Genome Based Taxonomic Classification

Journal:	<i>Genome</i>
Manuscript ID	gen-2018-0072.R4
Manuscript Type:	Article
Date Submitted by the Author:	15-Dec-2018
Complete List of Authors:	Paul, Bobby; School of Life Sciences, Manipal Academy of Higher Education, Bioinformatics Dixit, Gunjan; School of Life Sciences, Manipal Academy of Higher Education Murali, Thokur Sreepathy; School of Life Sciences, Manipal Academy of Higher Education, Biotechnology Satyamoorthy, K.; School of Life Sciences, Manipal Academy of Higher Education , Biotechnology
Keyword:	Bacterial taxonomy, Bioinformatics, Nucleotide identity, Nomenclature
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

SCHOLARONE™
Manuscripts

Genome Based Taxonomic Classification

Bobby Paul, Gunjan Dixit, Thokur Sreepathy Murali, Kapaettu Satyamoorthy*

School of Life Sciences, Manipal Academy of Higher Education, Manipal 576104, INDIA

*Corresponding Author: Dr. K. Satyamoorthy

Address: School of Life Sciences,

Manipal Academy of Higher Education,

Manipal, Karnataka

INDIA 576104

Email: ksatyamoorthy@manipal.edu

Draft

Abstract

Bacterial populations are routinely characterized based on the microscopic examination, colony formation and biochemical tests. However, in recent past, bacterial identification, classification and nomenclature have been strongly influenced by genome sequence information. Advances in bioinformatics and growth in genome databases has placed genome-based metadata analysis in the hands of researchers who will require taxonomic experience to resolve intricacies. To achieve this, different tools are now available to quantitatively measure genome relatedness within members of the same species and genome-wide average nucleotide identity (gANI) is one such reliable tool to measure genome similarity. A genome assembly showed gANI score of <95% at intra-species level is generally considered indicative of a separate species. In this study, we have analysed 300 whole-genome sequences belonging to 26 different bacterial species available in the Genome database and calculated their similarity at intra-species level based on gANI score. At the intra-species level, nine bacterial species showed less than 90% gANI and more than 10% of unaligned regions. We suggest appropriate use of available bioinformatics resources after genome assembly to arrive at proper bacterial identification, classification and nomenclature in order to avoid erroneous species assignments and disparity due to diversity at the intra-species level.

Key words

Bacterial taxonomy; Bioinformatics; Nucleotide identity; Nomenclature

1 Introduction

2 Accurate bacterial classification is essential to identify the potential genes to understand genetic
3 mechanisms, disease diagnosis and provide better treatment especially during infectious disease
4 spread, outbreaks and progression (Fricke and Rasko 2014; Croucher and Didelot 2015; Kubicova
5 and Provaznik 2016). Bacterial classification is based on microbial techniques such as isolation,
6 culture and staining to determine morphological and biochemical features. Over the years, in
7 addition to morphological characters, phenotypic and genotypic methods have gained much
8 importance in bacterial identification and classification schema (Donelli et al. 2013). With the
9 rapid advancements in DNA sequencing techniques, genome-based methods are expected to play
10 a major role in delineating the bacterial species at genome level and provide valuable clues towards
11 their identification (Hugenholtz et al. 2016). Recently, metabarcoding and metagenomics have
12 become robust tools to study bacterial and archaeal diversity by comparative genomic analysis
13 (Srivathsan et al. 2015; Bowers et al. 2017). Increased reliability, decreased costs and time for
14 whole-genome sequencing and assembly has led to an unprecedented amount of information that
15 can provide significant insights towards drug-repositioning and development of robust molecular
16 diagnostic platforms (Fournier et al. 2014; Baltrus 2016; Hahnke et al. 2016; Hugenholtz et al.
17 2016; Jansen van Rensburg et al. 2016; Gosiewski et al. 2017).

18 It has been reported that there are millions of bacterial species of which only a fraction have been
19 identified and characterized (Eichorst et al. 2007; Kim et al. 2014). The global microbial species
20 richness is estimated to be anywhere between 10^7 to 10^9 (Schloss and Handelsman 2004). Till
21 December 2017, there were 17947 bacterial species with validly published names in Taxonomy
22 browser (<https://www.ncbi.nlm.nih.gov/taxonomy>). The development of whole genome
23 sequencing methods has resulted in the availability of large quantities of genomic data in public
24 repositories (Cook et al. 2016; Mukherjee et al. 2017). Over the past 20 years, more than 124,311
25 bacterial genome sequences have been deposited in the Genome database. However, the sequence
26 information needs critical evaluation before deciphering species descriptions, delineating related
27 species and assigning them into a reference genome database. Currently, bacterial species
28 delineation is based on phylogenetic analysis of marker genes such as 16S rRNA, 23S
29 rRNA, rpoB, gyrB, dnaK, etc. (Konstantinidis and Tiedje 2005, Liu et al. 2012). However, this
30 approach takes into consideration evolution of single genes for grouping the species or delineating
31 a species in a genus. Genome-level dissimilarity can happen due to the cross-species genome

contaminations (Merchant et al. 2014; Mukherjee et al. 2015; Kryukov and Imanishi 2016), lateral gene transfer (Gillings 2017), and insertion sequences (Vincent et al. 2016). In addition, nucleotide substitutions, gene insertions and deletions, gene duplications and rearrangements, which have significant impact on evolution of prokaryotic genome (Coenye et al. 2005), are also considered as critical determinants while defining a species. Horizontal gene transfer (HGT) plays an important role in genome adaptation and evolution. Acquisition of new genes by HGT allows for habitat-specific adaptation and ultimately results in a subsequent divergence of species or speciation (Ochman et al. 2000). Additionally, genomes of several symbiotic bacterial species show the existence of smaller genomes, codon reassignments and extreme biases in genome sequences (McCutcheon and Moran 2012). Hence, it is imperative to look at whole genome information rather than at selected genes for arriving at relatedness between species in a given genus.

To further add to the complexity, advances in bioinformatics and growth in genome databases requires researchers with taxonomic experience to perform genome-based metadata analysis in order to critically evaluate the sequence information of genomes before assigning them into a reference genome database. Though, a conclusive quantitative measure of genome relatedness is indeed difficult, different methods have been routinely used to measure the relatedness within members of the same species. The average nucleotide identity (ANI) is one such reliable measure aimed at long-standing, systematic and scalable bacterial species classification (Kim et al. 2014). Calculation of genome-wide ANI (gANI) score usually involves the fragmentation of genome sequences, followed by calculating the nucleotide similarity between sequences by pairwise alignment based only on the aligned regions. This is widely acknowledged as a robust measure of genome relatedness based on the studies carried out on 13,151 prokaryotic genomes assigned to 3032 species (Varghese et al. 2015). Therefore the present study was aimed at identifying the genome level similarity between the strains of selected bacterial species based on the data garnered from genome assemblies available at the public databases by using parameter such as gANI.

Materials and methods

To highlight the genome-wide diversity at the species level, we analyzed a total of 300 genome assemblies retrieved from the Genome database (<https://www.ncbi.nlm.nih.gov/genome/>), which comprised of 26 unique bacterial species (Table 1). The selection of bacterial species for the study

was based on two criteria: a) the organism must have whole genome sequence in the completed stage and b) the organism should have only one chromosome with about 10-15 available genome assemblies. The genome sequence(s) were retrieved in Fasta file format and features including genome size, %GC, number of genes and proteins were derived in tabular format. Using Microsoft Excel, we calculated the differences in genome size, G+C content, number of coding genes and non-coding genes at species level for all the short-listed strains. To determine the genome-wide average nucleotide identity (gANI) at species level, multiple genome assemblies of the same species were aligned with each other. Jspecies (<http://jspecies.ribohost.com/jspeciesws/>), a BLAST-based tool was used for the calculation of gANI score and genome coverage (Richter et al. 2015). Using R statistical package, boxplots for each species were generated from pairwise genome identity and coverage. The 16S rRNA and 23S rRNA sequences from the 10 *Flavobacterium psychrophilum* strains used in this study were extracted, aligned using ClustalW and a Maximum Likelihood tree (500 bootstrap; Tamura-Nei distance model) was constructed using MEGA X software (Kumar et al. 2018).

Results

We analyzed a total of 300 genome assemblies representing 26 unique bacterial species retrieved from the Genome database (Table 1). A total of 10-15 strains were selected for each bacterial species for the analysis. With respect to genome features, no significant difference was found in the G+C content among the different strains of 26 bacterial genome assemblies analyzed. However, genome assemblies for a single species, *Prochlorococcus marinus*, showed 20% difference in total GC% among the different strains. We also observed a difference of more than one Mb in genome size and a difference of more than 1000 coding genes within the strains belonging to three species namely *Flavobacterium psychrophilum*, *Prochlorococcus marinus* and *Pseudomonas fluorescens* (Table 1). The number of coding and non-coding genes observed was found to vary proportionally according to their genome size.

Comparative genomics is a fundamental key to understand the functional elements across the genomes and genetic similarity or diversity. We calculated the pairwise genome identity scores and percentage of aligned genomic regions between the genome assemblies belonging to the same species. A boxplot showing pairwise genome comparison identity (gANI) and percentage of aligned genome region was constructed for all the 26 bacterial species (Fig. 1). We found nine out

of the 26 bacterial species analyzed to comprise of several strains with a gANI score of less than 90% at intra-species level (between strains of the same species). The minimum gANI scores (Table 1) were found in strains belonging to *Prochlorococcus marinus* (64.92), *Flavobacterium psychrophilum* (69.26), *Buchnera aphidicola* (69.64), *Clostridium botulinum* (70.09), *Pasteurella multocida* (78.4), *Pseudomonas fluorescens* (79.36), *Aeromonas hydrophila* (85.68), *Stenotrophomonas maltophilia* (86.85) and *Fusobacterium nucleatum* (89.83).

As an example, to explore the conventional phylogenetic analysis using single marker genes for grouping the species or delineating a species in a genus, we performed further analysis of the 10 strains of *Flavobacterium psychrophilum*. The 16S rRNA and 23S rRNA regions were extracted from the whole genome sequences and compared. The gANI score between two *F. psychrophilum* strains namely Z1 and Z2 was shown to be 100% while, the gANI scores of these two strains were highly dissimilar from other strains studied (Table 2).

A BLAST based sequence analysis of 16S rRNA region also showed that *F. psychrophilum* Z1 and Z2 strains showed only 93% identity with the other eight strains of *F. psychrophilum* studied while, the 16S rRNA sequence shared more than 95% identity with other *Flavobacterium* species (*F. sp.* NW20 [99%], *F. hauense* [98%], *F. subsaxonicum* [97%]) and 99% identity with 16S rRNA sequences from uncultured bacterial species. Sequence homology of 23S rRNA region from Z1 and Z2 showed only 94% identity at intra-species level while, these regions showed a maximum identity of 96% with *F. arcticum*. There were no 23S rRNA sequences available for *F. sp.* NW20, *F. hauense* and *F. subsaxonicum* and hence they were not compared. Further, we performed a phylogenetic analysis of 16S and 23S rRNA sequences from the 10 strains of *Flavobacterium psychrophilum* strains (Fig. 2). The 16S rRNA and 23S rRNA sequence analyses showed that Z1 and Z2 strains form a distinct cluster compared to other strains.

Our analysis identified 20 bacterial species which showed <90% alignment against strains of the same species (Table 1). The lowest percentage of aligned regions was found in strains of *Prochlorococcus marinus* (16.37), *Clostridium botulinum* (21.62), *Buchnera aphidicola* (24.42) and *Flavobacterium psychrophilum* (25.41). The median gANI identity score was found to be less than 95% in six bacterial species (*B. aphidicola*, *C. botulinum*, *F. nucleatum*, *P. marinus*, *P. fluorescens*, *S. maltophilia*). This clearly suggested that at least 50% of the genome assemblies belonging to above indicated six species have significant variations in their genomes. A similar result was observed for analysis of trees which were plotted based on the whole genome sequences

for the six bacterial species (available in the NCBI Genome database; see <https://www.ncbi.nlm.nih.gov/genome/tree/170> for *B. aphidicola*) which showed distinct clusters. Similarly, only 12 species (*B. anthracis*, *C. ulcerans*, *G. bethesdensis*, *K. pneumoniae*, *M. avium*, *M. bovis*, *M. gallisepticum*, *M. haemolytica*, *P. acnes*, *P. multocida*, *R. prowazekii*, *R. rickettsii*) showed the median genome coverage of above 90% indicating that genome assemblies of the remaining 14 species (*A. hydrophila*, *B. breve*, *B. longum*, *B. aphidicola*, *C. difficile*, *C. botulinum*, *F. psychrophilum*, *F. nucleatum*, *L. delbrueckii*, *M. mycoides*, *P. marinus*, *P. fluorescens*, *R. anatipestifer*, *S. maltophilia*) analysed have large stretches of unaligned unique regions. These unique genomic regions might code for strain specific genes that might play a significant role in their adaptation and survival.

Discussion

Comparative genomics analysis plays a major role in understanding evolution at both species as well as genus level. By comparing all the genomes at intra-species level, we can now determine the degree of difference in a particular species. The average nucleotide identity calculated from pairwise alignment between two genomes has been proposed as a new measure for taxonomical classification of bacterial species (Zhang et al. 2014; Yoon et al. 2017). A putative ANI cut-off score of 95% or above is suggested to indicate related species while, a score below 95% usually suggests a 'new' species (Figueras et al. 2014; Han et al. 2016). For instance, the pairwise genome comparison between strains of *Pectobacterium parmentieri* resulted in >99% gANI, suggesting they belong to the same species; in contrast, *Pectobacterium parmentieri* and *Pectobacterium wasabiae* showed a gANI score of 94%, indicating that they belong to different species (Khayati et al. 2016).

In this study, we showed nine out of the 26 bacterial species were with remarkable diversity in genome content with a gANI score of less than 90% at intra-species level. The genome diversity/indels might have happened during evolution to maintain the relationship between genotypic-phenotypic features and to enable adaptation to ecological or host habitat. These may impart large diversity at species level, which is challenging for bioinformaticians to characterize based on the data generated from a wide range of microbes using culture-independent high-throughput metagenomics experiments.

We also observed that in 20 different bacterial species, a pairwise alignment showed unaligned regions contributed to more than 10% of the genome, which clearly indicated that these genomes contained several uncharacterized genes. Reasons for such a vast amount of unaligned regions may be due to plasmids found in bacterial species or contamination from host genome (Merchant et al. 2014). Studies indicate that the genetic diversity derived from the plasmids might have roles in bacterial survival at extreme habitats (Dziewit and Bartosik 2014). Inclusion of extrachromosomal genomic regions in pairwise alignment will again increase the extent of unaligned regions. Hence, it appears to be better to avoid such extra chromosomal regions for gANI based taxonomical classification.

Recently, several bacterial species have been reclassified based on the gANI analysis (Godoy et al. 2003; Stropko et al. 2014; Lawson et al. 2016; Vincent et al. 2016; Kook et al. 2017; Loveridge et al. 2017; Sant'Anna et al. 2017; Liu et al. 2018). In an earlier study, Biller et al. (2014) observed remarkable sequence diversity among 27 different *Prochlorococcus* strains from diverse oceanic regions with wide variation in their GC% (30.8-50.6), and these strains were further grouped under physiologically distinct phylogenetic clades. However, till date, only one (*Prochlorococcus marinus*) well classified species is available under the genus *Prochlorococcus* in the Taxonomy browser. Environmental factors are known to influence GC content and are suspected to be the major reason for the differences in GC content observed among closely related species that differ sharply in genome sizes (Wu et al. 2012). Host habitat also plays a major role in species delineation (Basharat et al. 2018) and in this context, comparison of gANI scores in relation to geographical localities from which the strains were isolated is considered critical. However, in our study, we could not classify the strains based on geographical location since this information is not available for the strains whose whole genome sequences were included in the study.

Advances in genome sequencing technologies and bioinformatics approach for genome assembly and annotations have resulted in accumulation of several bacterial genomes in public databases. However, this also has resulted in erroneous species assignments leading to further confusions. Hence, the Genomic Standards Consortium (<http://gensc.org>) has developed a set of standards for the bacterial genome analysis and reporting it into the public domain (Bowers et al. 2017). However, utilising available simpler bioinformatics resources will also provide more insights into genome assemblies. For instance, in addition to genome identity, genome coverage could also be considered as a major factor for genome classification and nomenclature. Similarly, inclusion of

1 authentic and reference gene sequences from the ‘same’ species during sequence analysis will
2 provide better clues on taxonomic identity of the isolate. Further, Genomic Standards Consortium
3 recommendations for the generation, analysis and reporting of bacterial single amplified genomes
4 should be strictly adhered to before depositing microbial whole genome sequences in the
5 international nucleotide sequence databases.

7 **Conclusion**

8 Number of whole genome sequencing projects has surged at an unprecedented momentum and is
9 expected to gather pace since the current number of complete genomes represent only a tiny
10 fraction of the total number of bacterial species on this planet. Over the years, number of studies
11 have highlighted large scale genetic variations at intra-species level and this has led to
12 reclassification of several bacterial species (Stropko et al. 2014; Lawson et al. 2016; Vincent et al.
13 2016; Kook et al. 2017; Loveridge et al. 2017; Liu et al. 2018). This scenario might arise if the
14 species delineation is based only on the traditional methodology followed for the species
15 identification or due to outsourcing the genome sequencing and assembly to other service
16 providers. A move towards implementation of bacterial classification schemes based upon whole
17 genomes has significant implications in taxonomic classification and maintain genome level
18 commonality across the species in genome databases. Therefore, genome based taxonomic
19 classification can be considered when there are sufficient whole genome sequences available in
20 the database. The present study showed that genome assemblies belonging to the nine bacterial
21 species (*P. marinus*, *F. psychrophilum*, *B. aphidicola*, *C. botulinum*, *P. multocida*, *P. fluorescens*,
22 *A. hydrophila*, *S. maltophilia* and *F. nucleatum*) have significant differences (<90% gANI) at
23 species level. To understand the reason for this genetic diversity, we further analyzed genome
24 assemblies of *F. psychrophilum* as an example. Phylogenetic analysis of 16S and 23S rRNA gene
25 sequences from 10 assemblies showed that *F. psychrophilum* Z1 and Z2 strains form a distinct
26 cluster compared to other strains and homology search showed that these two strains have
27 higher degree of sequence similarity at inter-species level than intra-species level. Hence, we
28 recommend that the following steps aimed at reducing the genome level diversity at species level.

- 29 1. Estimation of sequence homology of potential marker genes such as 16S rRNA, 23S rRNA,
30 RecA, RpoB, GyrB, IleS, FusA using BLASTN program to identify relatedness of the

organisms at inter and intra species level. Also, traditional phylogenetic analysis using these orthologous genes will provide information on relationships between taxa.

2. Calculation of genome wide average nucleotide identity (gANI) and percent unaligned genomic regions as a potential measure of intra-species similarity.
3. The GC content, which is similar across most of the bacterial genome assemblies within a species, codon usage and average amino acid identity (AAI) analysis will further help in more accurately classifying an organism.
4. We recommend that researchers should strictly follow the recent guidelines by the genomic standard consortium for genome submission and use available computational resources to reduce the huge genome variations within species level.

Acknowledgements

The authors would like to thank DBT-FIST, Government of India and Manipal Academy of Higher Education, Manipal for the support and facilities provided.

References

- Baltrus, D.A. 2016. Divorcing strain classification from species names. *Trends Microbiol.* **24**: 431–439. doi:10.1016/j.tim.2016.02.004.
- Basharat, Z., Tanveer, F., Yasmin, A., Shinwari, Z.K., He, T., and Tong, Y. 2018. Genome of *Serratia nematodiphila* MB307 offers unique insights into its diverse traits. *Genome.* **61**:469-476. doi:10.1139/gen-2017-0250.
- Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., and Awad, L., et al. 2014. Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Sci. Data* **1**: 140034. doi:10.1038/sdata.2014.34.
- Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., and Reddy, T.B.K., et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**: 725-731. doi:10.1038/nbt.3893.
- Coenye, T., Gevers, D., Van De Peer, Y., Vandamme, P., and Swings, J. 2005. Towards a prokaryotic genomic taxonomy. *FEMS Microbiol. Rev.* **29**: 147-167. doi:10.1016/j.femsre.2004.11.004.
- Cook, C.E., Bergman, M.T., Finn, R.D., Cochrane, G., Birney, E., and Apweiler, R. 2016. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res.* **44**: D20–D26. doi:10.1093/nar/gkv1352.
- Croucher, N.J., and Didelot, X. 2015. The application of genomics to tracing bacterial pathogen transmission. *Curr. Opin. Microbiol.* **23**: 62-67. doi:10.1016/j.mib.2014.11.004.
- Donelli, G., Vuotto, C., and Mastromarino, P. 2013. Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism. *Microb. Ecol. Health Dis.* **24**: 1–8. doi:10.3402/mehd.v24i0.20105.

- 1 Dziewit, L., and Bartosik, D. 2014. Plasmids of psychrophilic and psychrotolerant bacteria and
2 their role in adaptation to cold environments. *Front. Microbiol.* **5**: 596.
3 doi:10.3389/fmicb.2014.00596.
- 4 Eichorst, S.A., Breznak, J.A., and Schmidt, T.M. 2007. Isolation and characterization of soil
5 bacteria that define *Terriglobus* gen. nov., in the phylum Acidobacteria. *Appl. Environ.*
6 *Microbiol.* **73**: 2708–2717. doi:10.1128/AEM.02140-06.
- 7 Figueras, M.J., Beaz-Hidalgo, R., Hossain, M.J., and Liles, M.R. 2014. Taxonomic affiliation of
8 new genomes should be verified using average nucleotide identity and multilocus
9 phylogenetic analysis. *Genome Announc.* **2**: e00927-14-e00927-14.
10 doi:10.1128/genomeA.00927-14.
- 11 Fournier, P.E., Dubourg, G., and Raoult, D. 2014. Clinical detection and characterization of
12 bacterial pathogens in the genomics era. *Genome Med.* **6**: 114. doi:10.1186/s13073-014-
13 0114-2.
- 14 Fricke, W.F., and Rasko, D.A. 2014. Bacterial genome sequencing in the clinic: Bioinformatic
15 challenges and solutions. *Nat. Rev. Genet.* **15**: 49–55. doi:10.1038/nrg3624.
- 16 Gillings, M.R. 2017. Lateral gene transfer, bacterial genome evolution, and the Anthropocene.
17 *Ann. N. Y. Acad. Sci.* **1389**: 20–36. doi:10.1111/nyas.13213.
- 18 Godoy, F., Vancanneyt, M., Martinez, M., Steinbüchel, A., Swings, J., and Rehm, B.H.A. 2003.
19 *Sphingopyxis chilensis* sp. nov., a chlorophenol-degrading bacterium that accumulates
20 polyhydroxyalkanoate, and transfer of *Sphingomonas alaskensis* to *Sphingopyxis*
21 *alaskensis* comb. nov. *Int. J. Syst. Evol. Microbiol.* **53**: 473–477. doi:10.1099/ijs.0.02375-
22 0.
- 23 Gosiewski, T., Ludwig-Galezowska, A.H., Huminska, K., Sroka-Oleksiak, A., Radkowski, P., and
24 Salamon, D., et al. 2017. Comprehensive detection and identification of bacterial DNA in
25 the blood of patients with sepsis and healthy volunteers using next-generation sequencing
26 method - the observation of DNAemia. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**: 329–336.
27 doi:10.1007/s10096-016-2805-7.
- 28 Hahnke, R.L., Meier-Kolthoff, J.P., Garcia-Lopez, M., Mukherjee, S., Huntemann, M., and
29 Ivanova, N.N., et al. 2016. Genome-based taxonomic classification of Bacteroidetes. *Front.*
30 *Microbiol.* **7**: 2003. doi:10.3389/fmicb.2016.02003.
- 31 Han, N., Qiang, Y., and Zhang, W. 2016. ANItools web: a web tool for fast genome comparison
32 within multiple bacterial strains. *Database (Oxford)* **2016**: 1–5.
33 doi:10.1093/database/baw084.
- 34 Hugenholtz, P., Skarshewski, A., and Parks, D.H. 2016. Genome-based microbial taxonomy
35 coming of age. *Cold Spring Harb. Perspect. Biol.* **8**: a018085.
36 doi:10.1101/cshperspect.a018085.
- 37 Jansen van Rensburg, M.J., Swift, C., Cody, A.J., Jenkins, C., and Maiden, M.C.J. 2016.
38 Exploiting bacterial whole-genome sequencing data for evaluation of diagnostic assays:
39 *Campylobacter* species identification as a case study. *J. Clin. Microbiol.* **54**: 2882–2890.
40 doi:10.1128/JCM.01522-16.
- 41 Kim, M., Oh, H.S., Park, S.C., and Chun, J. 2014. Towards a taxonomic coherence between
42 average nucleotide identity and 16S rRNA gene sequence similarity for species
43 demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**: 346–351.
44 doi:10.1099/ijs.0.059774-0.
- 45 Konstantinidis, K.T., and Tiedje, J.M. 2005. Towards a genome-based taxonomy for prokaryotes.
46 *J. Bacteriol.* **187**: 6258–64. doi:10.1128/JB.187.18.6258-6264.2005.

- 1 Kook, J.K., Park, S.N., Lim, Y.K., Cho, E., Jo, E., and Roh, H., et al. 2017. Genome-based
2 reclassification of *Fusobacterium nucleatum* subspecies at the species level. *Curr.*
3 *Microbiol.* **74**: 1137–1147. doi:10.1007/s00284-017-1296-9.
- 4 Kryukov, K., and Imanishi, T. 2016. Human contamination in public genome assemblies. *PLoS*
5 *One* **11**: e0162424. doi:10.1371/journal.pone.0162424.
- 6 Kubicova, V., and Provaznik, I. 2016. Use of whole genome DNA spectrograms in bacterial
7 classification. *Comput. Biol. Med.* **69**: 298–307. doi:10.1016/j.compbio.2015.04.038.
- 8 Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular
9 Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**: 1547–
10 1549. doi:10.1093/molbev/msy096.
- 11 Lawson, P.A., Citron, D.M., Tyrrell, K.L., and Finegold, S.M. 2016. Reclassification of
12 *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938.
13 *Anaerobe* **40**: 95–99. doi:10.1016/j.anaerobe.2016.06.008.
- 14 Liu, W., Li, L., Khan, M.A., and Zhu, F. 2012. Popular molecular markers in bacteria. *Mol. Genet.*
15 *Microbiol. Virol.* **27**: 103–107. doi:10.3103/S0891416812030056.
- 16 Liu, Y., Lai, Q., and Shao, Z. 2018. Genome analysis-based reclassification of *Bacillus*
17 *weihenstephanensis* as a later heterotypic synonym of *Bacillus mycoides*. *Int. J. Syst. Evol.*
18 *Microbiol.* **68**: 106–112. doi:10.1099/ijsem.0.002466.
- 19 Loveridge, E.J., Jones, C., Bull, M.J., Moody, S.C., Kahl, M.W., and Khan, Z., et al. 2017.
20 Reclassification of the specialized metabolite producer *Pseudomonas mesoacidophila*
21 ATCC 31433 as a member of the *Burkholderia cepacia* complex. *J. Bacteriol.* **199**: e00125-
22 17. doi:10.1128/JB.00125-17.
- 23 McCutcheon, J.P., and Moran, N.A. 2012. Extreme genome reduction in symbiotic bacteria. *Nat.*
24 *Rev. Microbiol.* **10**: 13–26. doi:10.1038/nrmicro2670.
- 25 Merchant, S., Wood, D.E., and Salzberg, S.L. 2014. Unexpected cross-species contamination in
26 genome sequencing projects. *PeerJ* **2**: e675. doi:10.7717/peerj.675.
- 27 Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N.C., and Pati, A. 2015. Large-scale
28 contamination of microbial isolate genomes by illumina Phix control. *Stand. Genomic Sci.*
29 **10**: 18. doi:10.1186/1944-3277-10-18.
- 30 Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemskaya, O., and Isbandi, M., et
31 al. 2017. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements.
32 *Nucleic Acids Res.* **45**: D446–D456. doi:10.1093/nar/gkw992.
- 33 Ochman, H., Lawrence, J.G., and Grolsman, E.A. 2000. Lateral gene transfer and the nature of
34 bacterial innovation. *Nature* **405**: 299–304. doi:10.1038/35012500.
- 35 Richter, M., Rossello-Mora, R., Oliver Glockner, F., and Peplies, J. 2015. JSpeciesWS: A web
36 server for prokaryotic species circumscription based on pairwise genome comparison.
37 *Bioinformatics* **32**: 929–931. doi:10.1093/bioinformatics/btv681.
- 38 Sant'Anna, F.H., Ambrosini, A., de Souza, R., de Carvalho Fernandes, G., Bach, E., and
39 Balsanelli, E., et al. 2017. Reclassification of *Paenibacillus riograndensis* as a genomovar
40 of *Paenibacillus sonchi*: Genome-based metrics improve bacterial taxonomic
41 classification. *Front. Microbiol.* **8**: 1–13. doi:10.3389/fmicb.2017.01849.
- 42 Schloss, P.D., and Handelsman, J. 2004. Status of the microbial census. *Microbiol. Mol. Biol. Rev.*
43 **68**: 686–691. doi:10.1128/MMBR.68.4.686-691.2004.
- 44 Srivathsan, A., Sha, J.C.M., Vogler, A.P., and Meier, R. 2015. Comparing the effectiveness of
45 metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix*
46 *nemaeus*). *Mol. Ecol. Resour.* **15**: 250–261. doi:10.1111/1755-0998.12302.

- 1 Stropko, S.J., Pipes, S.E., and Newman, J.D. 2014. Genome-based reclassification of *Bacillus cibi*
2 as a later heterotypic synonym of *Bacillus indicus* and emended description of *Bacillus*
3 *indicus*. Int. J. Syst. Evol. Microbiol. **64**: 3804–3809. doi:10.1099/ij.s.0.068205-0.
- 4 Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., and Kyrpides,
5 N.C., et al. 2015. Microbial species delineation using whole genome sequences. Nucleic
6 Acids Res. **43**: 6761–6771. doi:10.1093/nar/gkv657.
- 7 Vincent, A.T., Trudel, M. V., Freschi, L., Nagar, V., Gagne-Thivierge, C., and Levesque, R.C., et
8 al. 2016. Increasing genomic diversity and evidence of constrained lifestyle evolution due
9 to insertion sequences in *Aeromonas salmonicida*. BMC Genomics **17**: 1–12.
10 doi:10.1186/s12864-016-2381-3.
- 11 Wu, H., Zhang, Z., Hu, S., and Yu, J. 2012. On the molecular mechanism of GC content variation
12 among eubacterial genomes. Biol. Direct **7**: 2. doi:10.1186/1745-6150-7-2.
- 13 Yoon, S.H., Ha, S. min, Lim, J., Kwon, S., and Chun, J. 2017. A large-scale evaluation of
14 algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek, Int. J. Gen.
15 Mol. Microbiol. **110**: 1281–1286. doi:10.1007/s10482-017-0844-4.
- 16 Zhang, W., Du, P., Zheng, H., Yu, W., Wan, L., and Chen, C. 2014. Whole-genome sequence
17 comparison as a method for improving bacterial species definition. J. Gen. Appl.
18 Microbiol. **60**: 75–78. doi:10.2323/jgam.60.75.
- 19

Draft

1 Table 1 Genome level differences for isolates of selected 26 bacterial species

Sl. No.	Organism	No. of genome assemblies	Genome size (Mb)		Predicted CDS		Noncoding genes		G+C content (%)		gANIb Identity (%)		Aligned regions (%)	
			Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1	<i>Aeromonas hydrophila</i>	10	4.52	5.27	4164	4902	165	269	60.5	62.0	85.68	100	66.25	99.90
2	<i>Bacillus anthracis</i>	15	5.22	5.22	4952	5513	128	371	35.4	35.4	99.95	100	99.18	99.97
3	<i>Bifidobacterium breve</i>	10	2.24	2.42	1820	1960	135	214	58.6	59.0	97.31	98.82	81.77	94.44
4	<i>Bifidobacterium longum</i>	13	2.26	2.83	1796	2443	127	266	59.4	60.3	93.46	98.95	56.53	90.45
5	<i>Buchnera aphidicola</i>	10	0.44	0.64	355	577	45	223	23.0	26.3	69.64	99.92	24.42	76.36
6	<i>Clostridioides difficile</i>	11	4.04	4.29	3505	4128	143	251	28.5	29.2	95.49	100	74.93	90.23
7	<i>Clostridium botulinum</i>	12	3.61	4.39	3072	3971	66	360	27.4	28.3	70.09	100	21.62	87.58
8	<i>Corynebacterium ulcerans</i>	10	2.43	2.60	2035	2291	96	360	53.4	54.4	90.08	99.99	88.65	99.78
9	<i>Flavobacterium psychrophilum</i>	10	2.68	4.14	2311	3442	18	192	32.4	39.0	69.26	100	25.41	99.61
10	<i>Fusobacterium nucleatum</i>	10	2.03	2.54	1826	2325	19	198	26.8	27.2	89.83	99.97	52.93	83.65
11	<i>Granulibacter thebesensis</i>	10	2.69	2.74	2338	2423	77	113	58.5	59.2	91.39	100	92.17	99.83
12	<i>Klebsiella pneumoniae</i>	15	5.08	5.52	4767	5194	206	318	57.0	58.0	93.32	99.98	82.66	99.73
13	<i>Lactobacillus delbrueckii</i>	13	1.85	2.26	1547	1950	251	387	49.1	50.1	97.03	99.36	73.77	94.20
14	<i>Mannheimia haemolytica</i>	10	2.50	2.70	2263	2651	151	226	41.0	41.1	98.40	100	82.60	99.12
15	<i>Mycobacterium avium</i>	12	4.78	5.62	4150	4955	212	406	68.8	69.3	97.17	99.99	80.63	97.89
16	<i>Mycobacterium bovis</i>	15	4.33	4.41	3895	4778	48	331	65.6	65.7	99.63	100	95.69	97.58
17	<i>Mycoplasma gallisepticum</i>	12	0.92	1.01	701	745	61	113	31.4	31.6	96.82	99.94	89.67	97.05
18	<i>Mycoplasma mycoides</i>	12	1.08	1.21	845	1051	35	164	23.9	24.0	94.32	100	56.88	76.08
19	<i>Pasteurella multocida</i>	14	2.07	2.41	1885	2202	89	255	39.7	40.4	78.40	99.99	78.49	99.72
20	<i>Prochlorococcus marinus</i>	11	1.64	2.68	1784	2997	48	139	30.8	50.7	64.92	96.58	16.37	93.38
21	<i>Propionibacterium acnes</i>	10	2.48	2.56	2257	2416	130	216	60.0	60.1	97.03	99.98	92.63	99.88
22	<i>Pseudomonas fluorescens</i>	14	5.84	7.11	5179	6256	167	302	58.8	61.3	79.36	99.96	52.06	98.73
23	<i>Rickettsia prowazekii</i>	10	1.10	1.11	830	871	46	95	29.0	29.0	99.85	99.99	95.28	96.77
24	<i>Rickettsia rickettsia</i>	11	1.25	1.27	1246	1272	263	281	32.4	32.5	99.53	100	95.70	97.78
25	<i>Riemerella anatipestifer</i>	10	2.15	2.42	1801	2238	68	270	35.0	35.2	95	99.99	74.79	98.50
26	<i>Stenotrophomonas maltophilia</i>	10	4.50	4.85	3964	4368	112	287	66.3	67.4	86.85	100	67.16	99.59

2

Table 2. Whole genome average nucleotide identity (gANI) between different strains of *Flavobacterium psychrophilum*. Values in square brackets show percentage of aligned genomic region between strains.

STRAIN	Z1	Z2	950106-1/1	FPG3	FPG101	CSF259-93	V4-33	V4-24	V3-5	JIP02/86
Z1	*	100.00 [99.61]	69.29 [25.42]	69.27 [25.58]	69.29 [25.47]	69.30 [25.48]	69.29 [25.48]	69.30 [25.41]	69.29 [25.48]	69.29 [25.47]
Z2	100.00 [99.62]	*	69.27 [25.45]	69.28 [25.59]	69.27 [25.46]	69.27 [25.51]	69.27 [25.51]	69.27 [25.44]	69.27 [25.51]	69.26 [25.49]
950106-1/1	69.87 [34.79]	69.87 [34.79]	*	99.25 [82.58]	99.78 [92.13]	99.84 [92.78]	99.97 [94.21]	99.96 [93.95]	99.97 [93.89]	99.93 [94.64]
FPG3	70.42 [35.77]	70.43 [35.74]	99.19 [84.67]	*	99.21 [87.51]	99.19 [87.41]	99.09 [85.41]	99.23 [84.65]	99.17 [85.32]	99.25 [87.34]
FPG101	70.08 [34.25]	70.08 [34.25]	99.71 [90.24]	99.22 [83.45]	*	99.89 [94.75]	99.65 [90.87]	99.74 [90.40]	99.71 [91.06]	99.83 [93.52]
CSF259-93	70.07 [33.49]	70.07 [33.49]	99.76 [89.31]	99.06 [82.22]	99.78 [93.72]	*	99.66 [89.98]	99.77 [89.39]	99.71 [90.01]	99.81 [93.03]
V4-33	69.76 [34.85]	69.76 [34.84]	99.97 [94.94]	99.32 [83.63]	99.79 [93.19]	99.86 [93.90]	*	99.96 [94.95]	99.97 [95.26]	99.94 [94.16]
V4-24	69.74 [35.29]	69.75 [35.28]	99.97 [95.51]	99.31 [84.22]	99.81 [93.44]	99.88 [94.17]	99.98 [95.67]	*	99.97 [95.55]	99.94 [94.46]
V3-5	69.77 [35.36]	69.77 [35.36]	99.95 [95.16]	99.28 [84.22]	99.78 [93.47]	99.87 [94.11]	99.96 [95.58]	99.96 [95.00]	*	99.93 [94.40]
JIP02/86	70.36 [33.68]	70.36 [33.68]	99.82 [91.86]	99.18 [83.06]	99.81 [92.63]	99.87 [93.41]	99.75 [90.99]	99.83 [90.47]	99.82 [91.06]	*

Figure 1 Boxplot showing percentage genome identity (A) and coverage percentage (B) at intra-species level for the bacterial species studied. The boxes represent the 25th-75th percentile values with the median score while the circles indicate outliers.

347x175mm (300 x 300 DPI)

Figure 2. Phylogenetic tree based on the nucleotide sequences of A)16S ribosomal RNA and B) 23S ribosomal RNA genes from 10 *Flavobacterium psychrophilum* strains selected in the study. The phylogenetic tree was constructed by the Maximum Likelihood method, using the MEGA X Software.

513x186mm (300 x 300 DPI)