scientific reports

OPEN

A revisit to universal single-copy genes in bacterial genomes

Saidi Wang[1,4], Minerva Ventolero[2,4], Haiyan Hu[1,3]✉ & Xiaoman Li[2]✉

Universal single-copy genes (USCGs) are widely used for species classification and taxonomic profiling. Despite many studies on USCGs, our understanding of USCGs in bacterial genomes might be out of date, especially how different the USCGs are in different studies, how well a set of USCGs can distinguish two bacterial species, whether USCGs can separate different strains of a bacterial species, to name a few. To fill the void, we studied USCGs in the most updated complete bacterial genomes. We showed that different USCG sets are quite different while coming from highly similar functional categories. We also found that although USCGs occur once in almost all bacterial genomes, each USCG does occur multiple times in certain genomes. We demonstrated that USCGs are reliable markers to distinguish different species while they cannot distinguish different strains of most bacterial species. Our study sheds new light on the usage and limitations of USCGs, which will facilitate their applications in evolutionary, phylogenomic, and metagenomic studies.

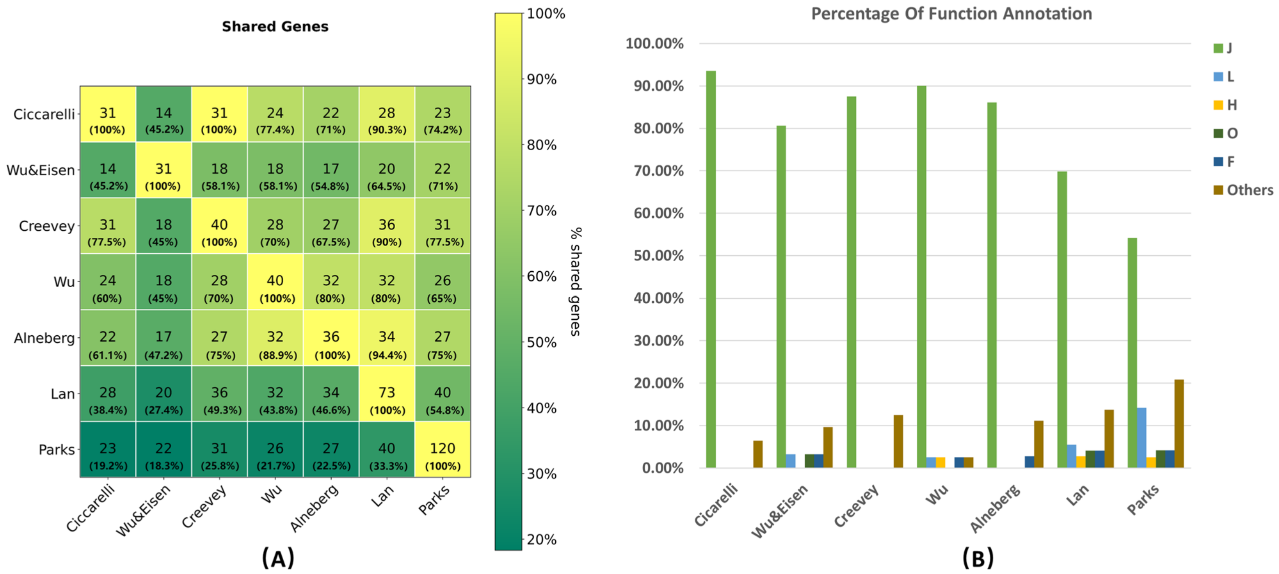
Universal single-copy genes (USCGs) are marker genes that occur once and only once in almost every genome[1]. Because of this property of ubiquitous existence and uniqueness in each genome, USCGs are widely used to study the evolution and classification of species[2–4]. In the past two decades, with the enormous amount of metagenomic data generated, USCGs are also routinely employed for taxonomic profiling of microbial species and completeness evaluation of metagenome-assembled genomes in shotgun metagenomic studies[5–9].

A widely used marker gene is the 16S rRNA gene, which is thought to be single-copy while shown otherwise and is thus not a perfect USCG[10]. 16S rRNA gene has been used as the gold standard in amplicon sequencing and demonstrated its power for taxonomic profiling in various metagenomic studies[4,11–14]. It is the most sequenced gene, with its sequences often the only ones we know about a species. Despite its indisputable value and popularity, the 16S rRNA gene may have multiple copies in an unknown genome[10]. Moreover, its sequences might be too conserved to distinguish certain species and/or correctly measure the species divergence[4]. It is thus natural to consider other USCGs in evolution and metagenomics.

Previous studies have identified USCGs other than the 16S rRNA gene[1–4,9,15–19]. For instance, Ciccarelli et al. obtained 31 USCGs to reconstruct the tree of life across all three domains[1]. Later, Creevey et al. inferred 40 USCGs from the same set of 191 complete genomes[16], which were then used for taxonomic profiling in metagenomics by the tools mOTUs and mOTUs2[5,6]. Wu and Eisen developed the AMPHORA pipeline for automatic phylogenomic analysis of microbial species based on 31 USCGs[4]. These 31 USCGs were combined with 104 archaeal USCGs for phylogenomic studies[19]. Later, the same group developed the phyEco gene sets that comprised 40 USCGs for "all bacteria and archaea" and 114 USCGs for "all bacteria"[3]. Alneberg et al. analyzed all clusters of orthologous groups in 525 genomes and identified 36 USCGs that occurred in > 97% of the 525 genomes with a frequency < 1.03 per genome[15]. Lan et al. discovered 73 USCGs in > 90% of 1897 genomes to classify prokaryotic species[2]. Parks et al. extracted 120 USCGs present in > 90% of genomes and single-copy in > 95% of their present genomes[9]. These sets of USCGs are inferred from different groups of sequenced genomes with different purposes. They are thus different, although they do share a fraction of USCGs.

With multiple USCGs, in addition to studying evolution, classifying and taxonomic profiling of species, several studies attempted to investigate the strain diversity[7,20–23]. For instance, Quince et al. developed tools to de novo reconstruct bacterial strain genomes from shotgun metagenomic reads with the aforementioned 36 USCGs[7,23]. The StrainPhlAn tool infers bacterial strains directly from shotgun reads based on ~ 200 clade-specific marker genes[22]. Nayfach et al. proposed a pipeline for strain profiling in metagenomic datasets with 15 USCGs[24]. Such studies of bacterial strains are important for understanding drug resistance, microbial diversity, and the cure of various complex diseases[25–30].

[1]Department of Computer Science, University of Central Florida, Orlando, FL, USA. [2]Burnett School of Biomedical Science, College of Medicine, University of Central Florida, Orlando, FL, USA. [3]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL, USA. [4]These authors contributed equally: Saidi Wang and Minerva Ventolero. ✉email: haihu@cs.ucf.edu; xiaoman@mail.ucf.edu



	Ciccarelli	Wu&Eisen	Creevey	Wu	Alneberg	Lan	Parks
Ciccarelli	31 (100%)	14 (45.2%)	31 (100%)	24 (77.4%)	22 (71%)	28 (90.3%)	23 (74.2%)
Wu&Eisen	14 (45.2%)	31 (100%)	18 (58.1%)	18 (58.1%)	17 (54.8%)	20 (64.5%)	22 (71%)
Creevey	31 (77.5%)	18 (45%)	40 (100%)	28 (70%)	27 (67.5%)	36 (90%)	31 (77.5%)
Wu	24 (60%)	18 (45%)	28 (70%)	40 (100%)	32 (80%)	32 (80%)	26 (65%)
Alneberg	22 (61.1%)	17 (47.2%)	27 (75%)	32 (88.9%)	36 (100%)	34 (94.4%)	27 (75%)
Lan	28 (38.4%)	20 (27.4%)	36 (49.3%)	32 (43.8%)	34 (46.6%)	73 (100%)	40 (54.8%)
Parks	23 (19.2%)	22 (18.3%)	31 (25.8%)	26 (21.7%)	27 (22.5%)	40 (33.3%)	120 (100%)

Figure 1. USCG sets and functional categories. (**A**) Overlap of the seven USCG sets. (**B**) The percentage of the functional categories of the USCGs in the seven sets. For each set, the percent of USCGs from the functional categories J, L, H, O, F, and others is shown in order.

Despite many studies on USCGs, our understanding of USCGs is somewhat outdated. For instance, how much agreement is there between the USCG sets from different studies? For a set of USCGs, how universal are these USCGs in the ever-increasing number of sequenced genomes? When defining the above sets of USCGs, to our knowledge, only the first two sets of USCGs were selected by requiring their single-copy occurrence in all 191 genomes. None of the USCG sets is tested on the latest set of complete genomes to show their universalism. Moreover, it is also not clear whether certain USCGs can tell two species apart better than others. In addition, it is unknown whether a set of USCGs such as the 40 USCGs from Creevey et al.[16] contain enough variations to distinguish different strains of a bacterial species.

To address these questions, especially the last one, in this study, we compared different sets of USCGs. We systematically studied how similar a USCG is in different species and strains with the latest set of complete bacterial genomes downloaded on October 27, 2021. We found that almost every USCG occurs multiple times in certain genomes, while more than 99.4% of USCGs are single-copy in a given genome. We also observed that USCGs together are good for separating species, while they cannot distinguish strains from the same bacterial species in general. Our study provides a more updated picture of USCGs and their potential applications in evolutionary and metagenomic studies.

Results

USCG sets are different in gene content while similar in gene function.

We compared USCG sets from seven studies[1–4,9,15,16] (Material and Methods, Supplementary Table S1). They were (1). 31 USCGs from Ciccarelli et al.; (2). 31 USCGs from Wu and Eisen; (3). 40 USCGs from Creevey et al.; (4). 40 USCGs from Wu et al. for “all” bacteria and archaea; (5). 36 USCGs from Alneberg et al.; (6). 73 USCGs from Lan et al.; and (7). 120 USCGs from Parks et al. These USCG sets require different universalism across all genomes and uniqueness in individual genomes. For instance, USCGs in the last two sets occurred in only > 90% of the genomes, while that was at least > 97% in the third to the fifth sets.

We found that even the USCG sets from the same research group could be very different (Fig. 1A). For instance, only 18 (77%) of the 31 USCGs from Wu and Eisen were shared with the 40 USCGs from Wu et al. Both sets were inferred by the same research group, with the latter inferred from a much larger number of genomes[3,4]. Their difference thus implied the significant effect of the genomes used to infer these USCGs. On the other hand, the USCG sets from the same research group could be highly consistent as well. For instance, the USCG set from Creevey et al. was a superset of the USCG set from Ciccarelli et al., as the same research group inferred them with the same genomes and a slightly different strategy. However, the additional nine USCGs from Creevey et al. also demonstrated the effect of such different strategies on defining USCGs.

The USCG sets from different research groups differed more (Supplementary Table S2). For instance, only 14 (45%) of the 31 USCGs from Ciccarelli et al. were shared with those from Wu and Eisen. Again, the difference in these two sets corroborated the effect of different strategies and different genomes. When we considered the two sets refined later by the corresponding research groups, the third and fourth sets, 28 (70%) of the 40 USCGs in these two sets were shared. Although the USCGs in these two sets were still drastically different, they shared much more USCGs than their earlier versions, likely because the number of genomes was large enough to choose a more representative set of genomes by Wu et al. Compared with the sets from Alneberg et al., Lan et al., and Parks et al., the USCGs from Creevey et al. was at least comparable with those from Wu et al., if not better. Overall, the USCG set from Creevey et al. is likely as reliable as any other set, if not more reliable, because

	Mean			SD			Minimum			Median			Maximum		
	C	W	A	C	W	A	C	W	A	C	W	A	C	W	A
fetchMG	99.0	98.6	99.0	0.5	2.7	0.8	97.7	83.9	95.9	99.3	99.3	99.3	99.6	99.7	99.6
BLAST	96.1	88.9	94.2	3.0	7.0	4.5	87.1	58.9	76.7	96.7	90.9	96.0	99.7	96.8	99.0
universal	100	99.8	99.9	3.4	9.4	5.3	99.7	93.4	99.0	100	99.9	100	100	100	100
unique	98.5	97.3	98.2	1.0	3.9	2.2	95.4	82.6	89.0	99.0	99.1	99.1	99.5	99.6	99.5

Table 1. Universalism and uniqueness of USCGs. The subcolumns with the name C, W, and A refer to the corresponding percentage for the USCGs from Creevey et al., Wu et al., and Alneberg et al., respectively.

of its stricter requirement of universalism and uniqueness, the similar evolutionary trajectory of these USCGs, and likely more representative genomes (the resulted USCG set has a good overlap with other sets that used much more genomes).

Despite the difference in the number and members of USCGs, the function of the USCGs in each of the above sets was quite consistent (Fig. 1B)[31–35]. By checking the functional annotations at https://www.ncbi.nlm.nih.gov/research/cog, we found that the vast majority of USCGs in every set are annotated with the functional category J (Translation, ribosomal structure, and biogenesis). The remaining USCGs were annotated with other translation and metabolism related categories, such as L (Replication, recombination, and repair), H (Coenzyme transport and metabolism), F (Nucleotide transport and metabolism), O (Posttranslational modification, protein turnover, chaperones), etc. Each category other than J was annotated with much fewer USCGs, from one to a handful of USCGs, compared with several dozen for J. Overall, the USCG sets are enriched with functions related to translation and metabolism, no matter which USCG set is concerned about. For instance, among the 40 USCGs from Wu et al., 35 are involved in the translation process, while the remaining five are related to the cellular metabolic process[3].

USCGs are indeed universal across species and unique in individual genomes.

To our knowledge, the number of genomes used to infer the above USCG sets was no more than 2000 except those from Parks et al. With more than 25,000 complete bacterial genomes at National Center for Biotechnology Information (NCBI), it was unclear whether the USCGs were still universal and unique. We thus studied the occurrence of the USCG sets in the latest set of complete genomes available on October 27, 2021 (Material and Methods). We focused on the sets from Creevey et al., Wu et al., and Alneberg et al. below because they are widely used in metagenomics and more updated than their previous versions (the first and third set). Moreover, they have more stringent criteria for universalism and uniqueness than the sets from Lan et al. and Parks et al.[2,9]. In the following, we presented our study on the above three sets.

We found that the 40 USCGs from Creevey et al. were universally distributed in almost each of the 25,271 genomes (Table 1, Supplementary Table S3, Material and Methods). For every USCG, it occurred in at least 97.7% of the 25,271 complete genomes, with the USCG Signal recognition particle GTPase occurring in the least genomes. For genomes without a copy of a USCG, we applied BLAST with an arbitrary cutoff of 1E-15 to search for this USCG. BLAST could identify a copy with this cutoff in most of these missed genomes. A tiny fraction of the genomes still did not have a USCG copy, likely due to the cutoff, the quality of the genomes, and the imperfectness of the gene retrieval methods used. Despite these limitations, each USCG occurred in at least 99.7% of the genomes. Of note, the number of USCGs occurring in each genome also showed the universal distribution of these USCGs (Supplementary Figure S1A). The mean number of USCGs occurring in each genome was 39.6. In other words, almost each USCG occurred in every genome, indicating the universalism of the USCGs from Creevey et al.

We also observed that each of the 40 USCGS from Creevey et al. indeed had only one copy in almost every genome (Table 1). The mean and median percentage of the 25,271 genomes with only one copy of a USCG were 98.5% and 99.0%, respectively. For individual USCGs, 24 (60%) of the 40 USCGs occurred more than once in $<0.2\%$ of the genomes, while several other USCGs occurred more than once in $>2.0\%$ of the genomes. Interestingly, these USCGs that occurred more than once in $>2\%$ of the genomes were often not identified in other USCG sets. For instance, for the top ten USCGs occurring more than once, only one of them was also identified in the set from Wu et al., suggesting that these USCGs might have occurred multiple times in other studies and were thus filtered by those studies. Overall, the average number of USCGs with only one copy in a genome was 39.4. This average number suggests that almost all USCGs have only one copy in almost all complete bacterial genomes. Note that rarely is a genome with multiple copies of a USCG for multiple USCGs (Supplementary Figure S1B).

We further studied the universalism and uniqueness of the USCG sets from Wu et al. and Alneberg et al. (Supplementary Table S3). Compared with the set from Creevey et al., the other two sets had a similar median (Fig. 2). For instance, the median of the universalism was 99.3% for all three sets, and the corresponding median of the uniqueness was 99.0%, 99.1%, and 99.1%. However, there were more individual USCGs in the other two sets that were not so universal and unique as those from Creevey et al. It is thus evident that the above 40 USCGs from Creevey et al. might be a better USCG set.

USCGs separate different species well, but not strains of most species.

With the USCGs indeed universal and unique in bacterial genomes, we investigated how well they together could distinguish two bac-

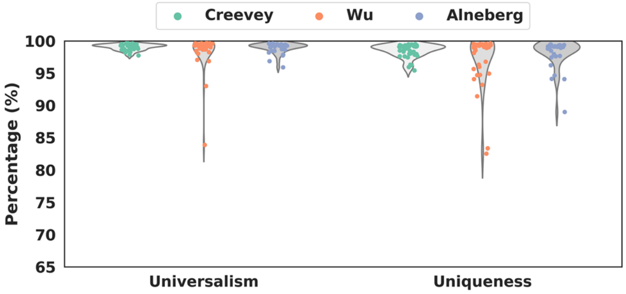


Figure 2. The universalism and uniqueness of the three USCG sets. Universalism means how many percent of genomes has at least a copy of a USCG. Uniqueness refers to the percentage of genomes with only one copy of a USCG.

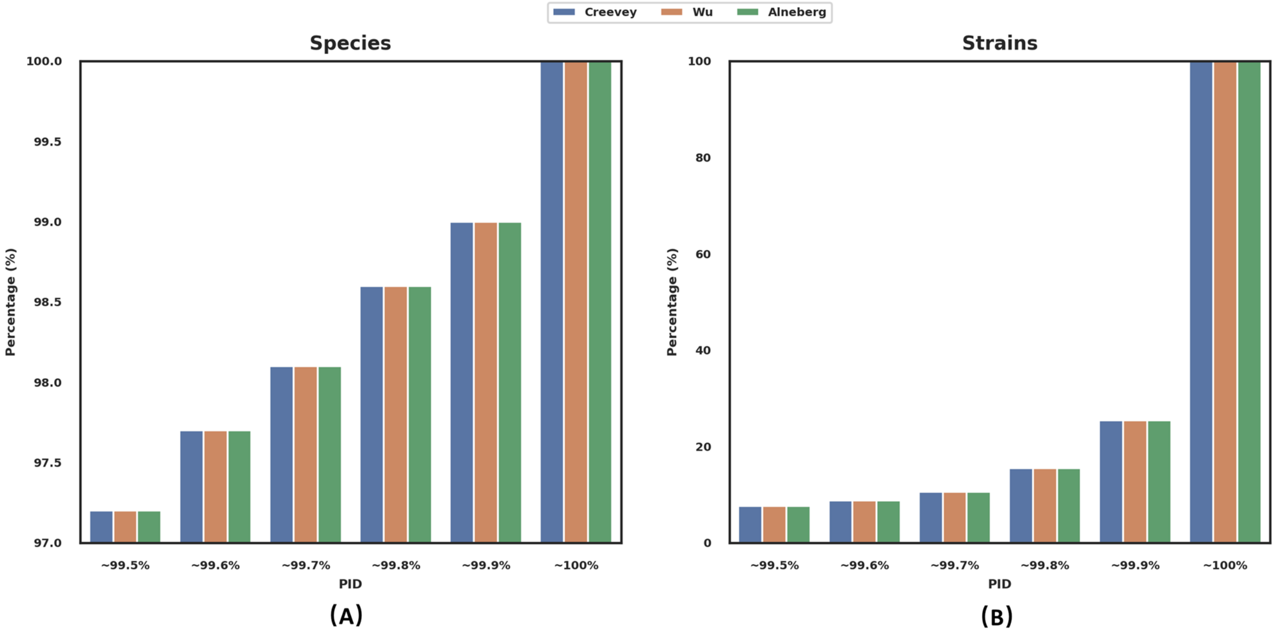


Figure 3. The cumulative distribution of the PID of species pairs and strain pairs. (**A**) Species pairs. (**B**) Strain pairs.

terial species in the same genus and whether they together are enough to set two bacterial strains of the same bacterial species apart. To address these questions, we studied all 440 genera with at least two sequenced species genomes and all 747 bacterial species with at least two sequenced strain genomes in the 25,271 complete genomes (Material and Methods). We found that at least the 40 USCGs from Creevey et al. together were enough to distinguish almost all pairs of species in the same genus. However, they could not separate pairs of strains from most species (Fig. 3).

With the Creevey et al. USCG set, we studied the similarity of 40 pairs of corresponding USCGs in every two species from the same genus for all 440 genera with at least two complete genomes (Supplementary Table S4, Figure S2A). We measured the similarity by the percentage of identity (PID) in the alignment of the 40 pairs of USCGs from a pair of species because the PID determines how similar the shotgun metagenomics reads from these USCGs are. The PID had a mean and median of 90.2% and 90.6%. However, about 2.3% of species pairs had a PID larger than 99.5%, which was the limit that the current computational tools could distinguish two genomes[7,36–40]. Given that the total nucleotide length of the 40 USCGs is about 33,852 base pairs, the 99.5% PID means there are at least 170 varied positions in the 40 USCGs required to separate the two species. Therefore, the 40 USCGs could tell most pairs of species apart, which may leave about 2.3% of the pairs of species indistinguishable.

With the same USCG set, we also studied the similarity of 40 pairs of corresponding USCGs in every two strains from the same species for all 747 species with at least two complete genomes (Fig. 3, Supplementary Table S5, Figure S2B). The PID in a pair of strains had a mean of 99.8% and a median of 99.9%. About 91.2% of the PID scores between pairs of strains from the same species were already larger than 99.5% (Fig. 3). In other words, the 40 USCGs alone may have difficulty in separating more than 91.2% of pairs of bacterial strains.

We also studied how well the USCG set from Wu et al. and from Alneberg et al. separated species and strains (Fig. 3). Compared with the USCG set from Creevey et al., these two sets had a slightly lower performance, especially in terms of separating strains of the same species. Many more pairs of species and strains had close to 100% PID based on the Wu et al. set compared with the other two sets. Overall, the USCG set from Creevey et al. is likely to perform better in distinguishing species and strains.

Discussion

We systematically studied the occurrence of USCGs in the most updated list of complete bacterial genomes. We showed that USCGs were universally distributed and uniquely present in almost every bacterial species. The 40 USCGs from Creevey et al. together could distinguish different species from the same genus and different strains from the same species well. Barely was there one individual USCG that could do so alone.

There are inevitably mistakes in the annotation of the complete genomes. For instance, the MSHR1132 genome was annotated as the genome of a *Staphylococcus aureus* strain previously[41] and is now classified as the genome of a *Staphylococcus argenteus* species in the current NCBI. Such an incorrect annotation may change the similarity of individual species and strain pairs. However, it does not affect our conclusions about universalism, uniqueness, and species and strain pair similarity, because universalism and uniqueness are based on individual genomes, and the species and strain pair similarity are based on the upper bound of the similarities. The incorrect annotation is likely to affect low similar pairs. Moreover, the incorrect number of annotations is too small to affect the conclusions made in this study.

There are many other sets of USCGs not compared here. Wrighton et al. used 30 USCGs to measure the completeness of the reconstructed 49 bacterial genomes[18]. Haroon et al. employed the Amphora2 marker genes, including the 31 USCGs for bacteria from Wu and Eisen and 104 archaeal genes, to measure the completeness of the assembled genomes[17]. Rinke et al. estimated the completeness of the assembled genomes based on 139 bacterial and 162 archaeal conserved marker genes[42]. Wu et al. reported 114 USCGs for "all" bacteria in addition to the 40 USCGs compared here[3]. We did not compare these sets of USCGs because they were more for other purposes and not as universal and unique as those compared in this study.

We found that three of the seven sets of USCGs were universal and unique. Other sets of USCGs we studied, except the one from Ciccarelli et al., were unlikely to be as universal and unique as the three sets because their selection was not as strict. For instance, Lan et al. and Parks et al. required USCGs to occur in only more than 90% of the genomes instead of all genomes[2,9]. The Ciccarelli et al. set was already included in the set from Creevey et al.

For the complete genomes, we demonstrated that USCGs could distinguish different species but not strains. This was based on the fact that the PID of 99.5% is the limit of current tools to separate similar sequences. When the strain similarity is higher, say more than 99.5%, we are left with only a few dozen variable loci to distinguish different strains. In this case, with the available USCG sets, it may be challenging to distinguish them, even if possible. Note that all aforementioned USCG sets are not created to distinguish bacterial species and strains in microbiomes. In the future, one may hope to generate new USCG sets for this purpose by developing novel methods and integrating different sources of information[43–49]. Alternatively, one may consider overlapping the assembled contigs or metagenome-assembled-genomes with the USCGs so that more polymorphic sites are available to distinguish strains[7,23].

Material and methods

Seven sets of USCGS.

We studied seven sets of USCGs: the 31 set from Ciccarelli et al.[1], the 31 set from Wu and Eisen[4], the 40 set from Creevey et al.[16], the 40 USCGs from Wu et al.[3], the 36 USCGs from Alneberg et al.[15], the 73 USCGs from Lan et al.[2], and the 120 USCGs from Parks et al.[9]. We focused on these sets for the separation of similar strains, not for the completeness of the assembled genomes in metagenomics. We converted the gene name of each USCG into its clusters of orthologous groups (COG) name with the information from https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/cog-20.def.tab. In this way, we measured the overlap of two USCG sets by the number of their shared COG gene names.

The complete bacterial genomes.

We studied the complete bacterial genomes at NCBI. We retrieved these bacterial genomes on October 27, 2021, from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/) by selecting the filtering criteria as the 'Bacteria' Kingdom and the 'Complete' Assembly level. In total, we obtained 25,271 complete bacterial genomes. We also downloaded the protein sequences in each genome from the NCBI FTP site at https://ftp.ncbi.nlm.nih.gov/genomes/all/.

Identification of USCGs in a genome.

To identify the corresponding sequences of a USCG in a given genome, we used the tool fetchMG for USCGs from Ciccarelli et al. and Creevey et al.[6]. This tool is based on hidden Markov models to identify the occurrence of a USCG in input sequences. It was used previously to identify and quantify the microbial species in assembled contigs from shotgun metagenomic reads. We used the following command to fetch the USCG sequence in a genome: ./fetchMG.pl -m extraction -p proteins.fasta -o output. Since we input the aforementioned protein sequences in a genome to fetchMG, the fetchMG output includes the full-length protein sequences that are likely the copies of USCGs in this genome. The fetchMG tool might be imperfect, miss certain copies of the USCGs from these two sets in a genome, and does not work for USCGs not in these two sets but in other sets. Therefore, we also applied BLAST to retrieve copies of USCGs that were not from the above two sets or USCGs from the two sets but without an identified copy in a genome. The command we used was: psiblast -in_msa marker_aligned.fasta -db protein.fasta -evalue 1E-15 -num_alignments num_keep. When we applied BLAST, the multiple sequence alignments used as input (the parameter -in_msa marker_aligned.fasta) was the sequences downloaded from https://ftp.ncbi.nih.gov/pub/COG/COG20

20/data/fasta/ and aligned with the mafft tool[50] for all other USCG set except the Wu et al. set. The input multiple sequence alignment for the Wu et al. set was downloaded from their paper[3]. The E-value cutoff used was 1E-15 as previously[3,4,19]. We considered the fetched protein sequences as the corresponding USCG sequences in the genomes. Note that neither fetchMG nor BLAST is perfect for identifying copies of USCGs in a genome, as they sacrifice accuracy for speed. However, the identified copies of USCGs are likely to be reliable. We obtained the functional categories of each USCG from https://www.ncbi.nlm.nih.gov/research/cog.

The USCG similarity measurement. We aligned the extracted USCG protein sequences for every USCG with the tool MAFFT[50]. In other words, we did multiple alignments of the USCG sequences extracted from each genome under consideration for each USCG. In order to assess the similarity of the USCG sequences, we calculated the score based on the blosum62 matrix, the matched number, the mismatch number, the indel number, and the PID. We chose the PID to evaluate the similarity of each pair of USCG sequences because it directly relates to the shotgun metagenomic reads mapped to different USCG sequences. To measure the PID, we obtained the two corresponding sequences in the alignment and removed the loci with an indel versus an indel. We then calculated the PID as the ratio of the number of matched positions to the number of all remaining aligned positions.

Data availability

The 25,271 complete bacterial genomes downloaded on October 27, 2021 are from https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/, by selecting the filtering criteria as the 'Bacteria' Kingdom and the 'Complete' Assembly level. The full-length protein sequences in each of these genomes are from https://ftp.ncbi.nlm.nih.gov/genomes/all/. The seven sets of USCGs are from the corresponding publications[1–4,9,15,16] and are listed in Table S1.

Received: 28 June 2022; Accepted: 18 August 2022
Published online: 25 August 2022

References

1. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**(5765), 1283–1287 (2006).
2. Lan, Y., Rosen, G. & Hershberg, R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* **4**(1), 18 (2016).
3. Wu, D., Jospin, G. & Eisen, J. A. 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* **8**(10), e77033 (2013).
4. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**(10), R151 (2008).
5. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**(12), 1196–1199 (2013).
6. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**(1), 1014 (2019).
7. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol* **18**(1), 181 (2017).
8. Ventolero, M.F., et al., *Computational analyses of bacterial strains from shotgun reads.* Brief Bioinform., 2022. **23**(2).
9. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**(11), 1533–1542 (2017).
10. Vetrovsky, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**(2), e57923 (2013).
11. Wang, Y., Hu, H. & Li, X. MBMC: an effective markov chain approach for binning metagenomic reads from environmental shotgun sequencing projects. *OMICS* **20**(8), 470–479 (2016).
12. Eisen, J. A. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol* **5**(3), e82 (2007).
13. Brooks, J. P. *et al.* The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15**, 66 (2015).
14. Wang, Y. & Qian, P. Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE* **4**(10), e7401 (2009).
15. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**(11), 1144–1146 (2014).
16. Creevey, C. J. *et al.* Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS ONE* **6**(8), e22099 (2011).
17. Haroon, M. F. *et al.* Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**(7464), 567–570 (2013).
18. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**(6102), 1661–1665 (2012).
19. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**(7), 1033–1034 (2012).
20. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**(8), 811 (2012).
21. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling (vol 12, pg 902, 2015). *Nat. Methods* **13**(1), 101–101 (2016).
22. Truong, D. T. *et al.* Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**(4), 626–638 (2017).
23. Quince, C. *et al.* STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol* **22**(1), 214 (2021).
24. Nayfach, S. *et al.* An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**(11), 1612–1625 (2016).
25. Forbes, N. S. Engineering the perfect (bacterial) cancer therapy. *Nat. Rev. Cancer* **10**(11), 785–794 (2010).
26. Hartstra, A. V. *et al.* Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* **38**(1), 159–165 (2015).
27. Jiang, C. *et al.* The gut microbiota and Alzheimer's disease. *J. Alzheimers Dis.* **58**(1), 1–15 (2017).
28. Ott, S. J. *et al.* Detection of diverse bacterial signatures in atherosclerotic lesions of patients with coronary heart disease. *Circulation* **113**(7), 929–937 (2006).
29. Wang, Y. *et al.* Prognostic cancer gene signatures share common regulatory motifs. *Sci. Rep.* **7**(1), 4750 (2017).

30. Zaky, A., et al., The role of the gut microbiome in diabetes and obesity-related kidney disease. *Int. J. Mol. Sci,* 2021. **22**(17).
31. Ding, J., et al., ChIPModule: systematic discovery of transcription factors and their cofactors from ChIP-seq data. In *Pac Symp Biocomput*, 2013: p. 320–31.
32. Harris, M. A. *et al.* The gene ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-61 (2004).
33. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
34. Young, M. D. *et al.* Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**(2), R14 (2010).
35. Zhao, C., Li, X. & Hu, H. PETModule: a motif module based approach for enhancer target gene prediction. *Sci Rep* **6**, 30043 (2016).
36. Li, X., H. Hu, and X. Li, *mixtureS: a novel tool for bacterial strain reconstruction from reads*. Bioinformatics, 2020.
37. Li, X. *et al.* BHap: a novel approach for bacterial haplotype reconstruction. *Bioinformatics* **35**(22), 4624–4631 (2019).
38. Pulido-Tamayo, S. *et al.* Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Res.* **43**(16), e105 (2015).
39. Smillie, C. S. *et al.* Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* **23**(2), 229 (2018).
40. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**(10), 1053–1060 (2015).
41. Chng, K. R. *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nat. Microbiol.* **1**(9), 16106 (2016).
42. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**(7459), 431–437 (2013).
43. Chen, I. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**(D1), D666–D677 (2019).
44. Federhen, S., *The NCBI Taxonomy database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D136-43.
45. Langr, J. and V. Bok, *GANs in action : deep learning with generative adversarial networks*. 2019, Shelter Island, New York,: Manning Publications. xxiii, 214 pages.
46. Li, X. *et al.* Integrative analyses shed new light on human ribosomal protein gene regulation. *Sci. Rep.* **6**, 28619 (2016).
47. Shi, J. Q., Choi, T. & Gaussian process regression analysis for functional data.,. *Boca Raton* 196 (CRC Press. xix, 2011).
48. Talukder, A. *et al.* EPIP: a novel approach for condition-specific enhancer-promoter interaction prediction. *Bioinformatics* **35**(20), 3877–3883 (2019).
49. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**(5338), 631–637 (1997).
50. Katoh, K. *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**(14), 3059–3066 (2002).

Acknowledgements

This work was supported by the National Science Foundation (1661414, 2015838, 2120907).

Author contributions

H.H. and X.L. conceived the idea. S.W. and M.V. implemented the idea and generated results. S.W., M.V., X.L. and H.H. analyzed the results and wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18762-z.

Correspondence and requests for materials should be addressed to H.H. or X.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.