W282–W288 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 14 June 2018*
doi: 10.1093/nar/gky467

The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level

Luis M. Rodriguez-R[1], Santosh Gunturu[2], William T. Harvey[1], Ramon Rosselló-Mora[3], James M. Tiedje[2,4], James R. Cole[2] and Konstantinos T. Konstantinidis[1,5,*]

[1]School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr. NW, Atlanta, GA 30332, USA, [2]Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA, [3]Grup de Microbiologia Marina, Institut Mediterrani d'Estudis Avançats (IMEDEA), Universitat de les Illes Balears (UIB) and Consejo Superior de Investigaciones Científicas (CSIC), C/Miquel Marques 21, 07190 Esporles, Illes Balears, Spain, [4]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA and [5]School of Biological Sciences, Georgia Institute of Technology, 310 Ferst Dr. NW, Atlanta, GA 30332, USA

Received March 12, 2018; Revised May 08, 2018; Editorial Decision May 09, 2018; Accepted May 24, 2018

ABSTRACT

The small subunit ribosomal RNA gene (16S rRNA) has been successfully used to catalogue and study the diversity of prokaryotic species and communities but it offers limited resolution at the species and finer levels, and cannot represent the whole-genome diversity and fluidity. To overcome these limitations, we introduced the Microbial Genomes Atlas (MiGA), a webserver that allows the classification of an unknown query genomic sequence, complete or partial, against all taxonomically classified taxa with available genome sequences, as well as comparisons to other related genomes including uncultivated ones, based on the genome-aggregate Average Nucleotide and Amino Acid Identity (ANI/AAI) concepts. MiGA integrates best practices in sequence quality trimming and assembly and allows input to be raw reads or assemblies from isolate genomes, single-cell sequences, and metagenome-assembled genomes (MAGs). Further, MiGA can take as input hundreds of closely related genomes of the same or closely related species (a so-called 'Clade Project') to assess their gene content diversity and evolutionary relationships, and calculate important clade properties such as the pangenome and core gene sets. Therefore, MiGA is expected to facilitate a range of genome-based taxonomic and diversity studies, and quality assessment across environmental and clinical settings. MiGA is available at http://microbial-genomes.org/.

MiGA CLASSIFICATION OF A QUERY GENOMIC SEQUENCE

The small subunit ribosomal RNA gene (16S) has been successfully used to catalogue and study the diversity of prokaryotic species and their communities for over thirty years. However, genome-based methods are needed to better resolve microbial communities at the species and finer levels, which cannot be efficiently assessed by 16S (1), and to catalogue whole-genome diversity and fluidity. One such genomic method is the Average Nucleotide Identity or ANI (2,3). ANI represents the average nucleotide identity of all orthologous genes shared between any two genomes and offers robust resolution between strains of the same or closely related species (i.e. showing 80–100% ANI). The ANI measurement does not strictly represent core genome evolutionary relatedness, as orthologous genes can vary widely between pairs of genomes compared. Nevertheless, it closely reflects the traditional microbiological concept of DNA-DNA hybridization (DDH) relatedness for genomically circumscribing species (2,3), as it takes into account the fluid nature of the bacterial gene pool and hence, implicitly considers shared function. Accordingly, ANI has been recognized internationally for its potential to replace DDH as the standard measure of relatedness, as it is easier to estimate and represents portable and reproducible data (4,5).

The Microbial Genomes Atlas (MiGA) project offers robust taxonomic classification of a query genome or assembled contig sequences based on ANI or, for more divergent (deep-branching) sequences, Average Amino Acid Identity (AAI; (6)) values against a reference genome database. The reference database could be NCBI's RefSeq, which encompass 1927 high-quality closed genomes from 1865 species (as of May 2018; updated bimonthly), or the NCBI Genome

[*]To whom correspondence must be addressed. Tel: +1 404 385 3628; Fax: +1 404 894 8266; Email: kostas@ce.gatech.edu

Database, Prokaryotic section (henceforth NCBI_Prok), encompassing 11 487 genomes from 3921 species (updated monthly). MiGA identifies the best matching reference genome for the query sequence based on ANI/AAI values and subsequently, assesses if the query sequence should be assigned to the same taxonomic rank (e.g. species, genus, *etc.*) as its best match or instead, if it represents a new/novel taxon at that rank. Additional information is provided for the best-matching sequence (and all reference genomes), including sequence statistics, genome quality and type material (when available). For the assignment step, MiGA compares the AAI value between the query sequence and its best match against empirical distributions of AAI values for each taxonomic rank derived from (pre-computed) AAI comparisons among the RefSeq genomes (Figure 1) to estimate the probability of not sharing the same rank or not being novel (empirical *P*-values of taxonomic classification and taxonomic novelty, respectively; see Supplementary Methods for details). For example, genomes of the same species typically share more than 95% ANI (Supplementary Figure S2). Hence, a query genome with a best match ANI value <95% most likely represents a novel species, and the confidence (probability) for the latter classification (as novel species) would depend on how different its best ANI value is from the distribution of ANI values among all genomes assigned to the same species in RefSeq (Figure 3). Similarly, MiGA can determine the level of novelty of a query genome, e.g. if it represents a novel species within a known/described genus, or a novel genus of a known family *etc*., based on the AAI value of the best match. We tested MiGA's classification accuracy using genomes of known taxonomy from NCBI_Prok and classifying them against the RefSeq database. MiGA's accuracy was 90% or higher when its *P*-value for the assignment was 0.05–0.01 (Supplementary Figure S1). False positive or negative calls were of low frequency and almost always associated with inconsistently named species; e.g. species encompassing genome pairs with ANI values much lower than 95% (see also below).

Further, the query genomic sequence can be searched against unclassified genomes of isolates as well as selected metagenome-assembled genomes (MAGs) and single-cell amplified genome (SAGs) collections to identify the closest relatives in the total available genome sequence space. General statistics as well as quality assessments, including estimates of completeness, contamination, and taxonomically biased regions (see below), are provided both for the query sequence as well as for all reference MAGs and SAGs, so users can evaluate the significance of the best matches on their own. Note that in this case no direct classification to species, or other low taxonomic levels, is typically possible since the matching reference genomes are not officially classified yet, and the only available classification is that transferred by MiGA based on the (pre-computed) best matches of the reference MAGs and SAGs to NCBI_Prok. However, this analysis could provide answers to several ecological questions such as 'has the query genome been found elsewhere?' and 'how similar at the gene content is it to its best match?' and facilitate future novel taxon descriptions. In this case, the reference genomes are organized in so-called Projects, which are distinct reference genome databases from RefSeq and NCBI_Prok within MiGA. For instance, genomes are grouped into projects based on the habitat they originated from (see also below). We are currently expanding this collection with recently published collections of MAGs, and welcome submissions from external users of MAGs with sequence data registered in member databases of INSDC (7) and metadata registered in the Digital Protologue Database in order to capture key metadata and make them available to all users (8). Therefore, MiGA can also help cataloging the uncultivated microbial majority, which represents a great and urgent need (9), especially if external users are willing to (freely) share their genomic sequences with the scientific community through MiGA.

Clade projects

Finally, Clade Projects involve the analysis of tens to hundreds of genome sequences assigned to the same or closely related species, using the ANI approach outlined above, in order to assess gene content diversity and genetic relatedness among the genomes. MiGA computes the full AAI and ANI values among the sequences of a Clade Project, produces the ANI-based phylogenetic tree (see also below), and saves, in addition, the matches of individual genes contributing to the AAI calculation. The latter information is used to identify all orthologous groups of proteins (defined as reciprocal best matching proteins), and provide descriptive statistics on the distribution of genes within the clade such as the size of the core gene set and the pangenome (*i.e.*, total, non-redundant genes among all input genomes). Thus, Clade Projects are ideal for microdiversity and epidemiological studies, including of plant and animal pathogens. For instance, our analysis of ∼400 *Bacillus anthracis*, one of the least diverse and thus, most challenging to resolve species known (ANI values among *B. anthracis* genomes >99.7%) revealed that the ANI-based tree recovered all previously known sub-clades of the species. The ANI tree also provided higher resolution compared to canonical SNP (Single Nucleotide Polymorphosms) or MLVA (Multi-Locus VNTR Analysis; VNTR: Variable Number of Tandem Repeats; Figure 2), which are commonly used to genotype *B. anthracis* genomes. MiGA Online offers an expanding collection of Clade Projects to browse, including all (complete and draft) genomes of *Bacillus cereus sensu lato*, *Marinobacter*, *Pelagibacter ubique*, and *Thaumarchaeota* (updated bimonthly). Developing new Clade Projects by external users is possible in the standalone MiGA implementation and the cloud computing platforms, but not in the online MiGA webserver at the time of writing. Additional species can be requested in the public roadmap of MiGA Online (http://roadmap.microbial-genomes.org/).

Additional features and utilities

MiGA also offers the possibility to search any project dataset or genome sequence database (*e.g.*, NCBI_Prok or RefSeq) by species name or metadata such as taxonomy at any registered rank, type material, and genome quality. In addition, users can browse all reference datasets on each Project by either AAI clustering or taxonomy, and explore the distributions of hAAI (see below), AAI and ANI values

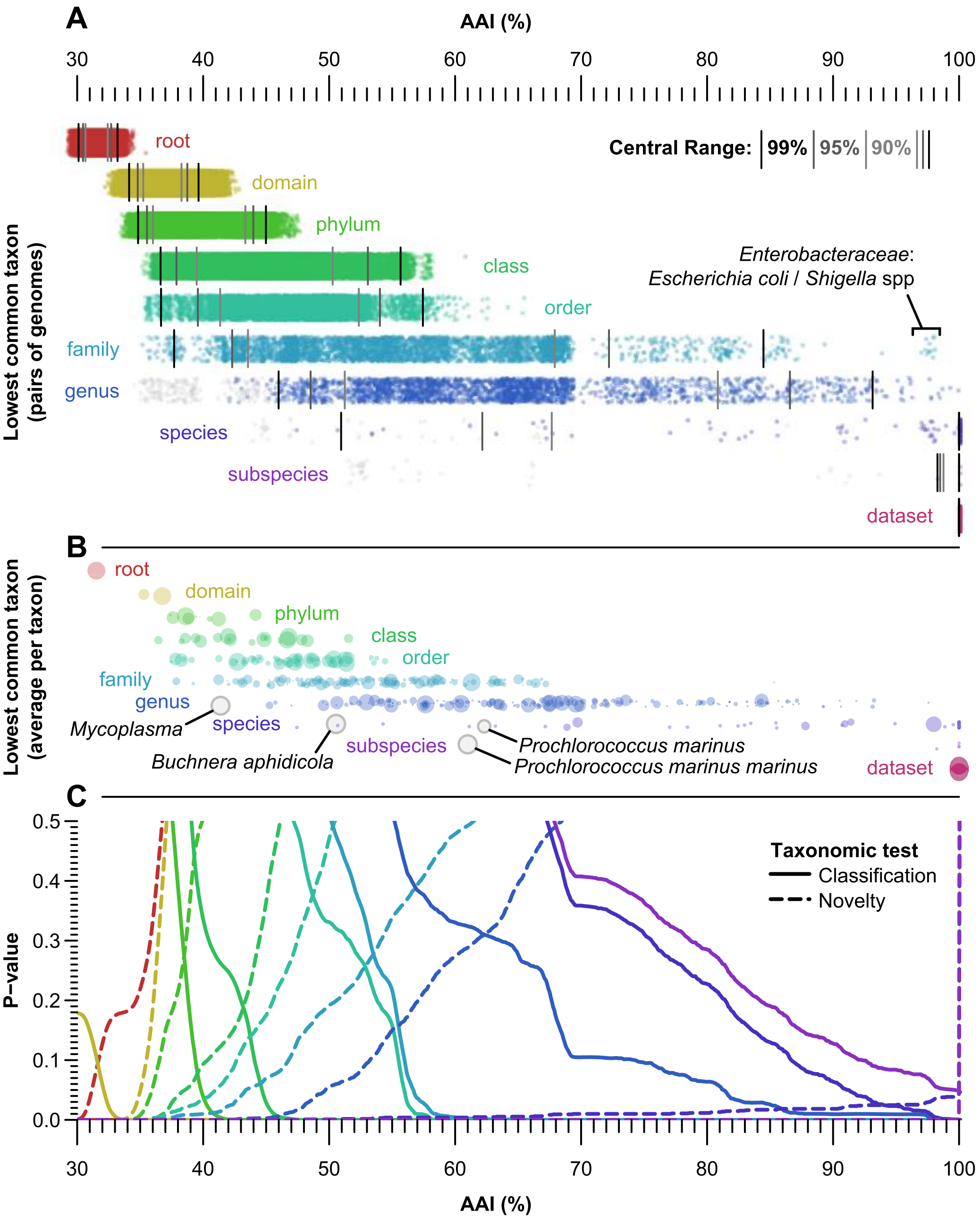


Figure 1. Estimate the confidence of taxonomic assignments based on AAI values. (**A**) Distributions of AAI values for genome pairs in the RefSeq database per rank of lowest common taxon. The central 90%, 95% and 99% ranges are shown on each rank. Outlier taxa, listed in supplementary methods, are presented here as gray dots. Note that values between *Escherichia coli* and *Shigella* spp significantly deviate from the central tendency, but are not excluded because the lowest common taxon (family *Enterobacteriaceae*) is not significantly affected. (**B**) Average AAI values per taxon by lowest common taxon. Within each row, dot sizes linearly reflect the size of the taxon. Outlier taxa (excluded from the empirical distributions) are highlighted. (**C**) Empirical *P*-values (per AAI to best match) for the alternative hypotheses that (i) the classification of query and reference genomes are the same at the given rank (taxonomic classification test; solid lines) or (ii) that the query genome represents a novel taxon with respect to the database (taxonomic novelty test; dashed lines). Colors correspond to the evaluated rank (as in A and B).

within each Project. Additional features available for examining the user-submitted query genome sequences include: (i) estimates of the completeness and contamination levels of a query sequence, (ii) detection of regions of the query sequence that may have different phylogenetic origin than its best matching reference genome due to recent horizontal gene transfer or miss-assembly (chimeric sequences; see also MyTaxa scan below), (iii) 16S analysis using the RDP classifier (10) when 16S gene sequences are present in the query genome and (iv) general sequence statistics such as G+C% content, coding density, and assembly length and N50.

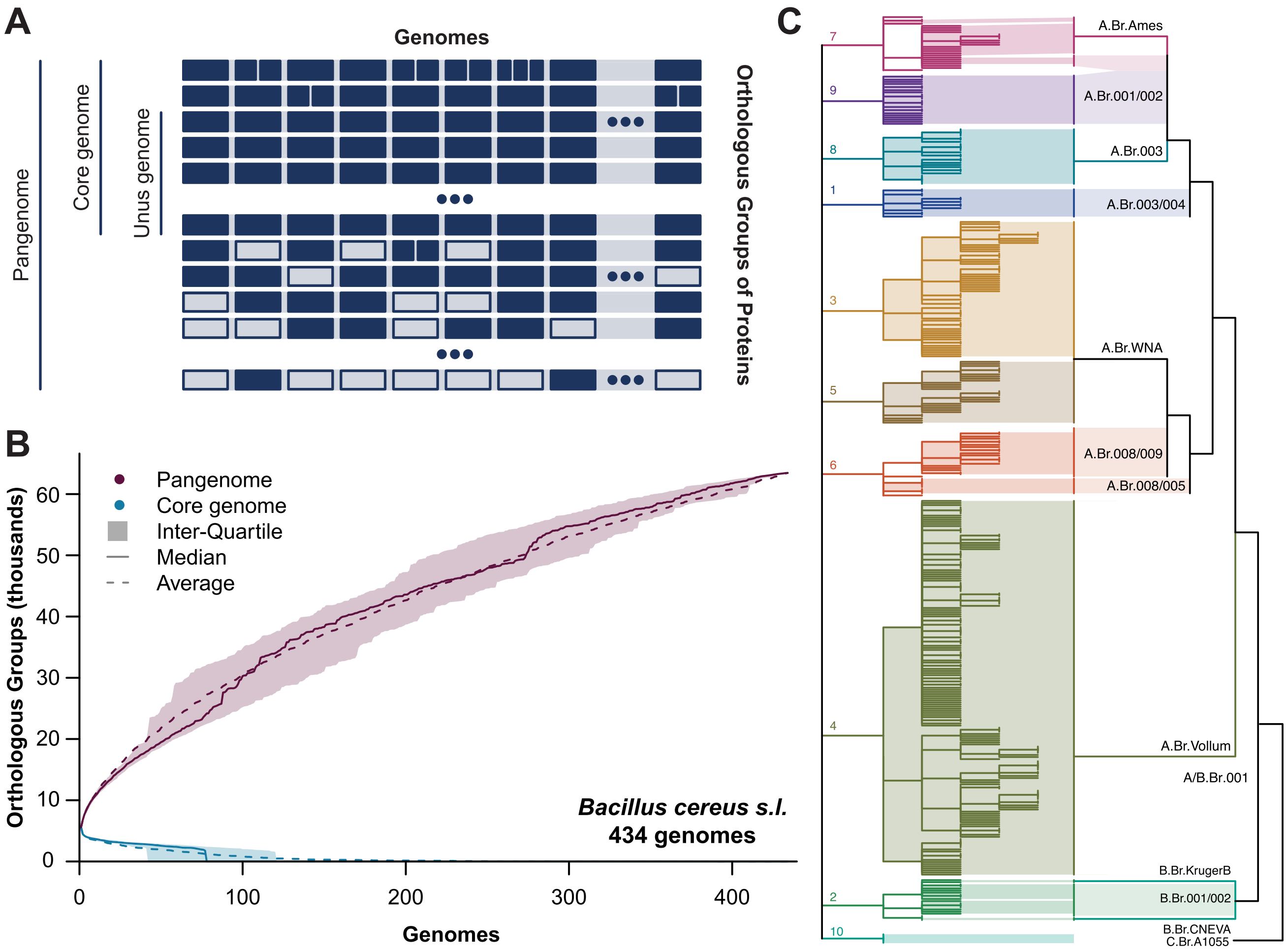


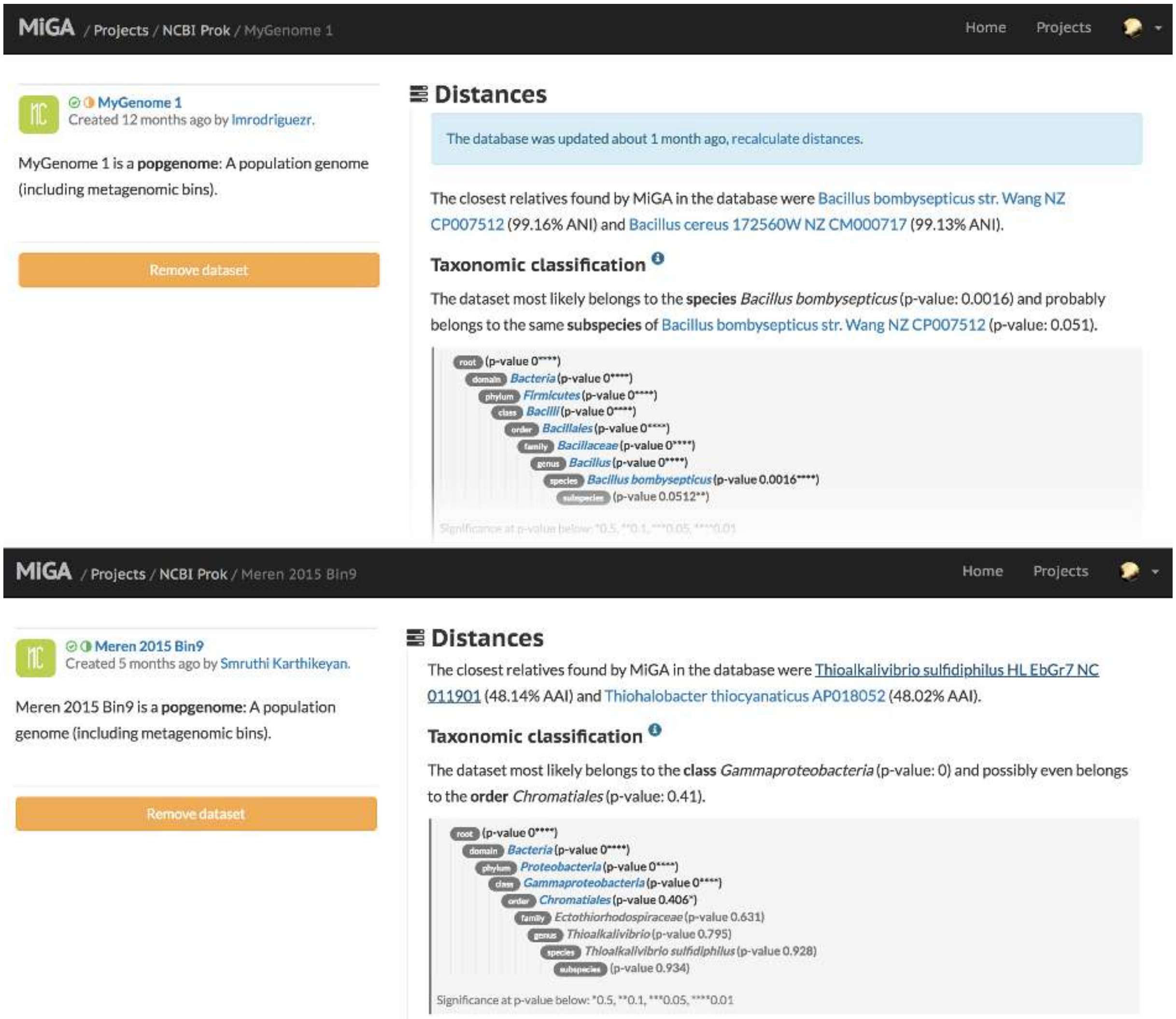
Figure 2. Clade Project on *Bacillus cereus sensu lato* (*s.l*). Collection of publicly available genomes from *B. cereus*, *B. thuringensis*, *B. mycoides*, *B. pseudomycoides* and *B. anthracis* species, a clade of highly related species collectively known as *B. cereus s.l.* (**A**) Schematic indicating the various definitions used in Clade Projects. The matrix represents genes organized by groups of orthology (OG; rows) per genome (columns). Absent genes (genomes missing a given OG) are indicated with empty boxes, while solid boxes represent present genes. Multiple boxes in the same cell indicate multiple copies (*i.e.*, internal paralogs). The collection of all non-redundant OGs is termed pangenome, the subset of OGs present in all genomes is termed core genome, and the subset of OGs present in single copy in all genomes is termed Unus genome. (**B**) Rarefied OG counts in the Pangenome and the Core genome per sampled genomes. This graphical output is directly generated by MiGA. (**C**) Comparison between the ANI clustering (left cladogram) and canonical SNP scheme (right cladogram) for the *B. anthracis* genomes in the collection. Note that both techniques produce the same large groupings (colors), but ANI offers higher resolution, with subclades defined within each clade up to four degrees.

MiGA workflow: ANI/AAI distances

To efficiently search query sequences against all these reference genome sequences (e.g. over 11 000 for NCBI_Prok), MiGA employs several heuristic approximations. First, MiGA estimates the genome-based relatedness between each query sequence and all reference medoid genomes in the database. Medoid genomes represent single representatives of AAI clusters of reference genomes, or distinct subclusters within each AAI cluster (see Supplementary Methods); medoids are used so that not all genomes of highly populated clusters are included in the search, which results in substantial reduction in computing time (Supplementary material). For each pair of genomes, relatedness is first assessed by heuristic AAI (hAAI), based on only 110 universal genes (see supplementary methods). If this value cannot be calculated (*e.g.*, there are less than ten common universal gene matches in the two genomes) or is close to saturation (hAAI ≥ 90%), MiGA estimates the complete AAI (6) using aai.rb (11). If this value is also close to saturation (AAI ≥ 90%), MiGA estimates the ANI of the genome pair (3) using ani.rb (11). For each query genome, MiGA applies this hierarchical approach (hAAI, next AAI, and finally ANI; Supplementary Figure S2) against medoid reference datasets. Once the medoid with the highest identity to the query is identified, MiGA proceeds to compare against the medoids in the related subclusters. This process is recursively applied until a cluster without subclusters is found, at which point MiGA calculates the similarities against all members of that cluster and thus, identifies the best match member.

MiGA workflow: Data preprocessing

MiGA implements a genomic data management and processing system that integrates best practices in genomic analyses, and with presets that allow input sequences to


MiGA / Projects / NCBI Prok / MyGenome 1
Home
Projects
Remove dataset
MiGA / Projects / NCBI Prok / Meren 2015 Bin9
Home
Projects
Remove dataset

MyGenome 1
Created 12 months ago by lmrodriguezr.

MyGenome 1 is a **popgenome**: A population genome (including metagenomic bins).

Distances

The database was updated about 1 month ago, recalculate distances.

The closest relatives found by MiGA in the database were Bacillus bombysepticus str. Wang NZ CP007512 (99.16% ANI) and Bacillus cereus 172560W NZ CM000717 (99.13% ANI).

Taxonomic classification

The dataset most likely belongs to the **species** *Bacillus bombysepticus* (p-value: 0.0016) and probably belongs to the same **subspecies** of Bacillus bombysepticus str. Wang NZ CP007512 (p-value: 0.051).

```
root (p-value 0****)
  domain Bacteria (p-value 0****)
    phylum Firmicutes (p-value 0****)
      class Bacilli (p-value 0****)
        order Bacillales (p-value 0****)
          family Bacillaceae (p-value 0****)
            genus Bacillus (p-value 0****)
              species Bacillus bombysepticus (p-value 0.0016****)
                subspecies (p-value 0.0512**)
Significance at p-value below: *0.5, **0.1, ***0.05, ****0.01
```

Meren 2015 Bin9
Created 5 months ago by Smruthi Karthikeyan.

Meren 2015 Bin9 is a **popgenome**: A population genome (including metagenomic bins).

Distances

The closest relatives found by MiGA in the database were Thioalkalivibrio sulfidiphilus HL EbGr7 NC 011901 (48.14% AAI) and Thiohalobacter thiocyanaticus AP018052 (48.02% AAI).

Taxonomic classification

The dataset most likely belongs to the **class** *Gammaproteobacteria* (p-value: 0) and possibly even belongs to the **order** *Chromatiales* (p-value: 0.41).

```
root (p-value 0****)
  domain Bacteria (p-value 0****)
    phylum Proteobacteria (p-value 0****)
      class Gammaproteobacteria (p-value 0****)
        order Chromatiales (p-value 0.406*)
          family Ectothiorhodospiraceae (p-value 0.631)
            genus Thioalkalivibrio (p-value 0.795)
              species Thioalkalivibrio sulfidiphilus (p-value 0.928)
                subspecies (p-value 0.934)
Significance at p-value below: *0.5, **0.1, ***0.05, ****0.01
```

Figure 3. Classification examples in MiGA Online. Two examples of the typical output of MiGA classification analysis are shown. The top panel shows a genome classified as *Bacillus bombysepticus*, displaying high ANI values against its best match and consequently high classification confidence. The bottom panel shows a query genome with only distant relatives available in the database, with a maximum AAI of 48% classified to order level (*Chromatiales*) with low confidence, and with high confidence only to class level (*Gammaproteobacteria*).

represent raw reads or assembled sequences from isolate genomes, single-cell sequences, and metagenome-derived bins or contigs (Figure 4). Datasets initialized from raw short reads are trimmed, stripped from typically used adaptor sequences, and quality-checked using SolexaQA++ with Phred quality score threshold of 20 and minimum length of 50 bp (12), Scythe with default parameters (13), and FastQC (14), respectively. Trimmed and clipped datasets are next assembled using IDBA UD with default parameters (15). Genome binning of the resulting assemblies, if desired, must be currently done outside MiGA currently.

Next, coding sequences are predicted using Prodigal with default parameters (16). Genes frequently observed to be present in single-copy in bacterial and archaeal genomes are next identified as described in (17), and implemented in HMM.essential.rb from the Enveomics collection (11) including a quality report of redundancy and completeness. Redundancy is defined as the fraction of the total single-copy genes that have more than one copy in the query sequence; completeness represents the fraction of the total single-copy genes recovered in the query sequence. In genome sequences (isolate genomes, MAGs or SAGs), MiGA identifies potentially problematic regions like large horizontally transferred regions, contamination, erroneous assembly, *etc.*, using MyTaxa scan (Supplementary Figure S3). MyTaxa scan traverses the genome sequence in windows of ten genes and compares the taxonomic affiliations predicted by MyTaxa (18), flagging regions with unexpectedly large Hellinger distances to the genome-wide distribution (19). This analysis is performed for all reference genomes that are part of MiGA's genome databases (e.g. NCBI_Prok) and Projects, and can be launched (optionally) for any query dataset.

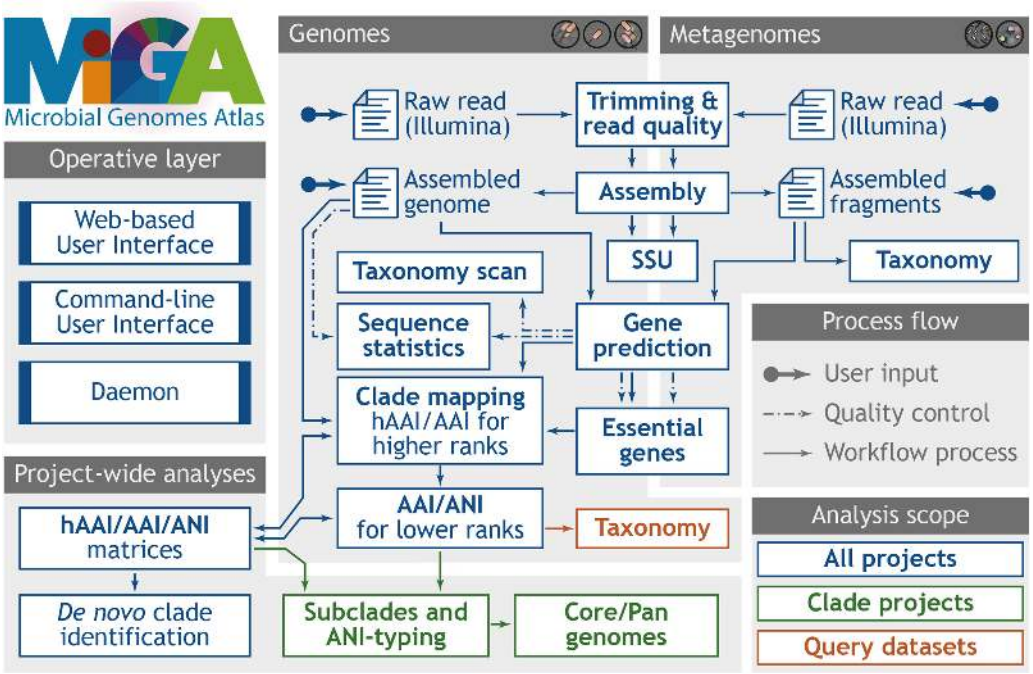


Figure 4. MiGA Workflow. MiGA users can initialize their analyses from raw reads or assemblies, either for genomic datasets (isolate genomes, single-cell amplified genomes, or metagenome-assembled genomes) or metagenomic (microbial metagenomes or viral enrichment metagenomes). After basic pre-processing, MiGA will query the resulting sequences against its reference genome sequences to identify the closest relatives using a hierarchical hAAI/AAI/ANI scheme, and determine the best-supported taxonomic assignment.

MiGA INFRASTRUCTURE AND ONLINE WEB-SERVER

The main infrastructure of MiGA is an object-oriented application written in Ruby with minimal additional requirements, and ancillary shell code is used to implement analysis tasks. Data is stored in a variety of formats while metadata is stored in JSON files. MiGA features Application Programmatic, Command Line, and Web-based User Interfaces (API, CLI and Web). For further details on the infrastructure, see the manual available at http://manual.microbial-genomes.org/. The entire system, including the processing core and the webserver, can be replicated from http://code.microbial-genomes.org/, available under the Artistic License 2.0. The MiGA infrastructure has been successfully deployed for the autonomous processing of various input data types as described above. MiGA Online makes all its internal Projects and genome databases publicly available to browse, search, and query. Short videos on how to upload data, which Project to choose for specific research questions, understand the different outputs and interpret the results are also available through the homepage. MiGA is also available in preconfigured virtual machines for cloud computing through the Amazon Web Services and Google Cloud platforms.

CONCLUSIONS

We introduce MiGA Online, a webserver featuring a processing workflow for genomic and metagenomic analyses, including novel developments in whole-genome-based taxonomy and classification. Genome-wide comparisons in MiGA, including ANI/AAI and the implementation of heuristic approximations, enable the classification of query genomes against the 11 487 genomes (NCBI_Prok) currently included in its database. MiGA offers several Projects to browse or search query sequences against in order to suit different research needs, with NCBI_Prok being the most general and widely used. Therefore, MiGA fulfills a critical need of contemporary research, as no similar web resource is currently available, and is expected to greatly facilitate genome-level classification and gene diversity studies across the fields of environmental and clinical microbiology. Classification of genome sequences has been attempted based on the phylogenetic analysis of a few singe-copy universally conserved genes (20, https://www.biorxiv.org/content/early/2018/01/31/256800), such as 16S rRNA and ribosomal protein-encoding genes, but these genes typically show higher sequence conservation than the genome average (e.g. ANI). Consequently, analysis of universal genes is not always feasible (depending on completeness), provide lower resolution than whole-genome comparisons near or below the species level, and has frequently resulted in lack of clear sequence/genetic relatedness boundaries between species (e.g. 16S rRNA), which is rarely observed with ANI (1, https://www.biorxiv.org/content/early/2017/11/27/225342). Thus, MiGA provides significant advantages over these alterative methods for classification analysis, especially at the species level and below. In addition, MiGA uniquely provides information on gene content diversity and the degree of taxonomic novelty of query sequences, and is equally applicable to complete as well incomplete genomes that may be missing universal genes (e.g. the genes are not assembled) or include contamination, substantially expanding the microbial genome diversity that can be catalogued.

SUPPLEMENTARY MATERIAL

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mehmet Belgin from the Office of Information Technology at the Georgia Institute of Technology for his advice and technical assistance, John Quensen for his support on the MiGA Online documentation, and several participants of the MiGA workshops for helpful suggestions regarding the layout of the MiGA webserver.

FUNDING

U.S. National Science Foundation [1356288, 1356380]. Funding for open access charge: U.S. National Science Foundation [1356288, 1356380].

Conflict of interest statement. None declared.

REFERENCES

1. Rodriguez-R,L.M., Castro,J.C., Kyrpides,N.C., Cole,J.R., Tiedje,J.M. and Konstantinidis,K.T. (2018) How much do rRNA gene surveys underestimate extant bacterial diversity? *Appl. Environ. Microbiol.: AEM*, **84**, e00014–18.
2. Konstantinidis,K.T. and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2567–2572.
3. Goris,J., Konstantinidis,K.T., Klappenbach,J.A., Coenye,T., Vandamme,P. and Tiedje,J.M. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, **57**, 81–91.
4. Richter,M. and Rosselló-Móra,R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 19126–19131.
5. Kim,M., Oh,H-S., Park,S-C. and Chun,J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **64**, 346–351.
6. Konstantinidis,K.T. and Tiedje,J.M. (2005) Towards a Genome-Based taxonomy for prokaryotes. *J. Bacteriol.*, **187**, 6258–6264.
7. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and and Sequence Database Collaboration IN (2016) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
8. Rosselló-Móra,R., Trujillo,M.E. and Sutcliffe,I.C. (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of Archaea and Bacteria. *Syst. Appl. Microbiol.*, **40**, 121–122.
9. Konstantinidis,K.T., Rosselló-Móra,R. and Amann,R. (2017) Uncultivated microbes in need of their own taxonomy. *ISME J.*, **11**, 2399–2406.
10. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
11. Rodriguez-R,L.M. and Konstantinidis,K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*, **4**, e1900v1.
12. Cox,M.P., Peterson,D.A. and Biggs,P.J. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.
13. Buffalo,V. (2014) Scythe—a Bayesian adapter trimmer (version 0.994 BETA). https://github.com/vsbuffalo/scythe (29 June 2015, date last accessed).
14. Andrews,S. FastQC A Quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (29 June 2015, last date accessed).
15. Peng,Y., Leung,H.C.M., Yiu,S.M. and Chin,F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.
16. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
17. Albertsen,M., Hugenholtz,P., Skarshewski,A., Nielsen,K.L., Tyson,G.W. and Nielsen,P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.
18. Luo,C., Rodriguez-R,L.M. and Konstantinidis,K.T. (2014) MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.*, **42**, e73.
19. Hellinger,E. (1909) Neue begründung der theorie quadratischer formen von unendlichvielen Veränderlichen. *J. Für. Reine Angew. Math. Crelles J.*, **1909**, doi:10.1515/crll.1909.136.210.
20. Sunagawa,S, Mende,D.R., Zeller,G., Izquierdo-Carrasco,F., Berger,S.A., Kultima,J.R., Coelho,L.P., Arumugam,M., Tap,J., Nielsen,H.B. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196–1199.