 organisms encompassing 47 Archaea, 288 Bacteria, and 23 Eukarya with both free-living and non-free-living (i.e., parasitic/symbiotic) lifestyles. ToLs reconstructed from the *total* dataset highlighted the bias associated with the inclusion of the functionomes of organisms that were not free-living (and interact with hosts) as most of these taxa occupied the most basal positions in the tree (red squares in Supplementary Fig. S1). These taxa included notable parasites such as *Nanoarchaeum equitans* (archaeal parasite), *Guillardia theta*, (nucleomorph with a highly reduced genome; Cryptophyta, marine plankton, Eukarya), *Candidatus Carsonella rudii* ( $\gamma$ -Proteobacteria) and *Candidatus Sulcia muelleri* ( $\beta$ -Proteobacteria) (both symbionts of sap-feeding insects) (Nasir et al. 2011). In addition, functionomes from a number of bacterial parasites were clustered at the base of the bacterial group including Actinobacteria, Bacteroidetes, Chlamydia, Firmicutes, Fusobacteria, Spirochetes, and various proteobacterial symbionts (Supplementary Fig. S1). Moreover, functionomes from Eukarya did not form a distinct superkingdom but rather appeared as a subgroup derived from Bacteria, highlighting important inaccuracies in the topology of this tree.

The link between parasitism and genome reduction has been explained previously (Wang et al. 2007; Nasir et al. 2011; Kim and Caetano-Anollés 2012) and is known to bias tree reconstructions and affect the topology of ToLs (Kim and Caetano-Anollés 2011). For example, organisms

that engage in obligate parasitism can lose nearly all of their metabolic genes and depend upon the host for survival (Nasir et al. 2011). These idiosyncratic host-mediated tendencies of genome reduction (scattered in parasitic lineages throughout the ToL) affect the functional make up of proteomes and complicate phylogenetic reconstruction. They also add a bias to our evolutionary model, which based on the principle of continuity assumes that ancestral functionomes had a simpler repertoire of MFs that progressively became richer. Because parasitic/symbiotic organisms harbor highly reduced genomes, our model favored their placement at basal positions of the tree. To avoid these biases, we examined the lifestyles of the 358 organisms of the *total* dataset and excluded 109 organisms with parasitic/symbiotic lifestyles. The remaining 249 organisms harbored a functional repertoire of 2,039 GO<sub>TMF</sub> terms (Table S1 for sampled taxa). This *free-living* dataset included functionomes from 45 Archaea, 183 Bacteria, and 21 Eukarya.

ToLs reconstructed from the *free-living* dataset, now free from the effects of problematic taxa, supported the division of living organisms into three distinct superkingdoms: Archaea, Bacteria, and Eukarya (Supplementary Fig. S2). Archaeal lineages rooted the tree paraphyletically and made up the most ancient superkingdom (read below). In turn, Bacteria and Eukarya formed monophyletic groups that shared a common ancestor separated from Archaea by 89 % BS (Supplementary Fig. S2). We note that BS values depend on number of taxa and are generally expected to be low in ToLs of these sizes. This fact should be taken in consideration when evaluating the significance of phylogenetic relationships. We also note that genome reduction is not restricted to only parasitic and symbiotic organisms. Gene loss may also occur in free-living cells, albeit at lower levels. Robustness of our phylogenetic methodology against these cases is supported by the phylogenetic positions of *Pelagibacter ubique* (marine  $\alpha$ -Proteobacteria) and *Prochlorococcus marinus* (Cyanobacteria), both well-documented examples of genome reduction in free-living organisms (Dufresne et al. 2005; Giovannoni et al. 2005). Previous phylogenetic studies based on gene sequences showed that *P. ubique* and *P. marinus* were the closest to *Zymomonas mobilis* and *Synechococcus* sp., respectively. Unlike *P. ubique* and *P. marinus* (genome sizes ca. 1.3 and 1.7 Mbp, respectively), *Z. mobilis* and *Synechococcus* sp. have larger genomes (ca. 2 and 2.5 Mbp, respectively) and are free from genome reduction. Nevertheless, *P. ubique* and *P. marinus* are the closest to *Z. mobilis* and *Synechococcus* sp., respectively, in the ToLs reconstructed from both the *free-living* (Supplementary Fig. S2) and the *non-HGT* datasets (see Fig. 2). This strongly supports the claim that ToLs reconstructed using genomic abundance are robust against inclusion of reduced free-living functionomes. In fact,

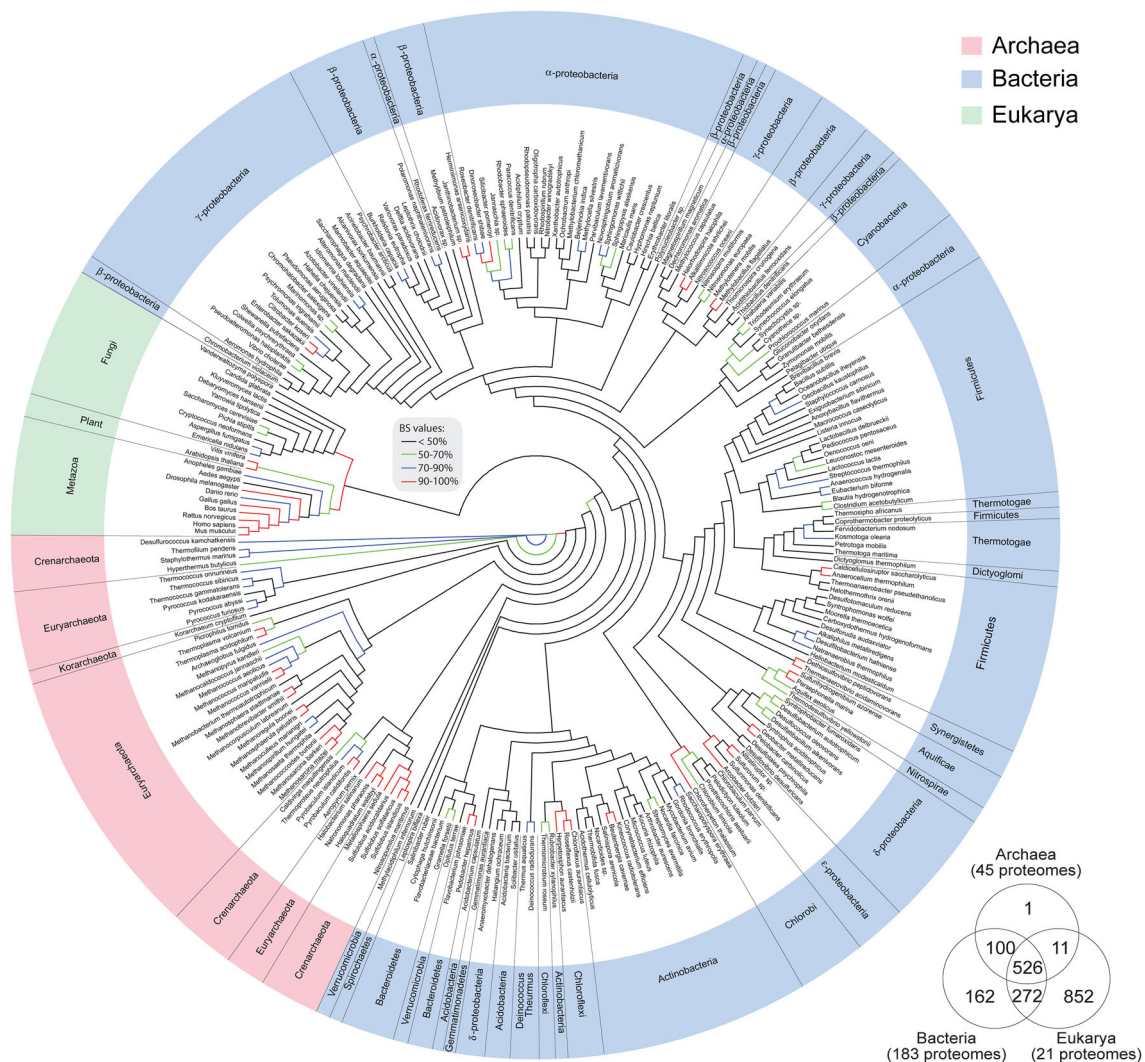
genome reduction in free-living organisms is mostly limited to auxiliary genes, still allowing most of essential genes to encode core MFs. Since functionally important genes largely represent the genomic abundance of a functionome, genome reduction of free-living organisms may result in a small decrease of their genomic abundance. Consequently, ToLs reconstructed using genomic abundance would only be marginally affected by the inclusion of reduced free-living functionomes.

#### Identification of GO<sub>TMF</sub> Terms Associated with Horizontally Transferred Proteins

To better resolve phylogenomic relationships, problematic characters that are acquired via HGT and contribute to homoplasy must be also excluded (Kim and Caetano-Anollés 2011, 2012). HGT is believed to have played an important role in microbial evolution, especially in Bacteria (Jain et al. 1999). Because the *free-living* dataset included a large number of bacterial functionomes (73 %), ToLs built from this set could also lead to confounding results. HTPs do not contribute to ‘shared and derived’ GO<sub>TMF</sub> terms, which are the backbone of vertical phylogenetic signatures, and can only add phylogenetic noise. Their exclusion is thus justified at the expense of reducing phylogenetic accuracy. To define GO<sub>TMF</sub> terms that were significantly associated with HTPs, we evaluated the enrichment of HTPs for individual GO<sub>TMF</sub> terms using the hypergeometric distribution, which was already successfully applied to evolutionary studies of this kind (Forslund et al. 2008; Kim and Caetano-Anollés 2011). We identified HTPs in 72 out of 249 free-living organisms that were cross-listed in the HGT-DB (Garcia-Vallve et al. 2003) and extracted their GO associations (Supplementary Table S2). We then compared the enrichment of these GO<sub>TMF</sub> terms (sample) to the enrichment of the rest of the GO<sub>TMF</sub> terms in the 72 functionomes (background) and evaluated statistical significance at 95 % confidence level. A total of 115 out of the 2,039 GO<sub>TMF</sub> terms were significantly associated with HTPs ( $P < 0.05$ ; Supplementary Table S3). Exclusion of these terms from the *free-living* dataset resulted in 1,924 phylogenetic characters. This new *non-HGT* dataset was used to reconstruct a ToL that described the evolution of functionomes from 249 free-living organisms and minimized the effect of HGT (Fig. 2). The new tree was mostly congruent to the tree reconstructed from the *free-living* dataset (Supplementary Fig. S2; also read below).

#### Phylogenomic Patterns

The optimized ToL generated from the *non-HGT* dataset supported the tripartite nature of the living world and monophyletic



**Fig. 2** Tree of cellular life derived from the *non-HGT* functionome dataset. One optimal most parsimonious tree describing the evolution of 249 free-living functionomes built from the genomic census of 1,924 terminal GO terms that were not influenced by HGT (1,843 parsimony informative characters; 87,897 steps; CI = 0.1342;

RI = 0.5798;  $g_1 = -0.839$ ). Terminal leaves of Archaea, Bacteria, and Eukarya were labeled in pink, blue, and green, respectively, while major phyla and domains are also identified. Edges were colored according to BS values. The Venn diagram at the bottom describes the sharing patterns of  $GO_{TMF}$  terms between the three superkingdoms

Bacteria and Eukarya, which were grouped as sister taxa (61 % BS) emerging from paraphyletic Archaea (Fig. 2). The ToL also uncovered notable phylogenomic patterns:

- (i) *A Hyperthermophilic Origin of Diversified Life in Archaea*. A closer examination of the basal archaeal lineages of the ToL with splits exhibiting 50–90 % BS showed that they embodied organisms belonging to the orders Desulfurococcales and Thermoproteales of Crenarchaeota. They included *Desulfurococcus kamchatkensis*, *Hyperthermus butylicus*, *Staphylothermus marinus*, and *Thermofilum pendens*. *Desulfurococcus* is a genus of thermophilic, organotrophic, and anaerobic archaea generally found in hyperthermic habitats

such as deep-sea thermal vents and subterranean hot springs (Ravin et al. 2009). *T. pendens* is a thermophilic and moderate acidophile archaeon isolated from a solfataric hot spring that uses sulfur and peptides as energy source (Anderson et al. 2008). *S. marinus* and *H. butylicus* are also hyperthermophile archaea belonging to the Desulfurococcales that can be sulfur reducing and live in deep-sea vents and hot solfataric floor habitats (Zillig et al. 1990; Anderson et al. 2009). While the hyperthermophilic origin of diversified life has been always associated to the rise of Bacteria, our finding that the root of the ToL lies in hyperthermophilic Archaea is very significant.

- (ii) *Cohesive archaeal orders but non-cohesive major archaeal groups.* Organisms in well-recognized archaeal orders were unified but with widely ranging supports, from well-supported clades in Halobacteria (100 % BS), Sulfolobales (98 % BS), Thermococci (89 % BS), Methanococci (82 % BS), to moderate support for the branch grouping of both Methanomicrobia and Methanobacteria (74 % BS), and Thermoplasmata (64 % BS). However, support for deeper branches unifying these orders was consistently low. We found that crenarchaeal organisms belonging to the order Sulfolobales were derived and appeared associated with Thaumarchaeota, while the rest of archaeons belonging to Euryarchaeota occupied intermediate basal positions in the tree, together with Korarchaeota.
- (iii) *A Non-Thermophilic Origin of Bacteria.* Groupings of phyla in the ToL favored the non-thermophilic origin of the bacterial superkingdom. The tree placed the anaerobic rod-shaped Bacteroidetes and some members of the PVC superphylum such as Verrucomicrobia in the most basal positions, linked to a more derived actinobacterial phylum. Some well-recognized bacterial phyla were strongly unified with good to moderate support, including the Chlorobi (100 % BS), Synergistetes (100 % BS), Chloroflexi (99 % BS),  $\epsilon$ -Proteobacteria (93 % BS), Cyanobacteria (69 % BS), and Aquificae (57 % BS), while other were unified with poor BS supports (<50 % BS), including Thermotogae,  $\delta$ -Proteobacteria and a large group of  $\gamma$ -Proteobacteria. Firmicutes appeared in more basal positions than  $\alpha$ -Proteobacteria,  $\beta$ -Proteobacteria, and  $\gamma$ -Proteobacteria, none of which formed cohesive groups. The thermophilic Aquificae and Thermotogae were quite derived when compared to organisms of the basal PVC group.
- (iv) *Monophyletic Relationships in Major Eukaryal Groups and Close Relationship Between Plants and Metazoa.* Eukarya formed a strong monophyletic group (100 % BS; Fig. 2). Metazoa, Plants, and Fungi were also monophyletic with taxa in the individual groups well positioned. Remarkably, the ToL of functionomes recovered a sister taxa relationship of Metazoa and Plants (59 % BS). At the time of the analysis, the functionomes of only two flowering plants (*Arabidopsis thaliana* and *Vitis vinifera*) with coverage of more than 50 % were available. While the sister relationship between Metazoa and Plants may be due to limited sampling of taxa, the close relationship

between the two groups was also recovered in previous evolutionary studies that focused on abundance of protein domains (Wang and Caetano-Anollés 2006; Wang et al. 2007). Recently, a ToL reconstructed from the abundance of conserved protein domains in 420 free-living organisms also identified a close relationship between Metazoa and Plants (Kim and Caetano-Anollés 2011). In this study, authors sampled a large number of eukaryal proteomes including 64 Metazoa, 44 Fungi, 16 Protista, and 9 Plants. The ToL revealed that Fungi was distant from both Metazoa and Plants, while the latter two were clustered in close proximity and separated by 5 animal-like protist proteomes. This suggests that inclusion of more eukaryal functionomes, especially of protists, can change existing deep phylogenetic relationships in Eukarya. However, the topological consistency between the functionome-based and the protein domain-based ToLs at least supports that Plants are a closer evolutionary relative of Metazoa than Fungi. It is, therefore, likely that plants and animals share physiological similarities and encode a functional apparatus that is quite similar. It would be interesting to validate this hypothesis in future studies.

#### Evaluating the Degree of Monophyly in Phylogenetic Trees

To quantify and compare the historical relationships among groups of organisms in the ToLs reconstructed from the *free-living* and *non-HGT* datasets, we calculated the degree of monophyly (GSI values) for individual groups consisting of at least five functionomes (Table 1). Six out of 14 groups of the ToL reconstructed from the *non-HGT* dataset (including Crenarchaeota, Actinobacteria, and all proteobacterial phyla) had larger GSI values than the one reconstructed from the *free-living* dataset. In turn, only two phyla (Euryarchaeota and Bacteroidetes) exhibited larger GSI values in the ToL reconstructed from the *free-living* dataset. In case of the remaining six groups (Chlorobi, Cyanobacteria, Firmicutes, Thermotogae, Fungi, and Metazoa), both trees showed the same degree of monophyly. Since HGT occurrences in proteobacterial genomes are very common (Kloesges et al. 2011), increased GSI values of proteobacterial phyla in the ToL derived from the *non-HGT* dataset indicated that the exclusion of HTPs characters increased significantly the accuracy of phylogenetic statements despite reducing cladistic information.

Because rRNA genes are highly conserved and commonly used in sequence-based phylogenies, we also



**Table 1** Measuring the degree of monophyly with the Genealogical Sorting Index (GSI)

	Superkingdom	Phylum (no. proteomes)	<i>free-living</i>	<i>non-HGT</i>	<i>info</i>	<i>rRNA</i>
The GSI values and significance levels with 10,000 permuted replicates were examined for phyla having at least five proteomes	Archaea	Crenarchaeota (16)	0.66**	0.80**	0.63**	0.60**
		Euryarchaeota (28)	0.86**	0.70**	0.77**	0.92**
	Bacteria	Actinobacteria (17)	0.83**	0.88**	0.78**	0.87**
		Bacteroidetes (6)	1.00**	0.83**	0.13*	0.28**
		Chlorobi (5)	1.00**	1.00**	1.00**	1.00**
		Cyanobacteria (6)	1.00**	1.00**	0.17**	1.00**
		Firmicutes (33)	0.82**	0.82**	0.49**	0.72**
		$\alpha$ -Proteobacteria (31)	0.62**	0.66**	0.69**	0.70**
		$\beta$ -Proteobacteria (18)	0.36**	0.48**	0.53**	0.87**
		$\gamma$ -Proteobacteria (27)	0.59**	0.74**	0.69**	0.41**
		$\delta$ -Proteobacteria (11)	0.48**	0.51**	0.13*	1.00**
		Thermotogae (5)	0.80**	0.80**	0.66*	1.00**
	Eukarya	Fungi (10)	1.00**	1.00**	0.68**	0.21*
		Metazoa (9)	1.00**	1.00**	0.79**	0.56**

\*  $P < 0.05$ ; \*\*  $P < 0.01$

compared the degree of monophyly of the *non-HGT* tree to the NJ tree reconstructed from 16S and 18S rRNA gene sequence alignment. A comparison of GSI values showed that groups in the *non-HGT* tree were generally better supported (Table 1). Overall, seven out of 14 groups had higher GSI values in the *non-HGT* tree compared to the rRNA tree including, Crenarchaeota (0.80 vs. 0.60), Actinobacteria (0.88 vs. 0.87), Bacteroidetes (0.83 vs. 0.28), Firmicutes (0.82 vs. 0.72),  $\gamma$ -Proteobacteria (0.74 vs. 0.41), Fungi (1.00 vs. 0.21), and Metazoa (1.00 vs. 0.56). These included both the very basal (e.g., Crenarchaeota) and derived (e.g., Fungi and Metazoa) branches of the ToL. In contrast, rRNA tree performed poorly in resolving the very derived branches of Fungi (GSI = 0.21) and Metazoa (GSI = 0.56). Five out of 14 groups had higher GSI values in the rRNA tree and included proteobacterial phyla [ $\alpha$ -Proteobacteria (0.70 vs. 0.66),  $\beta$ -Proteobacteria (0.87 vs. 0.48),  $\delta$ -Proteobacteria (1.00 vs. 0.51)], Thermotogae (1.00 vs. 0.80), and Euryarchaeota (0.92 vs. 0.70). Chlorobi and Cyanobacteria had GSI value of 1.00 in both trees.

This exercise revealed that the *non-HGT* tree performed superior to the rest of the reconstructed trees and that the use of GO definitions as phylogenetic characters served better in resolving monophyletic relationships. We argue that trees built from the entire functional toolkit (e.g., *free-living*, *non-HGT*) are more powerful in charting organismal relationships than those built from limited character sets (e.g., *info* tree; read below) or a single molecule (rRNA tree), especially when considering that the entire functional apparatus of an organism approximates the physiology of that organism and truly depicts a ToL. In contrast, rRNA represents only one of the three classes of rRNA molecules that make structural components of ribosomes and does not represent the entire evolutionary history of an organism

(see Discussion). Therefore, inferences regarding entire systems (i.e., organisms) should include all the individual components that make up that system (i.e., protein domains, functional definitions) rather than only a single (albeit ancient and central) molecule. Hence, from hereinafter, we will only focus on elaborating phylogenies resulting from the census of MFs as they allow to make systemic comparisons among organisms and enable the evolutionary study of organisms as biological systems.

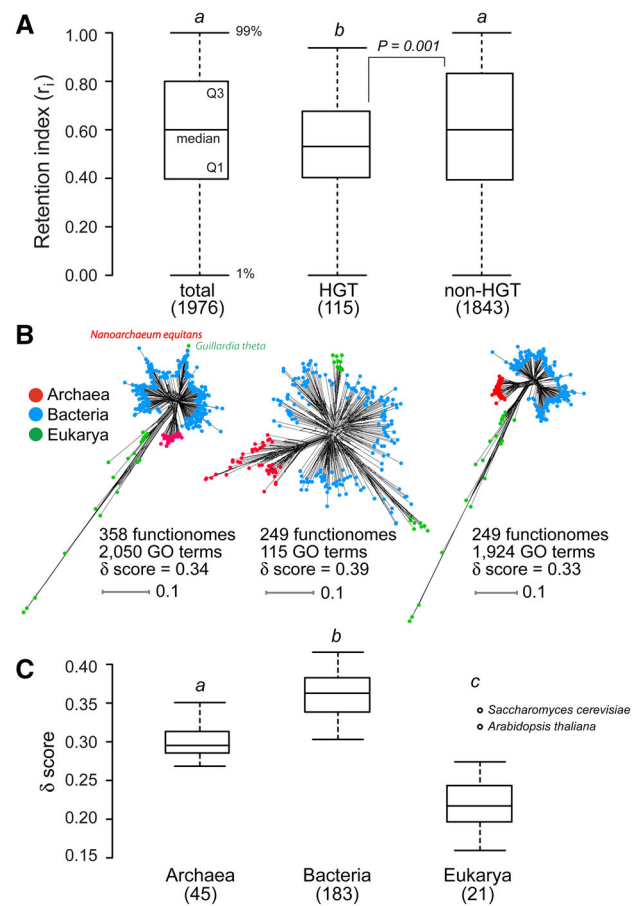
#### Exclusion of Problematic Taxa and Horizontally Acquired Characters Improved Phylogenetic Reconstructions

We inspected the reliability of phylogenetic trees recovered from the census of MFs by selecting only 120 GO<sub>TMF</sub> terms that were involved in informational processes, including transcription and translation. This character set was used to build a new ToL. It has been proposed that information-related genes are refractory to the effects of HGT (Jain et al. 1999). A single most parsimonious tree reconstructed from the limited set of informational GO<sub>TMF</sub> terms in the 249 free-living functionomes was largely congruent with the ToL reconstructed from the *non-HGT* dataset (Supplementary Fig. S3). This tree also favored the groupings of organisms into three superkingdoms and was rooted paraphyletically in Archaea. While Korarchaeota clustered with the eukaryal clade, the tree fared well in terms of overall groupings among phyla (Supplementary Fig. S4). The number of monophyletic phyla recovered was, however, lower than the number recovered in the *non-HGT* ToL (Supplementary Fig. S4). Furthermore, only three phyla of the tree of information processes (i.e., Euryarchaeota,  $\alpha$ -Proteobacteria,  $\beta$ -Proteobacteria) had larger

GSI values than the *non-HGT* ToL (Table 1). One explanation is the lesser number of phylogenetic characters used to reconstruct the tree (120 versus 1,924). In general, using large number of characters improve phylogenetic reconstruction (Dopazo et al. 2004; Delsuc et al. 2005). To test this, we extracted 1,000 random samples each consisting of 120 GO<sub>TMF</sub> terms from the 1,843 parsimoniously informative non-HGT characters and generated 1,000 trees. We noted that most of the random *non-HGT* ToLs still had more monophyletic phyla compared to the tree of information processes (data not shown). It is, therefore, desirable to generate trees from the entire functional toolkit and not just a specific functional repertoire, as explained above.

To further investigate the reliability of phylogenetic reconstructions, we compared the  $r_i$  distributions of ToLs recovered from functionomic data (Fig. 3a). These included trees derived from the *total* (all taxa and characters included), *HGT* (only 115 HTP-derived GO terms included), and *non-HGT* (both the problematic taxa and characters excluded) datasets. In general, higher  $r_i$  values support better fit of phylogenetic characters to the phylogeny and thus lower probability of non-vertical inheritance. The boxplots indicated that the best trees were recovered using the *non-HGT* dataset (Fig. 3a), supporting previous results. In contrast, *HGT* trees indicated the worst fit and were on average distributed with the lowest  $r_i$  values. A comparison between the *HGT* and *non-HGT* trees was statistically significant ( $P < 0.05$ ) (as expected) (Fig. 3a) suggesting that any confounding effects resulting from HGT were controlled in the *non-HGT* trees.

Finally, we confirmed the validity of our MP-based ToLs and tested for any conflicts between our evolutionary model and phylogenomic trees by reconstructing phylogenomic networks. Our phylogenomic model assumes that functionomes became progressively more complex; i.e., we consider gene gain and loss, gene rearrangements, and gene duplications to be the major evolutionary forces shaping the functionomes of living organisms (Wang and Caetano-Anollés 2006, 2009; Kim and Caetano-Anollés 2012). When the phylogeny involves complex evolutionary processes, a more abstract network representation can be used to test any conflicts between the model and the tree (Bryant and Moulton 2004). Phylogenomic networks generated from the occurrence data (i.e., presence or absence of GO<sub>TMF</sub> terms) for the *total*, *HGT*, and *non-HGT* datasets validated the *non-HGT* dataset and highlighted important shortcomings of the *HGT* dataset (Fig. 3b). Phylogenomic networks generated from the *total* dataset included archaeal and eukaryal parasites (*Nanoarchaeum equitans* and *Guillardia theta*) that were clustered within Bacteria clearly suggesting a revision of the evolutionary model. In contrast, the *non-HGT* network supported the three-superkingdom classification system with no contamination of taxa (Fig. 3b). Finally, the *HGT*



**Fig. 3** Reliability of phylogenomic trees and the evolutionary model. **a** Boxplots comparing the fit of characters between trees reconstructed using various character sets. Mean, median, and quartiles are identified.  $P$  values are indicated for individual comparisons. Numbers in parenthesis represent total number of parsimony informative characters for which  $r_i$  values were available. Boxplots headed by different letter are significantly different. Statistical significance was evaluated using Student’s unpaired two-tailed  $t$  test at 95 % confidence level. **b** Phylogenomic networks generated for *total*, *HGT*, and *non-HGT* datasets. Terminal nodes of Archaea, Bacteria, and Eukarya were labeled in red, blue, and green, respectively. **c** Boxplots comparing the distribution of  $\delta$ -scores in the three superkingdoms. Outliers are labeled. Numbers in parenthesis indicate total number of free-living functionomes in each group. All the comparisons are significant at 0.05

network constructed from the 115 HTP-derived GO terms failed to re-enact a tree-like structure with true bacterial and eukaryal groupings and showed that the HTP-derived GO terms did not complicate archaeal relationships (Fig. 3b). This was a significant result and raised important questions. First, it questioned the existence of pervasive HGT within and between microbes. Second, it showed that the exclusion of HGT-derived GO terms significantly improved the phylogenies of *non-HGT* dataset. Third, it challenged the existence of fundamental organismal fusions used to explain evolutionary reticulation. All of these observations are significant and mandate future investigation.

To test if the poor resolution of the *HGT* network was not due to the limited number of phylogenetic characters used for its reconstruction, we randomly sampled 115  $GO_{TMF}$  characters from the *non-HGT* dataset and prepared 1,000 random files for network analysis. We discovered that the majority of the random networks partitioned the organisms into three unified groups and did not suffer from limited sampling (Supplementary Fig. S5). Thus in light of our results, the poor resolution of the *HGT* network should be considered significant. To identify taxa that were contributing to reticulation patterns in the networks, we calculated  $\delta$ -scores for individual phyla and superkingdoms. The  $\delta$ -distribution is shown for the *non-HGT* network that revealed interesting but expected patterns (Fig. 3c). Both microbial superkingdoms were distributed with high  $\delta$ -values with scores ranging from 0.27 to 0.35 in Archaea and 0.30–0.42 in Bacteria. In contrast, the contribution to genetic exchange of eukaryal functionomes appeared minimal (0.16–0.34) (Fig. 3c). All the comparisons were statistically significant at 95 % confidence level and suggested that the rates of non-vertical evolutionary processes or HGT varied significantly between superkingdoms. The degree of reticulation in superkingdoms increased in the order Eukarya, Archaea, and Bacteria (Fig. 3c), suggesting a similar trend for the HGT correlate. The lowest  $\delta$ -score averages were observed in mammals and primates ( $\delta = 0.16$ – $0.17$ ) in Eukarya, Methanococci, Methanobacteria, and Thermococci ( $\delta = 0.28$ – $0.29$ ) in Archaea, and Thermotogae and Dictyoglomi ( $\delta = 0.31$ – $0.32$ ) in Bacteria. A comparison of  $\delta$ -scores for the different bacterial groups confirmed that the majority of the major bacterial taxonomic groups (e.g., Gemmatimonadetes, Verrucomicrobia, Bacteroidetes, Acidobacteria, and others; Table 2) were the largest contributors to genetic exchange. In contrast, eukaryal superkingdoms appeared to be best supported in the ToLs with lowest  $\delta$ -scores. Finally, archaeal phyla were supported with intermediate values (Table 2). The overall  $\delta$ -score for the *non-HGT* network was 0.33, in comparison to 0.34 for the *total* network and 0.39 for the *HGT* network, clearly identifying *non-HGT* networks and trees to be best resolved.

These experiments revealed that the ToL derived from the *non-HGT* dataset reflected phylogenomic relationships most accurately. This dataset is free from the effects of parasitic organisms and is minimally affected by non-vertical evolutionary processes. We conclude by mentioning that our phylogenomic approach is robust against unequal sampling of proteomes per superkingdoms, which can lead to incorrect parsimonious trees due to long-branch attraction (Kim and Caetano-Anollés 2011). Therefore, the relatively large number of bacterial proteomes in the *non-HGT* dataset (once the HGT-

**Table 2** A comparison of average  $\delta$ -scores in major taxonomic groups of superkingdoms

Classification	Superkingdom	No. of taxa	$\delta$ -score
Euryarchaeota-Methanococci	Archaea	4	0.28
Euryarchaeota-Methanobacteria	Archaea	3	0.29
Euryarchaeota-Thermococci	Archaea	6	0.29
Crenarchaeota-Sulfolobales	Archaea	4	0.29
Crenarchaeota-Thermoproteales	Archaea	5	0.30
Euryarchaeota-Methanomicrobia	Archaea	9	0.30
Crenarchaeota-Desulfurococcales	Archaea	4	0.31
Euryarchaeota-Archaeoglobi	Archaea	1	0.31
Euryarchaeota-Thermoplasmata	Archaea	3	0.32
Euryarchaeota-Methanopyri	Archaea	1	0.32
Korarchaeota	Archaea	1	0.34
Euryarchaeota-Halobacteria	Archaea	3	0.34
Thaumarchaeota	Archaea	1	0.35
Thermotogae	Bacteria	5	0.31
Dictyoglomi	Bacteria	1	0.32
Synergistetes	Bacteria	2	0.33
Firmicutes	Bacteria	33	0.34
Nitrospirae	Bacteria	1	0.35
$\beta$ -Proteobacteria	Bacteria	18	0.35
Chlorobi	Bacteria	5	0.35
Aquificae	Bacteria	3	0.36
$\alpha$ -Proteobacteria	Bacteria	31	0.36
$\gamma$ -Proteobacteria	Bacteria	27	0.36
Deinococcus-Thermus	Bacteria	2	0.37
$\epsilon$ -Proteobacteria	Bacteria	4	0.37
Cyanobacteria	Bacteria	6	0.37
$\delta$ -Proteobacteria	Bacteria	11	0.38
Chloroflexi	Bacteria	4	0.38
Spirochetes	Bacteria	1	0.38
Actinobacteria	Bacteria	17	0.38
Acidobacteria	Bacteria	3	0.39
Bacteroidetes	Bacteria	6	0.39
Verrucomicrobia	Bacteria	2	0.39
Gemmatimonadetes	Bacteria	1	0.39
Chordata-Mammals	Eukarya	3	0.16
Chordata-Primates	Eukarya	1	0.17
Fungi-Basidiomycota	Eukarya	1	0.20
Chordata-Birds	Eukarya	1	0.22
Chordata-Fish	Eukarya	1	0.23
Arthropoda	Eukarya	3	0.23
Fungi-Ascomycota	Eukarya	9	0.24
Plants-Streptophyta	Eukarya	2	0.28

derived characters are excluded) is not expected to bias phylogenomic relationships, as reported previously (Kim and Caetano-Anollés 2011).

## GO Coverage Does Not Bias Phylogenetic Relationships

In this study, we included only organisms with functionomes that provided at least 50 % coverage of MFs. We note that many of the sampled functionomes were annotated in reference to the experimentally verified GO annotations in few model organisms. Thus, large GO coverage differences in functionomes could reflect the similarity of functionomes to model organisms and thus bias the phylogenetic relationships. However, the functionomes we sampled had a mean GO coverage of 59.23 % and a standard deviation of 5 % (Supplementary Fig. S6; Table S1). The small variance indicated that the distribution of GO coverage was quite even across functionomes. Furthermore, the coverage of most model organisms (e.g., *Homo sapiens* of 62 %, *A. thaliana* of 51 %, *Mus musculus* of 67 %, *Drosophila melanogaster* of 65 %, etc.) was quite similar to the mean and within the upper and lower whiskers of the GO distribution (Supplementary Table S1). There were only few outliers: *Saccharomyces cerevisiae* (82 % in GO coverage), *Rattus norvegicus* (78 %), and *Gallus gallus* (78 %) (Supplementary Fig. S6). These results indicate that the degree of GO annotation for non-model organisms is comparable to that for model organisms. In other words, the GO coverage of the functionomes we sampled shows that electronic GO annotations (mostly for non-model organisms) are quite saturated in comparison to experiment-based annotations (mostly for model organisms).

Although the GO coverage of most functionomes was close to the mean coverage, we observed that few taxonomic groups were associated with relatively large variance of coverage across the three superkingdoms. Remarkably, the functionomes of these taxonomic groups were still grouped together in the *non-HGT* tree (Fig. 2). For example, three *Pyrococcus* functionomes (*P. abyssi*, *P. furiosus*, and *P. kodakaraensis*) that had 63, 58, and 52 % coverage, respectively (Supplementary Table S1), were clustered monophyletically as a single genus (Fig. 2). A more extreme case in Eukarya was the monophyletic clade of *S. cerevisiae* and *Pichia stipitis* that belongs to Saccharomycetaceae. Although the GO coverage of the two species was significantly different (82 % for *S. cerevisiae* and 55 % for *P. stipitis*), they were still clustered together in the *non-HGT* tree (Fig. 2). In Bacteria, previous phylogenetic studies have supported the strong monophyly of Cyanobacteria. Remarkably, all six cyanobacterial functionomes with GO coverage ranging from 51 to 57 % grouped together. Based on the evidence from balanced distributions of GO coverage and phylogenetic groupings of closely related taxa with large variance GO coverage, we conclude that the extent of GO annotations did not

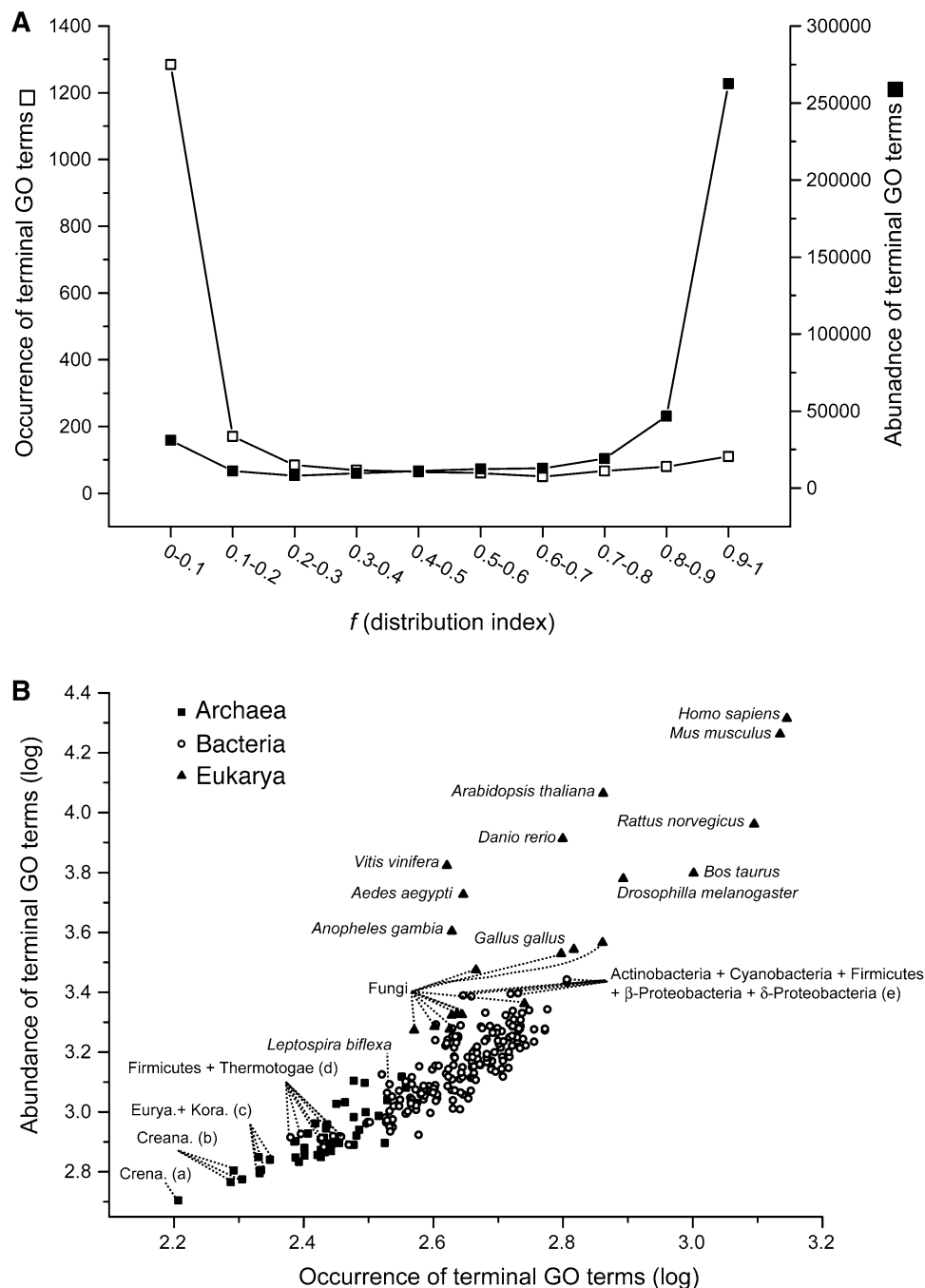
significantly affect positioning of organisms in the ToL. Instead, our previous phylogenetic experiments showed that tree topologies of MFs largely depend on how differently individual GO terms are assigned to a functionome but not on how many GO annotations are assigned to a functionome (Kim and Caetano-Anollés 2010). This implies that GO coverage had minimal effect on phylogenetic placements.

## Interplay Between Genomic Abundance and Occurrence

Different evolutionary forces are responsible for the accumulation of functions in genomic repertoires (Wang and Caetano-Anollés 2006; Wang et al. 2007), including gene duplications, gene rearrangements, and HGT. These events lead to a direct increase in the genomic abundance of genes and corresponding MFs (Kim and Caetano-Anollés 2012). Abundance is, therefore, a naturally occurring biological process that is valuable for reconstructing phylogenies (Caetano-Anollés et al. 2009; Caetano-Anollés and Nasir 2012). In contrast, occurrence-based approaches involve non-redundant representations of genes (and their functions) that generally result in more balanced topologies (Yang and Bourne 2009). We observed that both abundance and occurrence of  $GO_{TMF}$  terms were correlated and resulted in congruent classifications (Fig. 4). For instance, plotting occurrence and abundance of  $GO_{TMF}$  terms against their distribution in proteomes (distribution index or  $f$ -value = number of functionomes encoding a  $GO_{TMF}$  term/total number of functionomes) revealed interesting relationships (Fig. 4a).

The majority of the  $GO_{TMF}$  terms ( $\sim 1,300$  or  $>60$  %) were not conserved across taxa ( $f < 0.1$ ) but were distributed with low abundance values ( $\sim 200$ /functionome). These terms represent MFs that are relatively new to the functional toolkits of proteomes and are not universally distributed. They also correspond to organism-specific functions that have been acquired late in evolution. In contrast,  $GO_{TMF}$  terms that were universally present ( $0.9 < f < 1.0$ ) were very few in number ( $\sim 185$ ) but had the highest abundance values ( $\sim 25,000$ /functionome) (Fig. 4a). These terms represent ancient MFs that are vital for cellular life and are conserved across most taxa (e.g., ATPase activity, helicase activity). Excluding the two extremes (i.e., most recent and most ancient) resulted in both abundance and occurrence being evenly distributed and showed, there was no bias favoring one or the other. This analysis supported our choice to study the terminal terms that provided very high resolution for differentiation of organismal relationships.

When plotted individually for each functionome, we found a strong correlation between the two concepts



**Fig. 4** Relationship between abundance and occurrence. **a.** Abundance and occurrence values for terminal GO terms plotted against the distribution index ( $f$ ; number of functionomes encoding a  $GO_{TMF}$  term/total number of functionomes). **b.** Abundance and occurrence counts plotted against each other for a number of functionomes. Both values are positively correlated. Axes are in logarithmic scale. (a) Crenarchaeota (*Desulfurococcus kamchatkensis*); (b) Crenarchaeota (*Hyperthermus butylicus*, *Thermofilum pendens*, *Staphylothermus marinus*); (c) Eutyarchaeota (*Thermococcus onnurineus*, *Thermoplasma acidophilum*, *Thermoplasma volcanium*) and Korarchaeota

(*Korarchaeum cryptofilum*); (d) Firmicutes (*Anaerococcus hydrogenalis*, *Eubacterium bifforme*, *Pediococcus pentosaceus*, *Lactobacillus delbrueckii*, *Oenococcus oeni*, *Streptococcus thermophilus*, *Coprothermobacter proteolyticus*, *Leuconostoc mesenteroides*, *Macrococcus caseolyticus*) and Thermotogae (*Thermosiphon africanus*, *Kosmotoga olearia*, *Fervidobacterium nodosum*, *Thermotoga maritima*); (e) Actinobacteria (*Streptomyces avermitilis*, *Saccharopolyspora erythraea*), Cyanobacteria (*Anabaena variabilis*), Firmicutes (*Brevibacillus brevis*),  $\beta$ -Proteobacteria (*Ralstonia eutropha*), and  $\delta$ -Proteobacteria (*Haliangium ochraceum*)

(Fig. 4b). Organisms followed a trend from simplicity toward complexity in biological organization, beginning with the simplest functionomes of Archaea, closely followed by a diverse range of bacterial and eukaryal functionomes (Fig. 4b) and ending with the extraordinarily rich functionomes of *Homo sapiens*, *Mus musculus*, *A. thaliana*, *Rattus norvegicus*, *Danio rerio*, *Bos taurus*, and *Drosophila melanogaster*, which appeared as outliers but were distinguished by both abundance and occurrence parameters. This result is in line with a similar analysis of protein domain abundance and occurrence (Wang et al. 2007; Kim and Caetano-Anollés 2012). We conclude that genomic abundance and occurrence are positively correlated and that using abundance enhances deep phylogenetic signal (Caetano-Anollés and Nasir 2012) in the study of MFs.

## Discussion

### A New ToL with Taxa that Better-Depicts the Physiology of Organisms

Using an atypical application to a well-established cladistic methodology, here, we reconstructed rooted ToLs without the use of outgroups directly from a genomic census of biological functions. These trees are unprecedented. They describe the evolution of entire repertoires of MFs and have an evolutionary arrow built into their driving evolutionary model. This is highly significant. Thus far, ToLs are extrapolations of molecular trees that rest on the assumption that the essence of an organism can be appropriately depicted by a single molecule or a repertoire of molecules that are hopefully minimally affected by HGT (Pace 2009). In particular, the small subunit of rRNA has been used as gold standard despite of representing only one of three RNA subunits that typically, and together with dozens of ribosomal proteins, make up the ribosomal ensemble. The finding that rRNA coevolves with ribosomal proteins and that the ribosome is younger than tRNA and important enzymes (e.g., aminoacyl-tRNA synthetases) and has a protracted history (Harish and Caetano-Anollés 2012) complicates the arguments of the evolutionary centrality of one or a set of its components and the functional link between the ribosome and the organism. Instead, the functionome, suitably defined by ontological terms, approaches the entire collection of functions of an organism and is, therefore, unbiased by preconceptions on molecular biology and biochemistry. The abundance-based approach also shields deep phylogenomic relationships of functionomes from the effect of HGT and functional recruitment (Kim and Caetano-Anollés 2010). Functions that are laterally transferred or are recruited must be first

fixed and then amplified to high levels in genomic evolution if they are to have an impact on the deep branches of the ToLs. In other words, HGT or recruitment of functions that are abundant and are ancient will have little impact on the basal topologies of the trees. In contrast, small changes in genomic abundance of functions that are rare, of recent ancestry and specific to selected lineages can only significantly affect very derived branches of the trees. This and other properties of the new reconstruction method makes trees of functionomes excellent complements to trees of molecules derived from sequence analysis, which perform best when comparing closely related organisms.

We confirmed the validity of our phylogenomic statements by comparing the degree of monophyly with the canonical reference tree, building distance-based phylogenomic networks, excluding problematic taxa and HTP-linked characters, and evaluating phylogenetic reticulation due to non-vertical evolutionary processes such as HGT, endosymbiosis, and recruitment. Remarkably, we observed cohesiveness and robustness of Archaeal relationships in phylogenomic networks that question the idea that HGT between microbes (e.g., between Archaea and Bacteria) occurs at dramatically high levels (Doolittle 1999; Koonin et al. 2001) and challenges the fusion model for the origin of eukaryotes that attributes the origin of Eukarya to a primordial fusion event between archaeal and bacterial cells (Vesteg and Krajcovic 2008 and references therein). For example, reticulation measures in networks ( $\delta$ -score) showed minimal reticulation in Eukarya, intermediate levels in Archaea, and as expected, maximal reticulation impact in Bacteria (Fig. 3c). However, reticulation levels of some euryarchaeal (e.g., Methanococci and Methanobacteria) and crenarchaeal (e.g., Sulfolobales and Thermoproteales) orders in Archaea were not so far away from reticulation levels in plants, and reticulation of several bacterial orders such as Thermotogae, Firmicutes, and Chlorobi were comparable to average levels of archaeal reticulation (e.g., in crenarchaeal orders; Table 2). In particular,  $\gamma$ -Proteobacteria harbor species that exhibit unprecedented HGT levels, such as *Shewanella baltica*, which exchanges up to 20 % of their entire core and auxiliary genome in short time frames (Caro-Quintero et al. 2011). These processes of rapid adaptation through massive acquisition of genes, which are common in the ocean (McDaniel et al. 2010) and in other aquatic environments (Aminov 2011), are not reflected in the  $\delta$ -scores of the *Shewanella* genus (e.g., *S. putrefaciens*;  $\delta = 0.34$ ) or the  $\gamma$ -proteobacterial order ( $\delta = 0.36$ ), which are comparable to those of *A. thaliana* ( $\delta = 0.32$ ), *S. cerevisiae* ( $\delta = 0.34$ ), and other eukaryotes. All of these results challenge the perception that reticulation and its HGT correlate is rampant in the long-term evolution of microbes.

## The Early Thermophilic Origin of Archaea

ToLs generated from the genomic census of MFs supported the view that Archaea was the first cellular superkingdom to appear in evolution (Fig. 2; Supplementary Fig. S2). The archaeal rooting of the ToL has been recovered previously in a number of studies where the focus was on building reliable phylogenies using conserved structural information in protein and nucleic acid molecules (Wang et al. 2007; Kim and Caetano-Anollés 2012). ToLs built from proteomic abundance of domain structure and organization defined at different levels of structural conservation of the structural classification of proteins (SCOP) (Murzin et al. 1995) and CATH (Orengo et al. 1997) classifications consistently displayed a paraphyletic rooting in Archaea (Wang et al. 2007; Wang and Caetano-Anollés 2009; Kim and Caetano-Anollés 2011; Nasir et al. 2012; Bukhari and Caetano-Anollés 2013). Similar results were obtained when building trees of RNA molecules from nucleic acid structure in 5S rRNA (Sun and Caetano-Anollés 2009) and RNase P RNA (Sun and Caetano-Anollés 2010) and from nucleic acid sequence and structure in tRNA (Sun and Caetano-Anollés 2008a). More importantly, timelines of accretion of helical RNA substructures of tRNA (Sun and Caetano-Anollés 2008b) and 5S rRNA (Sun and Caetano-Anollés 2009) uncovered two accretion pathways, one specific to Archaea and the other common to Bacteria and Eukarya.

Timelines of accretion in RNase P RNA showed that the most ancient substructures were universal and harbored the core catalytic activities of the endonuclease (Sun and Caetano-Anollés 2010). However, the first RNase P RNA substructures that were lost were specific to Archaea and this episode occurred before molecules were accessorized with superkingdom-specific substructures (Sun and Caetano-Anollés 2010). Evolutionary timelines of protein domain appearance in the protein world also showed the early loss of domains in Archaea prior to the appearance of superkingdom-specific domain structures in the analysis of domain and domain interactome evolution (Wang et al. 2007; Wang and Caetano-Anollés 2009; Kim and Caetano-Anollés 2011; Nasir et al. 2012). In fact, a phylogenetic tree reconstructed using 1,924 GO<sub>TMF</sub> terms as taxa and 249 functionomes from free-living organisms as characters (*non-HGT* dataset) identified both the very ancient and derived GO<sub>TMF</sub> terms (Nasir et al. 2014b). In this study, most of the very ancient GO<sub>TMF</sub> terms were only detected in the bacterial and eukaryal functionomes, but were completely absent in Archaea. While it can be argued that loss of ancient GO<sub>TMF</sub> terms in Archaea could be a very recent event, the scenario does not seem very likely. This is because a single molecular activity is a product of multiple genes that have accumulated over the course of evolution. These genes multiply and increase

their abundance in cells with the progression of time. Thus, losing an *ancient* MF *late* in evolution is more costly than losing it earlier in evolution when genes have low abundance levels. In light of these considerations, we propose that genome reduction in thermophilic archaeal species was likely an ancient event that started very early in evolution and before the divergence of Bacteria and Eukarya. In comparison, the alternative scenario is not well supported by the distribution of conserved protein structures (Wang et al. 2007; Kim and Caetano-Anollés 2012) and MFs (Nasir et al. 2014b) in the proteomes and functionomes of contemporary organisms and is, therefore, less likely. Moreover, the paraphyletic archaeal root of the ToL has also been suggested by early studies of interparalog distances of tRNA paralogs (alloacceptors) and paralogous pairs of aminoacyl-tRNA synthetases, which depend on intraspecies comparisons and are, therefore, intrinsic to each species (Xue et al. 2003, 2005). These findings were further supported by additional polyphasic evidence (Di Giulio 2007; Wong et al. 2007) and have been recently reviewed (Wong 2014). It is striking that the study of tRNA and proteomes and their associated functions placed LUCA amongs the hyperthermophilic Archaea. Congruency provides strong backing to Archaea representing the first domain of diversified life. However, the primitivity of Crenarchaea is favored in the study of functionomes, whereas the primitivity of the junction between Euryarchaea and Crenarchaea in close proximity to *Methanopyrus* is favored in RNA studies.

The paraphyletic rooting of the ToL in Archaea is in striking disagreement with for example the canonical rooting in Bacteria that is achieved by the use of protein paralogs as mutual outgroups for central proteins such as aminoacyl-tRNA synthetases, elongation factors (e.g., EF-Tu/EFG), ATPases, carbamoyl phosphate synthetases, and signal recognition particle proteins (reviewed in Gogarten and Olendzenski 1999). These paralogous rootings, however, are considered weak because of a number of problems and artifacts of sequence analysis (e.g., long-branch attraction, mutational saturation, taxon sampling, HGT, hidden paralogy, historical segmental gene heterogeneity) and because they depend on the history of a small set of molecules out of the entire molecular repertoire of the cell. Distance-based approaches have also been used to build universal network trees from gene families defined by reciprocal best BLAST hits, which showed a midpoint rooting of the ToL between Bacteria and Archaea (Dagan et al. 2010). However, this rooting involves a complex optimization of path lengths in the split networks and critically assumes that lineages evolve at roughly similar rates. This diminishes the confidence of rootings of this kind, especially when considering the uncertainties of distances inferred from BLAST analyses and the fact that

domains in genes hold different histories and rates of change. In fact, current approaches to rooting of molecular sequences bring almost insurmountable complexities that require novel conceptual frameworks, such as critical analysis of major evolutionary transitions (e.g., ‘transition analysis’; Cavalier-Smith 2002) to establish polarity of change (Baptiste and Brochier 2004) or the analysis of genomic insertions and deletions that are rare in paralogous gene sets (Lake et al. 2009). However and as we commented above, the use of molecular sequence is problematic on many grounds, especially mutational saturation, violation of character independence by the mere existence of atomic structure, and different historical signatures in domains of multidomain proteins (Caetano-Anollés and Nasir 2012; Caetano-Anollés et al. 2014). Similarly, establishing the validity of evolutionary transitions in polarization schemes can also be problematic and requires well-grounded assumptions for each transition that is used (Baptiste and Brochier 2004). Remarkably, the assumptions of the intrinsic rooting scheme of MFs that we here present are supported by timely successions of major evolutionary transitions that increased biological complexity (Szathmari and Smith 1995) and information transmission (Jablonka and Lamb 2006) when these transitions are mapped along a timeline of MFs (see Fig. 4 in Kim and Caetano-Anollés 2010).

The rooting of the tree of cellular life in Archaea is paraphyletic and requires explanation. While paraphyly could result from loss of phylogenetic signal or from primordial homoplasy-generating processes operating during the early differentiation of superkingdoms, trees are particularly well supported at their base and the paraphyletic rooting is congruently obtained in different studies employing a diverse set of phylogenetic characters, from ontological terms to tRNA molecules. Thus, a more plausible explanation is that the early diversification of LUCA involved spatial colonization of uncharted environments that were ecologically unique to the individual primordial lineages (Sun and Caetano-Anollés 2009). This colonization was followed by selective reductive loss of genomic components (Wang et al. 2007) as the emerging archaeal lineages adapted to the different (initially auxinic) ocean and land environments of the late Archaeon. This divergence-by-isolation scenario explains patterns of loss and gain of molecular structures and their associated functions in evolutionary timelines (Kim and Caetano-Anollés 2010, 2012), which are for example responsible for delimiting the three evolutionary epochs proposed by Wang et al. (2007): (i) an early architectural diversification epoch in which ancient molecules and their functions emerged and accumulated in proteomes as cells of a communal global ancestor became modularized into individual entities, (ii) a superkingdom specification epoch in which many of

accumulating molecules and functions were preferentially lost in emerging archaeal lineages or preferentially accreted in the primordial emerging eukaryal-like lineages, and (iii) an organismal diversification epoch in which increasing numbers of lineage-specific variants of already existing molecules and functions appeared in an increasingly diversified tripartite world (Wang et al. 2007).

Our ToL showed that the most basal lineages belonged to crenarcheal hyperthermophiles of the orders Desulfurococcales and Thermoproteales. This observation supports the previously proposed thermophilic origin of the superkingdom (Gribaldo and Brochier-Armanet 2006) and extends it to diversified life. We note that the basal placement of Crenarchaeota was also recently recovered in phylogenomic analyses of fold family domains (Kim and Caetano-Anollés 2012), with roots that often included *T. pendens*. While clear groupings of recognized archaeal orders were evident in the tree, their relationships to each other were not so clear. The coherence of the Crenarchaeota and Euryarchaeota phyla originally identified using cultured strains on the basis of 16S rRNA (Woese 1987) has been questioned by further addition of cultivars and environmental samples and by analysis of other molecules (Pace 2009). In contrast with Crenarchaeota, the Euryarchaeota has failed to represent a phylogenetically coherent group and has biological signatures related to Korarchaeota. However, new biological signatures of the archaeal groups and more widely encompassing phylogenetic analyses promise more clear definitions (Brochier-Armanet et al. 2011). In our case, the ToL showed lack of coherence of both crenarchaeal and euryarchaeal microbes. However, it revealed groupings of archaeal orders, showed the postulated close links between Sulfolobales and Thaumarchaeota, and included Korarchaeota within the euryarchaeal groups (Brochier-Armanet et al. 2011).

#### The Non-Thermophilic Origin of Bacteria

Our ToLs failed to support a thermophilic origin for bacteria. This result is consistent with a number of recent studies (e.g., Brochier and Philippe 2002) and challenges the canonical reference tree derived from 16S rRNA. While there is no general consensus for the branching order of bacterial phyla, trees generated from conserved 16S rRNA sequences have been rooted in Thermotogae and Aquificae, both of which include thermophilic and hyperthermophilic bacteria (Woese et al. 1990; Rappe and Giovannoni 2003). The most convincing support for the ancestral nature of thermophilic bacteria is the presence of enzyme reverse gyrase that is found only in thermophiles (including thermophilic Archaea) (Forterre et al. 2000). This enzyme harbors two domains, a helicase-like domain and a Type I topoisomerase domain that is primarily responsible for



positively supercoiling the DNA (Confalonieri et al. 1993; Forterre et al. 1996). A recent evolutionary analysis of protein fold families, however, suggested a relatively late origin for the helicase-like domain (Kim and Caetano-Anollés 2012) while the Type I topoisomerase domain was proposed to have been acquired from Archaea via HGT (Forterre et al. 2000). Many other studies based on different sets of genes and proteins also failed to recover the thermophilic rooting and sister relationship between Thermotogae and Aquificae (Gupta 2000; Griffiths and Gupta 2004; Ciccarelli et al. 2006). Another study that focused on only the highly conserved and slow evolving sites of 16S rRNA revealed that both Thermotogae and Aquificae emerged later in evolution together with mesophiles (e.g., Fusobacteria), suggesting a secondary adaptation to life for the bacterial superkingdom (Brochier and Philippe 2002). The basal appearance in our ToLs of the anaerobic rod-shaped Bacteroidetes and some members of the PVC superphylum (Verrucomicrobia) is also compatible with the findings of Brochier and Philippe (2002). The phylum occupied deep positions in their tree, not far away from Planctomycetales, aquatic bacteria that often engage in parasitic relationships (and were excluded in our analysis). Remarkably, we found that the most basal orders of bacterial microbes in our ToL exhibited the highest level of reticulation that was observed ( $\delta = 0.39$ ), which were derived from network reconstructions (Table 2). This suggests that HGT-like processes may have been important determinants in the emergence of the bacterial superkingdom. We conclude that the ancestor of Bacteria was more likely a mesophile that adapted to warm but comfortable environments that were becoming common on Earth about 2.1 billion years ago (Kim and Caetano-Anollés 2011).

#### A Close Relationship Between Plants and Metazoa

Within the strong monophyletic Eukarya, groups exhibited minimal trends of reticulation ( $\delta = 0.16$ – $0.28$ ; Table 2) and main eukaryal kingdoms formed cohesive groups with taxa in the individual groups well positioned. Remarkably, the ToL of functionomes recovered again the close relationship of Metazoa and Plants that was obtained in previous phylogenomic analyses of domain structures (Wang et al. 2007) and domain interactomes (Wang and Caetano-Anollés 2006). The relationships of the fungal, plant, and animal groups are the object of ongoing controversy as these have been consistently poorly resolved in the sequence-based phylogenetic analyses (Pace 2009). This probably stems from a rather explosive radiation of eukaryotic crown taxa and phylogenetic reconstruction problems imposed by long-branch attraction and a ‘Felsenstein’s zone’ defined by short internal branches followed by long edges in trees derived from sequences

(Loytynoja and Milinkovitch 2001). The congruent and well-supported relationship of plants and animals identified in the phylogenomic study of entire functionomic repertoires is, therefore, very encouraging and challenges the proposed fungal-animal split.

#### Advantages and Limitations of GO Terms as Phylogenetic Characters

In this study, we introduce a novel way of reconstructing organismal phylogenies built directly from the genomic ontological annotations. The choice of GO<sub>TMF</sub> terms as phylogenetic characters carries several advantages over traditional phylogenies and few limitations that need to be addressed. The advantages include, but are not limited to (i) GO<sub>TMF</sub> terms portray organismal physiology and truly approximate the reconstruction of ToLs. (ii) GO<sub>TMF</sub> terms represent a class of molecular characters that are more robust than amino acid or nucleotide site characters in sequence alignments. Sequence sites are prone to substitutions and suffer from high mutation rates (Caetano-Anollés and Nasir 2012). In contrast, substitution of a MF into another function is rare. (iii) GO<sub>TMF</sub> terms serve as informative tools to describe both the very deep and very derived organismal relationships. For example, the ancient GO<sub>TMF</sub> terms that are evolutionarily conserved (e.g., ATP binding, structural constituent of ribosome) are highly abundant and widely distributed in living organisms (Kim and Caetano-Anollés 2010). This highlights the conserved nature of GO<sub>TMF</sub> terms and their power to reliably describe deep relationships. In contrast, recently acquired GO<sub>TMF</sub> terms by gene duplication or positive selection (e.g., diphosphokinase activity, coenzyme synthase activity) are less abundant and serve as useful tools to dissect the very derived branches of the ToL. Therefore, utilizing the genomic abundance of GO<sub>TMF</sub> terms as phylogenetic characters increases the resolution in both the very deep and derived branches of the ToL and enables reconstruction of reliable phylogenies. (iv) GO<sub>TMF</sub> terms empower phylogenetic analysis by considering functional conservation. For example, the *Ly49* gene family in mice and *KIR* family in humans are sequentially ‘non-homologous’ but both activate natural-killer cells of the immune system and trigger defensive mechanisms in a similar manner (Emes et al. 2003). This represents a case of functional conservation that cannot be studied with molecular sequences. GO<sub>TMF</sub> terms are advantageous in this regard as they account for the physiological responses of organisms and the genomic abundance value of MFs can be used to both group and differentiate organisms. (v) The impact of non-vertical evolutionary processes that can complicate traditional sequence-based phylogenies appears to be very minimal in our phylogenies.

With respect to limitations associated with the choice of  $GO_{TMF}$  terms as phylogenetic characters, we note that GO characters could well be interdependent. For example, a particular MF may be a consequence of another function and thus would require co-occurrence. However, this is a natural outcome of studying the evolution of entire systems (i.e., organisms), as individual parts in systems (GO terms in this case) are always dependent on other parts (Caetano-Anollés et al. 2014). This same problem exists, for example, when using gene, genome, or concatenated gene sequences to build ToLs. While we have not yet explored or quantified the effect of co-occurrence of MFs, our paper lays foundation for functionomic network studies. Another possible limitation that is shared with sequence phylogenies is that the accuracy of the ToLs reconstructed in this study can suffer from individual  $GO_{TMF}$  terms harboring different evolutionary histories, especially because evolution of MFs depends on functional constraints. While incompatibility between phylogenetic characters decreases the accuracy of a tree topology, many previous studies have shown that multi-gene phylogenies are more robust than single-gene phylogenies. This indicates that the use of a large number of genes increases the amount of phylogenetic signal and overwhelms the problem of phylogenetic heterogeneity (summarized for genes in Liu et al. 2008). Consequently, the ToLs that were reconstructed by analyzing all available MFs should be considered robust against phylogenetic noise resulting from  $GO_{TMF}$  term interdependency and heterogeneity.

Phylogenetic reconstruction is limited by the informativeness of the phylogenetic characters, which in turn depends on how fast characters change along the branches of the trees (the mutation rate). For sequences of proteins and nucleic acids, change is highly dynamic and mutational saturation problems can severely impact the retrodiction effectiveness of Markov models of sequence evolution (Sober and Steel 2002). Under such high mutation rates, time destroys historical information and sequence analysis is only useful for studying relatively recent divergences. A phase transition, however, occurs when mutation rates occur at lower pace; i.e., when characters are evolutionarily conserved. The transition was made mathematically evident on an ‘infinite state’ Poisson process model, in which character states are seldom revisited in the tree (Mossel and Steel 2004). This model is inapplicable to sequences but is compatible with genomic features that are conserved such as abundance of molecular structures and functions, including individual  $GO_{TMF}$  terms. Are overall levels of evolutionary change of  $GO_{TMF}$  terms slower than those of sequence? Will exponential loss of information in MFs limit retrodiction statements about deep divergences of the ToL? While these questions are always difficult to answer, there is growing knowledge of the level of conservation in

molecular structure and function. Molecular functions are embedded in 3-dimensional atomic structural cores that evolve 3–10 times slower than sequences (Illergård et al. 2009). The encoded functions of these cores involve strings of residues that are highly conserved and make up functional pockets or regions, which together with highly conserved allosteric structural networks define MFs of proteins (e.g., Reynolds et al. 2012). Growing knowledge suggests that MFs and interactions, even if defined at the terminal level of the GO DAG, are much more conserved than sequence. However, this does not mean that deep phylogenetic signal of MFs and interactions will stand the ‘erasing’ effects of time. Explicit and more encompassing exploration of information loss will be required. A mitigating aspect, however, is the fact that MFs and their associated molecular structural counterparts (which we here define as modules) accumulate in genomes by gene duplication and rearrangements and spread in biological networks by recruitment as molecules diffuse by mutation in sequence space (Caetano-Anollés et al. 2010). Since this accumulation and spread increase with time and populate in nested manner the nested branches of the evolving trees, the process of monotonic information decay in Markov chains is reversed and information about the past is strongest in the deepest divergences of the trees (see below). Operationally, ancient  $GO_{TMF}$  terms are highly abundant and widely distributed in organisms and provide more information about the deep branches of the ToL (Fig. 4a).

In this study, we used GO terms without reference to their evidence codes. As a result, our dataset included both manually and electronically curated GO terms. We have previously shown that tree topologies are robust against the difference of evidence codes and thus this should not significantly affect our interpretations (Kim and Caetano-Anollés 2010). Finally, we expect functional annotations of genes to undergo revisions as more genomes are being sequenced. Thus, it is possible that few  $GO_{TMF}$  terms sampled in this analysis are later classified as parent terms for some other terms. Therefore, our phylogenies and interpretations rest on GO definitions available at the time of analysis and caution the reader to focus on general trends in our data rather than specific numbers, which are expected to change. However, we assume that global patterns described in our study will remain unaffected with an increase in genomic data.

Finally, trees of functionomes only look at the history of the functional set embedded in the proteomes of cells. What about the set of functions embedded in RNA? This set plays central roles in translation (e.g., the ancient tRNA and rRNA), splicing, regulation, and epigenetics. Information about genomic abundance of RNA sequences or their structural modules and their associated functions should be incorporated in phylogenetic reconstructions.

What about the functional set of metabolites that support the chemical workings of the cell? The landscape of their molecular scaffolds constrains the evolution of enzymatic and transport processes of metabolism. A total evidence approach that combines the many repertoires of the cell should be sought. At the same time, character independence should be carefully monitored. Although GO<sub>TMF</sub> terms of tRNA sequences were absent (nor suitable analogs; Wong 2014), our datasets included tRNA-related MFs that are embedded in proteomes, such as MFs of aminoacyl-tRNA synthetases, transferases, ligases, and other tRNA-related proteins, many of which coevolve with tRNA (Caetano-Anollés et al. 2014). Consequently, protein-centric information of tRNA functions had been already considered in ToL reconstructions.

To conclude, while there is much yet to be learned from the use of GO terms in evolutionary studies, we expect our analysis to provide grounds for more and improved studies. One obvious hurdle is to quantify the impact of homoplasious processes on evolution of MFs. In this study, we carefully sampled both taxa and characters and performed statistical tests to identify and exclude problematic characters from the analysis. During all these steps, the likely effects of non-vertical evolution were minimized.

#### A Note on Retrodiction, Knowledge, and Belief

ToL reconstruction is subject to the technical and philosophical problems of tree building, including the choice of parametric (process) and non-parametric (non-process) models of evolution (Sober 2004), the strategy of rooting (outgroup versus model-based and other methods; Bryant and Wagner 2001), and the assumption of a nested hierarchy embedded in a bifurcating tree structure that accommodates or not reticulations (Lamarck versus Darwin; Merhej and Raoult 2012). Typical retrodiction methods use the trilogy of observations (data), tree, and evolutionary model to find the best solution within a space of possible trees defined by the number of leaves (taxa) that are being considered. Non-parametric methods typical of the cladistic approach focus on the optimization of tree topology, while parametric alternatives such as the frequentist and Bayesian approaches are more interested in the probability of observations along their branches. This brings a number of objections to the different approaches (Sober 2004) and raises important concerns about the epistemological relation of the present and the past. How much of it can be turned into knowledge and certainty and how much into justified belief? Sober and Steel (2013) make the strong mathematical argument that if there were a process connecting the past to the present, the present would provide strong evidence about the past. The emergence of modules of molecular structure and function and

their growth and widespread distribution in the living world represents an example of such a process (Caetano-Anollés et al. 2010). The use of module abundances as phylogenetic characters and simple process models of molecular accumulation that trace the most parsimonious (and likely) distribution of abundance changes in trees complies with Weston's generality criterion of nesting and additive phylogenetic change for rooting (Caetano-Anollés et al. 2014) and promises to offset the problem of vanishing phylogenetic signal of Markov chains with increasing time (Sober and Steel 2013). In this case, changes in molecular abundance take advantage of the 'data processing inequality' and the 'impact of branching' on information, both of which are known to counteract information loss imposed by Markov chain convergence (Sober and Steel 2013). Consequently, the approach we here describe puts forth retrodiction hypotheses that are trustworthy for building historical knowledge and less likely to contribute to the world of justified belief (i.e., apriorism that is typical of the ToL field of inquiry; Lienau and DeSalle 2010). We end by noting that the cladistics rationale seeks to maximize explanatory power and severity of test through reciprocal illumination of trees and character homologies (Farris 2008). An example of the application of Popper's formula to the retrodiction of the genetic code makes explicit the rationale of scientific inquiry in cladistics (Caetano-Anollés et al. 2013).

#### Conclusions

In this study, we introduce the reconstruction of trees of cellular life that describe the evolution of functionomes. These phylogenies are built directly from genomic ontological annotations that portray organismal physiology and truly approximate the construction of trees of organisms. Remarkably, our methodology recovered the tripartite nature of the living world heralded by the biological school of Carl Woese and the very ancient and thermophilic origin of Archaea embodied in multiple (paraphyletic) branching patterns of archaeal lineages appearing at the base of the ToL. The early rise of Archaea is not only compatible with several lines of molecular evidence we previously discussed but also supports paleobiological claims of early archaeal lipids and methanogenic activity linked to the fossil record (Chappe et al. 1979; Michaelis and Albrecht 1979; Schopf 1999) and the early archaeal role in biogeochemical processes (Blank 2009). The analyses also recovered a non-thermophilic origin for the bacterial superkingdom and a close relationship between Metazoa and Plants that excluded Fungi, dissecting a long-standing controversy associated with the trichotomy of crown eukaryotic taxa. Our results agree with a theoretical

framework in which lineages evolve unique trade-off solutions among three strategies, economy, flexibility, and robustness (Yafremava et al. 2013). This framework places evolving lineages in a ‘persistence triangle’ supported by protein domain structure and many other lines of evidence. Within the triangle, Archaea and Bacteria gravitate toward the triangle’s economy vertex and arise very early in evolution, with Archaea biased toward robustness mainly due to very early adaptations to the thermophilic habitats of early Earth. Protista in turn occupy a saddle manifold that separates akaryotic microbes from multicellular organisms. According to this framework, the manifold was historically defined by the viscosity of water, which sets a critical barrier to organism size (100 μm) and possible trade-off solutions that unfold toward the economy vertex in microbes and delimit positive feedback loops toward flexibility and robustness in higher organisms. In our study, we also evaluated the effects of parasitic taxa (reductive evolution) and the functions of HTP characters (HGT) and suggested that they should be excluded for reliable interpretations. We conclude by proposing that functionomic data are useful and reliable additions to the toolkit of molecular features used for phylogeny reconstruction. The new ToLs that describe the evolution of functionomes reveal deep phylogenetic relationships with considerable explanatory power for the deep evolutionary study of cellular species. The new methodology can also yield novel insights into the evolution of MFs in genomes, since phylogenetic characters describing potentially interesting MFs can be traced along the branches of the ToL.

**Acknowledgments** We are thankful to the members of GCA laboratory for fruitful discussions and anonymous reviewers who reviewed the paper and provided comments that significantly improved the manuscript. Research was supported by grants from the National Science Foundation (MCB-0749836 and OISE-1132791) and the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to GCA and grants from KRIBB Research Initiative Program and the Next-Generation BioGreen 21 Program, Rural Development Administration (PJ0090192013) to KMK.

## References

- Aminov RI (2011) Horizontal gene exchange in environmental microbiota. *Front Microbiol* 2:158
- Anderson I, Rodriguez J, Susanti D, Porat I, Reich C, Ulrich LE, Elkins JG, Mavromatis K, Lykidis A, Kim E, Thompson LS, Nolan M, Land M, Copeland A, Lapidus A, Lucas S, Detter C, Zhulin IB, Olsen GJ, Whitman W, Mukhopadhyay B, Bristow J, Kyrpides N (2008) Genome sequence of *Thermophilus pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction. *J Bacteriol* 190:2957–2965
- Anderson IJ, Dharmarajan L, Rodriguez J, Hooper S, Porat I, Ulrich LE, Elkins JG, Mavromatis K, Sun H, Land M, Lapidus A, Lucas S, Barry K, Huber H, Zhulin IB, Whitman WB, Mukhopadhyay B, Woese C, Bristow J, Kyrpides N (2009) The complete genome sequence of *Staphylothermus marinus* reveals differences in sulfur metabolism among heterotrophic Crenarchaeota. *BMC Genomics* 10:145
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25:25–29
- Bapteste E, Brochier C (2004) On the conceptual difficulties in rooting the tree of life. *Trends Microbiol* 12:9–13
- Blank CE (2009) Not so old Archaea—the antiquity of biogeochemical processes in the archaeal domain of life. *Geobiology* 7:495–514
- Brochier C, Philippe H (2002) Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* 417:244
- Brochier-Armanet C, Forterre P, Gribaldo S (2011) Phylogeny and evolution of the Archaea: one hundred genomes later. *Curr Opin Microbiol* 14:274–281
- Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* 21:255–265
- Bryant HN, Wagner G (2001) Character polarity and the rooting of cladograms. In: Wagner GP (ed) *The character concept in evolutionary biology*. Academic Press, New York, pp 319–338
- Buckley CD (2012) Investigating cultural evolution using phylogenetic analysis: the origins and descent of the southeast Asian tradition of warp ikat weaving. *PLoS One* 7:e52064
- Bukhari SA, Caetano-Anollés G (2013) Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. *PLoS Comput Biol* 9:e1003009
- Caetano-Anollés G, Caetano-Anollés D (2003) An evolutionarily structured universe of protein architecture. *Genome Res* 13:1563–1571
- Caetano-Anollés G, Nasir A (2012) Benefits of using molecular structure and abundance in phylogenomic analysis. *Front Gene* 3:172
- Caetano-Anollés G, Wang M, Caetano-Anollés D, Mitternthal J (2009) The origin, evolution and structure of the protein world. *Biochem J* 417:621–637
- Caetano-Anollés G, Yafremava L, Mitternthal JE (2010) Modularity and dissipation in evolution of macromolecular structures, functions and networks. In: Caetano-Anollés G (ed) *Evolutionary genomics and systems biology*. John Wiley & Sons, Hoboken, NJ, pp 431–449
- Caetano-Anollés G, Wang M, Caetano-Anollés D (2013) Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLoS One* 8:e72225
- Caetano-Anollés G, Nasir A, Zhou K, Caetano-Anollés D, Mitternthal JE, Sun FJ, Kim KM (2014) Archaea: the first domain of diversified life. *Archaea* 2014:590214
- Caro-Quintero A, Deng J, Auchtung J, Brettar I, Hofle MG, Klappenbach J, Konstantinidis KT (2011) Unprecedented levels of horizontal gene transfer among spatially co-occurring *Shewanella* bacteria from the Baltic Sea. *ISME J* 5:131–140
- Cavalier-Smith T (2002) The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52:7–76
- Chang CW, Lyu PC, Arita M (2011) Reconstructing phylogeny from metabolic substrate-product relationships. *BMC Bioinform* 12(Suppl 1):S27
- Chappe B, Michaelis W, Albrecht P, Ourisson G (1979) Fossil evidence for a novel series of archaeobacterial lipids. *Naturwissenschaften* 66:522–523
- Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287

- Confalonieri F, Elie C, Nadal M, de La Tour C, Forterre P, Duguet M (1993) Reverse gyrase: a helicase-like domain and a type I topoisomerase in the same polypeptide. *Proc Natl Acad Sci USA* 90:4753–4757
- Cummings MP, Neel MC, Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution* 62:2411–2422
- Dagan T, Roetger M, Bryant D, Martin W (2010) Genome networks root the tree of life between prokaryotic domains. *Genome Biol Evol* 2:379
- Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375
- Di Giulio M (2007) The tree of life might be rooted in the branch leading to Nanoarchaeota. *Gene* 401:108–113
- Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284:2124–2128
- Dopazo H, Santoyo J, Dopazo J (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics* 20(Suppl 1):i116–i121
- Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* 6:R14
- Emes RD, Goodstadt L, Winter EE, Ponting CP (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* 12:701–709
- Farris JS (1989) The retention index and homoplasy excess. *Syst Zool* 38:406–407
- Farris JS (2008) Parsimony and explanatory power. *Cladistics* 24:1–23
- Forslund K, Henricson A, Hollich V, Sonnhammer EL (2008) Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol* 25:254–264
- Forterre P, Philippe H (1999) Where is the root of the universal tree of life? *BioEssays* 21:871–879
- Forterre P, Bergerat A, Lopez-Garcia P (1996) The unique DNA topology and DNA topoisomerases of hyperthermophilic archaea. *FEMS Microbiol Rev* 18:237–248
- Forterre P, Bouthier De La Tour C, Philippe H, Duguet M (2000) Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. *Trends Genet* 16:152–154
- Garcia-Vallve S, Guzmán E, Montero M, Romeu A (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* 31:187–189
- Gerstein M (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* 33:518–534
- Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245
- Gogarten JP, Olenzki L (1999) Orthologs, paralogs and genome comparisons. *Curr Opin Genet Dev* 9:630–636
- Gough J (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics* 21:1464–1471
- Gribaldo S, Brochier-Armanet C (2006) The origin and evolution of Archaea: a state of the art. *Phil Trans R Soc B* 361:1007–1022
- Griffiths E, Gupta RS (2004) Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. *Int Microbiol* 7:41–52
- Gu X, Zhang H (2004) Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol* 21:1401–1408
- Gupta R (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev* 24:367–402
- Harish A, Caetano-Anollés G (2012) Ribosomal history reveals origins of modern protein synthesis. *PLoS One* 7:e32776
- Holland BR, Huber KT, Dress A, Moulton V (2002) Delta plots: a tool for analyzing phylogenetic distance data. *Mol Biol Evol* 19:2051–2059
- Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14:68–73
- Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R (2007) Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinform* 8:460
- Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* 77:499–508
- Jablonka E, Lamb MJ (2006) The evolution of information in the major transitions. *J Theor Biol* 239:236–246
- Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806
- Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425
- Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505
- Kim KM, Caetano-Anollés G (2010) Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Mol Biol Evol* 27:1710–1733
- Kim KM, Caetano-Anollés G (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evol Biol* 11:140
- Kim KM, Caetano-Anollés G (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. *BMC Evol Biol* 12:13
- Kloesges T, Popa O, Martin W, Dagan T (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol Biol Evol* 28:1057–1074
- Kolaczowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984
- Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127–136
- Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 55:709–742
- Lake JA, Skophammer RG, Herbold CW, Servin JA (2009) Genome beginnings: rooting the tree of life. *Phil Trans R Soc B* 364:2177–2185
- Lienau EK, DeSalle R (2010) Is the microbial tree of life verificationist? *Cladistics* 26:195–201
- Liolios K, Chen IA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38:D346–D354
- Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62:2080–2091
- Loytynoja A, Milinkovitch MC (2001) Molecular phylogenetic analyses of the mitochondrial ADP-ATP carriers: the Plantae/Fungi/Metazoa trichotomy revisited. *Proc Natl Acad Sci USA* 98:10202–10207
- Lundberg JG (1972) Wagner networks and ancestors. *Syst Biol* 21:398–413
- Marcet-Houben M, Puigbo P, Romeu A, Garcia-Vallve S (2007) Towards reconstructing a metabolic tree of life. *Bioinformatics* 2:135–144
- McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH (2010) High frequency of horizontal gene transfer in the oceans. *Science* 330:50

- Merhej V, Raoult D (2012) Rhizome of life, catastrophes, sequence exchanges, gene creations, and giant viruses: how microbial genomics challenges Darwin. *Front Cel Inf Microbiol* 2:113
- Michaelis W, Albrecht P (1979) Molecular fossils of Archaeobacteria in Kerogen. *Naturwissenschaften* 66:420–421
- Mossel E, Steel M (2004) A phase transition for a random cluster model on phylogenetic trees. *Math Biosci* 187:189–203
- Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540
- Nasir A, Caetano-Anollés G (2013) Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification. *Archaea* 2013:648746
- Nasir A, Naeem A, Khan MJ, Nicora HDL, Caetano-Anollés G (2011) Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms. *Genes* 2:869–911
- Nasir A, Kim KM, Caetano-Anollés G (2012) Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol Biol* 12:156
- Nasir A, Kim KM, Caetano-Anollés G (2014a) Global patterns of protein domain gain and loss in superkingdoms. *PLoS Comput Biol* 10:e1003452
- Nasir A, Kim KM, Caetano-Anollés G (2014b) A phylogenomic census of molecular functions identifies modern thermophilic Archaea as the most ancient form of cellular life. *Archaea* (in press)
- Orengo CA, Michie A, Jones S, Jones DT, Swindells M, Thornton JM (1997) CATH—a hierarchical classification of protein domain structures. *Structure* 5:1093–1109
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702
- Pace NR (2009) Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev* 73:565–576
- Penny D, Poole A (1999) The nature of the last universal common ancestor. *Curr Opin Genet Dev* 9:672–677
- Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MP, Silva A, O'Brien SJ, Pecon-Slattery J (2011) A molecular phylogeny of living primates. *PLoS Genet* 7:e1001342
- Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–D596
- Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394
- Ravin NV, Mardanov AV, Beletsky AV, Kublanov IV, Kolganova TV, Lebedinsky AV, Chernyh NA, Bonch-Osmolovskaya EA, Skryabin KG (2009) Complete genome sequence of the anaerobic, protein-degrading hyperthermophilic crenarchaeon *Desulfurococcus kamchatkensis*. *J Bacteriol* 191:2371–2379
- Reynolds KA, McLaughlin RN, Raganathan R (2012) Ho spots for allosteric regulation on protein surfaces. *Cell* 147:1564–1575
- Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. *Nat Rev Genet* 9:509–515
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci USA* 89:6575–6579
- Schopf JW (1999) Deep divisions in the Tree of Life—what does the fossil record reveal? *Biol Bull* 196:351–353 discussion 354–355
- Sober E (2004) The contest between parsimony and likelihood. *Syst Biol* 53:644–653
- Sober E, Steel M (2002) Testing the hypothesis of common ancestry. *J Theor Biol* 218:395–408
- Sober E, Steel M (2013) Time and knowability in evolutionary processes. arXiv 1301.6470 [q-bio.PE]
- Sun F, Caetano-Anollés G (2008a) Evolutionary patterns in the sequence and structure of transfer RNA: early origins of Archaea and viruses. *PLoS Comput Biol* 4:e1000018
- Sun F, Caetano-Anollés G (2008b) The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J Mol Evol* 66:21–35
- Sun F, Caetano-Anollés G (2009) The evolutionary history of the structure of 5S ribosomal RNA. *J Mol Evol* 69:430–443
- Sun F, Caetano-Anollés G (2010) The ancient history of the structure of ribonuclease P and the early origins of Archaea. *BMC Bioinform* 11:153
- Swofford DL (2002) *Phylogenomic Analysis Using Parsimony and Other Programs (PAUP\*)* Ver 4.0b10. Sinauer, Sunderland, MA
- Szathmary E, Smith JM (1995) The major evolutionary transitions. *Nature* 374:227–232
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41
- Vesteg M, Krajcovic J (2008) Origin of eukaryotic cells as a symbiosis of parasitic alpha-proteobacteria in the periplasm of two-membrane-bounded sexual pre-karyotes. *Commun Integr Biol* 1:104–113
- Vogel C, Chothia C (2006) Protein family expansions and biological complexity. *PLoS Comput Biol* 2:e48
- Wang M, Caetano-Anollés G (2006) Global phylogeny determined by the combination of protein domains in proteomes. *Mol Biol Evol* 23:2444–2454
- Wang M, Caetano-Anollés G (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17:66–78
- Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res* 17:1572–1585
- Wang M, Jiang Y, Kim KM, Qu G, Ji H, Mittenthal JE, Zhang H, Caetano-Anollés G (2011a) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol Biol Evol* 28:567–582
- Wang M, Kurland CG, Caetano-Anollés G (2011b) Reductive evolution of proteomes and protein structures. *Proc Natl Acad Sci USA* 108:11954–11958
- Warnow T (2012) Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Curr* 4:RRN1308
- Wichmann K, Holman EW, Rama T, Walker RS (2011) Correlates of reticulation in linguistic phylogenies. *Lang Dyn Change* 1:205–240
- Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271
- Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–5090

- Woese CR, Maniloff J, Zablen LB (1980) Phylogenetic analysis of the mycoplasmas. *Proc Natl Acad Sci USA* 77:494–498
- Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci* 87:4576–4579
- Wong J (2014) Emergence of life: from functional RNA selection to natural selection and beyond. *Front Biosci* 19:1117–1150
- Wong J, Chen J, Mat W, Ng S, Xue H (2007) Polyphasic evidence delineating the root of life and roots of biological domains. *Gene* 403:39–52
- Xue H, Tong K, Marck C, Grosjean H, Wong J (2003) Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene* 310:59–66
- Xue H, Ng S, Tong K, Wong J (2005) Congruence of evidence for a *Methanopyrus*-proximal root of life based on transfer RNA and aminoacyl-tRNA synthetase genes. *Gene* 360:120–130
- Yafremava LS, Wielgos M, Thomas S, Nasir A, Wang M, Mittenthal JE, Caetano-Anollés G (2013) A general framework of persistence strategies for biological systems helps explain domains of life. *Front Gene* 4:16
- Yang S, Bourne PE (2009) The evolutionary history of protein domains viewed by species phylogeny. *PLoS One* 4:e8378
- Zillig W, Holz I, Janekovic D, Klenk HP, Imsele E, Trent J, Wunderl S, Forjaz VH, Coutinho R, Ferreira T (1990) *Hyperthermus butylicus*, a hyperthermophilic sulfur-reducing archaeobacterium that ferments peptides. *J Bacteriol* 172:3959–3965