RESEARCH NOTE

REVISED Recapitulating phylogenies using *k*-mers: from trees to networks [version 2; referees: 2 approved]

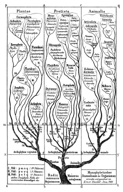
Guillaume Bernard, Mark A. Ragan, Cheong Xin Chan

Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

v2 **First published:** 29 Nov 2016, **5**:2789 (doi: 10.12688/f1000research.10225.1)
Latest published: 23 Dec 2016, **5**:2789 (doi: 10.12688/f1000research.10225.2)

Abstract

Ernst Haeckel based his landmark Tree of Life on the supposed ontogenic recapitulation of phylogeny, i.e. that successive embryonic stages during the development of an organism re-trace the morphological forms of its ancestors over the course of evolution. Much of this idea has since been discredited. Today, phylogenies are often based on families of molecular sequences. The standard approach starts with a multiple sequence alignment, in which the sequences are arranged relative to each other in a way that maximises a measure of similarity position-by-position along their entire length. A tree (or sometimes a network) is then inferred. Rigorous multiple sequence alignment is computationally demanding, and evolutionary processes that shape the genomes of many microbes (bacteria, archaea and some morphologically simple eukaryotes) can add further complications. In particular, recombination, genome rearrangement and lateral genetic transfer undermine the assumptions that underlie multiple sequence alignment, and imply that a tree-like structure may be too simplistic. Here, using genome sequences of 143 bacterial and archaeal genomes, we construct a network of phylogenetic relatedness based on the number of shared *k*-mers (subsequences at fixed length *k*). Our findings suggest that the network captures not only key aspects of microbial genome evolution as inferred from a tree, but also features that are not treelike. The method is highly scalable, allowing for investigation of genome evolution across a large number of genomes. Instead of using specific regions or sequences from genome sequences, or indeed Haeckel's idea of ontogeny, we argue that genome phylogenies can be inferred using *k*-mers from whole-genome sequences. Representing these networks dynamically allows biological questions of interest to be formulated and addressed quickly and in a visually intuitive manner.



This article is included in the Phylogenetics channel.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 23 Dec 2016		✓ report
version 1 published 29 Nov 2016	✓ report	? report

1 **Bernhard Haubold**, Max Planck Institute for Evolutionary Biology Germany

2 **Weilong Hao**, Wayne State University USA

Discuss this article

Comments (0)

Corresponding author: Cheong Xin Chan (c.chan1@uq.edu.au)

How to cite this article: Bernard G, Ragan MA and Chan CX. **Recapitulating phylogenies using *k*-mers: from trees to networks [version 2; referees: 2 approved]** *F1000Research* 2016, **5**:2789 (doi: 10.12688/f1000research.10225.2)

Grant information: We thank funding support from the Australian Research Council (DP150101875) awarded to MAR and CXC, and a James S. McDonnell Foundation grant awarded to MAR.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 29 Nov 2016, **5**:2789 (doi: 10.12688/f1000research.10225.1)

REVISED **Amendments from Version 1**

In this revision, we have rewritten part of the Abstract and Introduction to clarify that (a) phylogenetic approaches based on multiple sequence alignment do not exclude inference of a network, (b) multiple sequence alignment is computationally demanding, and that (c) phylogenetic inference is complicated by non-treelike evolutionary processes that shape microbial genomes. In Results and discussion, we have now provided justification for using the 143-genome dataset in this work, and have made explicit the scope of this study. We have also cited a number of additional publications in the areas of phylogenetic networks and alignment-free methods.

See referee reports

Introduction

Ernst Haeckel coined the term *Phylogenie* to describe the series of morphological stages in the evolutionary history of an organism or group of organisms[1]. In his Tree of Life published 150 years ago[2], Haeckel postulated that living organisms trace their evolutionary origin(s) along three distinct lineages (Plantae, Protista and Animalia) to a "common Moneran root of autogonous organisms". In some (but not all) later works (e.g. in 1868[3]) he allowed that different Monera may have arisen independently by spontaneous generation. Either way, these views accord with the Larmackian notion of a built-in direction of evolution from morphologically simple "lower" organisms to more-complex "higher" forms[4].

Haeckel through his "Biogenetic Law" advocated that "ontogeny recapitulates phylogeny"[2]: that the embryonic series of an organism is a record of its evolutionary history. Under this view, morphologies observed at different developmental stages of an organism resemble and represent the successive stages (including adult stages) of its ancestors over the course of evolution. Of course, he worked before the advent of genetics and the modern synthesis, and before it was appreciated that information on hereditary is carried by DNA and can be recovered by sequencing and statistical analysis. He could not have foreseen that these DNA sequences code for other biomolecules and control life processes, including his beloved developmental series and organismal phenotype, through vastly complex molecular webs of interactions. Nor could Haeckel have envisaged the scale of phylogenetic analysis that can be carried out today using these DNA sequences across multiple genomes, made possible by the advent of high-throughput sequencing and computing technologies.

Fast-forwarding 150 years, phylogenetic inference based on comparative analysis of biological sequences is now a common practice. The similarity among sequences is commonly interpreted as evidence of homology[5,6], i.e. that they share a common ancestry. From the earliest days of molecular phylogenetics, multiple sequences have been aligned[7,8] to display this homology position-by-position along the length of the sequences. That is, the residues are arranged relative to each other such that the best available hypothesis of homology is achieved at every position (column) of the alignment. By default, it is assumed that the best alignment can be achieved simply by displaying the sequences in the same direction, and inserting gaps where needed (to represent insertions and deletions). This assumption is largely valid when working with highly conserved orthologs of any source, and with exons or proteins of morphologically complex eukaryotes. However, microbial genomes are often affected by recombination and rearrangement[9], undermining the assumption of homology along adjacent positions, while lateral genetic transfer would not be represented by a common treelike process[10–13]. As Haeckel observed when he drew his Tree[2], biological evolution can be anything but straightforward, and these complications have become ever more-complicated[14,15].

Alternative approaches for inferring and representing phylogenies are available. An attractive strategy that addresses the issue of full-length alignability is to compute relatedness among a set of sequences based on the number or extent of k-mers (short sub-sequences of a fixed length k) that they share. Such approaches avoid multiple sequence alignment, and for this reason are termed *alignment-free*. As opposed to heuristics in multiple sequence alignment, these methods provide exact solutions. Various modifications are available, e.g. the use of degenerate k-mers, scoring match lengths rather than k-mer composition, and grammar-based techniques; see recent reviews[16,17] for more detail. Methods for inferring lateral genetic transfer have also been developed[18,19]. Importantly, evolutionary relationships can also be depicted as a network, with taxa and relationships represented respectively as nodes and edges[20–24], rather than as a strictly bifurcating tree. Using simulated and empirical sequence data, we recently demonstrated that alignment-free approaches can yield phylogenetic trees that are biologically meaningful[25–27]. We find that these approaches are more robust to genome rearrangement and lateral genetic transfer, and are highly scalable[25,26], a much-desired feature given the current deluge of sequence data facing the research community[28]. Here we extend the alignment-free phylogenetic approaches on 143 bacterial and archaeal genomes to generate a network of phylogenetic relatedness, and assess biological implications of this network relative to the phylogenetic tree. The phylogenetic relationships among these genomes have been carefully studied using the standard approach based on multiple sequence alignment[10] and an alignment-free approach[25]; this dataset thus provides a good reference for comparison.

Methods

Using 143 complete genomes of Bacteria and Archaea[25], we inferred the relatedness of these genome sequences using an alignment-free method based on the D_2^S statistic[29,30]. We computed a D_2^S distance, d for each possible pair of 143 genomes based on the presence of shared 25-mers using jD2Stat version 1.0 (http://bioinformatics.org.au/tools/jD2Stat/)[26] and following Bernard *et al.*[25]. Here the distance d is normalised based on genome sizes and the probabilities that corresponding k-mers occur in the compared sequences[29,30]; d ranges between 0.0 (i.e. two genomes are identical) and 15.5 (< 0.0001% 25-mers are shared between the two genomes). For a pair of genomes a and b, we transformed d_{ab} into a similarity measure S_{ab}, in which $S_{ab} = 10 - d_{ab}$. We ignore instances of $d > 10$, as these pairs of sequences share $\leq 0.01\%$ of 25-mers (i.e. there is little evidence of homology). To visualise the phylogenetic relatedness of these genomes, we adopted the D3 JavaScript library for data-driven documents (https://d3js.org/). In this network, each node represents a genome, and an edge connecting two nodes represents the qualitative evidence of shared *k*-mers between them. We set a threshold function *t* for which only edges with $S \geq t$ are displayed on the screen. Changing *t* dynamically changes the network structure. The resulting dynamic network is available at http://bioinformatics.org.au/tools/AFnetwork/.

Results and discussion

Figure 1 shows the phylogenetic tree of the 143 Bacteria and Archaea genomes that we previously inferred using an alignment-free method based on the D_2^S statistic[29,30]. In an earlier study[10], a supertree was generated for these genomes, summarising 22,432 protein phylogenies. Incongruence between the two trees was observed in 42% of the bipartitions, most of which are at terminal branches[25]. The alignment-free tree (Figure 1) recovers 13 out of the 15 "backbone" nodes[10], distinct clades of Archaea and Bacteria, a monophyletic clade of Proteobacteria, and the lack of resolution between gamma- and beta-Proteobacteria, in agreement with previously published studies; as such, this tree captures most of the major biological groupings of Bacteria and Archaea as presently understood.

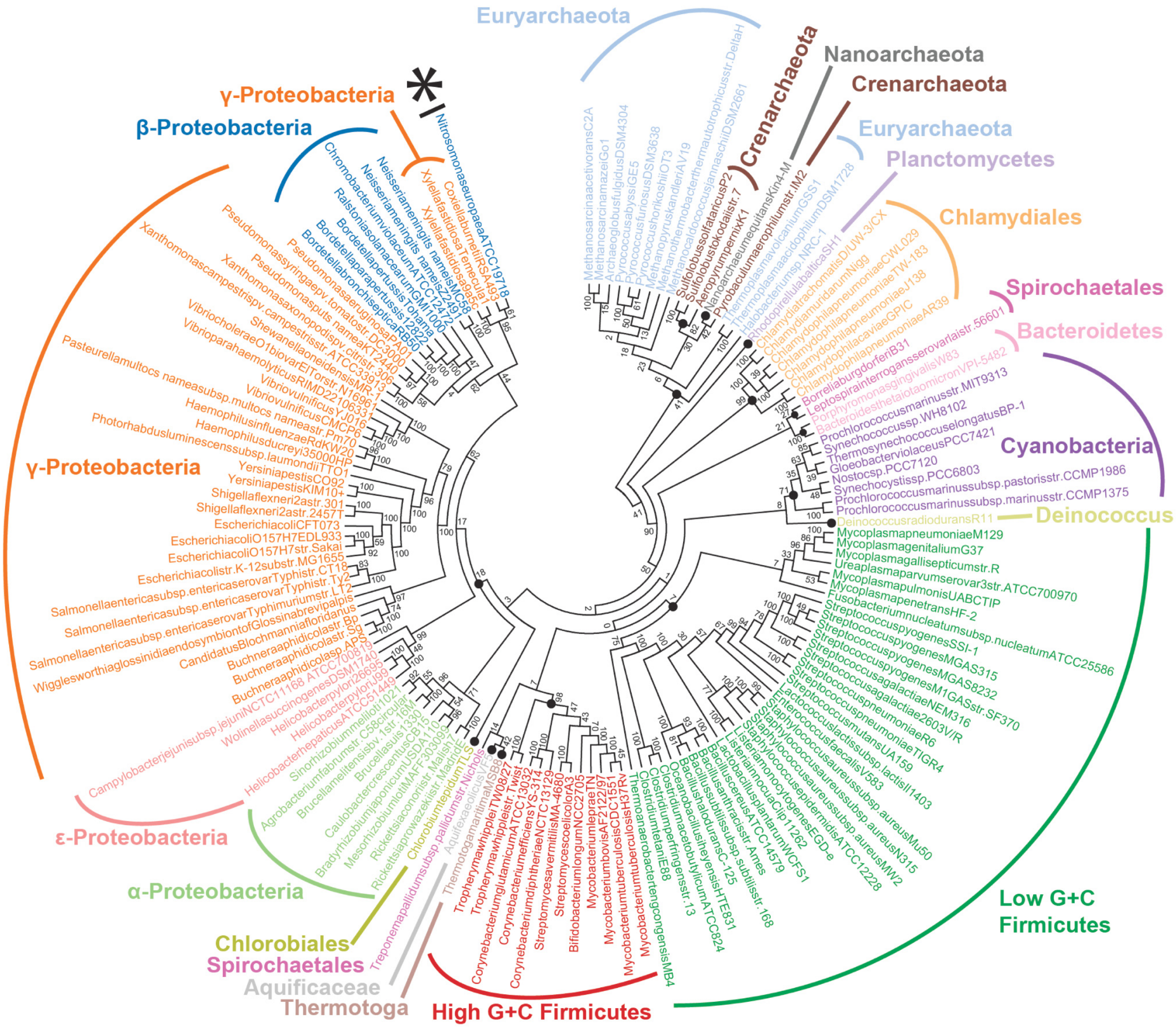


Figure 1. The alignment-free phylogenetic tree topology of the 143 Bacteria and Archaea genomes based on D_2^S statistic, modified based on the tree in Bernard *et al.*[25]; jackknife support at each internal node is shown. Each phylum is represented in a distinct colour, and the backbones identified in Beiko *et al.*[10] are shown on the internal node with black filled circles. The association of *Coxiella burnetii* and *Nitrosomonas europaea* is marked with an asterisk.

Figure 2 shows the network of phylogenetic relatedness of the same 143 genomes; a dynamic view of this network is available at http://bioinformatics.org.au/tools/AFnetwork/. As in our tree (Figure 1), Archaea and Bacteria form two separate paracliques; even at $t = 0$, we found only one archaean isolate (the euryarchaeote *Methanocaldococcus jannaschii* DSM 2661) linked to the bacterial groups Thermotogales and Aquificales[25]. Upon reaching $t = 3$, most of the 14 phyla have formed distinct densely connected subgraphs in our network, i.e. Cyanobacteria and Chlamydiales form cliques at $t = 1.5$ and all subgroups of Proteobacteria form a large paraclique with the Firmicutes at $t = 2$. Four *Escherichia coli* and two *Shigella* isolates, known to be closely related, form a clique up to $t = 8.5$. Interestingly, this network also showcases the extent that genomic regions are shared among diverse phyla, e.g. the high extent of genetic similarity among Proteobacteria *versus* the low extent between Chlamydiales and Cyanobacteria. Our observations largely agree with published studies[10,25], but also highlight the inadequacy of representing microbial phylogeny as a tree. For instance, in the tree *Coxiella burnetii*, a member of the gamma-Proteobacteria, is grouped with *Nitrosomonas europaea* of the alpha-Proteobacteria (marked with an asterisk in Figure 1); in the network, the strongest connection of *C. burnetii* is with *Wigglesworthia glossinidia*, a member of the gamma-Proteobacteria (marked with asterisk in Figure 2) at $t = 2$. Both *W. glossinidia* and *C. burnetii* are parasites; the *W. glossinidia* genome (0.7 Mbp) is highly reduced[31] and the *C. burnetii* genome (2 Mbp) is proposed to be undergoing reduction[32]. As both the tree (Figure 1) and network presented here were generated using the same alignment-free method, the contradictory position of *C. burnetii* is likely caused by the neighbour-joining algorithm used for tree inference[25]. In this scenario, the *C. burnetii* genome connects with *N. europaea* because it shares high similarity with *N. europaea* and *Neisseria* genomes of the beta-Proteobacteria (S between 1.43 and 1.68), second only to *W. glossinidia* ($S = 2.05$), and because it shares little or no similarity with other genomes of gamma-Proteobacteria that are closely related to *W. glossinidia*, i.e. *Buchnera aphidicola* isolates (average $S = 0.63$) and "*Candidatus* Blochmannia floridanus" ($S = 0$).

By changing the threshold t, we can dynamically visualise changes in the network structure. These changes are not random, but appear to correlate to the evolutionary history of the species. At $t = 0$, Archaea and Bacteria form two distinct paracliques, linked only by two edges, and the Planctomycetes isolate forms a singleton. When we increase t from 1 to 2, the Archaea and Bacteria paracliques quickly dissociate from each other; within the Bacteria, cliques of Chlamydiales and Cyanobacteria are formed and the Spirochaetales become isolated. Going from $t = 2$ to $t = 3$ we observe a scission between Firmicutes and Proteobacteria, and at $t > 3$ all classes of Proteobacteria start to form respective paracliques. The separation (as t is incremented) of a densely connected subgraph involving all representatives of a phylum, from the rest of the network mimics the divergence of this phylum from a common ancestor. Because the similarity measures do not have a unit (such as number of substitutions per site), it is not straightforward to interpret S as an evolutionary rate or divergence time. A comprehensive comparative analysis between our network here and one that is generated using multiple sequence alignment is beyond the scope of this work. However, our findings suggest that our alignment-free network yields snapshots of biologically meaningful evolutionary relationship among these genomes, and that increasing the threshold based on the proportion of shared k-mers recapitulates the progressive separation of genomic lineages in evolution.

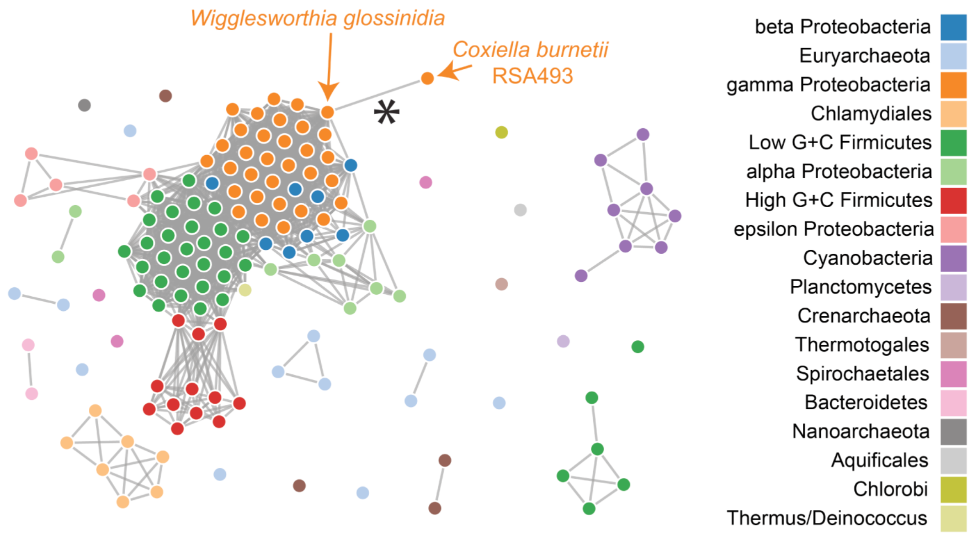


Figure 2. Alignment-free phylogenetic network of the 143 Bacteria and Archaea genomes based on D_2^S statistic using 25-mers, at $t = 2$. Each phylum is represented in a distinct colour, each node represents a genome and an edge represents a qualitative evidence of shared 25-mers between two genomes. The association between *Coxiella burnetii* and *Wigglesworthia glossinidia* is marked with an asterisk.

The alignment-free network reconstructed using whole-genome sequences thus recovers phylogenetic signals that cannot be captured in a binary tree. Using this approach, we generated the network in < 30 minutes; a whole-genome alignment of 143 sequences would have taken days, and even then, the alignment would be difficult to interpret given the genome dynamics in Bacteria and Archaea[9–13,33]. One can imagine inferring a network of thousands of microbial genomes in a few hours using distributed computing. More importantly, the network can be visualised dynamically, explored interactively and shared.

Other biological questions could be addressed by linking the *k*-mers to their genomic locations and annotated genome features, e.g. in a relational database[34]. For instance, we could use such a database to compare thousands of isolates and identify core gene functions for a specific phylum or genus, or exclusive *versus* non-exclusive functions in bacterial pathogens, in a matter of seconds. We can also use *k*-mers to quickly search for biological information e.g. functions relevant to lateral genetic transfer, recombination or duplications.

In contrast to Haeckel's "Biogenetic Law", *k*-mers used in this way recapitulate phylogenetic signal, not ontogeny. Alignment-free approaches generate a biologically meaningful phylogenetic inference, and are highly scalable. More importantly, representing alignment-free phylogenetic relationships using a network captures aspects of evolutionary histories that are not possible in a tree. As more genome data become available, Haeckel's goal of depicting the History of Life is closer to reality.

Data availability

The 143 Bacteria and Archaea genomes used in this work are the same dataset used in an earlier study[25], available at http://dx.doi.org/10.14264/uql.2016.908[35]. The dynamic phylogenetic network of these genomes is available at http://bioinformatics.org.au/tools/AFnetwork, with the source code available at http://dx.doi.org/10.14264/uql.2016.952[36]

Author contributions

GB, MAR and CXC conceived the study and designed the experiments. GB carried out the experiments. GB and CXC prepared the first draft of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

We thank funding support from the Australian Research Council (DP150101875) awarded to MAR and CXC, and a James S. McDonnell Foundation grant awarded to MAR.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Dayrat B: **The roots of phylogeny: how did Haeckel build his trees?** *Syst Biol.* 2003; **52**(4): 515–27.
PubMed Abstract | Publisher Full Text
2. Haeckel E: **Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenztheorie.** Bd. 1 und 2. Berlin: Reimer; 1866.
Publisher Full Text
3. Haeckel E: **Natürliche Schöpfungsgeschichte.** Berlin: Reimer; 1868.
Reference Source
4. Burkhardt RW Jr: **Lamarck, evolution, and the inheritance of acquired characters.** *Genetics.* 2013; **194**(4): 793–805.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Fitch WM: **Homology: a personal view on some of the problems.** *Trends Genet.* 2000; **16**(5): 227–31.
PubMed Abstract | Publisher Full Text
6. Hall BK: **Homology: the hierarchical basis of comparative biology.** San Diego: Academic Press; 1994.
Reference Source
7. Notredame C: **Recent progress in multiple sequence alignment: a survey.** *Pharmacogenomics.* 2002; **3**(1): 131–44.
PubMed Abstract | Publisher Full Text
8. Notredame C: **Recent evolutions of multiple sequence alignment algorithms.** *PLoS Comput Biol.* 2007; **3**(8): e123.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Darling AE, Miklós I, Ragan MA: **Dynamics of genome rearrangement in bacterial populations.** *PLoS Genet.* 2008; **4**(7): e1000128.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Beiko RG, Harlow TJ, Ragan MA: **Highways of gene sharing in prokaryotes.** *Proc Natl Acad Sci U S A.* 2005; **102**(40): 14332–7.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science.* 1999; **284**(5423): 2124–9.
PubMed Abstract | Publisher Full Text
12. Koonin EV: **Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions [version 1; referees: 2 approved].** *F1000Res.* 2016; **5**: pii: F1000 Faculty Rev-1805.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Puigbò P, Lobkovsky AE, Kristensen DM, *et al.*: **Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes.** *BMC Biol.* 2014; **12**: 66.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Adl SM, Simpson AG, Lane CE, *et al.*: **The revised classification of eukaryotes.** *J Eukaryot Microbiol.* 2012; **59**(5): 429–93.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Spang A, Saw JH, Jørgensen SL, *et al.*: **Complex archaea that bridge the gap between prokaryotes and eukaryotes.** *Nature.* 2015; **521**(7551): 173–9.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Bonham-Carter O, Steele J, Bastola D: **Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis.** *Brief Bioinform.* 2014; **15**(6): 890–905.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Haubold B: **Alignment-free phylogenetics and population genetics.** *Brief Bioinform.* 2014; **15**(3): 407–18.
PubMed Abstract | Publisher Full Text
18. Cong Y, Chan YB, Ragan MA: **A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF.** *Sci Rep.* 2016; **6**: 30308.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Domazet-Lošo M, Haubold B: **Alignment-free detection of local similarity among viral and bacterial genomes.** *Bioinformatics.* 2011; **27**(11): 1466–72.
PubMed Abstract | Publisher Full Text
20. Corel E, Lopez P, Méheust R, *et al.*: **Network-thinking: graphs to analyze microbial complexity and evolution.** *Trends Microbiol.* 2016; **24**(3): 224–37.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Dagan T: **Phylogenomic networks.** *Trends Microbiol.* 2011; **19**(10): 483–91.
PubMed Abstract | **Publisher Full Text**

22. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol.* 2006; **23**(2): 254–67.
PubMed Abstract | **Publisher Full Text**

23. Huson DH, Scornavacca C: **A survey of combinatorial methods for phylogenetic networks.** *Genome Biol Evol.* 2011; **3**: 23–35.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Kunin V, Goldovsky L, Darzentas N, *et al.*: **The net of life: reconstructing the microbial phylogenetic network.** *Genome Res.* 2005; **15**(7): 954–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Bernard G, Chan CX, Ragan MA: **Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer.** *Sci Rep.* 2016; **6**: 28970.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Chan CX, Bernard G, Poirion O, *et al.*: **Inferring phylogenies of evolving sequences without multiple sequence alignment.** *Sci Rep.* 2014; **4**: 6504.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. Ragan MA, Bernard G, Chan CX: **Molecular phylogenetics before sequences: oligonucleotide catalogs as *k*-mer spectra.** *RNA Biol.* 2014; **11**(3): 176–85.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Chan CX, Ragan MA: **Next-generation phylogenomics.** *Biol Direct.* 2013; **8**: 3.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Reinert G, Chew D, Sun F, *et al.*: **Alignment-free sequence comparison (I): statistics and power.** *J Comput Biol.* 2009; **16**(12): 1615–34.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Wan L, Reinert G, Sun F, *et al.*: **Alignment-free sequence comparison (II): theoretical power of comparison statistics.** *J Comput Biol.* 2010; **17**(11): 1467–90.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Akman L, Yamashita A, Watanabe H, *et al.*: **Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia.*** *Nat Genet.* 2002; **32**(3): 402–7.
PubMed Abstract | **Publisher Full Text**

32. Seshadri R, Paulsen IT, Eisen JA, *et al.*: **Complete genome sequence of the Q-fever pathogen *Coxiella burnetii.*** *Proc Natl Acad Sci U S A.* 2003; **100**(9): 5455–60.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

33. Dagan T, Martin W: **The tree of one percent.** *Genome Biol.* 2006; **7**(10): 118.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

34. Greenfield P, Roehm U: **Answering biological questions by querying k-mer databases.** *Concurr Comput Pract Exper.* 2013; **25**(4): 497–509.
Publisher Full Text

35. Bernard G, Chan CX, Ragan MA: **143 Prokaryote genomes.** Dataset. 2016.
Data Source

36. Bernard G, Chan CX, Ragan MA: **Alignment-free network of 143 prokaryote genomes.** Dataset. 2016.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓

Version 2

Referee Report 23 December 2016

doi:10.5256/f1000research.11322.r18754

✓

Weilong Hao
Department of Biological Sciences, Wayne State University, Detroit, MI, USA

The authors' responses are acceptable.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 13 December 2016

doi:10.5256/f1000research.11014.r18060

?

Weilong Hao
Department of Biological Sciences, Wayne State University, Detroit, MI, USA

The manuscript uses *k*-mers from whole-genome sequences to recapitulate phylogenetic relationships from trees to networks. The analyses seemed to be convincing, and of general interest. I just have some comments on the manuscript structure and some other minor suggestions.

The authors used Ernst Haeckel’s phylogeny and Biogenetic Law to start their manuscript. Although it is fun to read all these historical pieces, the link between Haeckel’s ideas and the construction of networks using *k*-mers was not made strong in the current version of the manuscript.

The authors compared alignment-free data against sequence alignments, and stated that the sequence alignment approach “ignores important evolutionary processes that are known to shape the genomes of microbes” followed by mentioning recombination, genome rearrangement, and lateral gene transfer. This is not accurate, as sequence alignments can also be used to reconstruct web-like phylogenetic relationships, which are sometimes called phylogenetic networks (e.g., Huson and Bryant 2006). I think it is important to carefully define and compare the networks mentioned in this manuscript and the phylogenetic networks mentioned by Huson and Bryant. Along this line, approaches based on sequence alignments might not all assume tree-like relationship. Furthermore, the authors mentioned evolutionary events, such as recombination, genome rearrangement, and lateral gene transfer, that are difficult to study using sequence alignments, but did not provide detailed evidence on whether *k*-mers can tackle them all. I suggest the authors to rather stay closer to their data and make more specific statements.

In the third introduction paragraph, “By default, it is assumed that the best alignment can be achieved simply by displaying the sequences in the same direction and inserting gaps where needed. This assumption is largely valid when working with exons or proteins of morphologically complex eukaryotes. However, in microbes this assumption is violated...” I feel the meaning of “assumption” in each of these sentences is a moving target. If they are talking about orthologous sequences, the analysis of orthologs should hold for both eukaryotes and prokaryotes. The key here, I guess, is the comparison of ortholgs, versus, the comparison of exenologs even non-homologs. Another minor point is the use of “microbes”, which can mean, bacteria, archaea, and small-eukaryotes. I don’t think it is a good word to use here.

The authors did not justify the use of the 143 genomes. It seemed that they were inherited from their previous study conducted some time ago, and likely skewed in terms of taxon-sampling. Since taxon-samping is important for tree-like phylogenetic analysis, it would be nice to address how the improved (or more balanced) taxon-sampling can benefit the network analyses.

The authors wrote “... in agreement with previously published studies; as such, this tree represents reality as presently understood, i.e., is biologically correct”. The use of words such as reality, biologically correct here, is inappropriate.

The data of Wigglesworthia, Coxiella and others are of potential interest. The readers would definitely appreciate some real data analyses to address them, which are currently lacking.

The cited references are relatively recent and skewed. Some of the older and more influential papers need to be added (for both networks and alignment free).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response (*Member of the F1000 Faculty*) 15 Dec 2016

Cheong Xin Chan, Institute for Molecular Bioscience, The University of Queensland, Australia

Thank you for these comments.

- **The link between Haeckel’s ideas and the construction of networks using *k*-mers was not made strong in the current version of the manuscript.**

The work we present here is a proof-of-concept for a biologically informative network based on *k* -mers extracted from whole-genome sequences. We hope to convince readers that dynamic visualization of such a network is intuitive for exploring and addressing biological questions, aiding discovery. The paper is part of a special collection of F1000Research articles in phylogenetics, commemorating the 150th anniversary of Ernst Haeckel’s Tree of Life published in 1866. Here we argue that by using *k*-mers we can recapitulate phylogenetic signal, somewhat in the same spirit as Haeckel famously argued that “ontogeny recapitulates phylogeny”. More precisely, our claim is that “increasing the threshold based on the proportion of shared *k*-mers recapitulates the progressive separation of genomic lineages in evolution”. Full consideration of Haeckel’s work in the context of

Darwinian evolution then and today is well beyond the scope of our brief paper, although we cite some key references.

- **The authors … stated that the sequence alignment approach “ignores important evolutionary processes that are known to shape the genomes of microbes” followed by mentioning recombination, genome rearrangement, and lateral gene transfer. This is not accurate, as sequence alignments can also be used to reconstruct web-like phylogenetic relationships, which are sometimes called phylogenetic networks (e.g., Huson and Bryant 2006). I think it is important to carefully define and compare the networks mentioned in this manuscript and the phylogenetic networks mentioned by Huson and Bryant. Along this line, approaches based on sequence alignments might not all assume tree-like relationship.**

We agree and have now rewritten part of the Abstract to stage our argument more clearly: genomic processes in microbes can undermine the assumptions that underlie multiple sequence alignment, hence phylogenetic inference as usually practiced. We have now cited other articles on phylogenetic networks in the text where appropriate, specifically Huson and Bryant[1] and Kunin *et al.*[2]. Comprehensive comparison of *k*-mer-based and (alignment-based) phylogenetic networks is important but, due to its complexity, beyond the scope of this paper; we have now clarified this in the revised text.

- **The authors mentioned evolutionary events, such as recombination, genome rearrangement, and lateral gene transfer … but did not provide detailed evidence on whether k-mers can tackle them all. I suggest the authors to rather stay closer to their data and make more specific statements.**

In Chan *et al.*[3] and Bernard *et al.*[4] we provided detailed evidence that alignment-free approaches based on *k*-mers, at multi-genome scale, can be robust to insertions/deletions, genome rearrangement and lateral genetic transfer; these articles are cited where appropriate.

- **In the third introduction paragraph, “By default, it is assumed that the best alignment can be achieved simply by displaying the sequences in the same direction and inserting gaps where needed. This assumption is largely valid when working with exons or proteins of morphologically complex eukaryotes. However, in microbes this assumption is violated...” I feel the meaning of “assumption” in each of these sentences is a moving target. If they are talking about orthologous sequences, the analysis of orthologs should hold for both eukaryotes and prokaryotes.**

We have now revised the text to make it clear that the main assumption underlying multiple sequence alignment, i.e. that the alignment columns display homology position-by-position along the length of the sequences, is largely valid when working with highly conserved orthologs of any source; and that the validity of this assumption is often undermined in the case of microbial genome sequences, due to recombination and rearrangement.

- **Another minor point is the use of “microbes”, which can mean, bacteria, archaea, and small-eukaryotes. I don’t think it is a good word to use here.**

We used the word “microbes” here specifically to include archaea, bacteria and microbial eukaryotes. Genomes of many microbial eukaryotes are known to be impacted by lateral genetic transfer, at frequencies sometimes nearly as large as in bacteria and archaea.

- **The authors did not justify the use of the 143 genomes. … Since taxon-sampling is important for tree-like phylogenetic analysis, it would be nice to address how the improved (or more balanced) taxon-sampling can benefit the network analyses.**

Here we used the 143-genome dataset because the phylogenetic relationships among these genomes have been studied using careful alignment-based methods[5] and by alignment-free approaches[4]; it thus provides a good reference for comparison. We have now clarified this in the text. In our alignment-free network, each edge represents the qualitative evidence of *k*-mers shared pairwise between two genomes. This evidence is not affected by other genomes present in (or absent from) the dataset. Therefore, our networks are not affected by taxon-sampling biases of the sort encountered in tree inference. Of course, the presence or absence of a critical node (genome) might affect the biological conclusion we draw from a network, but the same is true for any scientific analysis. We considered the effect of phyletic balance on the inference of lateral genetic transfer networks in another context[6].

- **The authors wrote "... in agreement with previously published studies; as such, this tree represents reality as presently understood, i.e., is biologically correct". The use of words such as reality, biologically correct here, is inappropriate.**

We agree and now state that "as such, this tree captures most of the major biological groupings of Bacteria and Archaea as presently understood".

- **The data of *Wigglesworthia*, *Coxiella* and others are of potential interest. The readers would definitely appreciate some real data analyses to address them, which are currently lacking.**

A follow-up analysis between *Wigglesworthia* and *Coxiella* would indeed be interesting, but is beyond the scope of this Research Note, the aim of which is to present limited findings in hopes of inspiring and encouraging others to explore this research area.

- **Some of the older and more influential papers need to be added (for both networks and alignment free).**

We have now cited older, relevant references in the text for both networks[1, 2] and alignment-free methods[7].

References

1. Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006; **23**(2): 254-67.
2. Kunin V, Goldovsky L, Darzentas N*, et al.*: The net of life: reconstructing the microbial phylogenetic network. *Genome Res*. 2005; **15**(7): 954-9.
3. Chan CX, Bernard G, Poirion O*, et al.*: Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep*. 2014; **4**: 6504.
4. Bernard G, Chan CX, Ragan MA: Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci Rep*. 2016; **6**: 28970.
5. Beiko RG, Harlow TJ, Ragan MA: Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A*. 2005; **102**(40): 14332-7.
6. Cong Y, Chan YB, Ragan MA: Exploring lateral genetic transfer among microbial genomes using TF-IDF. *Sci Rep*. 2016; **6**: 29319.
7. Domazet-Lošo M, Haubold B: Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics*. 2011; **27**(11): 1466-72.

Competing Interests: No competing interests were disclosed.

Referee Report 12 December 2016

doi:10.5256/f1000research.11014.r18402

✓ **Bernhard Haubold**

Department Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany

Phylogeny reconstruction is a classical research topic in bioinformatics. In this context the standard trade-off between speed and accuracy becomes a choice between slow but accurate sequence alignment on the one hand and fast but less accurate alignment-free methods on the other. Bernard *et al.* aim for speed and use an established alignment-free measure, D_2, to reconstruct the phylogeny of 143 Bacteria and Archaea from full genome sequences. D_2 is based on the number of shared *k*-mers, and the main contribution of the paper is the visualization of the D_2 distance matrix of the 143 taxa as a network rather than the traditional bifurcating tree. This visualization is dynamic in the sense that the user can choose a similarity threshold between 0 and 10, and watch as the taxa disintegrate from initially two clusters to essentially every taxon on its own. This is an innovative way of presenting large-scale evolutionary relationships, and the tool is fun to use. As the authors remark, it is unclear how the D_2 metric scales with more familiar measures of evolutionary time such as substitutions per site. It would thus be interesting to explored this in future work; for example by supplying a version of the visualization tool that allows users to upload their own sequences. I was also wondering how the networks generated by Bernard *et al.* compare to established methods of network-based evolutionary analysis such as SplitsTree and minimum spanning trees. I realize that these are both usually based on alignments, but it is always possible to analyze a given alignment using D_2, thereby allowing a direct assessment of the accuracy lost (if any) for the speed gained.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response (*Member of the F1000 Faculty*) 15 Dec 2016

Cheong Xin Chan, Institute for Molecular Bioscience, The University of Queensland, Australia

Thank you for these comments. Indeed, the correlation between D2 metrics and evolutionary distances is an interesting area, and a tool that allows users to upload their own datasets would be useful. A comparative analysis between a *k*-mer-based network and a phylogenetic network based on multiple sequence alignment, although doable, is not straightforward. We believe the adoption of alignment-free methods in phylogenetic inference is still in its infancy, and we hope that this work will inspire and encourage other researchers to pursue this approach.

Competing Interests: No competing interests were disclosed.