A new genomics drive taxonomy: Are we there, yet?

George M. Garrity
Then and now: a systematic review of the systematics of prokaryotes in the last 80 years

Aharon Oren · George M. Garrity

Fig. 1 Then and now. A comparison of the accepted taxonomies in use at the time of the first and current issues of Antonie van Leeuwenhoek. Printed with permission. Copyright, NamesforLife, LLC 2013. All rights reserved.
Dispelling a few misconceptions

Valid taxa, valid names and an official taxonomy
  – Only nomenclature is governed
  – The concept of validly published names

Taxonomic proposals and names
  – Hypotheses, not facts
  – Falsifiable

Defining species and higher taxa
  – The “Gold Standard” and rules of thumb
  – Polyphasic methods and 16S rRNA sequence
A few important terms and concepts

- Taxonomy
- Types and typification
- Circumscription and diagnosis
- “taxon calling”

- Metadata
- Knowledge management and knowledgebases
- Findability
- Scalability
Crowdsourcing the global prokaryotic taxonomy

The current 16S rRNA based taxonomy
In development for 35 years
Work product of > 17,500 authors
> 20,600 taxonomic description
12,195 effective publications
Earliest taxa > 175 yrs
Average age 16.8 yrs
13.8% synonyms
7.8% explicitly emended

Validly published named prokaryotes used in the analyses

<table>
<thead>
<tr>
<th>Rank</th>
<th>Validly Published</th>
<th>Coverage (16S)(^a)</th>
<th>Genomes</th>
<th>Outliers</th>
<th>Singleton</th>
<th>2 or 3</th>
<th>4 to 9</th>
<th>≥10</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>13.134</td>
<td>11,800</td>
<td>3842(^d)</td>
<td>274</td>
<td>13,134</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>genus</td>
<td>2,666</td>
<td>2422</td>
<td>1289</td>
<td>131</td>
<td>1,253</td>
<td>636</td>
<td>467</td>
<td>283</td>
</tr>
<tr>
<td>family</td>
<td>369</td>
<td>451(^b)</td>
<td>356</td>
<td>86</td>
<td>74</td>
<td>49</td>
<td>64</td>
<td>182</td>
</tr>
<tr>
<td>order</td>
<td>160</td>
<td>202(^b)</td>
<td>178</td>
<td>47</td>
<td>32</td>
<td>18</td>
<td>23</td>
<td>87</td>
</tr>
<tr>
<td>class</td>
<td>67</td>
<td>85(^b)</td>
<td>72</td>
<td>19</td>
<td>14</td>
<td>9</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>phylum</td>
<td>0</td>
<td>34(^c)</td>
<td>33</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>domain</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Source - NamesforLife, LLC, based on validly published named prokaryotic species, subspecies and higher taxa as of May 31, 2016; adjusted for synonyms, basonyms, orthographic corrections and rejected names arising from Judicial Opinions.
CRUCIAL QUESTIONS

Here are five questions that anyone conducting or evaluating this research should ask to keep from getting carried away by hype.

Can experiments detect differences that matter? Profiling a microbiome could produce a catalogue at the level of phyla, species or genes. Much work relies on analysis of 16S rRNA, an ancient gene that tolerates little variation and so is reliably found across the bacterial kingdom. But this allows only a coarse sorting. For example, microbiomes associated with obesity have been distinguished by different ratios of bacterial phyla, which encompass a staggering range of diversity. If this criterion were used to characterize animal communities, an aviary of 100 birds and 25 snails would be considered identical to an aquarium with 8 fish and 2 squid, because each has four times as many vertebrates as molluscs. Even within a single species, strains often differ greatly in the genes they contain.
“The greatest value of a picture is when it forces us to notice what we never expected to see. ”

John W. Tukey
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.

Procarvotic 16S sequences (RDP 7.0)

Scree plot for 16S sequences (n=4405)
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.

Archaea

Bacteria
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Recovering shapes from the data
Lesson learned
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Refinement of the underlying taxonomy

Improved version of the Self-Organizing Self Correcting Classifier
Smooth classification and protect nomenclatural integrity

Input
11,800 dereplicated HQ full-length 16S rRNA sequences
Aligned to

Output filtered to N4L-exemplar DOIs
Select validly published named species/subspecies
Identify species with > 3 members (within group distribution)
as represented in the lower triangle of each corresponding
sub-matrix (2σ)
Visualize output as phylograms (APE), heatmaps, and histograms.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
16S rRNA similarity of type strains of Bacteria and Archaea 1980
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
16S rRNA similarity of type strains of Bacteria and Archaea 1982
16S rRNA similarity of type strains of Bacteria and Archaea 1983
16S rRNA similarity of type strains of Bacteria and Archaea 1984
16S rRNA similarity of type strains of Bacteria and Archaea 1985
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
16S rRNA similarity of type strains of Bacteria and Archaea 1987
16S rRNA similarity of type strains of Bacteria and Archaea 1988
16S rRNA similarity of type strains of Bacteria and Archaea 1990
16S rRNA similarity of type strains of Bacteria and Archaea 1990
16S rRNA similarity of type strains of Bacteria and Archaea 1992
16S rRNA similarity of type strains of Bacteria and Archaea 1994
16S rRNA similarity of type strains of Bacteria and Archaea 1996
16S rRNA similarity of type strains of Bacteria and Archaea 1997
16S rRNA similarity of type strains of Bacteria and Archaea 1998
16S rRNA similarity of type strains of Bacteria and Archaea 1999
16S rRNA similarity of type strains of Bacteria and Archaea 2000
16S rRNA similarity of type strains of Bacteria and Archaea 2001
16S rRNA similarity of type strains of Bacteria and Archaea 2002
16S rRNA similarity of type strains of Bacteria and Archaea 2003
16S rRNA similarity of type strains of Bacteria and Archaea 2004
16S rRNA similarity of type strains of Bacteria and Archaea 2005
16S rRNA similarity of type strains of Bacteria and Archaea 2006
16S rRNA similarity of type strains of Bacteria and Archaea 2007
16S rRNA similarity of type strains of Bacteria and Archaea 2008
16S rRNA similarity of type strains of Bacteria and Archaea 2009
16S rRNA similarity of type strains of Bacteria and Archaea 2010
16S rRNA similarity of type strains of Bacteria and Archaea 2011
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.

16S rRNA similarity of type strains of Bacteria and Archaea 2012
16S rRNA similarity of type strains of Bacteria and Archaea 2013
16S rRNA similarity of type strains of Bacteria and Archaea 2014
16S rRNA similarity of type strains of Bacteria and Archaea 2015
The problem with heuristics
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Challenges

Taxonomic ranks in current use based on 16S rRNA similarity
  - Significantly overlap
  - Validly named taxa may be over-specified
  - Rank of order may not be justifiable by a single measure
  - Reclassification and re-naming of at least species may be justified

Additional measures could help resolve taxa above the genus level
  - Projection along orthogonal axes or planes

Obvious candidates

Methods that classify other genes, sequences, amino acids or proteins
  - Average nucleotide identity
  - Average amino acid identity
  - K-mers
  - Latent semantic analysis of DNA sequences
16S rRNA sequence subset

ANI, Oak Ridge National Laboratories

166 type strains, with conditioning of AAI signal

$\text{cor} = -0.46807$
16S rRNA sequence subset

AAI, Oak Ridge National Laboratories

166 type strains, with conditioning of AAI signal

\[ \text{cor} = -0.6631114 \]
16S rRNA sequence subset

Latent Semantic analysis, 50 words.
Univ. Nebraska and NamesforLife

166 type strains, with conditioning of LSA signal
$\text{cor} = -0.02849482$
Type strains reordered according to LSA classification

Type strain genomes reordered according to LSA
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
Results

Taxonomic resolving power of 16S rRNA
Taxonomic resolving power of ANI, AAI and LSA

Operational details
  Adequacy of metadata for downstream analyses

Surprises

Standards needed for:
  Naming and taxon calling
  Persistence and links to relevant literature
  Data quality  Weighting schemes
  Validation methods
  Models
  Software
Use this slide for a major point, takeaway, call out or quote – helps to accent the importance.
This work was funded through the Small Business technology transfer program of the United States Department of Energy under grants number DE-FG02-07ER86321 and DE-SC0006191. Funding for business development was provided through grants and loans from the Michigan Economic Development Corporation and the Michigan Universities Commercialization Initiative. NamesforLife semantic resolution technology is covered under US Patent 7,925,444 B2. The SOSCC systems and methods is covered under US Patent 8,036,997. Semiotic fingerprinting is covered under US 8,903,824. Semantic tagging, indexing, equivalency mapping, and latent semantic analysis of molecular sequence data are the subjects of pending US and EPO patent applications.