Do the human gut metagenomic species possess the minimal set of core functionalities necessary for life?

Matteo Soverini (matteo.soverini5@unibo.it)
University of Bologna https://orcid.org/0000-0002-3026-9460
Simone Rampelli
University of Bologna
Silvia Turroni
University of Bologna
Patrizia Brigidi
University of Bologna
Elena Biagi
University of Bologna
Marco Candela
University of Bologna

Research article

Keywords: Gut microbiome, metagenomic assembled genomes, uncultured metagenomic species, minimal bacterial genome

Posted Date: October 9th, 2019

DOI: https://doi.org/10.21203/rs.2.15864/v1

Version of Record: A version of this preprint was published on September 30th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07087-8.

Abstract

Background. Advances in bioinformatics recently allowed for the recovery of ‘metagenomes assembled genomes’ from human microbiome studies carried on with shot-gut sequencing techniques. Such approach has been proposed as a mean to discover new so-called ‘uncultured metagenomic species’. Results. In the present analysis we compare 400 genomes from isolates available on NCBI database and 400 human gut metagenomic species, screening all of them for the presence of a minimal set of core functionalities necessary, but not sufficient, for life. Even if a genome completeness up to 96% is reported in the original study, the metagenomic species resulted substantially depleted in genes encoding for essential functions to support autonomous bacterial life, with the 16S rRNA gene missing in 237 out of the 400 metagenomic species we analyzed. Conclusions. The relevant degree of lacking core functionalities that we observed in metagenomic species raises some concerns about the effective completeness of metagenome-assembled genomes, suggesting caution in extrapolating information about their metabolic propensity and ecological role in the human gastrointestinal tract.

Background

Integral to the human biology, the Gut Microbiome (GM) is a key determinant of our health and its dysbiotic variations have been associated with several inflammatory diseases (1). Species-level variation in GM has been indicated as an emergent factor to be considered both for a better understanding of the biology of the GM-host mutualism (2) and for a refined evaluation of the individual health risk (3). However, even if it is perceived as strategic in GM study, the capability of shotgun metagenomics to infer species-level taxonomic and functional information is traditionally limited by the relative paucity of reference genomes. Indeed, despite the important progresses in culturomics, the degree of unclassified GM diversity at the species level is still very high. An important step forward in this direction has been recently provided by genome-resolved metagenomics, which involves the simultaneous recovery of draft and complete genomes directly from sequenced metagenomes (4). In particular, this approach consists in a de novo assembly of shotgun metagenomic reads into contigs, which are binned on the basis of coverage and tetranucleotide frequency (5, 6). This strategy allows the recovery of thousands of new genomes, i.e. the so called ‘Metagenomes Assembled Genomes’ (MAGs), directly from metagenomic reads, considerably expanding the tree of life beyond the limits of cultivability (7). Very recently, Almeida et al. (8) provided a first extensive discovery campaign of MAGs from 13,133 human metagenomic samples. In particular, the Authors successfully characterized 1,175 nearly complete MetaGenomic Species (MGS), estimating a median completeness of 96.5% and 0.8% of contamination. Further, additional 893 medium quality MGS were also detected, with median completeness of 77.8% and 1.1% of contamination. The 94% of these MAG (1,952) did not match any bacterial isolate genome included in the Human-specific Reference (HR) (9) and RefSeq databases and thus indicated as new Uncultured MetaGenomic Species (UMGS). 74% of UMGS correspond to entirely novel genomes. 26% of the UMGS belonged to potential new families and 40% to new genera, thus expanding our current knowledge of human bacterial lineage by 281%. The Authors also performed an in-depth functional characterization of

2,505 human gut species, 1,952 UMGS and 553 isolates from the Human Gut Reference (HGR) database, i.e. gut-specific species from the HR database (8). Interestingly, UMGS resulted depleted in genes involved in antioxidant activities and redox functions, being conversely enriched in iron-sulfur and ion binding genes. Thus, the Authors concluded that the recovered UMGS corresponded to strict anaerobes, with a distinctive metabolic propensity, well adapted to specific niches of the gastrointestinal tract with particularly low oxygen tension and high iron concentration.

Several research projects have been carried out with the specific purpose to define a 'minimal genome' as a model for understanding the basic functions of life (10). This resulted in the identification of a set of core functionalities necessary for autonomous life, as a universal minimal gene set represented in all living systems (11). In order to explore the efficacy of genome resolved metagenomics in providing comprehensive biological information on the uncultured members of the human microbiome, here we wondered if UMGS, which now remain bioinformatic entities, possesses the minimal set of core genes necessary – even if not sufficient – for life. To this aim, two publicly available minimal genomes were used as reference to generate a Core gene set of Minimal Functions (CMF), apparently necessary – but not sufficient – for life. Then we attempted at a screening of both UMGS and isolated NCBI genomes for the presence of genes included in CMF, showing that UMGS were generally depleted in essential functionalities necessary for autonomous life.

Results

The aim of the present study was to provide a first screening of UMGS and isolates genomes for a minimal subset of genetic functions (CMF) necessary – but not sufficient - to sustain bacterial life. In order to generate the CMF, two publicly available minimal genomes were downloaded from NCBI website: JCVI-syn 3.0 genome generated by Hutchison et al. (11) and *C. Eth-2.0* genome generated by Venetz et al. (12). The two genomes were annotated and only the genes assigned with certainty and present in both genomes were retained and used as a reference set for the CMF. The CMF mostly includes genes involved in genetic information processing and cytosolic metabolism (Additional file 1). In particular, of the 190 genes included in CMF (Additional file 2), 143 were assigned by KEGG orthology to the genetic information processing pathways, with the functions involved in translation highly represented, including 115 genes among which 44 encode for ribosomal subunits, 20 for aminoacids-tRNA ligases, and 24 for tRNA. Replication and repair is another group of functions highly represented in the CMF list, including 16 genes encoding for DNA polymerases, gyrases, and topoisomerases among others. Conversely, 35 out of 190 genes are devoted to metabolic functions, including especially carbohydrate metabolic pathways (e.g. glycolysis and gluconeogenesis, galactose metabolism, starch and sucrose metabolism, etc), energy metabolism (including all subunits of ATP synthase), and metabolism of nucleotides. Only two genes included in CMF are exclusively devoted to environmental information processes, and other two to cellular processes. However, 7 of the 190 genes showed multiple functionalities according to their KEGG orthology; for instance, Enolase is involved in metabolism, genetic information processes and environmental information processes, Phosphoglycerate kinase is involved in both metabolism and

environmental information process, and two Protein translocase subunits (SecA and SecY) are involved in genetic information processes, environmental information processes and cellular processes.

Next, we scanned both UMGS and isolated NCBI genomes for the presence of genes included in CMF. To this aim,a total of 800 genomes were randomly selected and downloaded. In particular, 400 genomes of isolated species were randomly obtained from NCBI, covering a wide array of bacterial species, and 400 were selected from the UMGS in the database of Almeida et al. (8). For more information about the genomes included in this study, and the species included in the 400 randomly selected NCBI genomes, see Additional file 3. Each genome set was annotated, and for both the NCBI and UMGS genomes, the presence or the absence of each gene included in CMF was computed, generating a binary matrix of CMF presence/absence profiles. For each tested genome, the percentages of adherence to the CMF and the absolute amounts of missing entries were also computed. Our analysis revealed that the NCBI and the UMGS genomes are characterized by a different percentage of representativeness of the CMF ($P < 0.001$, Kruskall-Wallis test), with the NCBI genomes showing a higher average representativeness value and a lower standard deviation when compared to UMGS (93.2 % ± 2.9 and 78.2 % ± 11.5 SD for NCBI and UMGS genomes, respectively) (Figure 1A). In Figure 1B the overall profile of the missing CMF in NCBI and UMGS genomes is reported. The CMF was found generally less represented in UMGS, with a total of 45 genes lacking in more than 200 analyzed genomes, with respect to the NCBI isolates. Interestingly, among the missing genes, the 16S rRNA gene, fundamental for bacterial life, has not been retrieved in 237 out of 400 UMGS. Clustering analysis and PCA of the presence/absence profiles of CMF genes in NCBI and UMGS genomes showed a segregation between the two groups of genomes (Figure 2). Clustering analysis cuts the genomes batch in two parts, neatly separating the two types of genomes ($P < 0.001$, Fisher's exact test), with UMGS grouped on the left side of the heatmap (Figure 1A). In the same graphics, it is possible to notice how UMGS genomes systematically lack in more genes when compared to NCBI genomes (87.1 ± 89.0 and 27.1 ± 54.2 genomes missing for a single CMF gene in UMGS and NCBI set, respectively). Finally, the PCA analysis carried out using the Euclidean metric showed a separation of the genomes in the two-dimensional plan ($P < 0.001$, permutation test with pseudo-F ratio), with NCBI genomes less disperse if compared to UMGS, indicating a greater homogeneity in the representativeness of the CMF genes inside the NCBI group (Figure 1A).

Conclusions

The present report provide a first attempt at screening both newly-proposed “uncultured metagenomic species” and isolated genomes for the presence of a minimal set of core functionalities necessary for life. Our results showed that UMGS were substantially depleted in such essential genetic functions, including the 16S rRNA gene, which is missing in more than half of the analyzed UMGS. Even if a median completeness of up to 96.5% has been reported for UMGS (8), our data suggest that a relevant fraction of these genomes is missing genes encoding for essential functionalities to support life, thus raising concerns about the effective degree of their completeness. The recurrence and amount of missed functionalities in at least some of the UMGS suggest caution when interpreting their metabolic and ecological propensity on the basis on their peculiar profile of gene relative abundance. Our analysis also

points out the need to confirm the recovery of UMGS on metagenomic datasets obtained using third generation sequencing platforms providing longer reads, which have shown to aid genome completeness in *de novo* assembly and preserve more genomic information useful for species-level taxonomic assignment, such as operon structures (13).

Methods

The minimal genomes JCVI-syn 3.0 (11) and *C. Eth-2.0* (12) were downloaded from NCBI website and annotated using prokka 1.13.3 (14) in a Unix CentOS environment. Genes assigned with certainty were retained for of CMF. A total of 400 NCBI and 400 UMGS genomes have been randomly downloaded from the “Assembly” page of NCBI (the parameters “Complete” and “Representative” were selected) and from the European Nucleotide Archive under study ID PRJEB26432, respectively, and annotated as reported above. For each, the presence/absence profyle of CMF was obtained. Clustering of the CMF presence/absence profiles of NCBI and UMGS genomes was performed in R studio (version 1.2.1355 - R version 3.5.1 (15)), using the binary distance and the Ward's minimal variance clustering method (packages ‘stats’ V3.6.0 (15) and ‘gplot’ V3.0.1.1 (16)). Finally, the Euclidean distances between the CMF presence/absence profiles of UMGS and NCBI genomes were calculated and a multivariate analysis was carried out using the vegan package (V2.5-5 (17)). The separation between the NCBI genomes and the UMGS in the two-dimensional space was verified using a permutation test with pseudo-F ratio ('adonis' function of the ‘vegan’ package).

Abbreviations

CMF: Core gene set of Minimal Functions

GM: Gut Microbiota

HGR: Human Gut Reference

HR: Human-specific Reference

MAGs: Metagenome Assembled Genomes

PCA: Principal Components Analysis

UMGS: Unknown MetaGenomic Species

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Availability of data and material

The datasets analyzed during the current study are available in the ENA repository, https://www.ebi.ac.uk/ena/data/view/PRJEB26432&portal=wgs_set, and in the NCBI repository, https://www.ncbi.nlm.nih.gov/assembly/?term.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

MS and MC conceived the concept. MS developed the different scripts and performed the bioinformatics and statistical analysis. MS, MC, EB, and SR wrote the manuscript. PB, ST and EB revised and edited the draft. All authors discussed the results and have read and approved the manuscript.

Acknowledgements

Not applicable.

References

1. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. Nature. 2019;569:641-8.
2. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe. 2019;25:656-67.
3. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, et al. Structural variation in the gut microbiome associates with host health. Nature. 2019;568:43-8.
4. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol. 2019;4:693-700.

5. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 2016;7:13219.
6. Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol. 2018;3:804-13.
7. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533-42.
8. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568:499-504.
9. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019;37:186-92.
10. Martínez-García E, de Lorenzo V. The quest for the minimal bacterial genome. Curr Opin Biotechnol. 2016;42:216-24.
11. Hutchison CA 3rd, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351:aad6253.
12. Venetz JE, Del Medico L, Wölfle A, Schächle P, Bucher Y, Appert D, et al. Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. Proc Natl Acad Sci U S A. 2019;116:8070-9.
13. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience. 2019;8.
14. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068-9.
15. R Core Team. R: A language and environment for statistical computing. 2017. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/
16. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1.1. https://CRAN.R-project.org/package=gplots.
17. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.5-5. https://CRAN.R-project.org/package=vegan.

Figures

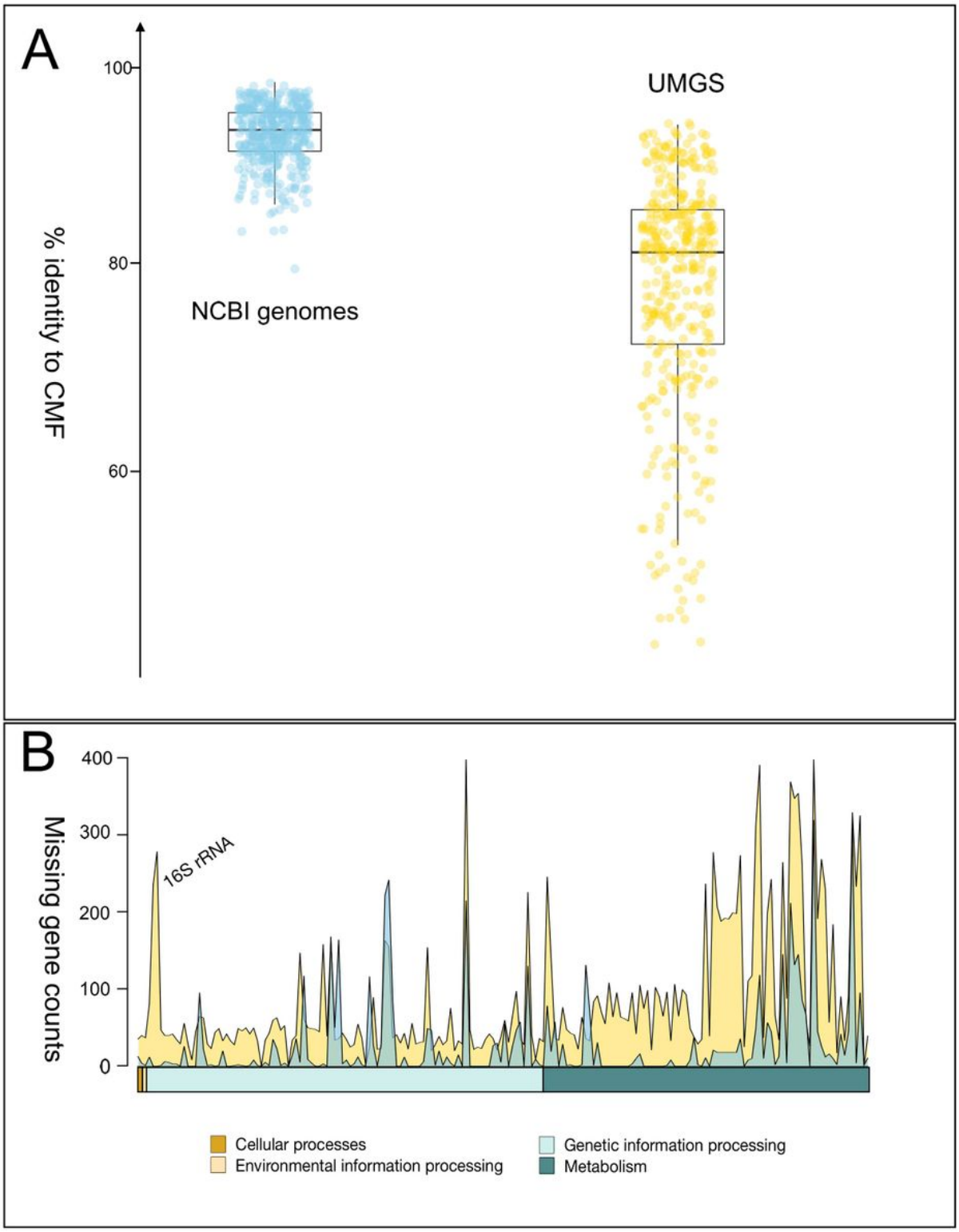

Figure 1

A) Percentage of genes in NCBI (skyblue) and UMGS (gold) genomes that were included in CMF. Distribution of genomes in each group is superimposed to the boxplot representation, with median value indicated by the line inside the box. NCBI genomes show a significantly greater adherence to CMF ($P < 0.001$, Wilcoxon test). B) Superimposed distribution of missing CMF genes in NCBI (skyblue) and UMGS

(gold) genomes. For each gene included in CMF, the number of genomes lacking the correspondent function is plotted. Genes in CMF are clustered according to the functional classes, as in Additional file 1.

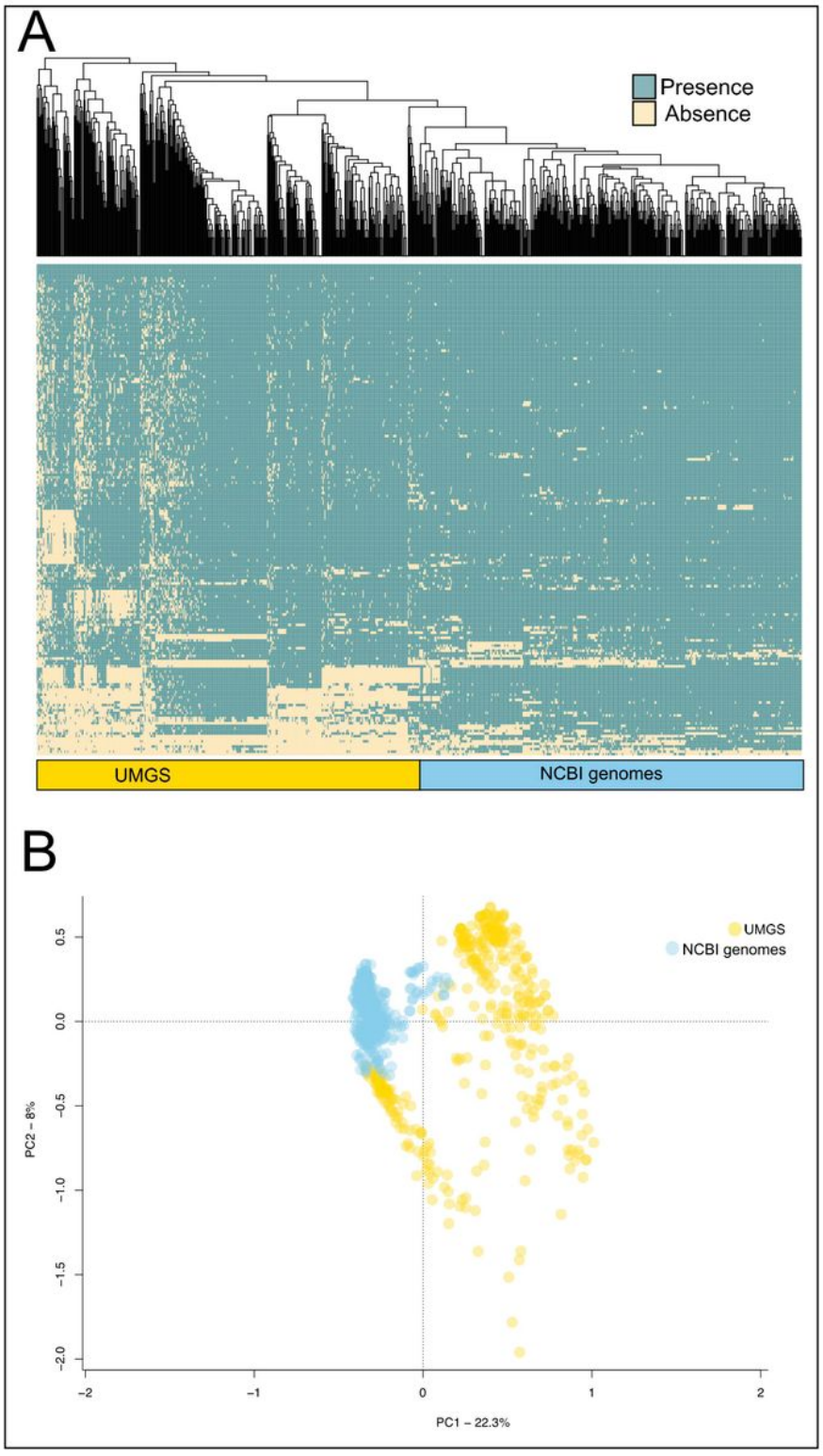

Figure 2

A) Genomes clustering based on the presence/absence profile of CMF genes. The two generated clusters are highlighted by gold and skyblue underlying vectors for UMGS and NCBI genomes, respectively. The separation between the two groups is statistically significant ($P < 0.001$, Fisher's exact test). B) PCA

based on Euclidean distances showing a significant separation between NCBI (skyblue) and UMGS (gold) genomes according to the presence/absence profile of CMF genes ($P < 0.001$, permutation test with pseudo-F ratio).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.pdf
- Additionalfile3.pdf