Tian and Imanian *Microbiome (2023)11:247*
https://doi.org/10.1186/s40168-023-01705-9

Microbiome

SOFTWARE

Open Access

VBCG: 20 validated bacterial core genes for phylogenomic analysis with high fidelity and resolution

Renmao Tian[1] and Behzad Imanian[1,2*]

Abstract

Background Phylogenomic analysis has become an inseparable part of studies of bacterial diversity and evolution, and many different bacterial core genes have been collated and used for phylogenomic tree reconstruction. However, these genes have been selected based on their presence and single-copy ratio in all bacterial genomes, leaving out the gene's 'phylogenetic fidelity' unexamined.

Results From 30,522 complete genomes covering 11,262 species, we examined 148 bacterial core genes that have been previously used for phylogenomic analysis. In addition to the gene presence and single-copy rations, we evaluated the gene's phylogenetic fidelity by comparing each gene's phylogeny with its corresponding *16S rRNA* gene tree. Out of the 148 bacterial genes, 20 validated bacterial core genes (VBCG) were selected as the core gene set with the highest bacterial phylogenetic fidelity. Compared to the larger gene set, the 20-gene core set resulted in more species having all genes present and fewer species with missing data, thereby enhancing the accuracy of phylogenomic analysis. Using *Escherichia* coli strains as examples of prominent bacterial foodborne pathogens, we demonstrated that the 20 VBCG produced phylogenies with higher fidelity and resolution at species and strain levels while *16S rRNA* gene tree alone could not.

Conclusion The 20 validated core gene set improves the fidelity and speed of phylogenomic analysis. Among other uses, this tool improves our ability to explore the evolution, typing and tracking of bacterial strains, such as human pathogens. We have developed a Python pipeline and a desktop graphic app (available on GitHub) for users to perform phylogenomic analysis with high fidelity and resolution.

Keywords Phylogenomics, Bacterial core genes, Phylogenetic tree, Pathogen typing

Background

Bacterial pangenomes consist of the core genome, a set of genes present in all genomes within a clade, and the accessory genome, a set of genes found in one or more but not all members of the clade [1]. While the core genome of a clade can be, and often is, different from that of another clade, it invariably contains highly conserved genes essential for the vital cellular functions of the members of that clade.

The core genome reflects the vertical accumulation of mutations on the conserved set of genes, and thus it can provide phylogenetic signals for assigning bacterial strains to their corresponding populations [2]. Over the past decade, there has been a significant increase in research on bacterial core genes and their applications in phylogenomic analysis [3–6]. Bacterial core genes are

*Correspondence:
Behzad Imanian
bimanian@iit.edu
[1] Institute for Food Safety and Health, Illinois Institute of Technology, Bedford Park, IL 60501, USA
[2] Food Science and Nutrition Department, Illinois Institute of Technology, Chicago, IL 60616, USA

widely used for phylogenomic analysis due to a) their high level of conservation; and b) their widespread distribution among bacterial genomes. With the increased availability of massive amounts of genomic data and improved phylogenetic methods in recent years, phylogenomic analysis has become a popular approach to study bacterial evolution, diversity and ecology. Phylogenomic analysis using core genes [7] has also been shown to be more accurate and robust than the traditional phylogenetic methods that use a single gene (e.g. *16S rRNA*) or multiple genes (e.g. MLST).

With different number of marker genes, often within a core gene set, different teams of researchers have developed tools and pipelines to simplify and streamline core gene phylogenomic analysis. In 2008, Wu et al. used 31 predefined marker genes to build a phylogenomic pipeline named AMPHORA [8, 9]. These 31 marker genes included those involved in glycolysis, DNA replication, translation etc. In 2011, Creevey et al. identified 40 bacterial core genes from 191 species [10] and indicated that they had constant rate of horizontal transfer. In 2012, Dupont et al. investigated metagenome-assembled genomes (MAGs) of SAR86 (an abundant marine bacterial lineage) and estimated the genome completeness and contamination [11] using 107 single-copy genes that were single-copy in > 95% of bacterial genomes in the Comprehensive Microbial Resource database. In 2013, Wu et al. identified 114 bacterial core genes and multiple lineage-specific core gene sets (100 s – 1000 s) from 666 genomes by examining the genes' universality, evenness and uniqueness [12]. Later in 2016, Markus et al. developed a bioinformatic pipeline called bcgTree that used 107 essential single-copy core genes, identified by Dupont et al., as mentioned above, to reconstruct their phylogenetic history [6]. This pipeline automatically extracted these genes using hidden Markov models and performed a partitioned maximum-likelihood analysis. In 2018, another group presented an up-to-date bacterial core gene set, named UBCG, which consisted of single-copy, homologous genes present in most known bacterial species [4]. The 92 UBCG gene set was selected from 1,429 species covering 28 phyla with the criteria of 95% presence ratio and 95% single-copy ratio. The method was successfully used to infer phylogenomic relationships in *Escherichia* and related taxa, and it could be used at any taxonomic level for Bacteria. In 2021, the same group updated their bacterial core genes with more extensive genome data to UBCG2 [5], which included 81 genes from 3,508 species spanning 43 phyla with the same criteria of 95% presence ratio and 95% single-copy ratio.

All the previous studies have utilized the gene presence and single-copy ratios to screen for the conserved, single-copy genes to build their phylogenies. However, this has led to an obvious shortcoming: the selected genes were not examined for their fidelity in reconstructing the phylogenetic tree, and thus they may carry incongruent evolutionary signals. A core genome phylogeny is reconstructed, like other multiple-gene trees, based on concatenated sequences of a set of selected genes. The inclusion of a gene, whose tree shows discordance with the phylogenies of the others genes in the gene set, lowers the evolutionary signal and thus the accuracy of topology and resolution of the resulting phylogeny. We, thus, propose that in order to select the genes for a core genome to reconstruct an accurate phylogeny, in addition to high presence and single-copy ratios, the genes should be examined for phylogenetic fidelity, a measure of congruity of the phylogenies of the genes within a gene set, as well. Here, we used the 16 s rRNA gene trees to evaluate and compare the phylogenetic fidelity of the candidate core genes and identified 20 high fidelity genes for bacterial phylogenomic analysis. We have developed a pipeline, VBCG, which uses the core gene set to build phylogenomic trees automatically with input of genomic sequences.

Implementation

Test genomes and candidate core genes

Genome information of prokaryotes was accessed Through the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/) on Oct 11, 2022. In total, 30,552 finished bacterial genomes covering 11,262 species were selected and the protein sequences and rRNA genes were downloaded using a custom Python script with multiple processing. The rRNA files were then filtered to remove those with multiple *16S rRNA* genes that were less than 99% identical. For those with multiple copies that were > 99% identical, the first copy was used as a representative. The *16S rRNA* genes were then clustered using CD-HIT (version 4.8.1) [13] with a similarity threshold of 0.99 and an alignment coverage of 0.9 (-c 0.99 -G 0 -aL 0.9). The representative sequences (5506 sequences) were then randomly divided into 100 groups for further analysis.

We used a total of 148 genes collected from UBCG [4], UBCG2 [5] bac120 [14] and bcgTree [6] as candidate genes, which have been shown to have high presence and single-copy ratios in bacterial genomes. To acquire the HMM model files, PGAP HMM files and a PFAM file were downloaded from an NCBI website (https://ftp.ncbi.nlm.nih.gov/hmm/current/) and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) website (https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/), respectively, and

the corresponding HMM models were retrieved using a custom Python script.

To annotate the candidate core genes, hmmscan of the package HMMER [15, 16] was used to search all the protein sequences of the 5506 representative genomes against the acquired HMM models (with the trusted score cutoffs, –cut_tc), in parallel using Python package multiprocessing. The output file was parsed to assign the core genes to the proteins.

Ubiquity and uniqueness of the core genes

The ubiquity and uniqueness of the core genes were evaluated by calculating the presence ratio and single-copy ratio of the core genes in the 5506 representative genomes. The presence ratio was defined by the number of genomes with each core gene divided by the total number of genomes. The single-copy ratio was defined by the number of genomes with only one copy of each core gene divided by total number of genomes. Only the core genes with both presence ratio and single-copy ratio > 95%, respectively, were used for further analysis.

Fidelity test for the candidate core genes

For each candidate core gene, all the proteins were divided into 50 groups corresponding to the grouping of *16S rRNA* genes. For each group, a phylogenetic tree was reconstructed for the *16S rRNA* gene and the core gene, respectively. The sequences were first aligned using MUSCLE (version 3.8.1551) [17]. The multiple sequence alignment was then trimmed to remove terminal gaps. The trimmed alignment was then filtered to select conserved blocks using Gblocks (version 0.91b) [18] with the minimum length of a block as three (-b4 = 3) and the maximum number of allowed gap positions as half (-b5 = h). The resulted alignment was then converted into format Phylip and was then fed to FastTree (version 2.1.10) [19, 20] for phylogenetic tree reconstruction. The models gtr + gamma and lg + gamma were used for the *16S rRNA* gene and core gene protein sequences, respectively, with a bootstrap number of 100. The output trees in Newick format of *16S rRNA* gene and core genes of each group were compared using Dendropy (version 4.5.2) [21] calculating the Robinson Foulds (RF) distance [22]. The 20 core genes with the highest fidelity, also ensuring that > 80% of the genomes include the whole gene set, were chosen as the validated bacterial core gene (VBCG) set.

Comparison of multiple gene sets in fidelity

The 50 groups of genomes were used to compare the fidelity of the gene set VBCG with other gene sets. For each group, the core genes were aligned using MUSCLE and the terminal gaps were removed from the multiple alignments. Gblocks was then used to select conserved blocks. The processed alignments of the core genes were then concatenated, removing any taxa with > 1 gene missing. The concatenated alignments were converted into Phylip files and fed to FastTree for phylogenetic tree reconstruction. The trees were then compared to the corresponding *16S rRNA* gene trees by calculating the RF distances.

Results

Ribosomal protein-encoding genes have high ubiquity and uniqueness in the Kingdom Bacteria

In total 30,522 rRNA gene and their corresponding protein sequences were downloaded from RefSeq database. After filtering those with multiple divergent copies, 29,781 sequences of *16S rRNA* genes were remained. After clustering the sequences at 99%, 5,506 representative genes were observed, which were then randomly divided into 50 groups (Fig. 1).

The presence and single-copy ratios of the 148 core genes in the 5506 representative genomes were calculated (Fig. 2A). Interestingly, the 47 ribosomal proteins had much higher ubiquity (median presence ratios: 98.9%) and uniqueness (median single-copy ratios: 98.7%) than those of the other core genes (95.7% and 94.2%). Among the 50 proteins with top presence ratios, 39 were ribosomal proteins. We then selected the core genes (Fig. 2B, Table S1) with both presence and single-copy ratios > 95% for the fidelity and resolution tests (85 genes).

The validated core gene tree showed the highest concordance with the 16S rRNA gene phylogeny

The 5,506 representative genes were randomly divided into 50 groups (~ 110 taxa in each group, Fig. 1). The fidelity test compared the trees of the 50 groups for each of the 85 core genes with the corresponding *16S rRNA* gene trees based on Robinson-Foulds (RF) distance. The tests ranked the fidelity for each of the core genes (Fig. 3A). The core gene phylogenies with the closest and farthest distances to their corresponding *16S rRNA* gene trees ranged from 115.9 ± 14.8 RF and 173.8 ± 9.1 RF, respectively, with the maximum being 50% higher the minimum. We then chose the top 20 genes with the highest fidelity (with lowest RF distances) as our candidate validated bacterial core gene (VBCG) set (Table S2). A close examination of these 20 genes revealed that they were mostly involved in transcription (e.g. RNA polymerase) and translation (e.g. ribosomal proteins, translation initiation, elongation and termination).

We then compared the phylogenetic fidelity of the top 20 gene (VBCG) set and that of all the 85 genes, by concatenating the genes and calculating the RF distances of the resulting trees from the corresponding *16S rRNA*

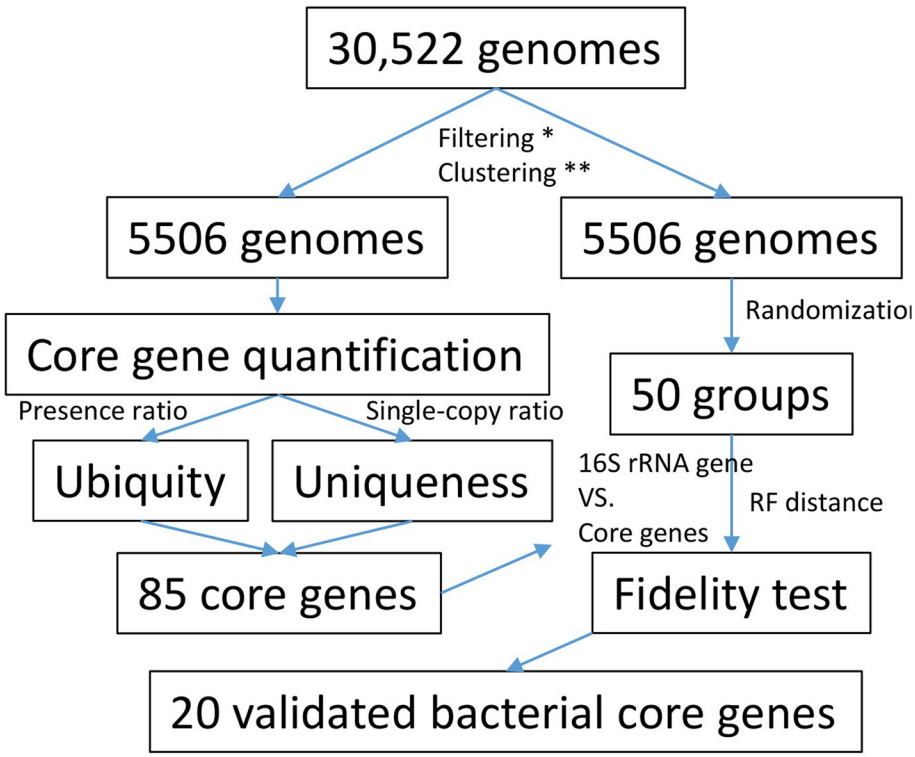

Fig. 1 The workflow of the data analysis in this study. The filtering step indicated by an asterisk removes genomes with > 2 copies of *16S rRNA* genes that are < 99% identical. The clustering step indicated by two asterisks clusters *16S rRNA* genes using a cutoff of 99%

gene trees of the 50 groups of genomes. As a result, the VBCG gene set had significantly lower RF distance from the corresponding *16S rRNA* gene trees (Paired Samples T test, $p < 1e\text{-}5$, Fig. 3B) than those of the 85 gene set, which was 20% higher.

The VBCG gene set resulted in lower missing data rate in the tree compared to that of all genes

The rationale behind selecting the top 20 genes is that these genes yield an acceptable missing data rate in the species for the phylogenetic tree. Because all the core genes have a presence ratio between 95 and 100%, only certain taxa will have all the genes of a given gene set. An advantage of using the 20-gene core set is its higher species data completeness. When comparing the 20-gene and 85-gene sets, 80.7% of species possess all 20 core genes, while only 19.3% have at least one missing gene (Fig. 3B). In contrast, for the larger gene set, 43% of species exhibit missing data. Given that increased missing data has been shown to reduce the accuracy of phylogenomic trees [23, 24], the 20-gene set offers enhanced accuracy in phylogenomic analysis due to its greater gene data presence.

The VBCG gene set phylogeny had better resolution than that of the 16S rRNA gene

We compared the discriminative capability of VBCG set and the *16S rRNA* gene in tree reconstruction on the Genus *Escherichia* with an outgroup species from the genus *Salmonella*, both having member species that are commonly implicated in foodborne disease outbreaks. Our results showed that the *16S rRNA* gene tree had polytomies and could hardly discriminate between the species of *Eschericha* mainly because the *16 s rRNA* genes in the main group had a very high sequence similarity (> 99.0% pairwise identities) (Fig. 4A). The *16 s rRNA* tree also suffered from misclassification of the *E. coli* Str. IMT2125, which was mistakenly positioned outside the *Eschericha* genus. Various strains of the species *E. coli,* and *E. fergusonii* grouped together and were not well separated by species. Some of them are at the same branch, implying that they had not diverged. However, the VBCG tree showed a well-defined grouping by species, with different species nicely separated in different clades. The *E. coli* Str. IMT2125 was correctly positioned inside the genus. Even within each species clade, the strains were well separated with different branch lengths. In the VBCG tree, *E. coli* Str.

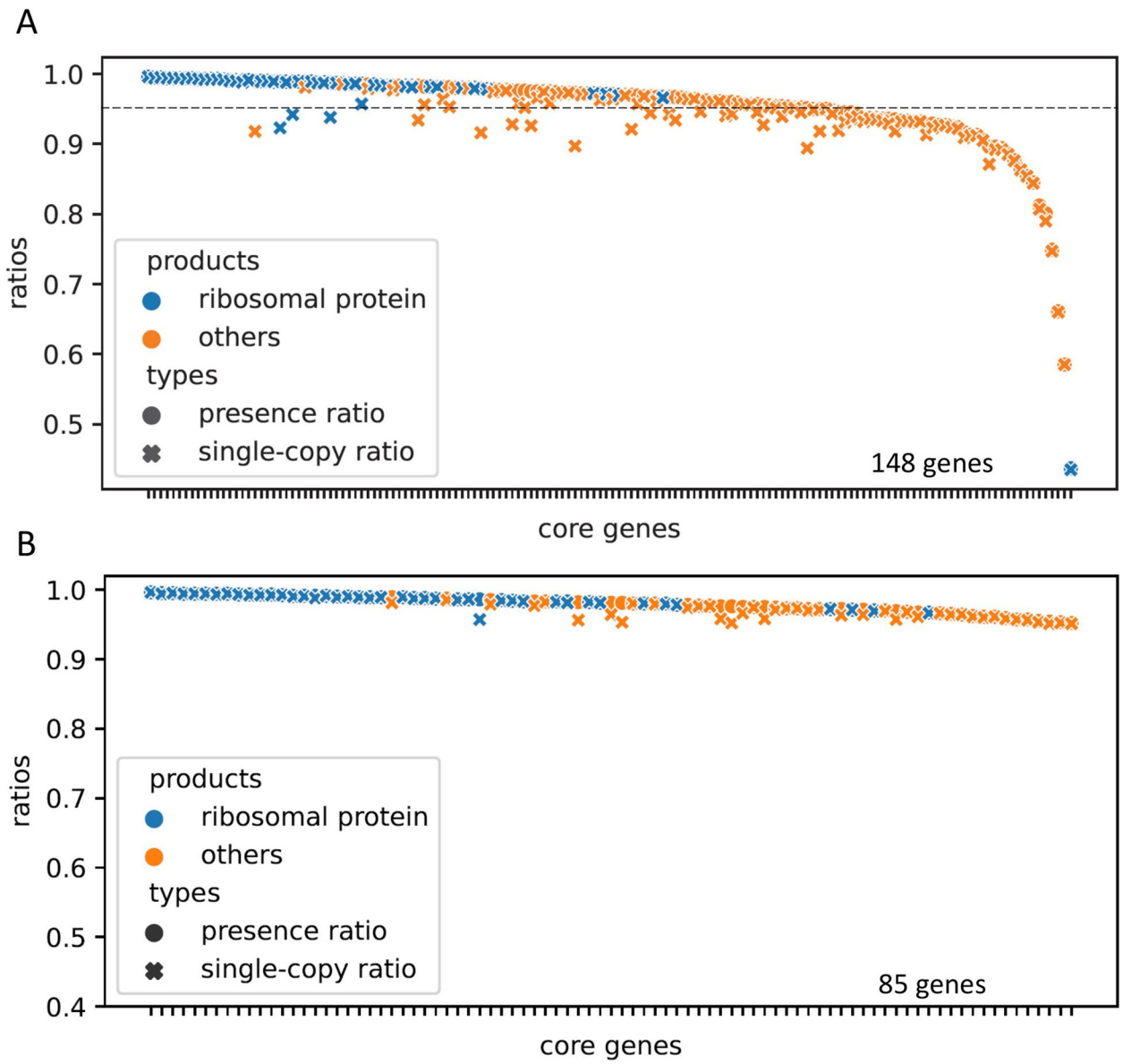

Fig. 2 The presence and single-copy ratios of the candidate core genes in the Kingdom Bacteria. In the total core genes (**A**), the 47 ribosomal proteins had much higher ubiquity (median presence ratios: 98.9%) and uniqueness (median single-copy ratios: 98.7%) than those of the other core genes (95.7% and 94.2%). The dashed line demarcates the 95% ratio. In total 85 core genes (**B**) with both presence and single-copy ratios > 95% were retained for the fidelity and resolution tests

2020CK-00188 had branched within the *E. fergusonii* clade. While we cannot completely dismiss that this might be a mistake within the VBCG tree that begs further investigations, the likely explanation is that *E. coli* Str. 2020CK-00188 has been misclassified and it is not an *E. coli* species at all. Overall, our results indicated that a phylogeny based on the 20 VBCG genes resulted in higher resolutions (even at the strain level) compared with the corresponding tree based on *16 s rRNA* genes.

A pipeline and a desktop app for reconstructing phylogenomic trees automatically

We have built a pipeline to reconstruct phylogenomic trees with the input of assembled genomic sequences (Figure S1). The pipeline first predicts the gene and protein sequences of the input genomes using Prodigal. The protein sequences are then used to identify the 20 VBCG genes with HMMER. Subsequently, the resulting VBCG genes of all the genomes are retrieved and each gene

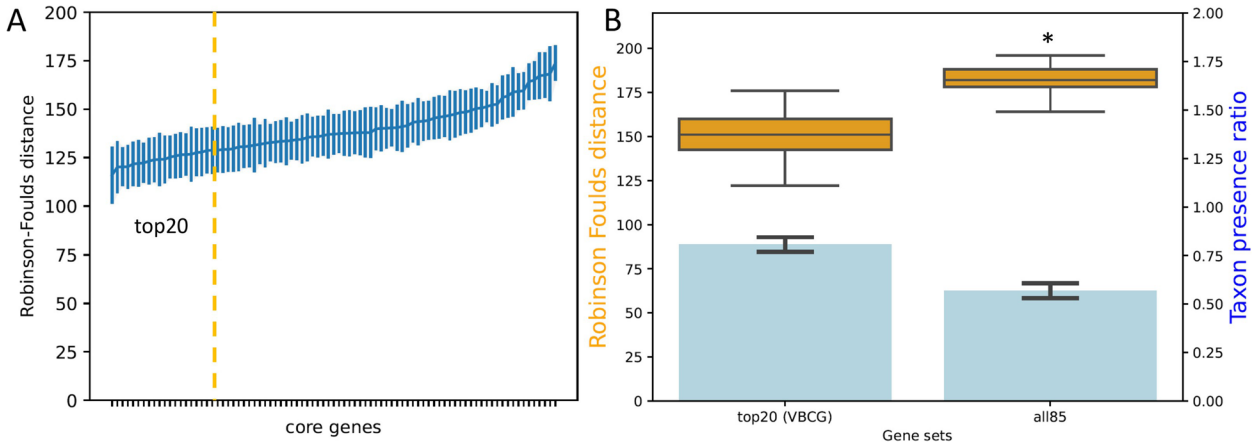

Fig. 3 The fidelity test of the validated core genes. The fidelity tests (**A**): The RF distances of the core genes trees from the corresponding *16S rRNA* gene trees. Each data point indicates the means and standard deviations of 50 trees (110 taxa each). The yellow dashed line indicates the top 20 genes with the highest fidelities in the left. The fidelity test and comparisons for phylogenies of multiple gene sets (**B**): The RF distances (brown boxes) of the trees based on the concatenated genes (top 20-gene set and all 85 gene set) from the corresponding *16S rRNA* gene trees of the 50 groups of genomes. The asterisk indicate significant difference from the top 20-gene set. The blue bars and right side y-axis show the taxon presence ratio of each gene set, which indicates the percentage of taxa with all members of the gene sets

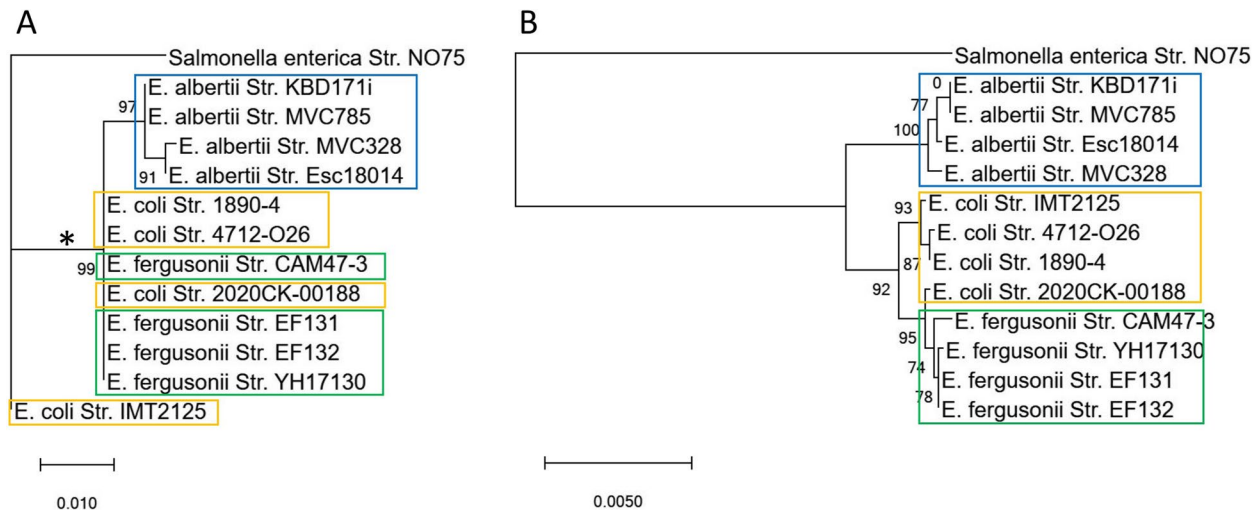

Fig. 4 Comparison between *16S rRNA* gene tree (**A**) and the VBCG tree (**B**) of concatenated core genes. The tree of the VBCG genes were reconstructed using the concatenated 20 genes of each genome. The asterisk in the *16S rRNA* gene tree indicates the main group of *Escherichia* species with pairwise identities > 99.0%

from all taxa is placed in its own single file. The genes are aligned using Muscle and the terminal gaps are removed from the multiple alignments. Gblocks is then used to select conserved blocks. The processed alignments of the core genes are then concatenated, removing any taxa with a certain number of missing genes (adjustable from one to six with four as the default). Any missing characters are replaced with question marks. The concatenated alignments are converted into Phylip files and fed into FastTree or RAxML for phylogenetic tree reconstruction.

The pipeline is written in Python 3.9, with the following dependencies: Bio > = 1.5.3, Muscle = = 3.8, Pandas, Prodigal and HMMER. The command line-based pipeline can be run in Linux operating systems. We have also built a desktop graphic user interface (GUI) app (Figure S2) on Windows for the users who have no access to a Linux system and a high-performance computer (HPC).

Discussion

With the increasing availability of genomic sequences in recent years, phylogenomic analysis has become a powerful critical tool for studying bacterial diversity and evolution. In the past, many marker genes and various core gene sets have been used for phylogenetic tree reconstruction [4, 5, 8, 9, 25, 26]. However, these genes have been selected based only on their presence and single-copy ratios in all bacterial genomes. The evolution of the genes within a genome is not uniform, and different genes evolve at different rates. Thus, single gene phylogenies, even from the same genome, could be quite different. The inclusion of the genes that are evolving at heterogeneous rates can weaken the phylogenetic signal, introduce noise, produce incorrect topologies, and eventually lower the resolution of the tree. In the previous methods and tools, the gene phylogenetic fidelity has not been considered, compared or tested in order to screen and select the highly concordant marker genes to include in a gene set (e.g. core genome set or its subsets) for phylogenomic analysis. Here, we have identified 20 validated bacterial core genes (VBCGs) that exhibit high fidelity and concordance in phylogeny with their corresponding *16S rRNA* gene trees. Moreover, we demonstrated that the VBCGs could produce superior resolution in discriminating species and strains compared to the widely used *16S rRNA* genes or other core gene sets. We have developed a Python pipeline and a desktop graphic app that enable users to perform phylogenomic analysis using the VBCGs with high fidelity and resolution.

In addition to a better resolution, the 20 VBCG also results in lower missing data rate and faster run-time in the phylogenomic analysis. Our 85 candidate bacterial core genes based upon the presence and single-copy ratios filtration are consistent with the UBCG2 [5], which contains 81 core genes using the same ratios for filtration. However, our results have shown that the 85 core gene set without fidelity screening resulted in higher RF distances from the corresponding *16 s rRNA* trees than the 20 VBCG gene set. In terms of taxon coverage, the 85 core genes resulted in the inclusion of only 57% of all the species in the concatenated sequences, whereas the VBCG gene set included 80% of the species with all the 20 genes. Thus, VBCG can effectively reduce the missing data due to absence of one or more core genes in the phylogenomic tree reconstruction. In addition, the fewer number of core genes in the VBCG tool lowers the running time significantly compared to those with the larger gene sets.

Phylogenomic analysis has emerged as a powerful tool, for example, in tracking the origins of foodborne disease outbreaks. This technique involves the sequencing of whole genomes of bacterial isolates recovered from infected individuals, foods and related environments. These sequences are then used in computational analyses to reconstruct a phylogenomic tree that reveals the evolutionary relationships between these isolates. One of the main benefits of these analyses is their ability to accurately trace the spread of a particular bacterial strain across different geographical locations and over time. Phylogenomic analyses has become an indispensable tool in the hands of epidemiologists as well. A phylogenomic tree of different clinical isolates, for example, can help researchers identify the common ancestors of the pathogens and determine the direction and timeline of the transmission of the disease and the pathogen between individuals within a population. In this study, we examined the existing phylogenomic analysis tools and their related core genome gene sets, and we added a phylogenomic fidelity criterion in selecting the core genes. We have developed a new pipeline and tool that relies on a new gene set, the VBCG. The phylogenomic analysis based on the VBGC gene set retains higher number of species, achieves higher speed, and higher resolution than those reliant on other core gene sets or the *16S rRNA* genes as demonstrated in the phylogenomic trees reconstructed for the common foodborne pathogens *Escherichia coli*.

Conclusion

In conclusion, our research used a refined selection process that takes into account gene presence, single-copy ratios, and phylogenetic fidelity, and revealed that the 20 validated bacterial core genes (VBCG) provide high phylogenetic fidelity and resolution for phylogenomic analysis, enhancing our understanding of bacterial diversity and evolution. We have developed a streamlined tool capable of producing more accurate and reliable phylogenies, even at the strain level. Our findings highlight the importance of VBCG in the efficient typing and tracking of bacterial pathogens, a valuable tool in pathogenic studies. We have made our Python pipeline and graphic desktop app available on GitHub to ensure accessibility for other researchers in the field.

Availability and requirements

Project name: Validated Bacterial Core Genes (VBCG).

Project home page: https://github.com/tianrenmaogithub/vbcg

Operating system(s): Linux, Windows.

Programming language: Python.

License: GNU GPL v2.0.

Abbreviations

VBCG	Validated Bacterial Core Genes
NCBI	National Center for Biotechnology Information

EMBL-EBI	European Molecular Biology Laboratory's European Bioinformatics Institute
RF	Robinson Foulds
GUI	Graphic User Interface
HPC	High-performance computer

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01705-9.

Additional file 1.

Acknowledgements
The authors acknowledge the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) for the support.

Authors' contributions
B.I. and R.T. conceptualized and designed the project; R.T. conducted the data analysis, code writing and pipeline development. R.T. and B.I. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
This publication is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) (Grant No. 5U19FD005322) as part of an award totaling $3,856,000 with 0% financed with nongovernmental sources. The funding body did not play a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official views of, nor endorsement by, the FDA, HHS, U.S. Government or Illinois Institute of Technology. For more information, please visit https://www.fda.gov/.

Availability of data and materials
The source code the pipeline VBCG can be accessed on GitHub. In addition, there is a graphic user interface (GUI) version of the pipeline for Windows users (https://github.com/tianrenmaogithub/vbcg).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 June 2023 Accepted: 18 October 2023
Published online: 08 November 2023

References

1. Segerman B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol. 2012;2. https://doi.org/10.3389/fcimb.2012.00116. Cited 27 Feb 2023.
2. Chung M, Munro JB, Tettelin H, Dunning Hotopp JC. Using core genome alignments to assign bacterial species. mSystems. 2018;3:e00236-18.
3. Shakya M, Ahmed SA, Davenport KW, Flynn MC, Lo C-C, Chain PSG. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Sci Rep. 2020;10:1723.
4. Na S-I, Kim YO, Yoon S-H, Ha S-M, Baek I, Chun J. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J Microbiol. 2018;56:280–5.
5. Kim J, Na S-I, Kim D, Chun J. UBCG2: Up-to-date bacterial core genes and pipeline for phylogenomic analysis. J Microbiol. 2021;59:609–15.
6. Ankenbrand MJ, Keller A. bcgTree: automatized phylogenetic tree building from bacterial core genomes. Genome. 2016;59:783–91.
7. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends Genet. 2006;22:225–31.
8. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008;9:R151.
9. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28:1033–4.
10. Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PLoS ONE. 2011;6:e22099.
11. Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Alexander Richter R, Valas R, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 2012;6:1186–99.
12. Wu D, Jospin G, Eisen JA. Systematic identification of gene families for use as "Markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE. 2013;8:e77033.
13. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
14. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, 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.
15. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29-37.
16. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.
17. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
18. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
19. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490.
20. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50.
21. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 2010;26:1569–71.
22. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53:131–47.
23. Xi Z, Liu L, Davis CC. The impact of missing data on species tree estimation. Mol Biol Evol. 2016;33:838–60.
24. Wiens JJ. Missing data and the design of phylogenetic analyses. J Biomed Inform. 2006;39:34–42.
25. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol. 2021;6:354–65.
26. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11.

Publisher's Note