A revisit to universal single-copy genes in bacterial genomes

Saidi Wang[1, *], Minerva Ventolero[2, *], Haiyan Hu[1,3, ¶], Xiaoman Li[2, ¶]

[1]Department of Computer Science, University of Central Florida, Orlando, Florida, United States of America,

[2]Burnett School of Biomedical Science, College of Medicine, University of Central Florida, Orlando, Florida, United States of America.

[3]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, Florida, United States of America.

[*]These authors contributed equally to this work.

[¶]Corresponding Authors:

Email: xiaoman@mail.ucf.edu, haihu@cs.ucf.edu

Abstract

Universal single-copy genes (USCGs) are widely used for species classification and taxonomic profiling. Despite many studies, our understanding of USCGs in bacterial genomes might be out of date, especially how different the USCGs are in different studies, how well a set of USCGs can distinguish two bacterial species, how capable each USCG is in discriminating species, whether USCGs can separate different strains of a bacterial species, and so on. To fill the gaps, we studied USCGs in the most updated sequenced 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 and different strains of a bacterial species. The USCG ArgS is shown to be more powerful than other USCGs to tell different species in the same genus and different strains from the same species apart. Our study shed new light on the evolution of USCGs, which will facilitate their applications in phylogenomic and metagenomic studies.

Introduction

Universal single-copy genes (USCGs) are 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 [5-8].

A well-known USCG is the 16S rRNA gene. It has been used as the gold standard in amplicon sequencing and demonstrated its power for reliable taxonomic profiling in various metagenomic studies [4, 9-12]. It is so far 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 might have multiple copies in an unknown genome [13]. 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, phylogenomics, and metagenomics.

Previous studies have identified USCGs other than the 16S rRNA gene [1-4, 14-18]. 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 [15], which were then used for taxonomic profiling of microbial species in metagenomics in 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 in AMPHORA were further combined with 104 Archaeal USCGs for phylogenomic studies [18]. Later, the same group developed the phyEco gene sets that comprised of 40 USCGs for “all bacteria and archaea” and 114 USCGs for “all bacteria” [3]. Alneberg et al. analyzed the occurrence of all clusters of orthologous groups in 525 genomes and identified 36 USCGs to measure the completeness of the metagenome-assembled genomes [14]. Lan et al. created the POGO-DB database that contained 73 USCGs to classify prokaryotic species [2]. These sets of USCGs are inferred from different groups of sequenced genomes with different purposes. They are thus different, although a fraction of USCGs are shared among them.

The study of bacterial strains is important for the understanding of drug resistance, microbial diversity, and the cure of various complex diseases [19-24]. With multiple USCGs, in addition to study evolution, classification and taxonomic profiling of species, several studies attempted to examine the strain diversity [7, 25-28]. For instance, Quince et al. developed tools to de novo reconstruct bacterial strain genomes from shotgun metagenomic reads with the aforementioned 36 USCGs [7, 28]. The StrainPhlAn tool infers bacterial strains directly from shotgun reads based on 200 or so clade-specific marker genes [27]. Nayfach et al. proposed a pipeline for strain profiling in metagenomic datasets with 15 USCGs [29].

Despite many studies on USCGs, the understanding of USCGs is somewhat out of date. For instance, how disparate are 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 sequenced 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. [15] 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 also systematically studied how similar a USCG is in different species and strains with two sets of sequenced genomes. We found that almost every USCG occurs multiple times in certain genomes, while more than 99.4% of USCGs are single-copied in a given genome. The USCG ArgS seems more capable of distinguishing different species and strains than other USCGs. Finally, USCGs together are good for separating species and strains, while they individually barely can do so. Our study provides a more updated picture of USCGs and their potential applications in metagenomic studies.

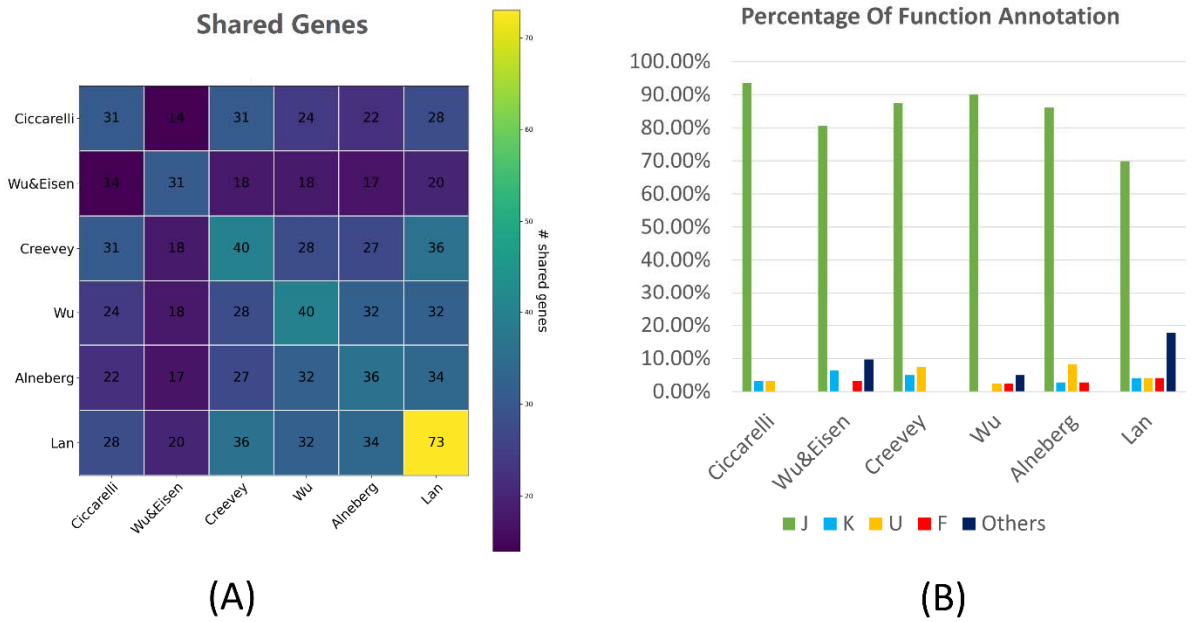
Results

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

We compared USCG sets from six studies [1-4, 14, 15] (Material and Methods). 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.; and (6). 73 USCGs from Lan et al. We focused more on the third to the fifth USCG sets instead of others because others did not emphasize the universalism across all genomes and uniqueness in individual genomes as much as these three sets.

We found that even the USCG sets from the same research group could be very different (Figure 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. Note that 30 (97%) of the 31 USCGs from Wu and Eisen did occur in the 114 USCGs for "all" bacteria defined by Wu et al. [3]. 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. completely included the USCG set from Ciccarelli et al., as the same research group inferred them with the same group of genomes and a slightly different strategy. However, the additional nine USCGs from Creevey et al. also demonstrated the critical effect of such different strategies on defining USCGs.

Figure1. USCG sets and functional categories. (A) Overlap of the six USCG sets. (B) The percentage of the functional categories of the USCGs in the six sets.



The USCG sets from different research groups were more different. 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 corresponding two sets refined later by the corresponding research groups, the aforementioned 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. and Lan et al., the USCGs from Creevey et al. was at least comparable with those from Wu et al., if not better. The USCGs from Alneberg et al. were more for the measurement of genome completeness, which were required to occur in >97% of the 525 genomes with a frequency smaller than 1.03 per genome. The Lan et al. set was even looser, which required the USCGs to occur in more than 90% of the 1897 genomes used. Overall, the

USCG set from Creevey et al. are likely as reliable as any other set if not more reliable because of its stricter requirement of universalism and uniqueness, similar evolutionary trajectory of these USCGs, and likely more representative genomes (consistent results with other sets that used much more genomes).

Despite the difference in the number and member USCGs, the function of the USCGs in each of the above sets was quite consistent (Figure 1B) [30-34]. By checking the functional annotations at https://www.ncbi.nlm.nih.gov/research/cog, we found that almost all USCGs in every set are annotated with the functional category J (Translation, ribosomal structure and biogenesis). There are a couple of USCGs annotated with the categories K (transcription), U (Intracellular trafficking, secretion, and vesicular transport), and F (Nucleotide transport and metabolism) in multiple USCG sets. There is also one USCG annotated with the category O, G, or E in individual sets. It is thus evident that USCG sets are enriched with functions related to the 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. 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 two sources of sequenced genomes (Material and Methods). We focused on the sets from Creevey et al., Wu et al., and Alneberg et

al. here and below because these sets are widely used in metagenomics and represent more updated versions than other sets. The set from Lan et al. was not considered also because of its obvious less universal criteria [2]. In the following, we presented our study on the three sets.

We found that the 40 USCGs from Creevey et al. were indeed universally distributed in almost each of the 34821 genomes (Table 1). For every USCG, it occurred in at least 97.7% of the 25271 NCBI complete genomes and 90.5% of the 11862 mOTUs2 genomes. Although mOTUs2 genomes were retrieved from NCBI, they were not necessarily complete genomes, and they could be archaeal genomes, which partially explained the slightly lower percentage of the mOTUs2 genomes. For those genomes without a copy of a USCG, we applied BLAST to search this USCG in these genomes. BLAST could identify a copy with E-value <1E-15 in most of these missed genomes. A tiny fraction of the genomes still did not have a USCG copy, likely due to the quality of the genomes and the imperfectness of the gene retrieval methods used. Overall, each USCG occurred in at least 99.7% of the NCBI genomes and 98.8% of the mOTUs2 genomes. Of note, the number of USCGs occurring in each genome also showed the universal distribution of these USCGs. The mean number of USCGs occurring in each NCBI genome was 39.6, while that was 39.0 for each mOTUs2 genome. In other words, almost each of the 40 USCGs occurred in every genome from the two sources, indicating the universalism of the USCGs from Creevey et al.

Table 1. Universalism and uniqueness of USCGs.

resources	mean	standard deviation	minimum	median	maximum

NCBI	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
	universalism	100, 99.8, 99.9	3.4, 9.4, 5.3	99.7, 93.4, 99.0	100, 99.9, 100	100, 100, 100
	uniqueness	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
mOTUs2	fetchMG	97.5, 96.9, 97.4	2.1, 4.2, 2.3	90.5, 76.0, 90.5	98.1, 98.5, 98.4	99.6, 99.7, 99.6
	BLAST	98.9, 95.4, 98.3	2.2, 6.3, 2.3	87.7, 66.1, 92.1	99.5, 97.8, 99.4	100, 100, 100
	universalism	100, 99.9, 100	4.2, 10.2, 4.5	98.8, 91.8, 99.2	100, 100, 100	100, 100, 100
	uniqueness	96.7, 95.0, 96.5	2.5, 6.1, 3.3	90.3, 72.8, 86.4	97.5, 97.2, 98.0	99.5, 99.4, 99.5

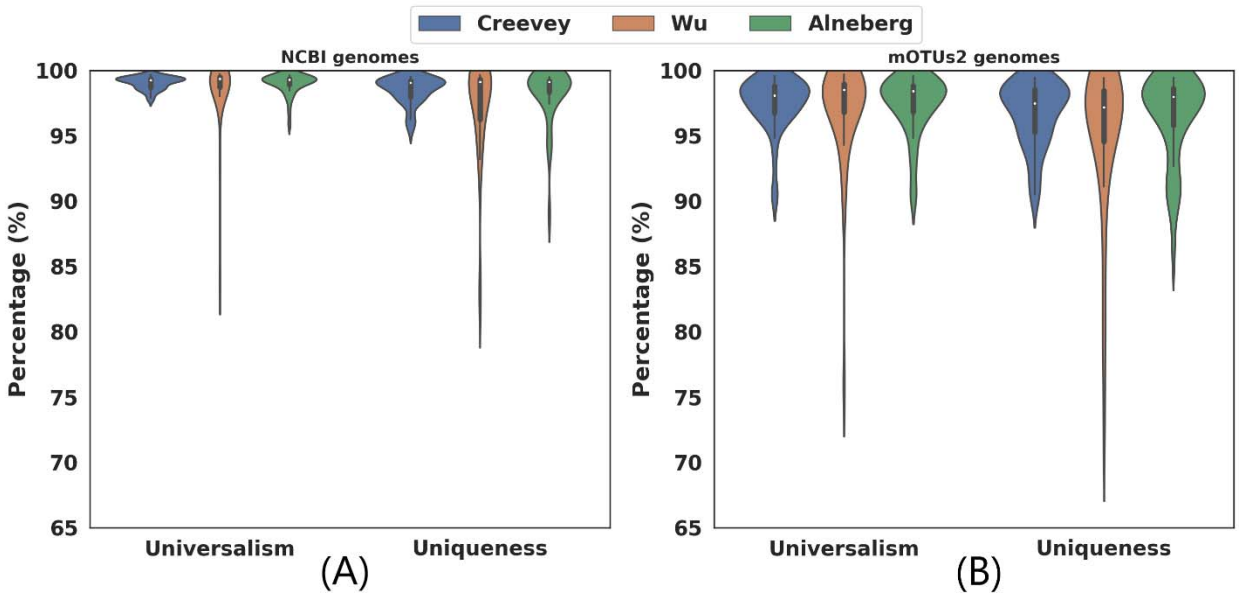
The three numbers in order in each entry are the percentage for the USCGs from Creevey et al., Wu et al., and Alneberg et al.

We also observed that each of the 40 USCGS from Creevey et al. indeed had only one copy in almost every genome (Table 1). For the 25271 NCBI genomes, the mean and median percentage of the genomes that had only one copy of a USCG was 98.5% and 99.0%, respectively. The corresponding percentage of the mOTUs genomes were 96.7% and 97.5%, respectively. Again, the NCBI genomes were more complete and might be thus more reliable, which resulted in higher percentages here. Although 24 (60%) of the 40 USCGs occurred more than once in few than 0.2% of the NCBI genomes, certain USCGs occurred more than once in more than 2.0% of the NCBI genomes, implying that these USCGs were not as unique as the 24 USCGs. Interestingly, these USCGs that occurred more than once in more than 2% of the NCBI

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 in those studies. Overall, the average number of USCGs with only one copy in a genome was 39.4 and 38.7 for the NCBI genomes and the mOTUs2 genomes, respectively. This average number for the NCBI genomes suggests that almost all USCGs have only one copy in almost all complete bacterial genomes. The lower number for mOTUs2 genomes is likely due to the lower quality of these mOTUs2 genomes and the difference between archaeal genomes and bacterial genomes.

We further studied the universalism and uniqueness of the USCG sets from Wu et al. and Alneberg et al. Compared with the set from Creevey et al., the other two sets had a similar median (Figure 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 not so universal and unique as those from Creevey et al., especially in the 25271 NCBI complete bacterial genomes. It is thus evident that the above 40 USCGs from Creevey et al. might be better choices for evolutionary studies.

Figure2. The universalism and uniqueness of the three USCG sets. Comparison of universalism and uniqueness in (A) NCBI genomes and (B) mOTUs2 genomes.



USCGs well separate different species and strains of the same species

With the USCGs indeed universal and unique in bacterial genomes, we investigated how well they together could distinguish two bacterial 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 649 genera with at least two sequenced genomes and all 766 bacterial species with at least two sequenced strain genomes from the two sources of genomes (Material and Methods). We found that at least the 40 USCGs from Creevey et al. together were enough to distinguish every pair of species in the same genus and every pair of strains from the same species (Table 2). In the following, we presented the analysis from the 40 Creevey et al. USCGs and its comparison with the Wu et al. USCG set and the Alneberg et al. USCG set.

Table 2. **PID at the species and strain levels**.

resources		mean	standard deviation	minimum	median	maximum
NCBI	species	87.2, 89.5, 88.7	7.1, 6.3, 6.7	38.5, 39.0, 3.9	87.6, 90.2, 89.0	99.2, 99.1, 99.3
	strain	97.7, 98.6, 98.6	2.0, 1.2, 1.2	51.8, 61.7, 53.2	99.1, 98.8, 99.1	99.2, 99.2, 99.3
mOTUs2	species	83.6, 86.7, 85.0	9.2, 6.6, 8.9	41.0, 45.0, 37.1	85.7, 88.2, 87.2	97.2, 98.0, 97.8
	strain	81.3, 89.9, 81.8	15.9, 10.0, 15.8	50.2, 59.5, 49.5	89.5, 93.3, 90.7	96.8, 97.8, 97.7

The three numbers in each entry are the PID for the USCG set from Creevey et al., Wu et al., and Alneberg et al., in order.

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 at NCBI. The percentage of identity (PID) in a pair of species had a mean and median of 87.2% and 87.6%. In other words, across all aligned positions of the 40 USCGs, about 13.0% of amino acids were on average different between species within the same genus. The most similar pair of species had a PID of 99.2%. Given the fact that the total nucleotide length of the 40 USCGs was about 33,852 and the sequences were less conserved at the nucleotide level than at the protein level, there would be at least 271 positions in the 40 USCGs with different nucleotides even in the two most similar species. Based on previous studies [7, 35-38], existing computational methods could distinguish them with enough sequencing coverage of these two species.

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 at NCBI. The PID in a pair of strains had a mean of 97.7% and a median of 99.1%. The 75% of the quartile of the PID between pairs of strains was already 99.0%, much larger than 93.4%, the corresponding PID for pairs of species. The most similar pair of strains had a PID of 99.2%, the same to that for the most similar pair of species. Similarly, there would be at least 271 positions across the 40 USCGs with different nucleotides in any two strains, which can be separated by existing computational methods given enough sequencing reads from the two strains in shotgun sequencing [35-38]. We examined this pair of most similar strains, Leclercia sp. Colony189 (GCA_009740165.1) and Leclercia sp. LSNIH1 (GCA_002902985.1), and found that there were 364 varied nucleotides across the 40 pairs of USCGs.

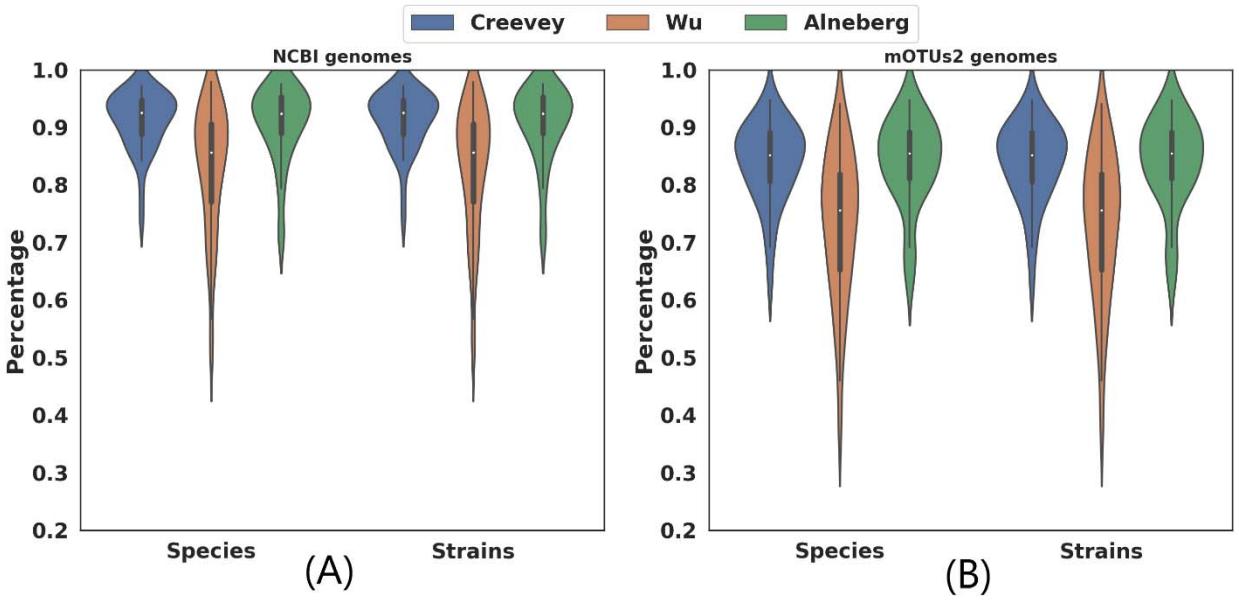
We further studied the similarity of 40 pairs of corresponding USCGs in pairs of strains and species with the mOTUs2 genomes. The PID was lower for both pairs of strains and pairs of species compared with NCBI complete genomes (Table 2), which might be the lower quality of the mOTUs2 genomes. The maximum PID for a pair of species was 98.0%, while the maximum PID for a pair of strains was 97.8%, which was again smaller than the corresponding number for the NCBI genomes. Therefore, with enough sequencing coverage, the 40 USCGs could distinguish any pairs of species and strains we considered here.

Although the 40 USCGs together are good markers to separate pairs of species and strains in metagenomics, all individual USCGs except ArgS are unable to distinguish them. For instance, for the USCG CysS, which has an aligned length of 1,308 base pairs long, may have a PID of 99.6% between two species from the same genus. In other words, there is about five nucleotide difference between orthologous genes of this USCG, which is not enough to distinguish different species and strains. For the USCG ArgS, which shows the lowest PID between species and between strains in terms of minimum, mean, median and maximum, it has

the largest PID of 97.3% between strains and an aligned length of 1,530 base pairs long, implying at least 41 nucleotide difference between orthologous copies across strains and species. The 41 nucleotide difference might be enough to distinguish different species and strains in amplicon sequencing, which might be still challenging to do so in the shotgun metagenomic sequencing.

We also studied how well the USCG set from Wu et al. and from Alneberg et al. separated species and strains (Figure 3). Compared with the USCG set from Creevey et al., these two sets had slightly lower performance, especially on the NCBI bacterial genomes. There were much more pairs of species and strains that were close to 100% PID based on the Alneberg et al. set compared with the other two sets on both sources of sequenced genomes. Overall, the USCG set from Creevey et al. is likely to perform better in distinguishing species and strains in metagenomics.

Figure3. PID of species and strains. PID of species within the same genera and strains within the same species based on three USCG sets from Creevey et al., Wu et al., and Alneberg et al., respectively.



Discussion

We systematically studied the occurrence of USCGs in the most updated list of sequenced microbial 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. However, the USCG ArgS might serve this purpose and was shown to perform better than other USCGs.

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 [17]. 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 [16]. Rinke et al. estimated the completeness of the assembled genomes based on 139 bacterial and 162 archaeal conserved marker genes [39]. 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 at least three of the six sets of USCGs were universal and unique. Other sets of USCGs, 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. required USCGs to occur in only more than 90% of the genomes instead of all genomes [2]. The Ciccarelli et al. set was already included in the set from Creevey et al.

For the two sources of microbial genomes, we demonstrated that USCGs could distinguish different species and strains. This was based on the fact that the PID of the orthologous USCGs in two strains was smaller than 99.6%. Because the number of the sequenced bacterial strains is still limited, one should be aware that when the strain similarity is higher, say more than 99.9%, we are left with only a couple of dozen variable loci to distinguish different strains. In this case, with the available USCG sets, it may be challenging to distinguish them even if it is possible. Note that all aforementioned USCG sets are not created for the purpose of distinguishing 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 [40-46].

Material and Methods

Six sets of USCGS

We studied six 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. [15], the 40 USCGs from Wu et al. [3], the 36 USCGs from Alneberg et al. [14], and the 73 USCGs from Lan et al. [2]. We focused on the sets for the evolutionary and phylogenetic studies, not for the completeness of the assembled genomes in metagenomics. We ended up with the six sets except the one from Alneberg et al. We included the Alneberg et al. set because it was used in several strain studies [7, 8, 28].

Two sources of genomes

We studied the sequenced genomes from two sources. One was the complete bacterial genomes from NCBI. There were 25271 genomes. The other was the reference genomes used by mOTUs2 [6]. mOTUs2 chose these reference genomes from NCBI, which included bacterial and archaeal genomes. Moreover, these reference genomes may be incomplete, which means that some genomes are still in the scaffold status. By including mOTUs2 genomes, we could consider more sequenced species and strains, genomes of different quality, archaeal genomes, and so on. We obtained the functional categories of each USCG from https://www.ncbi.nlm.nih.gov/research/cog.

Identification of USCGs in a genome

To identify the corresponding sequences of a USCG in a given genome, we used the tool fetchMG [6]. This tool 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. Because fetchMG might be imperfect and might miss certain USCG sequences in a genome, we also applied BLAST with the following command to retrieval USCG sequences that were not

identified in a genome: psiblast -in_msa marker_aligned.fasta -db db -evalue 1E-15 -num_alignments num_keep. The E-value cutoff used was 1E-15 as previously [3, 4, 18]. We considered the fetched sequences were the corresponding USCG sequences in the genomes.

The USCG similarity measurement

We aligned the extracted USCG protein sequences for every USCG by the tool MAFFT [47]. In other words, for a given USCG, we did multiple alignments of its USCG sequences extracted from each genome under consideration. 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.

References

1. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science (New York, NY. 2006;311(5765):1283-7. Epub 2006/03/04. doi: 10.1126/science.1123061. PubMed PMID: 16513982.
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. 2016;4(1):18. Epub 2016/05/04. doi: 10.1186/s40168-016-0162-5. PubMed PMID: 27138046; PubMed Central PMCID: PMCPMC4853863.
3. 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(10):e77033. Epub 2013/10/23. doi: 10.1371/journal.pone.0077033. PubMed PMID: 24146954; PubMed Central PMCID: PMCPMC3798382.
4. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome biology. 2008;9(10):R151. Epub 2008/10/15. doi: 10.1186/gb-2008-9-10-r151. PubMed PMID: 18851752; PubMed Central PMCID: PMCPMC2760878.
5. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nature methods. 2013;10(12):1196-9. Epub 2013/10/22. doi: 10.1038/nmeth.2693. PubMed PMID: 24141494.
6. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10(1):1014. Epub 2019/03/06. doi: 10.1038/s41467-019-08844-4. PubMed PMID: 30833550; PubMed Central PMCID: PMCPMC6399450.

7. 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):181. doi: 10.1186/s13059-017-1309-9. PubMed PMID: 28934976; PubMed Central PMCID: PMC5607848.
8. Ventolero MF, Wang S, Hu H, Li X. Computational analyses of bacterial strains from shotgun reads. Briefings in bioinformatics. 2022;23(2). Epub 2022/02/10. doi: 10.1093/bib/bbac013. PubMed PMID: 35136954.
9. Wang Y, Hu H, Li X. MBMC: An Effective Markov Chain Approach for Binning Metagenomic Reads from Environmental Shotgun Sequencing Projects. Omics : a journal of integrative biology. 2016;20(8):470-9. doi: 10.1089/omi.2016.0081. PubMed PMID: 27447888; PubMed Central PMCID: PMC4982950.
10. Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS biology. 2007;5(3):e82. PubMed PMID: 17355177.
11. Brooks JP, Edwards DJ, Harwich MD, Jr., Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15:66. Epub 2015/04/17. doi: 10.1186/s12866-015-0351-6. PubMed PMID: 25880246; PubMed Central PMCID: PMCPMC4433096.
12. Wang Y, Qian PY. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PloS one. 2009;4(10):e7401. Epub 2009/10/10. doi: 10.1371/journal.pone.0007401. PubMed PMID: 19816594; PubMed Central PMCID: PMCPMC2754607.
13. Vetrovsky T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PloS one. 2013;8(2):e57923. Epub 2013/03/06. doi: 10.1371/journal.pone.0057923. PubMed PMID: 23460914; PubMed Central PMCID: PMCPMC3583900.
14. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nature methods. 2014;11(11):1144-6. doi: 10.1038/nmeth.3103. PubMed PMID: 25218180.
15. 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(8):e22099. Epub 2011/08/19. doi: 10.1371/journal.pone.0022099. PubMed PMID: 21850220; PubMed Central PMCID: PMCPMC3151239.
16. Haroon MF, Hu S, Shi Y, Imelfort M, Keller J, Hugenholtz P, et al. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. Nature. 2013;500(7464):567-70. Epub 2013/07/31. doi: 10.1038/nature12375. PubMed PMID: 23892779.
17. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science (New York, NY. 2012;337(6102):1661-5. Epub 2012/09/29. doi: 10.1126/science.1224041. PubMed PMID: 23019650.
18. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics (Oxford, England). 2012;28(7):1033-4. Epub 2012/02/15. doi: 10.1093/bioinformatics/bts079. PubMed PMID: 22332237.
19. Forbes NS. Engineering the perfect (bacterial) cancer therapy. Nat Rev Cancer. 2010;10(11):785-94. Epub 2010/10/15. doi: 10.1038/nrc2934. PubMed PMID: 20944664; PubMed Central PMCID: PMCPMC3756932.
20. Hartstra AV, Bouter KE, Backhed F, Nieuwdorp M. Insights into the role of the microbiome in obesity and type 2 diabetes. Diabetes Care. 2015;38(1):159-65. Epub 2014/12/30. doi: 10.2337/dc14-0769. PubMed PMID: 25538312.
21. Jiang C, Li G, Huang P, Liu Z, Zhao B. The Gut Microbiota and Alzheimer's Disease. J Alzheimers Dis. 2017;58(1):1-15. Epub 2017/04/05. doi: 10.3233/JAD-161141. PubMed PMID: 28372330.
22. Ott SJ, El Mokhtari NE, Musfeldt M, Hellmig S, Freitag S, Rehman A, et al. Detection of diverse bacterial signatures in atherosclerotic lesions of patients with coronary heart disease. Circulation.

2006;113(7):929-37. Epub 2006/02/24. doi: 10.1161/CIRCULATIONAHA.105.579979. PubMed PMID: 16490835.
23. Wang Y, Goodison S, Li X, Hu H. Prognostic cancer gene signatures share common regulatory motifs. Sci Rep. 2017;7(1):4750. Epub 2017/07/08. doi: 10.1038/s41598-017-05035-3. PubMed PMID: 28684851; PubMed Central PMCID: PMCPMC5500535.
24. Zaky A, Glastras SJ, Wong MYW, Pollock CA, Saad S. The Role of the Gut Microbiome in Diabetes and Obesity-Related Kidney Disease. Int J Mol Sci. 2021;22(17). Epub 2021/09/11. doi: 10.3390/ijms22179641. PubMed PMID: 34502562; PubMed Central PMCID: PMCPMC8431784.
25. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods. 2012;9(8):811-+. doi: 10.1038/Nmeth.2066. PubMed PMID: WOS:000307015700022.
26. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling (vol 12, pg 902, 2015). Nature methods. 2016;13(1):101-. doi: 10.1038/nmeth0116-101b. PubMed PMID: WOS:000367463600036.
27. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome research. 2017;27(4):626-38. doi: 10.1101/gr.216242.116. PubMed PMID: WOS:000398058600012.
28. Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, et al. STRONG: metagenomics strain resolution on assembly graphs. Genome biology. 2021;22(1):214. Epub 2021/07/28. doi: 10.1186/s13059-021-02419-7. PubMed PMID: 34311761; PubMed Central PMCID: PMCPMC8311964.
29. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome research. 2016;26(11):1612-25. doi: 10.1101/gr.201863.115. PubMed PMID: WOS:000386609300014.
30. Ding J, Cai X, Wang Y, Hu H, Li X. ChIPModule: systematic discovery of transcription factors and their cofactors from ChIP-seq data. Pac Symp Biocomput. 2013:320-31. Epub 2013/02/21. PubMed PMID: 23424137.
31. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;32(Database issue):D258-61. Epub 2003/12/19. doi: 10.1093/nar/gkh036. PubMed PMID: 14681407; PubMed Central PMCID: PMCPMC308770.
32. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC bioinformatics. 2003;4:41. Epub 2003/09/13. doi: 10.1186/1471-2105-4-41. PubMed PMID: 12969510; PubMed Central PMCID: PMCPMC222959.
33. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology. 2010;11(2):R14. Epub 2010/02/06. doi: 10.1186/gb-2010-11-2-r14. PubMed PMID: 20132535; PubMed Central PMCID: PMCPMC2872874.
34. Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep. 2016;6:30043. Epub 2016/07/21. doi: 10.1038/srep30043. PubMed PMID: 27436110; PubMed Central PMCID: PMCPMC4951774.
35. Li X, Hu H, Li X. mixtureS: a novel tool for bacterial strain reconstruction from reads. Bioinformatics (Oxford, England). 2020. Epub 2020/08/18. doi: 10.1093/bioinformatics/btaa728. PubMed PMID: 32805048.
36. Li X, Saadat S, Hu HY, Li XM. BHap: a novel approach for bacterial haplotype reconstruction. Bioinformatics (Oxford, England). 2019;35(22):4624-31. doi: 10.1093/bioinformatics/btz280. PubMed PMID: WOS:000501728500011.
37. Pulido-Tamayo S, Sanchez-Rodriguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, et al. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic acids research. 2015;43(16):e105. doi: 10.1093/nar/gkv478. PubMed PMID: 25990729; PubMed Central PMCID: PMC4652744.

38. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell host & microbe. 2018;23(2):229-+. doi: 10.1016/j.chom.2018.01.003. PubMed PMID: WOS:000425281900013.
39. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431-7. Epub 2013/07/16. doi: 10.1038/nature12352. PubMed PMID: 23851394.
40. Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic acids research. 2019;47(D1):D666-D77. Epub 2018/10/06. doi: 10.1093/nar/gky901. PubMed PMID: 30289528; PubMed Central PMCID: PMCPMC6323987.
41. Federhen S. The NCBI Taxonomy database. Nucleic acids research. 2012;40(Database issue):D136-43. Epub 2011/12/06. doi: 10.1093/nar/gkr1178. PubMed PMID: 22139910; PubMed Central PMCID: PMCPMC3245000.
42. Langr J, Bok V. GANs in action : deep learning with generative adversarial networks. Shelter Island, New York,: Manning Publications; 2019. xxiii, 214 pages p.
43. Li X, Zheng Y, Hu H, Li X. Integrative analyses shed new light on human ribosomal protein gene regulation. Sci Rep. 2016;6:28619. Epub 2016/06/28. doi: 10.1038/srep28619. PubMed PMID: 27346035; PubMed Central PMCID: PMCPMC4921865.
44. Shi JQ, Choi T. Gaussian process regression analysis for functional data. Boca Raton, FL: CRC Press; 2011. xix, 196 p. p.
45. Talukder A, Saadat S, Li X, Hu H. EPIP: a novel approach for condition-specific enhancer-promoter interaction prediction. Bioinformatics (Oxford, England). 2019;35(20):3877-83. Epub 2019/08/15. doi: 10.1093/bioinformatics/btz641. PubMed PMID: 31410461; PubMed Central PMCID: PMCPMC7963088.
46. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science (New York, NY. 1997;278(5338):631-7. doi: 10.1126/science.278.5338.631. PubMed PMID: 9381173.
47. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002;30(14):3059-66. Epub 2002/07/24. doi: 10.1093/nar/gkf436. PubMed PMID: 12136088; PubMed Central PMCID: PMCPMC135756.