Microbial species delineation using whole genome sequences

Authors: Nikos Kyrpides[1], Neha Varghese[1], Supratim Mukherjee[1], Natalia Ivanova[1], Konstantinos Konstantinidis[2], Kostas Mavrommatis[1], Amrita Pati[1]

[1] U.S. Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA
[2] Georgia Tech - Atlanta, GA

** To whom correspondence may be addressed. Nikos Kyrpides, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA* - nckyrpides@lbl.gov

October, 2014

ACKNOWLEDGMENTS:

Work 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.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Microbial species delineation using whole genome sequences

Neha J. Varghese, Supratim Mukherjee, Natalia N. Ivanova , Konstantinos T. Konstantinidis, Kostas Mavrommatis, Nikos C. Kyrpides and Amrita Pati*

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

OVERVIEW

Species assignments in prokaryotes use a manual, polyphasic approach utilizing both phenotypic traits and sequence information of phylogenetic marker genes. With thousands of genomes being sequenced every year, an automated, uniform and scalable approach exploiting the rich genomic information in whole genome sequences is desired, at least for the initial assignment of species to an organism.

We have evaluated pairwise genome-wide Average Nucleotide Identity (gANI) values and alignment fractions (AFs) for nearly 13,000 genomes using our fast implementation of the computation, identifying robust and widely applicable hard cut-offs for species assignments based on AF and gANI.

Using these cutoffs, we generated stable species-level clusters of organisms, which enabled the identification of several species mis-assignments and facilitated the assignment of species for organisms without species definitions.

OBJECTIVES

Between two genomes, the

ALIGNMENT FRACTION (AF)

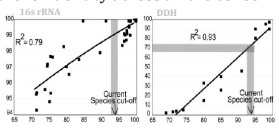
The fraction of genes between two genomes that are orthologous.

+

AVERAGE NUCLEOTIDE IDENTITY (gANI)

Sequence level identity across all the conserved genes

||



Konstantinidis & Tiedje (2005),

accurately reflects the degree of evolutionary distance

AF,gANI should be used as the primary guide for taxonomic species assignment, supplementing the existing polyphasic approach.

$$AF_{G1 \mapsto G2} = \frac{\text{lengths of BBH genes}}{\sum \text{length of genes in genome 1}}$$

$$gANI_{G1 \mapsto G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

METHOD

***Step 1* : Select high quality genomes**

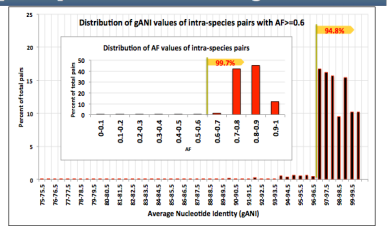
~13,000 genomes passed QC filters as of Feb 2014

***Step 2* : Compute gANI,AF for all vs all.**

~86 million pairwise computations (~190,000 core hours)

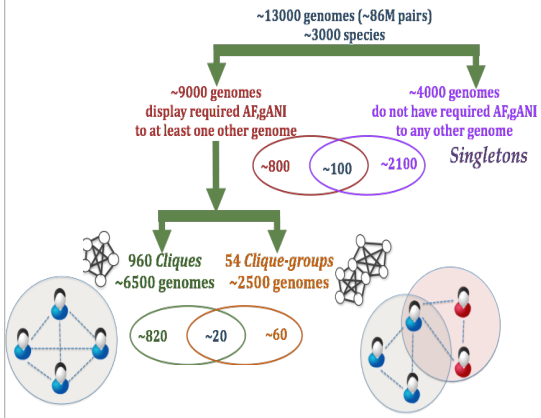
***Step 3* : Identity intra-species pairs**

***Step 4* : Use intra-species pairs to determine AF, gANI threshold**



***Step 5* : Use thresholds to cluster all the genomes into meaningful groups**

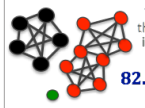
Maximal Clique Enumeration (MCE)to all genome pairs



RESULTS AND CONCLUSIONS

Species distributions within the generated clusters:

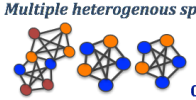
Single homogenous species



The taxonomic assignments of the genomes in these species are in complete agreement with the AF/gANI-based classification.

82.3%

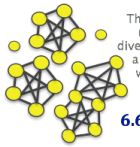
Multiple heterogenous species



0.5%

These species may suffer both from misclassification of individual genomes and errors of species delineation

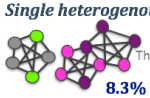
Multiple homogenous species



These genomes may have undergone significant divergence, possibly forming a core of a new species, which has escaped the existing taxonomy

6.6%

Single heterogenous species



These genomes may have been misidentified

8.3%

For the first time, gANI,AF was applied across all available sequenced prokaryotic genomes and we have shown that this fast implementation of gANI provides an objective and robust measure of genetic relatedness.

The gANI-based groups have been compared with "named" species, similarity of 16S rDNA, and similarity of conserved core pMGs and the superiority our method has been demonstrated.

The examination of cliques highlights several anomalies in the traditional classification of strains into species based on the polyphasic approach.

The clique-based approach also provides valuable insight into the evolutionary dynamics of prokaryotes and can be used to explore central questions such as whether microorganisms form a continuum of genetic diversity, or distinct species represented by distinct genetic signatures.

CONTACT

For questions or comments , please email

njvarghese@lbl.gov

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