Puigbò *et al*. *BMC Biology* 2013, **11**:46
http://www.biomedcentral.com/1741-7007/11/46

10th anniversary

BMC Biology

ANNIVERSARY UPDATE

Open Access

Seeing the Tree of Life behind the phylogenetic forest

Pere Puigbò, Yuri I Wolf and Eugene V Koonin*

In the article entitled 'Search for a Tree of Life in the thicket of the phylogenetic forest', published in 2009 in *Journal of Biology* [1] (see also the accompanying comment [2]), we presented evidence that the traditional Tree of Life (TOL) can and should be replaced with a statistical central trend in the genome-wide compendium of phylogenetic trees that reflects the coherence between the evolutionary histories of different genes and was later denoted the Statistical Tree Of Life (STOL) [3]. Since Darwin's day, the TOL is the dominant icon of evolutionary biology [4,5], the basis of taxonomy and an essential framework for evolutionary reconstructions. In the late 1970s, ribosomal (r)RNA was introduced as a universal phylogenetic marker, primarily through the work of Carl Woese and colleagues [6,7], and the rRNA tree, complemented with trees for other universal genes such as the large RNA polymerase subunits, became the standard model for TOL study.

Technical difficulties notwithstanding, progress in genome sequencing combined with advances in phylogenetic analysis seemed to put a well-resolved TOL within reach [8,9]. However, as soon as a reasonable number of complete genome sequences of bacteria and archaea became available, phylogenomics - genome-wide phylogenetic analysis of individual gene trees - hopelessly marred this neat picture by showing that the trees of different genes generally had different topologies. The topological inconsistencies between gene trees were far too extensive to be dismissed as phylogenetic artifacts, leading to the realization that no single gene tree, including those for universal genes such as rRNA, could represent the evolution of genomes in its entirety. Hence the concepts of horizontal genomics or a 'net of life' were brought about to replace the simple notion of the TOL [10,11]. In the extreme, several influential studies proposed to dispense with 'tree thinking' altogether as an artificial construct having little to do with actual evolution, at least as far as bacteria and archaea are concerned [12-15].

The concept of 'horizontal genomics' involves an internal contradiction because the notion of horizontal gene transfer (HGT) inherently implies the existence of a standard of vertical, tree-like evolution, and most of the existing methods for HGT detection are based on the comparison of gene trees to a standard 'species tree', in practice often the rRNA tree [16,17]. If the vertical standard does not exist, the concept of HGT becomes effectively meaningless, so all we can talk about is a network of life, with nodes corresponding to genomes and edges reflecting gene exchange [18]. The stakes here are high because replacement of the TOL with a network graph would change our entire perception of the process of evolution and invalidate all evolutionary reconstruction based on a species tree. However, the tree representation is by no means superfluous to the description of evolution because the very process of the replication of genetic information implies a bifurcating graph - in other words, a tree [19]. Thus, the key question is [1,20]: in the genome-wide compendium of phylogenetic trees, that we denoted the Forest Of Life (FOL), can we detect any order, any preferred tree topology (branching order) that would reflect a consensus of the topologies of other trees?

We set out to address the above question as objectively as possible, first of all dispensing with any pre-selected standard of tree-like evolution. The analyzed FOL consisted of 6,901 maximum likelihood phylogenetic trees that were built for clusters of orthologous genes from a representative set of 100 diverse bacterial and archaeal genomes [1]. The complete matrix of topological distances between these trees was analyzed using the Inconsistency Score, a measure that we defined specifically for this purpose that reflects the average topological (in)consistency of a given tree with the rest of the trees in the FOL (for the details of the methods employed in this analysis, see [21]). Although the FOL includes very few trees with exactly identical topologies, we found that the topologies of the trees were far more congruent than expected by chance. The 102 Nearly Universal Trees

*Correspondence: koonin@ncbi.nlm.nih.gov
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

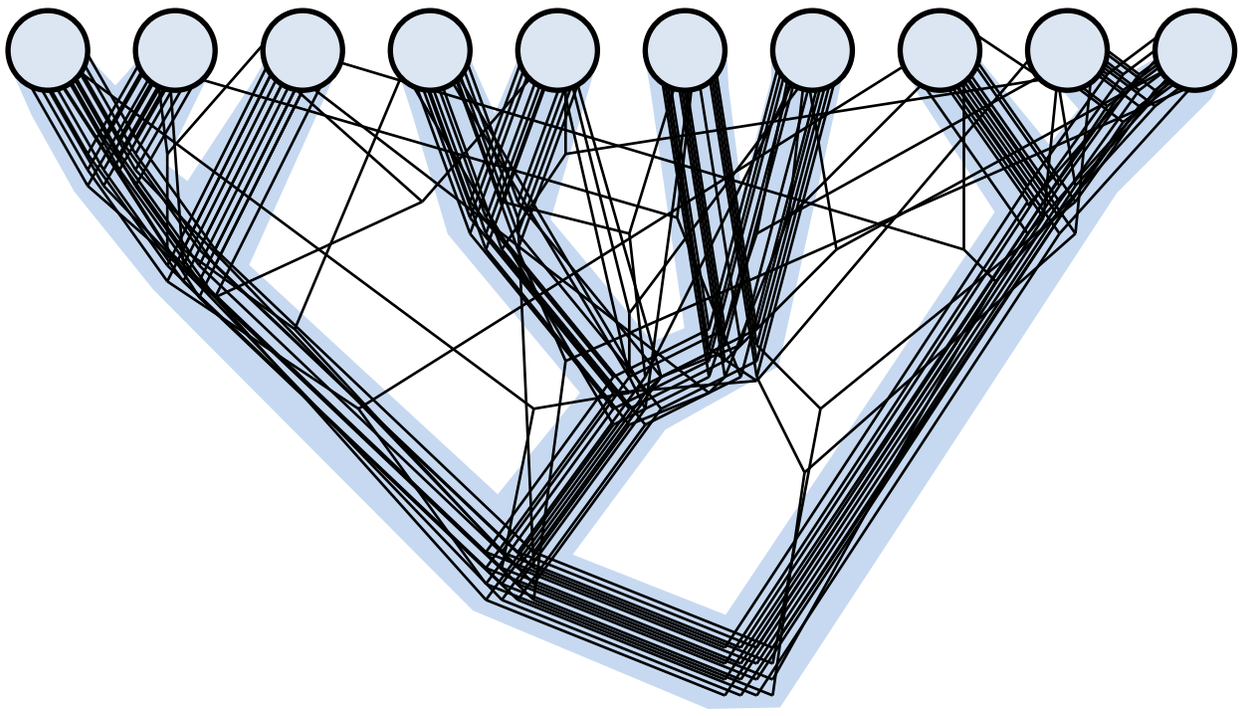

Figure 1. The central tree-like trend in the phylogenetic forest of life. The circles show genomes of extant species and the grey tree in the background shows the statistical central trend in the data. For the purpose of illustration, the figure shows an 'FOL' made of 16 trees with 20 deviations from the central tree-like pattern.

(NUTs; that is, the trees for genes that are represented in all or nearly all archaea and bacteria), which include primarily genes for key protein components of the translation and transcription systems, showed particularly high topological similarity to the other trees in the FOL. Although the topologies of the NUTs are not identical, apparently reflecting multiple HGT events, these transfers appeared to be distributed randomly. In other words, there seem to be no prominent 'highways' of HGT that would preferentially connect particular groups of archaea and bacteria. Thus, although the NUTs cannot represent the FOL completely, they appear to reflect a significant central trend, an attractor in the tree space that could be equated with the STOL (Figure 1).

The set of 6,901 phylogenetic trees that comprise the FOL has become a launching pad for several new studies addressing various aspects of prokaryote evolution and general questions of evolutionary biology. In our own hands, the sequel to the original FOL study involved quantitative dissection of the evolution of prokaryotes into tree-like and web-like components [22]. We applied the approach known as quartet analysis to quantify the contributions of these two distinct modes of evolution [21] and found that, although diverse routes of net-like evolution collectively dominate the FOL, the pattern of tree-like evolution that reflects the generally consistent topologies of the NUTs is the most prominent coherent trend [22]. Thus, the ubiquity of HGT notwithstanding, this central tree-like trend reflects a major aspect of genome evolution and hence has a legitimate claim to represent the STOL.

Having established the validity of the STOL, we employed it to reassess a fundamental aspect of evolutionary theory, the molecular clock (MC) model under which genes evolve at approximately constant gene-specific rates [23]. Using the supertree of the NUTs as a proxy for the STOL, we compared the fits of approximately 3,000 largest trees (that is, the trees with the largest number of species) from the FOL to the supertree that was constrained either under the MC assumption or according to another, more general model that we denoted Universal PaceMaker (UPM) of genome evolution [24]. Under the UPM model, the rate of evolution changes synchronously across genome-wide sets of genes in each evolving lineage (the genes accelerate or decelerate their evolution in sync), thus explaining the universal distribution of evolutionary rates of orthologous genes from diverse life forms [25,26]. However, unlike the MC model, the UPM model does not assume conservation of absolute gene-specific evolutionary rates. We showed that the UPM model fits the data substantially better than the MC model, with the implication that the MC should

be replaced by a more general constraint on the evolutionary process under which only the relative evolutionary rates of the genes are conserved [24,27]. Others have also employed the FOL to test new approaches for tree comparison and 'harvest' different kinds of evolutionary signals, in particular those that reflect HGT between diverse bacteria and archaea with similar life styles [28].

The study of the interplay between the vertical and horizontal trends in the evolution of prokaryotes continues, stimulated by the rapid accumulation of diverse archaeal and bacterial genome sequences. For example, a new and potentially promising twist of this theme is the use of shared HGT events to refine and root the species trees for prokaryotic phyla [29]

A key fact established by comparative genomics is that we already know all the NUTs: no new (nearly) universal genes can possibly be discovered, and it is equally unlikely that a substantial fraction of the NUTs will lose the 'nearly universal' status [30,31]. Thus, given that the coherent topologies of the NUTs seem to adequately represent a central statistical trend in the FOL, the STOL appears to be here to stay and could become a solid foundation for a genome-based classification of bacteria and archaea [32], and perhaps even more importantly, a robust framework for evolutionary reconstruction.

This article is part of the *BMC Biology* tenth anniversary series. Other articles in this series can be found at http://www.biomedcentral.com/bmcbiol/series/tenthanniversary.

Published: 15 April 2013

References

1. Puigbo P, Wolf YI, Koonin EV: **Search for a Tree of Life in the thicket of the phylogenetic forest.** *J Biol* 2009, **8:**59.
2. Swithers KS, Gogarten JP, Fournier GP: **Trees in the web of life.** *J Biol* 2009, **8:**54.
3. O'Malley MA, Koonin EV: **How stands the Tree of Life a century and a half after The Origin?** *Biol Direct* 2011, **6:**32.
4. Darwin C: *On the Origin of Species*. London: Murray; 1859.
5. Haeckel E: *The Wonders of Life: A Popular Study of Biological Philosophy*. De Young Press; 1997.
6. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci U S A* 1977, **74:**5088-5090.
7. Woese CR, Magrum LJ, Fox GE: **Archaebacteria.** *J Mol Evol* 1978, **11:**245-251.
8. Woese CR: **Bacterial evolution.** *Microbiol Rev* 1987, **51:**221-271.
9. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci U S A* 1990, **87:**4576-4579.
10. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2129.
11. Doolittle WF: **Lateral genomics.** *Trends Cell Biol* 1999, **9:**M5-8.
12. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19:**2226-2238.
13. Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF: **Do orthologous gene phylogenies really support tree-thinking?** *BMC Evol Biol* 2005, **5:**33.
14. Doolittle WF, Bapteste E: **Pattern pluralism and the Tree of Life hypothesis.** *Proc Natl Acad Sci U S A* 2007, **104:**2043-2049.
15. Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupré J, Dagan T, Boucher Y, Martin W: **Prokaryotic evolution and the tree of life are two different things.** *Biol Direct* 2009, **4:**34.
16. Snel B, Huynen MA, Dutilh BE: **Genome trees and the nature of genome evolution.** *Annu Rev Microbiol* 2005, **59:**191-209.
17. Lapierre P, Lasek-Nesselquist E, Gogarten JP: **The impact of HGT on phylogenomic reconstruction methods.** *Brief Bioinform* 2012 [Epub ahead of print].
18. Dagan T: **Phylogenomic networks.** *Trends Microbiol* 2011, **19:**483-491.
19. Koonin EV, Wolf YI: **The fundamental units, processes and patterns of evolution, and the Tree of Life conundrum.** *Biol Direct* 2009, **4:**33.
20. Koonin EV, Puigbo P, Wolf YI: **Comparison of phylogenetic trees and search for a central trend in the "forest of life".** *J Comput Biol* 2011, **18:**917-924.
21. Puigbo P, Wolf YI, Koonin EV: **Genome-wide comparative analysis of phylogenetic trees: the prokaryotic forest of life.** *Methods Mol Biol* 2012, **856:**53-79.
22. Puigbo P, Wolf YI, Koonin EV: **The tree and net components of prokaryote evolution.** *Genome Biol Evol* 2010, **2:**745-756.
23. Bromham L, Penny D: **The modern molecular clock.** *Nat Rev Genet* 2003, **4:**216-224.
24. Snir S, Wolf YI, Koonin EV: **Universal pacemaker of genome evolution.** *PLoS Comput Biol* 2012, **8:**e1002785.
25. Grishin NV, Wolf YI, Koonin EV: **From complete genomes to measures of substitution rate variability within and between proteins.** *Genome Res* 2000, **10:**991-1000.
26. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ: **The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.** *Proc Natl Acad Sci U S A* 2009, **106:**7273-7280.
27. Muers M: **Evolution: Genomic pacemakers or ticking clocks?** *Nat Rev Genet* 2012, **14:**81.
28. Schliep K, Lopez P, Lapointe FJ, Bapteste E: **Harvesting evolutionary signals in a forest of prokaryotic gene trees.** *Mol Biol Evol* 2011, **28:**1393-1405.
29. Abby SS, Tannier E, Gouy M, Daubin V: **Lateral gene transfer as a support for the tree of life.** *Proc Natl Acad Sci U S A* 2012, **109:**4962-4967.
30. Koonin EV: **Comparative genomics, minimal gene-sets and the last universal common ancestor.** *Nat Rev Microbiol* 2003, **1:**127-136.
31. Charlebois RL, Doolittle WF: **Computing prokaryotic gene ubiquity: rescuing the core from extinction.** *Genome Res* 2004, **14:**2469-2477.
32. Klenk HP, Goker M: **En route to a genome-based classification of Archaea and Bacteria?** *Syst Appl Microbiol* 2010, **33:**175-182.

doi:10.1186/1741-7007-11-46
Cite this article as: Puigbò P, *et al*.: Seeing the Tree of Life behind the phylogenetic forest. *BMC Biology* 2013, **11:**46.