Impact of Large-Scale Genomics on Microbial Systematics

Hans-Peter Klenk
Newcastle University
Microbial Systematics – Analytical Methods and Regulations

Since decades a core duty performed at collections of microbial type material.

Characterisation - Classification - Nomenclature

Based on the analysis of phenotypic features, such as morphology, physiology, biochemistry ... DNA, RNA, proteins, chemotaxonomic markers such as fatty acids, quinones, peptidoglycans, polar lipids and polyamines

- laborious
- time consuming
- technically demanding
- standardization required
- not suitable for all organisms

→ since late 1980s triumph of 16S rRNA gene sequences as Gold Standard

Chemotaxonomy:
Application of analytical methods to collect information on various chemical constituents of the cells

Vandamme et al. 1996
1995: Dawn of Genomic Era – Promises to Overcome the Lack of Sequences that Limit Progress ...

Limitations of 16S rRNA, the dominant marker molecule – but also for RNAPs ...

- limited sequence space
- Inter-operon differences 0-9%
- seq. identity ≠ strain identity
... to coordinate an international effort to construct draft genome sequences of each of the roughly 6500 (bacterial) type strains deposited in public strain collections.

... continue with the collection of genome sequences as new species and type strains are described.

... revise and use phenotypic tests to consider genomics data for an improved species definition.

Recommendations of the American Academy of Sciences Colloquium on Reconciling Microbial Systematics & Genomics (2006)

Reconciling Microbial Systematics & Genomics

Buckley & Roberts, 2006
Availability of Microbial Genome Sequences at the Beginning of Systematic Genome Sequencing [2007]

Much higher coverage required as basis for systematics

pale blue: no genome sequenced

green: only few representatives sequenced

red: plenty of genome sequences known

Proteobacteria
Firmicutes
Actinobacteria

N.C. Kyrpides & P. Hugenholtz
The Primer, 2007
Growth in the Number of Validly Published Species Names [1980-2013] - Only About 1% of Expected Diversity Covered

Over the last few years, the no. of newly described type strains surpassed the no. of sequenced type strain genomes...
Aim: Use of Whole-Genome-Sequences (Drafts) to Infer the Phylogenetic Position of Organisms for Taxonomic Assessment of Species all Over Diversity

A systematic, genomic exploration of all species of Bacteria and Archaea with validly published names extended by diversity of fungi
Only a complete Catalogue of bacterial and archaeal genomes will enable us to navigate microbial diversity as safely as we can navigate the oceans by following the stars. Let’s aim for this!
Key Project Requirements for Success

• Sufficient financial resources for draft sequence generation
• Access to production facilities for sequence drafts
• Adequate procedures for draft assembly and automated annotation
• Open access databases for data storage (INSDC)
• Taxonomic journals to request genome sequences for description of sp. nov.
• Standardised procedures for efficient exchange of genomic data (GSC)
• Tools for inference of high quality whole genome phylogenies
• Reliable tools for strain discrimination (ANI/dDDH)
• Platform for dissemination of research results (‘SIGS‘)
• Access to cultures of all type stains (bottleneck)
Meeting report: GenBank microbial genomic taxonomy workshop
(12–13 May, 2015)

Scott Federhen1*, Ramon Rossello-Mora2, Hans-Peter Klenk3, Brian J. Tindall4, Konstantinos T. Konstantinidis5, William B. Whitman6, Daniel Brown7, David Labeda8, David Ussery9, George M. Garrity10, Rita R. Colwell11, Nur Hasan12, Joerg Graf13, Aidan Parte14, Pablo Yarza15, Brittany Goldberg16, Heike Sichtig16, Ilene Karsch-Mizrachi1, Karen Clark1, Richard McVeigh1, Kim D. Pruitt1, Tatiana Tatusova1, Robert Falk1, Sean Turner1, Thomas Madden1, Paul Kitts1, Avi Kimchi1, William Klimke1, Richa Agarwala1, Michael DiCuccio1 and James Ostell1

Abstract

Many genomes are incorrectly identified at GenBank. We developed a plan to find and correct misidentified genomes using genomic comparison statistics together with a scaffold of reliably identified genomes from type. A workshop was organized with broad representation from the bacterial taxonomic community to review the proposal, the GenBank Microbial Genomic Taxonomy Workshop, Bethesda MD, May 12–13, 2015.

Keywords: GenBank, Genomic taxonomy, Misidentified sequence entries
Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,1,* Aharon Oren,2 Antonio Ventosa,3 Henrik Christensen,4 David Ruiz Arahal,5 Milton S. da Costa,6 Alejandro P. Rooney,7 Hana Yi,8 Xue-Wei Xu,9 Sofie De Meyer10 and Martha E. Trujillo11,*

Abstract

Advancement of DNA sequencing technology allows the routine use of genome sequences in the various fields of microbiology. The information held in genome sequences proved to provide objective and reliable means in the taxonomy of prokaryotes. Here, we describe the minimal standards for the quality of genome sequences and how they can be applied for taxonomic purposes.

OGRI: a form of similarity or distance were coined as the overall genome related index
Workflow for Genome-based Classification at Species Level – Including OGRI


1. Genome assembly
2. Obtain full length 16S rRNA gene sequence
3. Search
4. Find the phylogenetic neighbors
5. If 16S sequence similarity is ≤ 98.7%:
   - New species
6. If 16S sequence similarity is ≥ 98.7%:
   - Calculate OGRI
   - If 95–96% ANI or 70% dDDH ≤ 95–96% ANI or 70% dDDH:
     - Identified as a known species

Phylogenomic tree required
Genome Sequence-based Methods for Species Discrimination
Superior to wet lab procedure

70% criterion for wet lab DDHs

Beginning in 2005 with Konstantinidis & Tiedje’s publication of ANI and continued with Auch et al.’s 2010 dDDH and the recent AF, gANI paper by Vargese et al. (2016) modern digital procedures surpass the accuracy of long established wet lab procedures for species discrimination e.g. by providing confidence intervals for distance functions (Maier-Kolthoff et al. 2013).
Within-species difference in G+C content <1% calculated from genome sequences

Confirmed in 9279 genome-sequence pairs
Phylogenomics Analysis Pipeline
Complex Bioinformatics to Infer Highly-reliable Phylogenies

Legend:
- Intermediate result
- Method
- Final result

automated data handling pipeline
Phylogenomic Trees with Extremely High Bootstrapping Support [GBDP procedure]

phylogenies from genome sequences and proteomes - generated by automated data handling pipelines [TYGS-Server]

Insights into the diversity of catabolic metabolism from 10 haloarchaeal genomes
Anderson et al. (2011) PLoS ONE 6:e120237
Dissemination of Results: Metadata! [2009-yesterday]

The Genome Report Series in Stand Genomic Sci

Table 1. Classification and general features of A. ferrooxidans ICP based on MGS recommendations [11]

<table>
<thead>
<tr>
<th>MGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Domain: Bacteria</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phylum: Acidobacteria</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class: Acidobacteria</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Order: Acidimicrobiales</td>
<td>HIR [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Suborder: Acidimicrobiales</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Family: Acidimicrobiae</td>
<td>IAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genus: Acidimicrobium</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Species: Acidimicrobium ferrooxidans</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type strain: ICP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gram stain: positive</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell shape: rod shaped</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motility: mobile</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sporulation: nonsporulating</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Temperature range: moderate thermophile, 45-50°C</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimum temperature: 48°C</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Salinity: not reported</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oxygen requirement: aerobic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carbon source: CO2, autotrophic, yeast extract (heterotrophic)</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy source: autotrophic: oxidation of ferric iron with oxygen as the electron acceptor</td>
<td>TAS [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-6: Habitat: free living</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-13: Biotic relationship: none</td>
<td>tas [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-14: Pathogenicity:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biosafety level: 1</td>
<td>NL  [I]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Isolation: hot springs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-4: Geographic location:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keflavik geothermal area, Iceland</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sample collection time: before 1993</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-4: Altitude: not reported</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-4.2: Latitude - Longitude:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>63.95, -22.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-4.3: Depth: not reported</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MGS-4.4: Altitude: not reported</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Genome sequencing project information

<table>
<thead>
<tr>
<th>MGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Finished</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Library: Illumina HiSeq SMRT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sequencing read length:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assembly: no assembly</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gene calling method:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Database: IMG-ER</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Source material identifier:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Project reference:</td>
<td></td>
</tr>
</tbody>
</table>

Ecological data - IDA: Inferred from Direct Aisys (first time in publication); TAS: Traceable Author Statement (i.e., a direct report exists in the literature); IAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence. These evidence codes are from the GenBank Codex project [3]). If the evidence code is IAS the property was directly observed for a living isolate by one of the authors or an expert mentioned in the acknowledgments.

Indexed in PubMedCentral, ISI, SCOPUS (2011)
Impact Factor (2014) 3.17
Since 3/2014 @ BMC
Impact from Pilot Projects: Genome-Based Taxonomic Classification of Bacteroidetes [2016]

About 90 emendations from 189 novel genome sequences
Impact from Projects: Genome-Based Taxonomic Classification of Actinobacteria [2018]

Genome-Based Taxonomic Classification of the Phylum Actinobacteria

Imen Nouiou1, Lorena Carro1, Marina García-López2, Jan P. Meier-Kolthoff2, Tanja Woyke3, Nikos C. Kyriakos3, Friedger Pukall1, Hans-Peter Knief1, Michael Goodfellow1 and Markus Göker2

Based on ~1,030 Actinobacteria genomes of the Genomic Encyclopaedia and from GenBank

2 ord. nov., 10 fam. nov., 17 gen. nov., >150 comb. nov.
Emendations of 2 classes, 5 orders, 19 families, 11 genera
Whole-genome phylogenies proved to be much better resolved than 16S rRNA phylogenies
Genome-based Taxonomic Classification of Actinobacteria [2018]

Phylum
- Actinobacteria-Phylum
- Chloroflexi

Class
- Acidimicrobia
- Actinobacteria
- Anaerolineae
- Ardenticatenia
- Caldilineae
- Chloroflexia
- Coriobacteria
- Dehalococcoidia
- Ktedonobacteria
- Nitriliruptoria
- Rubrobacteria
- Thermoleophilia
- Thermomicrobia

Order
- Actinobacteria
Genome-Based Taxonomic Classification of Actinobacteria [2018]

Applications of the Improved Classification of Actinobacteria

The results of this comparative phylogenetic study provide a much improved framework for the classification of the phylum Actinobacteria, one of the largest taxonomic lineages recognized within the domain Bacteria. The improved classification provides a sound basis for future studies on actinobacteria, not least on those of agricultural, biotechnological, clinical, and ecological interest. In particular, the clarification of relationships within the family that, until now, have confounded the best efforts of actinobacterial systematists (Cattley et al., 2012) will be of especial interest to those involved in developing biotechnological applications of strains belonging to taxa classified in emended families such as the Gordoniales, Micromonosporales, Nocardioidae, and Streptomycetidae. Similarly, improvements in the classification of taxa assigned to emended families, notably Actinomycetales and Corynebacteriales, will provide valuable leads to those working on organisms relevant to human and veterinary medicine. Indeed, it can be concluded that comparative genomic analyses not only provide help the "taxonomic null" (Dekn. and Goker, 2013; Stutliff et al., 2012), as exemplified in this study, but also provide invaluable insights into actinobacterial biology in its entirety thereby helping to revalidate prokaryotic systems as a fundamental scientific discipline.

Taxonomic Consequences: New Taxa

Description of Skirbatalis, ord. nov.

Cryptosporangiales (N.L. neut. n. Cryptosporangium, type genus of the order, -ales, ending to denote an order; N.L. gen. n. Cryptosporangium, the Cryptosporangium group). The description is given for Cryptosporangiales (Chet et al., 2009; Srinivasan et al., 2014). The type and only genus of the order is Cryptosporangium (Tamura et al., 1998).

Description of Phyllosporales, ord. nov.

Cryptosporales (N.L. neut. n. Cryptosporus, type genus of the order, -ales, ending to denote an order; N.L. gen. n. Cryptosporus, the Cryptosporus group). The description is given for Cryptosporales (Chet et al., 2009). The type and only genus of the order is Cryptosporus (Tamura et al., 1998).

119 pages (printed) – including Emendations – all validated in IJSEM
High Impact Through Validation of Names in Taxonomic Journal [2018]

List of new names and new combinations previously effectively but not validly published

Aharon Oren¹² and George M. Garrity³⁴

The purpose of this announcement is to effect the valid publication of the following effectively published new names and new combinations under the procedure described in the International Code of Nomenclature of Prokaryotes (2008 Revision). Authors and other individuals wishing to have new names and/or combinations included in future lists should send an electronic copy of the published paper to the IJSEM Editorial Office for confirmation that all of the other requirements for valid publication have been met. It is also a requirement of IJSEM and the ICSP that authors of new species, new subspecies and new combinations provide evidence that types are deposited in two recognized culture collections in two different countries. It should be noted that the date of valid publication of these new names and combinations is not the date of publication of this list, nor the date of the original publication of the names and combinations. The authors of the new names and combinations are as given below. Inclusion of a name on these lists validates the publication of the name and thereby makes it available in the nomenclature of prokaryotes. The inclusion of a name on this list is not to be construed as taxonomic acceptance of the taxon to which the name is applied. Indeed, some of these names may, in time, be shown to be synonyms, or the organisms may be transferred to another genus, thus necessitating the creation of a new combination.

<table>
<thead>
<tr>
<th>Name/Authors</th>
<th>Proposed as</th>
<th>Nomenclatural type¹</th>
<th>Priority²</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdnibacterota Tahan et al. 2018, 288</td>
<td>class. nov.</td>
<td>Abdnibacterota</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Abdnibacterota Tahan et al. 2018, 288</td>
<td>fam. nov.</td>
<td>Abdnibacterota</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Abdnibacterota Tahan et al. 2018, 288</td>
<td>ord. nov.</td>
<td>Abdnibacterota</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Abdnibacterota Tahan et al. 2018, 288</td>
<td>gen. nov.</td>
<td>Abdnibacterota</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Abdnibacterota salinae Tahan et al. 2018, 288</td>
<td>sp. nov.</td>
<td>Abdnibacterota salinae</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Transfer of *Hoyosella* and *Tomitella* to *Nocardiaceae*
Genome BLAST Distance Phylogeny [GBDP procedure] an also be Applied for Fungi

Figure 4: GBDP-based phylogeny of 15 Basidiomycota genomes reconstructed with FastML. The following GBDP settings were used: BLAST* with default settings, trimming algorithm and distance formula d4. The root was set via mid-point rooting [Farris et al. 1972]. The boxed species names are conflicting with the NCBI classification, but only with respect to taxa of uncertain position.

Comparative genomics of biotechnologically important yeasts

One Thousand Phylogenetically-selected Genomes: Dramatic Progress in Phylogenetic Coverage
Current status: the database of microbial type strain genome sequences has too many gaps and therefore is difficult to navigate for taxonomists.