**Title ：Comparison between MAGs characteristics and genomes of isolates**

**Summary：** With the rapid increase in the number of environmental metagenomes, binning algorithms, and genomes assembled from metagenomes, it is easy to obtain a large amount of data, but how to check the reliability of the data is still uncertain. If this unreliability is ignored in practice, the resulting phylogenetic trees may have some peculiar branches, overestimated niche boundaries, and unique singletons in the whole genome. To address these issues, we defined a new metric to test the reliability of genomes by comparing their characteristics (e.g. genome size and GC-content distribution.) with those obtained from strictly related genomes of isolates (NCBI databases). We hope to provide a reference method for solving the limitation of single-copy gene evaluation.

**Availability and implementation**

The script code and related instructions involved in this research have been uploaded to GitHub (https://github.com/artur-sannikov/contact); the Python version used in the software design process is 3.8.6.

**1.Introduction**

Genomes from isolate sequencing and metagenomics assembly are being generated at higher and higher rates, and they are all available correspondingly more and more through public resources, which provides an invaluable understanding of the overall characteristics of the microbial diversity that affect the environment. There are tens of thousand microbial genomes in public databases and this made it impossible to resume characteristics of taxonomic groups manually. In addition, in spite of bacterial and archaeal single-copy core genes are useful proxies to estimate the level of 'completion' and 'contamination' of metagenome-assembled genomes, they cannot guarantee the correct results, especially in highly fragmented metagenome-assembled genomes (MAGs). If this unreliability is ignored in practice, due to hidden contamination, the resulting phylogenomic tree may have some peculiar branches, overestimated niche boundaries, and unique singletons in the whole genome. With the avalanche of data of microbial genomes generated, there is a lack of methods that can easily test the reliability of genomes.

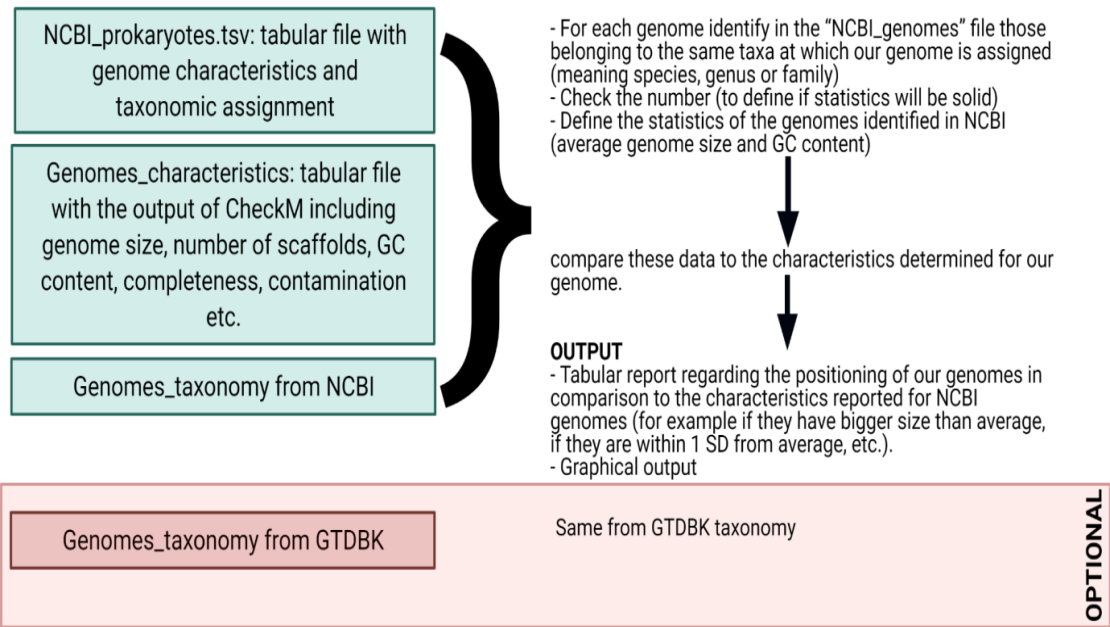
Recovering metagenome-assembled genomes from metagenomics data has recently become a common task in the field of microbial and environmental research. Some researchers have been developing software to verify the reliability of MAGs data representing natural populations. Identified core genes within MAGs or groups of MAGs of the same species offer computationally tractable cut-offs to gain quick insights into MAGs to identify mutations(Luo et al., 2015; Quince et al., 2017). Furthermore, Meziti et al. (Meziti et al., 2021) compared MAGs against Isolate Genomes Derived from the Same Fecal Sample to identify the quality of MAGs. For reference, the characteristics of the genome can be used as a factor in the quality of the test results. Through the comparison of genomic features with those of strictly related species and the identification of putative incomplete genomes, it can provide a promising strategy for verifying results.

Based on the above description, this research aimed to define a new metric to test the reliability of genomes by comparing their characteristics with those obtained from strictly related genomes of isolates (NCBI database). The microbial gene data was downloaded from the NCBI database. The genome characteristics studied mainly include: GC-content, genome size, number of scaffolds, completeness and contamination. In addition, the final results are presented in statistical tabulle and box plot formats.

**2Method and software features**

**2.1 research method**

1. the first input file will be the NCBI tabular file including genomes characteristics and taxonomic assignment (including species names) (NCBI\_prokaryotes.csv). this file contains many controllable filter factors (such as the completeness level of the assembly)；
2. the second input file will be represented by the genomic characteristics of the genomes under investigation (tabular file including genome size, number of scaffolds, GC content, completeness, contamination level, etc.) (checkM output file) ；
3. the third input file will be the taxonomic assignment of the genomes under investigation (genomes\_taxonomy). The specific workflow is shown in Fig.1. For each genome to be investigated (at least assigned to the genus level), we have identified those genomes that belong to the same taxa (ie species, genus) in which our genome is located in the NCBI file. In addition, we define the recognition in NCBI Statistics of the genome (average genome size and GC-content) and compare these data with the characteristics determined for our genome.



**Fig.1. Feature comparison pipeline of genome**

**2.2 Script functions**

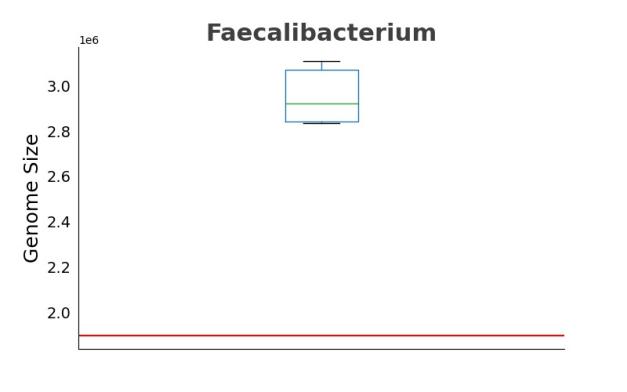
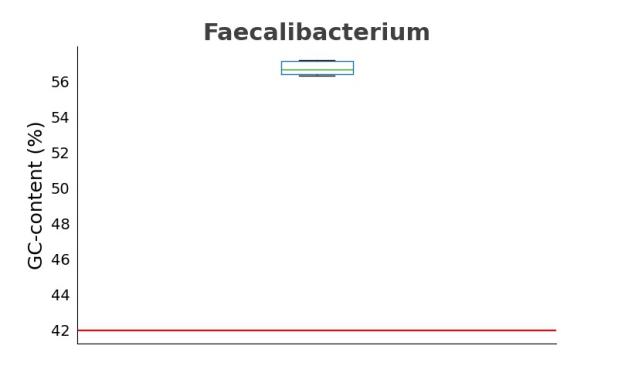
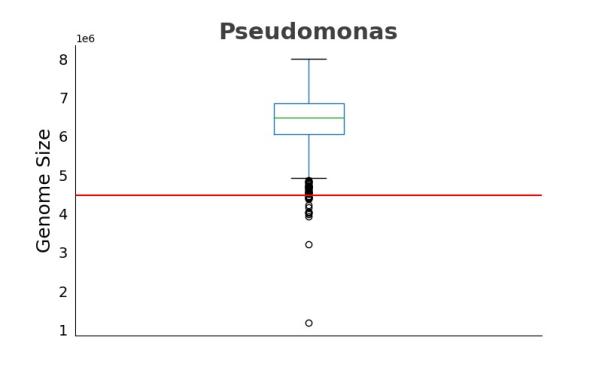
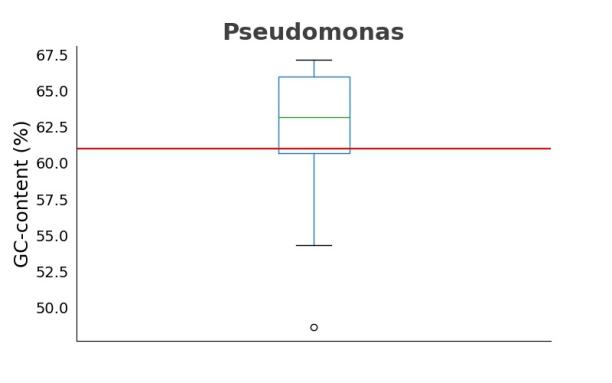
**Overview of script functions:**

1. Our script extracts genomes that are classified at least to the genus level in the investigated genome for later research;
2. The argparse module allows users to freely enter the specify enter the arguments to be tested ;
3. Filtered out the genomes of "uncompleteness" in the NCBI database;
4. Genome characteristic factors used for comparison are: "Genome size (bp)", "GC-content",
5. The output file compares the average value, standard deviation of GC-content and genome size between NCBI\_genome and the input genomes. In addition, the difference between different metric is also output;
6. The output also includes two columns that contain values indicating the number of standard deviations from the mean value of GC-content and genome size of input genomes.
7. The final results are displayed in table and box plots .

**3. Results:**

**Table.1** shows the comparison results of the characteristics of 115 speciesable (the script was ran with the default parameters), containingsome features of input genomes such as genuses they belong to, the Bin Id assigned by \*checkM\*, genome completeness, contamination, and the number of scaffolds. It also contains the difference in means of GC-content and genome size between input and reference genommes at the genus level. The last two columns contain the numbers that indicate if a GC-content or a genome size is within a certain number of standard deviations. Both stardard deviations were computed using the NCBI database and then the mean value of both statistics of input genomes was divided by the corresponding standard deviation.

The box plots in **Fig.2** show this more intuitively. This section selects two well-represented genus (*Pseudomonas* and *Faecalibacterium*) to display the results. The red line in the figure represents the mean value of GC-content and genome size of the studied genomes. the box plots represent the GC content distribution of the genome from the NCBI database. It can be seen from the figure that *Pseudomonas* has a strong similarity with the reference one, while *Faecalibacterium* and the data in the gene bank have significant differences (standard deviations difference are 34.3 and 8.2, respectively), and it is worthwhile to further explore the reasons.



**Fig.2. Box plot of GC content(%) and gene size comparison**

Std:8.2

Std:34.3

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **genus** | **Bin Id** | **Completeness** | **Contamination** | **# scaffolds** | **GC\_diff** | **genome\_size\_diff** | **GC\_std** | **genome\_size\_std** |
| Acetobacter | METABAT\_AS27yjCOA\_120 | 97.81 | 2.9 | 232 | -0.48543 | -396565 | 0.26419399 | 1.248824067 |
| Acholeplasma | METABAT\_AS08sgBPME\_69 | 94.67 | 0.12 | 214 | -1.73516 | -584147 | 0.702632781 | 3.240656952 |
| Acholeplasma | METABAT\_AS23ysBPME\_113 | 98 | 1.33 | 199 | -1.73516 | -584147 | 0.702632781 | 3.240656952 |
| Acinetobacter | METABAT\_AS04akNAM\_10 | 99.45 | 0 | 34 | 1.895374 | -1218754 | 1.161396518 | 3.559908136 |
| Acinetobacter | METABAT\_AS10tlH2TH\_60 | 92.6 | 4.19 | 584 | 1.895374 | -1218754 | 1.161396518 | 3.559908136 |
| Actinomyces | METABAT\_AS08sgBPME\_431 | 91.32 | 3.08 | 281 | -10.7676 | -1014498 | 1.559717919 | 1.899096587 |
| Arcanobacterium | METABAT\_AS4LglBPNY\_23 | 99.71 | 0.86 | 161 | 0.28 | -26409 | 0.085891995 | 0.085956298 |
| Arcobacter | METABAT\_AS07pgkLD\_148 | 97.02 | 4.94 | 243 | 1.066667 | -1063654 | 1.3014383 | 2.369657174 |
| Arcobacter | METABAT\_AS15tlH2ME\_226 | 90.45 | 1.02 | 304 | 1.066667 | -1063654 | 1.3014383 | 2.369657174 |
| Bacillus | METABAT\_AS04akNAM\_110 | 97.08 | 3.52 | 138 | -3.792525 | -647945 | 0.8060915 | 0.810212563 |

**Table.1 The difference in characteristics between the genome under investigation and the NCBI\_genomes(head 10)**

**References:**

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nature biotechnology 2015; 33: 1045-1052.

Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: Insights from comparing MAGs against isolate genomes derived from the same fecal sample. Applied and Environmental Microbiology 2021; 87.

Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome biology 2017; 18: 1-22.