## Genome Based Taxonomic Classification

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

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

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

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

**Conclusion**

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

1. Estimation of sequence homology of potential marker genes such as 16S rRNA, 23S rRNA, 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:  

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-  

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., and Awad, L., et  
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.


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  

Croucher, N.J., and Didelot, X. 2015. The application of genomics to tracing bacterial pathogen  

Donelli, G., Vuotto, C., and Mastromarino, P. 2013. Phenotyping and genotyping are both essential  
doi:10.3402/mehd.v24i0.20105.


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

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Organism</th>
<th>No. of genome assemblies</th>
<th>Genome size (Mb)</th>
<th>Predicted CDS Min</th>
<th>Predicted CDS Max</th>
<th>Noncoding genes Min</th>
<th>Noncoding genes Max</th>
<th>G+C content (%) Min</th>
<th>G+C content (%) Max</th>
<th>gANIb Identity (%) Min</th>
<th>gANIb Identity (%) Max</th>
<th>Aligned regions (%) Min</th>
<th>Aligned regions (%) Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aeromonas hydrophila</td>
<td>10</td>
<td>4.52</td>
<td>5.27</td>
<td>4164</td>
<td>4902</td>
<td>165</td>
<td>269</td>
<td>60.5</td>
<td>62.0</td>
<td>85.68</td>
<td>100</td>
<td>66.25</td>
</tr>
<tr>
<td>2</td>
<td>Bacillus anthracis</td>
<td>15</td>
<td>5.22</td>
<td>5.22</td>
<td>4952</td>
<td>5513</td>
<td>128</td>
<td>371</td>
<td>35.4</td>
<td>35.4</td>
<td>99.95</td>
<td>100</td>
<td>99.18</td>
</tr>
<tr>
<td>3</td>
<td>Bifidobacterium breve</td>
<td>10</td>
<td>2.24</td>
<td>2.42</td>
<td>1820</td>
<td>1960</td>
<td>135</td>
<td>214</td>
<td>58.6</td>
<td>59.0</td>
<td>97.31</td>
<td>98.82</td>
<td>81.77</td>
</tr>
<tr>
<td>4</td>
<td>Bifidobacterium longum</td>
<td>13</td>
<td>2.26</td>
<td>2.83</td>
<td>1796</td>
<td>2443</td>
<td>127</td>
<td>266</td>
<td>59.4</td>
<td>60.3</td>
<td>93.46</td>
<td>98.95</td>
<td>56.53</td>
</tr>
<tr>
<td>5</td>
<td>Buchnera aphidicola</td>
<td>10</td>
<td>0.44</td>
<td>0.64</td>
<td>355</td>
<td>577</td>
<td>45</td>
<td>223</td>
<td>23.0</td>
<td>26.3</td>
<td>69.64</td>
<td>99.92</td>
<td>24.22</td>
</tr>
<tr>
<td>6</td>
<td>Clostridoides difficile</td>
<td>11</td>
<td>4.04</td>
<td>4.29</td>
<td>3505</td>
<td>4128</td>
<td>143</td>
<td>251</td>
<td>28.5</td>
<td>29.2</td>
<td>95.49</td>
<td>100</td>
<td>74.93</td>
</tr>
<tr>
<td>7</td>
<td>Clostridium butulinum</td>
<td>12</td>
<td>3.61</td>
<td>4.39</td>
<td>3072</td>
<td>3971</td>
<td>66</td>
<td>360</td>
<td>27.4</td>
<td>28.3</td>
<td>70.09</td>
<td>100</td>
<td>21.62</td>
</tr>
<tr>
<td>8</td>
<td>Corynebacterium ulcerans</td>
<td>10</td>
<td>2.43</td>
<td>2.60</td>
<td>2035</td>
<td>2291</td>
<td>96</td>
<td>360</td>
<td>53.4</td>
<td>54.4</td>
<td>90.08</td>
<td>99.99</td>
<td>88.65</td>
</tr>
<tr>
<td>9</td>
<td>Flavobacterium psychrophilum</td>
<td>10</td>
<td>2.68</td>
<td>4.14</td>
<td>2311</td>
<td>3442</td>
<td>18</td>
<td>192</td>
<td>32.4</td>
<td>39.0</td>
<td>69.26</td>
<td>100</td>
<td>25.41</td>
</tr>
<tr>
<td>10</td>
<td>Fusobacterium nucleatum</td>
<td>10</td>
<td>2.03</td>
<td>2.54</td>
<td>1826</td>
<td>2325</td>
<td>19</td>
<td>198</td>
<td>26.8</td>
<td>27.2</td>
<td>89.83</td>
<td>99.97</td>
<td>52.93</td>
</tr>
<tr>
<td>11</td>
<td>Granulibacter bethesdensis</td>
<td>10</td>
<td>2.69</td>
<td>2.74</td>
<td>2338</td>
<td>2423</td>
<td>77</td>
<td>113</td>
<td>58.5</td>
<td>59.2</td>
<td>91.39</td>
<td>100</td>
<td>92.17</td>
</tr>
<tr>
<td>12</td>
<td>Klebsiella pneumoniae</td>
<td>15</td>
<td>5.08</td>
<td>5.52</td>
<td>4767</td>
<td>5194</td>
<td>206</td>
<td>318</td>
<td>57.0</td>
<td>58.0</td>
<td>93.32</td>
<td>99.98</td>
<td>82.66</td>
</tr>
<tr>
<td>13</td>
<td>Lactobacillus delbrueckii</td>
<td>13</td>
<td>1.85</td>
<td>2.26</td>
<td>1547</td>
<td>1950</td>
<td>251</td>
<td>387</td>
<td>49.1</td>
<td>50.1</td>
<td>97.03</td>
<td>99.36</td>
<td>73.77</td>
</tr>
<tr>
<td>14</td>
<td>Mannheimia haemolytica</td>
<td>10</td>
<td>2.50</td>
<td>2.70</td>
<td>2263</td>
<td>2651</td>
<td>151</td>
<td>226</td>
<td>41.0</td>
<td>41.1</td>
<td>98.40</td>
<td>100</td>
<td>82.60</td>
</tr>
<tr>
<td>15</td>
<td>Mycobacterium avium</td>
<td>12</td>
<td>4.78</td>
<td>5.62</td>
<td>4150</td>
<td>4955</td>
<td>212</td>
<td>406</td>
<td>68.8</td>
<td>69.3</td>
<td>97.17</td>
<td>99.99</td>
<td>80.63</td>
</tr>
<tr>
<td>16</td>
<td>Mycobacterium bovis</td>
<td>15</td>
<td>4.33</td>
<td>4.41</td>
<td>3895</td>
<td>4778</td>
<td>48</td>
<td>331</td>
<td>65.6</td>
<td>65.7</td>
<td>99.63</td>
<td>100</td>
<td>95.69</td>
</tr>
<tr>
<td>17</td>
<td>Mycoplasma gallisepticum</td>
<td>12</td>
<td>0.92</td>
<td>1.01</td>
<td>701</td>
<td>745</td>
<td>61</td>
<td>113</td>
<td>31.4</td>
<td>31.6</td>
<td>96.82</td>
<td>99.94</td>
<td>89.67</td>
</tr>
<tr>
<td>18</td>
<td>Mycoplasma mycoides</td>
<td>12</td>
<td>1.08</td>
<td>1.21</td>
<td>845</td>
<td>1051</td>
<td>35</td>
<td>164</td>
<td>23.9</td>
<td>24.0</td>
<td>94.32</td>
<td>100</td>
<td>56.88</td>
</tr>
<tr>
<td>19</td>
<td>Pasteurella multocida</td>
<td>14</td>
<td>2.07</td>
<td>2.41</td>
<td>1885</td>
<td>2202</td>
<td>89</td>
<td>255</td>
<td>39.7</td>
<td>40.4</td>
<td>78.40</td>
<td>99.99</td>
<td>78.49</td>
</tr>
<tr>
<td>20</td>
<td>Prochlorococcus marinus</td>
<td>11</td>
<td>1.64</td>
<td>2.68</td>
<td>1784</td>
<td>2997</td>
<td>48</td>
<td>139</td>
<td>30.8</td>
<td>50.7</td>
<td>64.92</td>
<td>96.58</td>
<td>16.37</td>
</tr>
<tr>
<td>21</td>
<td>Propionibacterium acnes</td>
<td>10</td>
<td>2.48</td>
<td>2.56</td>
<td>2257</td>
<td>2416</td>
<td>130</td>
<td>216</td>
<td>60.0</td>
<td>60.1</td>
<td>97.03</td>
<td>99.98</td>
<td>92.63</td>
</tr>
<tr>
<td>22</td>
<td>Pseudomonas fluorescens</td>
<td>14</td>
<td>5.84</td>
<td>7.11</td>
<td>5179</td>
<td>6256</td>
<td>167</td>
<td>302</td>
<td>58.8</td>
<td>61.3</td>
<td>79.36</td>
<td>99.96</td>
<td>52.06</td>
</tr>
<tr>
<td>23</td>
<td>Rickettsia prowazekii</td>
<td>10</td>
<td>1.10</td>
<td>1.11</td>
<td>830</td>
<td>871</td>
<td>46</td>
<td>95</td>
<td>29.0</td>
<td>29.0</td>
<td>99.85</td>
<td>99.99</td>
<td>95.28</td>
</tr>
<tr>
<td>24</td>
<td>Rickettsia rickettsia</td>
<td>11</td>
<td>1.25</td>
<td>1.27</td>
<td>1246</td>
<td>1272</td>
<td>263</td>
<td>281</td>
<td>32.4</td>
<td>32.5</td>
<td>99.53</td>
<td>100</td>
<td>95.70</td>
</tr>
<tr>
<td>25</td>
<td>Riemerella anatipestifer</td>
<td>10</td>
<td>2.15</td>
<td>2.42</td>
<td>1801</td>
<td>2238</td>
<td>68</td>
<td>270</td>
<td>35.0</td>
<td>35.2</td>
<td>95</td>
<td>99.99</td>
<td>74.79</td>
</tr>
<tr>
<td>26</td>
<td>Stenotrophomonas maltophilia</td>
<td>10</td>
<td>4.50</td>
<td>4.85</td>
<td>3964</td>
<td>4368</td>
<td>112</td>
<td>287</td>
<td>66.3</td>
<td>67.4</td>
<td>86.85</td>
<td>100</td>
<td>67.16</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>STRAIN</th>
<th>Z1</th>
<th>Z2</th>
<th>950106-1/1</th>
<th>FPG3</th>
<th>FPG101</th>
<th>CSF259-93</th>
<th>V4-33</th>
<th>V4-24</th>
<th>V3-5</th>
<th>JIP02/86</th>
</tr>
</thead>
</table>
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.
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.