## Title

The evolution of microbial species - a view through the genomic lens



## Authors

Varghese, Neha
Mukherjee, Supratim
ivanova, Natalia
et al.

## Publication Date

2014-03-19

# The evolution of microbial species – a view through the genomic lens

**Neha Varghese[1*], Supratim Mukherjee[1], Natalia Ivanova[1], Konstantinos Mavromatis[3], Konstantinos Konstantinidis[2], Nikos Kyrpides[1] and Amrita Pati[1]**

[1] LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

[2] Georgia Institute of Technology, Atlanta, GA USA

[3]Celgene Corporation, San Francisco, CA USA

**To whom correspondence should be addressed*: Email: Neha Varghese: njvarghese@lbl.gov, Supratim Mukherjee: supratimmukherjee@lbl.gov, or Amrita Pati:apati@lbl.gov

March 21, 2014

**ACKNOWLEDGMENTS:**

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was conducted under the LDRD grant YLD012: Computational, Data Management and Analysis Methods for the Study of a Rapidly Expanding Genome and Metagenome Sequence Data Space (YLD012). Development was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 and Los Alamos National Laboratory under contract DE-AC02-06NA25396.

# The evolution of microbial species – a view through the genomic lens

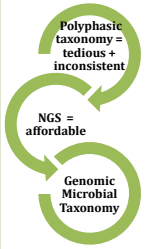
Neha J. Varghese[1], Supratim Mukherjee[1], Natalia N. Ivanova[1], Konstantinos Mavromatis[3], Konstantinos T. Konstantinidis[2], Nikos C. Kyrpides[1*] and Amrita Pati[1*]

*(1) DOE Joint Genomic Institute, Walnut Creek, CA (2) Georgia Institute of Technology, Atlanta, GA (3) Celgene Corp, San Francisco, CA*

## ABSTRACT

For a long time prokaryotic species definition has been under debate and a constant source of turmoil in microbiology. This has recently prompted the ASM to call for a scalable and reproducible technique, which uses meaningful commonalities to cluster microorganisms into groups corresponding to prokaryotic species. Whole-genome Average Nucleotide Identity (gANI) was previously suggested as a measure of genetic distance that generally agrees with prokaryotic species assignments based on the accepted best practices (DNA-DNA hybridization and 16S rDNA similarity). In this work, we prove that gANI is indeed the meaningful commonality based on which microorganisms can be grouped into the aforementioned clusters. By analyzing 1.76 million pairs of genomes we find that identification of the closest relatives of an organism via gANI is precise, scalable, reproducible, and reflects the evolutionary dynamics of microbes. We model the previously unexplored statistical properties of gANI using 6,000 microbial genomes and apply species-specific gANI cutoffs to reveal anomalies in the current taxonomic species definitions for almost 50% of the species with multiple genome sequences. We also provide evidence of speciation events and genetic continuums in 17.8% of those species. We consider disagreements between gANI-based groupings and "named" species and demonstrate that the former have all the desired features to serve as the much-needed "natural groups" for moving forward with taxonomy. Further, the groupings identified are presented in detail at http://ani.jgi-psf.org to facilitate comprehensive downstream analysis for researchers across different disciplines

## OBJECTIVES




Whole-genome sequencing (WGS) launched microbial taxonomy into the new era of genomic microbial taxonomy, with the possibility of establishing systematics on the basis of information retrieved from complete genomes.

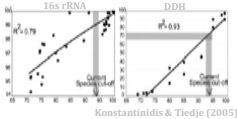
It is plausible to consider that microbial taxonomy will be steadily more dependent on genome sequences than relying on the classic polyphasic taxonomy, including phenotypic characterization using time-consuming laborious laboratory tests.

**ALIGNMENT FRACTION (AF)**

**The fraction of genes between two genomes that are orthologous.**

**AVERAGE NUCLEOTIDE IDENTITY (gANI)**

**Sequence level identity across all the conserved genes**

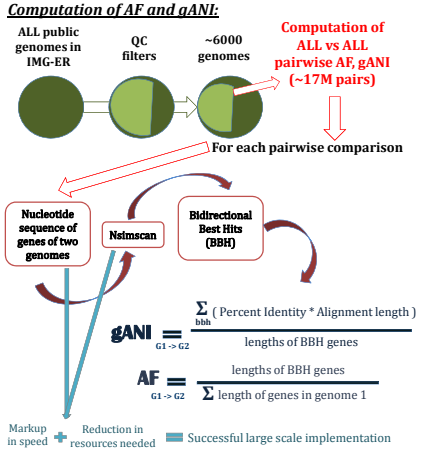



**reflects the degree of evolutionary distance**

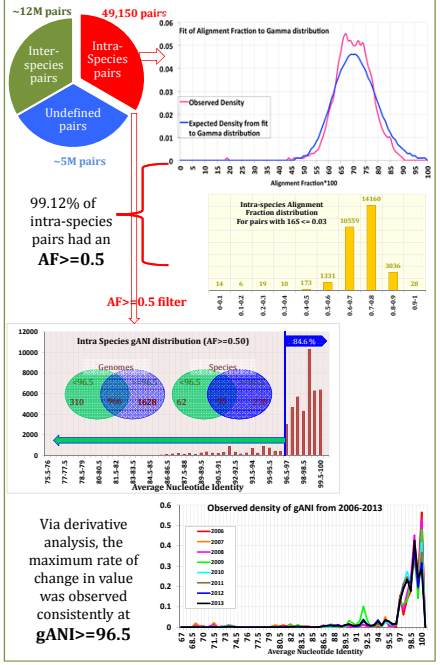
In this work, we propose whole-genome Average Nucleotide Identity (**gANI**) together with Alignment Fraction (**AF**) as a robust and reliable method for grouping of microbial genomes towards the goal of species classification.

## DATA SOURCES AND METHODS

***Computation of AF and gANI:***

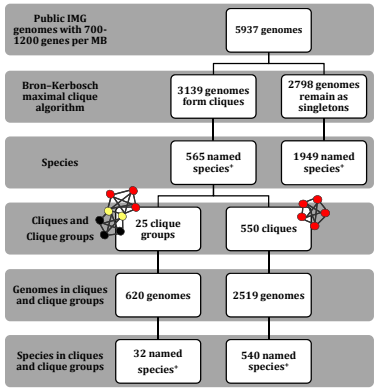



***Determination of gANI, AF thresholds that most closely model existing species assignments***

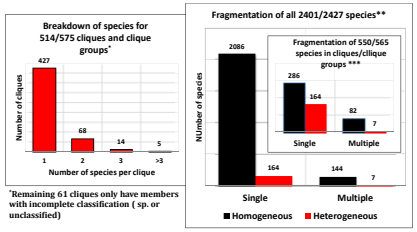



## RESULTS

We used the thresholds of **gANI>=96.5** and **AF>=0.5** to group the genomes using **M**aximal **C**lique **E**numeration **(MCE).**




+Named species are those with complete classification (not sp. or unclassified)




*Remaining 61 cliques only have members with incomplete classification ( sp. or unclassified)

** Remaining 26 and ***15 (inset) species are in combinations of single/multiple and homogeneous/heterogeneous

**Single homogeneous species**

*These species are concluded to be in agreement with current taxonomy. (e.g. Lactobacillus inners, Glaciecola agarilytica, Saccharospirillum impatiens)*

*The genomes of these species are captured in the first stage of speciation. (e.g. Lactobacillus johnsonii , Campylobacter coli, Enterococcus faecalis)*

**Single heterogeneous species**

*The names/classification of these species should be reconsidered. (e.g. Escherichia coli, Shigella flexneri, Shigella boydii, Shigella sonnei, Burkholderia mallei, Burkholderia pseudomallei)*

Each color represents a species and each circle a genome of that species.

**Multiple heterogeneous species**

*The names/classification of these species should be reconsidered (e.g. Bacillus mycoides, Salinispora arenicola, Rickettsia rickettsii)*

**Multiple homogeneous species**

*Such species are considered to be captured in the second stage of speciation (e.g. Bacillus amyloliquefaciens, Bacillus subtilis, Citrobacter freundii, Enterococcus casseliflavus,Enterococcus faecium,Lactobacillus jensenii)*

## CONCLUSIONS

Several reports have already illustrated that microbial taxonomic assignments are inconsistent with emerging genetic, systematic, and phenotypic information for a large number of species. According to a recent ASM report, "*in moving forward with microbial taxonomy, it is critical to determine whether microorganisms cluster in groups with meaningful commonalities or to determine what commonalities may be best used to cluster microorganisms into meaningful groups*".

For the first time, **gANI was applied across all available sequenced prokaryotic genomes** and its potential to cluster microorganisms into such "meaningful groups" was explored. We demonstrate that gANI, which maximally utilizes the commonalities between microbial genomes, **is a robust measure of genetic relatedness for establishing accurate evolutionary relationships.** The gANI-based cliques were validated by comparisons with "named" species, similarity of 16S rDNA, and similarity of conserved core pMGs. They were then used to address **central questions** such as whether microorganisms form a continuum of genetic diversity, or distinct species represented by distinct genetic signatures. Thus gANI-based cliques not only **provide insights into the evolutionary dynamics of prokaryotes**, but also significantly **assist in the refinement of the current taxonomy.**

## REFERENCES

1. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102:2567–2572.
2. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A.106:19126–19131
3. Konstantinidis KT, Ramette A & Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**: 1929–1940.
4. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, ZhaoX, Dubchak I, Hugenholtz P, Anderson I, et al. The integrated microbial genomes (IMG) system. Nucleic Acids Res.2006;34:D344-D348.

## ACKNOWLEDGMENTS

This work was conducted under the LDRD grant YLD012: Computational, Data Management and Analysis Methods for the Study of a Rapidly Expanding Genome and Metagenome Sequence Data Space (YLD012). Development was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 and Los Alamos National Laboratory under contract DE-AC02-06NA25396. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract DE-AC02-05CH11231.

## CONTACT

For questions or comments or more detailed information, please contact:

Neha Varghese <njvarghese@lbl.gov>

Supratim Mukerjee <supratimmukherjee@lbl.gov>

Amrita Pati <apati@lbl.gov>

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.