Lan *et al. Microbiome* (2016) 4:18
DOI 10.1186/s40168-016-0162-5

Microbiome

RESEARCH

Open Access

Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains

Yemin Lan[1], Gail Rosen[2†] and Ruth Hershberg[3*†]

Abstract

Background: The 16s rRNA gene is so far the most widely used marker for taxonomical classification and separation of prokaryotes. Since it is universally conserved among prokaryotes, it is possible to use this gene to classify a broad range of prokaryotic organisms. At the same time, it has often been noted that the 16s rRNA gene is too conserved to separate between prokaryotes at finer taxonomic levels.

Results: In this paper, we examine how well levels of similarity of 16s rRNA and 73 additional universal or nearly universal marker genes correlate with genome-wide levels of gene sequence similarity. We demonstrate that the percent identity of 16s rRNA predicts genome-wide levels of similarity very well for distantly related prokaryotes, but not for closely related ones. In closely related prokaryotes, we find that there are many other marker genes for which levels of similarity are much more predictive of genome-wide levels of gene sequence similarity. Finally, we show that the identities of the markers that are most useful for predicting genome-wide levels of similarity within closely related prokaryotic lineages vary greatly between lineages. However, the most useful markers are always those that are least conserved in their sequences within each lineage.

Conclusions: Our results show that by choosing markers that are less conserved in their sequences within a lineage of interest, it is possible to better predict genome-wide gene sequence similarity between closely related prokaryotes than is possible using the 16s rRNA gene. We point readers towards a database we have created (POGO-DB) that can be used to easily establish which markers show lowest levels of sequence conservation within different prokaryotic lineages.

Keywords: Less-conserved genes, Lineage-specific, Marker genes, Genome-wide similarity

Background

One key aim of microbiome studies is to characterize the genomic diversity of prokaryotic species present within environments of interest. However, addressing the genomic diversity can be highly challenging for several reasons. These reasons include technical limitations, such as short read length of high-throughput sequencing, as well as biological challenges that result from prevalent horizontal gene transfer and variation in rates of evolution between different genomes, and within genomes between different genes [1].

In studies where resolving the taxonomical composition of a microbiome is the main goal, it is common practice to amplify and sequence a marker gene and infer the microbiome composition based on its sequence. In most such studies, the 16s rRNA gene serves as the marker of choice [2–5]. The 16s rRNA gene is universally present in prokaryotes and is rarely affected by horizontal gene transfer (HGT) [6, 7]. The 16s rRNA gene contains several hyper-variable regions, allowing for the distinction of

* Correspondence: ruthersh@tx.technion.ac.il
†Equal contributors
[3]Rachel & Menachem Mendelovitch Evolutionary Processes of Mutation & Natural Selection Research Laboratory, Department of Genetics and Developmental Biology, the Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, 31096 Haifa, Israel
Full list of author information is available at the end of the article

different prokaryotes [8]. At the same time, outside of the hyper-variable regions, the sequence of 16s rRNA is relatively very highly conserved across prokaryotes. This allows for the design of universal primers that can be used to amplify 16s rRNA from a very large fraction of prokaryotes. These special characteristics of the 16s rRNA gene make it a useful marker for taxonomical classification and separation, giving rise to various 16s identification tools such as the Ribosomal Database Project [9], Greengenes [10], and SILVA [11].

While 16s rRNA has been demonstrated to be highly useful for taxonomical classification and separation in a great number of microbiome studies, its limitations have also been noted [12, 13]. Despite its hyper-variable regions, the sequence of 16s rRNA tends to be relatively conserved. This makes this gene very useful for studying levels of variation between distantly related prokaryotes. However, it may make 16s rRNA less useful for distinguishing between closely related prokaryotes. In a recent attempt to evaluate species-level identification using the 16s rRNA gene, it was shown that the trade-off between precision and recall could not be well balanced [14]. More recently, oligotyping has been introduced as a method to detect closely related lineages using the 16s rRNA gene [15]. While the method provides finer taxonomic breakdown of the sequenced microbiomes, it provides little information to resolve the relationship between taxa. It was also shown that the 16s rRNA gene has variable explanatory power between prokaryotic lineages, explaining as little as 28 % of the variance among *Enterobacteria* genomes and as much as 70 % of the variance among *Bacteroidetes* [16]. Additionally, some studies suggest that, contrary to previous assumptions, the 16s rRNA gene can be horizontally transferred [17]. Finally, the 16s rRNA gene has varying number of copies in approximately 80 % of all fully sequenced prokaryotic genomes, which may greatly skew estimates of the prokaryotic abundances in a community.

To overcome these limitations, the use of alternative marker genes has been suggested. In several studies, an essential housekeeping gene, such as *rpoB*, *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM*, was used to determine taxonomical relationships for lineages of interest [12, 18–22]. For example, the *recA* gene was used to provide unambiguous identification of *Lactobacillus* strains [21]. The *rpoB* gene was used to estimate biodiversity in a soil sample [12] and reconstruct taxonomical relationships among strains belonging to the *Halobacteriales* lineage [19]. The *rpoB* gene was also shown to be more useful than 16s rRNA at discriminating closely related organisms [20]. The chaperonin-60 universal target was shown to be a particularly useful marker as it was demonstrated to predict similar bacterial genome relatedness to whole genome sequence alignments over a broad range of taxa [22]. More recently, taxonomical inference dependent on whole genome sequencing and combined use of a variety of markers has been repeatedly proposed, where the definition of markers ranged from housekeeping genes to representative homologous gene groups and exclusive genetic episodes [23–26]. In these studies, housekeeping genes have been suggested to be useful for discriminating lineages, as they are a major component of the core genes for a lineage [25] and are thought to be subject to less environmental pressure than other genes [27]. Wu et al. identified 31 housekeeping genes from 100 genomes that can be used to classify and separate prokaryotes [23]. These markers have been used to speed up the taxonomic classification based on metagenomic shotgun data [24]. Ciccarelli et al. have used a concatenation of 31 marker genes to reconstruct the tree of life for 191 species with their whole genome sequenced at the time [28]. Another comprehensive study has surveyed 32 protein-coding genes that are widely distributed among bacterial genomes and demonstrated the usefulness of single-gene alignments in predicting genome relatedness in specific lineages [29].

In the meantime, whole genome sequencing (WGS) has become more feasible and more prevalent in microbiome studies [30–32]. In comparison with the sequencing of a single marker gene, WGS uncovers both the microbial composition and functional composition of the microbiomes of interest. Hence, various tools are developed over the past few years to facilitate analysis of WGS data, such as WGSQuikr [33] and MetaPhlAn [34]. In addition to taking advantage of the full microbiome sequenced from WGS, some approaches resort to a small subset of widely conserved marker genes mined from WGS data, such as PhyloSift [35], AMPHORA2 [36], MetaPhyler [24], EMIRGE [37], and PhylOTU [38]. Consequently, in various scenarios, many widely conserved genes other than the 16s rRNA gene have proven valuable towards better assessment of interspecies relationships [19–21, 24, 26, 39]. While WGS has become cheaper and more popular in microbiome studies, use of a single marker gene similar to that of the 16s rRNA gene still has its advantage in effectiveness and efficiency. Despite the well-understood benefits of utilizing marker genes for classifying and separating prokaryotes, a major challenge remains in determining which marker genes are most useful for different lineages and various scenarios. Therefore, a systematic analysis of the utility of different marker genes and the development of criteria for choosing the correct marker genes should yet prove beneficial.

Here, we conducted genome-wide comparisons of levels of gene similarity of ~2000 fully sequenced genomes. For each genome pair compared, we also estimated levels of similarity of the 16s rRNA gene as well as of 73 additional marker genes that are each present within at least 90 % of all prokaryotes. We show that levels of similarity of 16s rRNA are very good predictors of genome-wide levels of similarity for distantly related prokaryotes. At the same time, many other marker genes are much more useful than 16s rRNA for predicting genome-wide levels of similarity for more closely related prokaryotes. The identity of the most useful markers varies between prokaryotic lineages. However, within each lineage, the markers that are most useful for predicting genome-wide levels of similarity are those markers that have the lowest levels of sequence conservation within that lineage. Our results indicate that it should be possible to obtain far better separation of closely related strains in lineage-specific prokaryotic studies by using markers that are less conserved in their sequences.

Methods

Extracting whole genome and 16s rRNA gene sequences

The complete genomes of 2013 prokaryotic strains were downloaded from the NCBI database (in July, 2012). Genes annotated as "16s rRNA gene" were extracted from each strain. A total of 1897 genomes were retained with 16s rRNA genes of legitimate length from 1000- to 1800-bp nucleotides [8].

Identification of universal or nearly universal marker genes

Single-copy and universally distributed genes in the COG database [40] were considered as potential markers (Additional file 1: Table S1). Copies of these genes in each genome were recognized from BLAST reciprocal best hits, using the copies of these genes in *Escherichia coli K12 W3110* (NCBI Genome unique ID 161931) as references. Each of the 73 potential markers was found to be present in more than 90 % of the prokaryotic genomes. A total of 1204 genomes harbor all marker genes.

Calculating marker gene percent identity

For each pair of genomes, all 16s rRNA gene copies were pairwise-aligned using the Needleman-Wunsch algorithm [41]. The maximal 16s rRNA identity was recorded for each genome pair. Likewise, the percent identity of each marker gene between a pair of genomes was obtained by aligning the nucleotide sequences using the pairwise Needleman-Wunsch alignment algorithm.

Calculating average AAI

The average amino acid identity (AAI) was computed for all genome pairs whose maximal 16s rRNA gene identity was at least 80 %. Towards this end, protein-coding sequences in each genome were obtained from the NCBI Genome database [42]. Homologs between a pair of genomes were identified by reciprocal BLAST comparisons [43] and re-aligned using the Smith-Waterman algorithm [44]. To be defined as homologs, proteins had to be aligned over at least 70 % of the shorter sequence and the AAI had to be above 30 % [45]. The average AAI was computed using all identified homologs between two genomes, if they had at least 200 homologs (out of 717,861 pairwise comparisons, 2556 were discarded because less than 200 homologs were identified).

Average ranking of marker genes

Focusing only on genomes that contain all 74 marker genes (including the 16s rRNA gene), we calculated for each genome pair the relative rank of each marker gene by their percent identities within that genome pair. The marker gene that was least similar between the two genomes (i.e., had the lowest percent identity) received a rank of 74, and the marker gene that was most similar (i.e., had the highest percent identity) received a rank of 1. The average ranking of a gene within a certain lineage was then calculated based on its ranking for each pair of genomes from that lineage. At the end of this process, lower ranks (closer to 1) correspond to more conserved genes.

Generating trees based on average AAIs and based on marker gene percent identities

Trees were generated using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm, where the pairwise distance matrix values were one minus the AAIs or one minus the percent identities of a specific marker gene. Split distance (equivalent to the Robinson-Foulds distance) between trees was calculated using TOPD/FMTS software [46].

Results and discussion

The 16s rRNA gene is an ambiguous marker for inferring extent of similarity of closely related prokaryotes

The percent identity of the 16s rRNA gene is considered to be indicative of the taxonomical relationship between prokaryotes. For example, sequences with a percent identity of 97 % or higher are considered by many as coming from the same species [47] and a percent identity of 94 % or higher is thought to indicate that the sampled prokaryotes belong to the same genus [16]. However, as is discovered often and especially at finer taxonomic levels, relying on the 16s rRNA alone may give faulty inferences [12, 13].

To systematically address this issue, we examined the relationship between the percent identity of the 16s rRNA gene and taxonomical relationship, according to NCBI classifications (Fig. 1). As expected, a general trend holds that higher 16s rRNA identity is indicative of closer taxonomic relationships. However, the inference of the level of taxonomic relatedness based on 16s

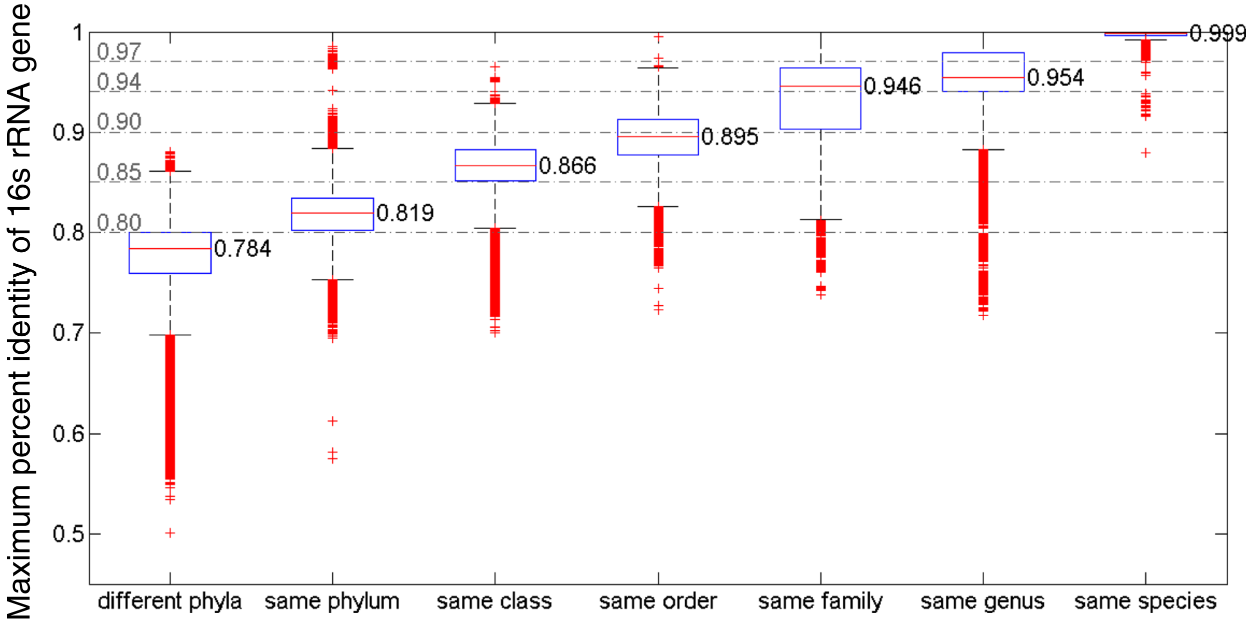

Fig. 1 Box and whisker plot of 16s rRNA percent identities between genomes of the same taxa (species/genus/etc.). The *whiskers* represent boundaries within 1.5 interquartile range of the lower and upper quartiles. Although the 16s rRNA genes are more similar at finer taxonomic levels, there are still large overlaps between taxonomic classification levels and outliers within each level. Therefore, the 16s rRNA gene is sometimes an ambiguous marker in taxonomic classification

percent identities is difficult, as the range of 16s identity overlaps between different taxonomic relationships. For example, sequences from the same species can be less than 94 % identical while sequences from different classes can be 97 % identical. It is of note though that due to missing taxa, inconsistent taxonomy, and the presence of numerous informal names, the fact that we use NCBI taxonomy may also contribute to the observed noisiness.

To overcome the limitations of the NCBI taxonomic classifications, we used an alternative approach for estimating the evolutionary distance between different prokaryotes. Namely, we calculated the average AAI between genomes using all homologous protein-coding genes they shared. Calculating AAI is computationally expensive and possible only for fully sequenced genomes. However, we assume that it is more accurate to estimate the evolutionary distance between two genomes based on all homologs, rather than based on a single locus such as the 16s rRNA gene. AAI likely does not provide a perfect estimation of relatedness, as it can also be somewhat affected by HGT. However, because it combines information across a very large number of loci, AAI likely provides us with the most reliable available estimation of the extent of relatedness of two genomes.

We compared every two genomes with a 16s identity higher than 80 %, and a total of 717,861 pairs of genomes were analyzed. In general, as expected, the percent identity of the 16s rRNA gene correlates well with AAI (rho = 0.7487 and p value ≪0.0001, according to a Spearman test, Fig. 2a). However, as the 16s rRNA genes get more similar, the range of AAI becomes broader (Fig. 2a). In other words, the percent identity of the 16s rRNA gene becomes more ambiguous at predicting AAI when genomes are more closely related.

The analysis described above was carried out by comparing all genomes to each other in a pairwise manner. Due to this, Fig. 2a might be greatly affected by individual genome outliers. For example, it is possible that one genome has mutated abnormally faster than other genomes of its species but retained the almost identical 16s rRNA gene. This may result in many data points with high 16s identity but low AAI, while the cause is merely one outlier genome. To reduce the impact of abnormal data points caused by individual outlier genomes, we performed 20 random samplings of 100 genome pairs and calculated the 95 % confidence intervals for AAI. In these samplings, a genome could only be selected once. This was done for each 16s rRNA gene identity level with an interval of 2 %. The result is shown in Fig. 2b, where the range of confidence interval increases as genomes become more similar. For example, genome pairs with 80–82 % 16s rRNA identity are 95 % likely to have an AAI between 40 and 46 %. In comparison, genome pairs with a 98–100 % 16s rRNA identity can have an AAI anywhere between 70 and 100 %. This again shows that 16s rRNA gene percent identities become more ambiguous predictors of AAI as genome pairs become more similar.

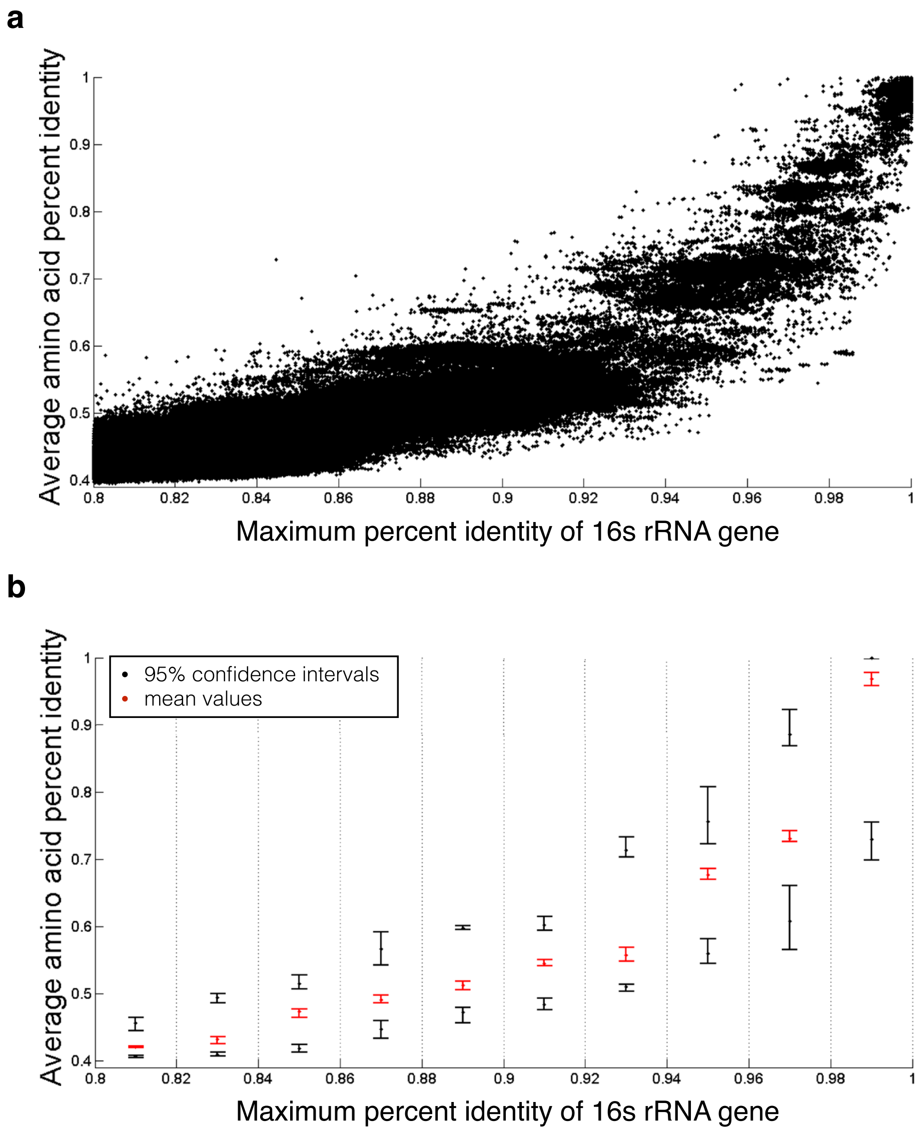

Fig. 2 The 16s rRNA percent identity compared with the average AAI for all pairs of genomes (**a**) and with random samplings (**b**). The latter shows the 95 % confidence interval of AAI for each 2 % range of 16s rRNA percent identity, based on 20 random samplings of 100 genome pairs to assure that no genome was considered twice. While there is a general trend that higher 16s rRNA similarity indicates higher AAI, the range of AAI becomes broader for prokaryotes with more similar 16s rRNA sequences. For example, pairs of prokaryotes with 98 % 16s rRNA identity can have an AAI of anywhere between 50 and 100 %. The range of average AAI values for each bin is plotted in *red*, with the 95 % confidence interval upper and lower boundaries plotted in *black*

When considering closely related prokaryotes, the percent identity of many marker genes better predicts AAI than the percent identity of 16s rRNA

The ambiguity of 16s rRNA in predicting genome-wide levels of similarity (as measured by AAI), especially for closely related prokaryotes, led to the question of whether other housekeeping genes might be more useful for this purpose. To find a potentially better marker gene, we started with the genes that were common to all genomes in the COG database and trimmed them down to 73 that were present in over 90 % of fully sequenced prokaryote genomes ("Methods" section, Additional file 1: Figure S1).

Figure 3 shows the Spearman correlation between the percent identity of each marker gene and AAI, for genome pairs with AAI values lower or greater than 95 % (see Additional file 1: Figure S2 for Spearman's correlation between each marker gene and AAI for all genomes altogether). For the more distantly related genomes with AAIs lower than 95 %, the 16s rRNA percent identity correlates very well with AAI. In fact,

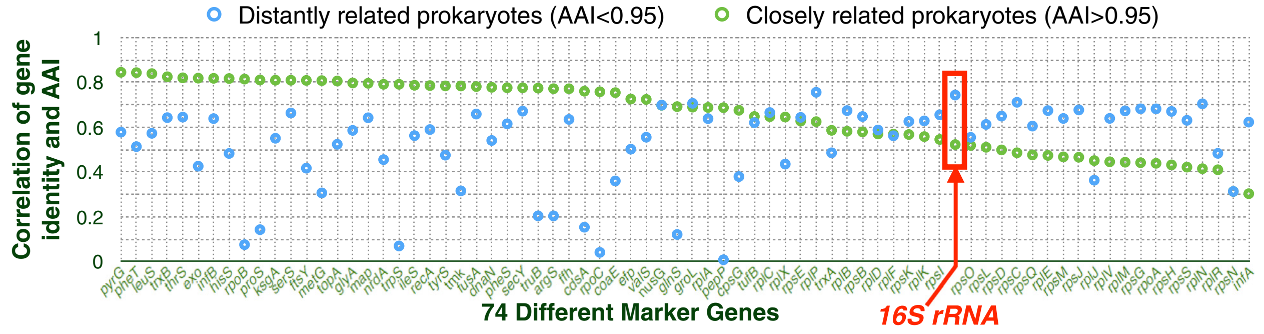

Fig. 3 Spearman's correlation between each marker gene and the average AAI for distantly related prokaryotes whose AAIs are lower than 95 % and closely related prokaryotes whose AAIs are greater than 95 %. Genes are ordered by decreasing correlation coefficient for closely related prokaryotes

there is only one other marker gene (*rplP*) that yields a slightly better correlation than 16s rRNA for these genome pairs. However, when it comes to closely related genomes, with AAIs over 95 %, most markers outperform 16s rRNA (Fig. 3). Among the markers whose percent identities correlate better with AAI, for genomes with AAI over 95 %, is *rplP*. This means that *rplP* provides better correlations with AAI for both closely and distantly related genomes. It is interesting to note that *rpoB* and *recA* are among the many marker genes that outperform 16s rRNA for closely related genomes. This likely explains why these markers were previously shown to be more useful for classifying strains within specific lineages [19–21]. It is also worth noting the *groL* gene (often referred to as *groEL* or cpn60 (see Additional file 1: Table S2 for alternative names of surveyed marker genes)). Fitting with previous suggestions that *groL* may be a particularly well-performing marker [48, 49], we found that the percent identity of *groL* correlates equally well with AAI for both closely and distantly related prokaryotes (Fig. 3).

The better correlation with AAI observed for the percent identities of various marker genes suggests that some of them should be more useful than the 16s rRNA gene in classifying and separating strains within closely related lineages. While we cannot know the true species trees of any bacterial lineage, we again relied on an assumption that the best trees available to us will be generated by considering levels of similarity between all shared orthologs or, in other words, by considering AAI. We therefore reconstructed predicted species trees of three bacterial lineages, using AAI or each of the 74 marker genes including 16s rRNA. The three bacterial lineages for which we reconstructed such trees were *Escherichia*/*Shigella* (Additional file 1: Figure S3), *Streptococcus* (Additional file 1: Figure S4), and *Bacillus* (Additional file 1: Figure S5). The trees reconstructed based on each of the marker genes (including the 16s rRNA gene) were compared by how well they resemble the tree reconstructed based on AAI (Additional file 1: Table S3).

Manual examination of the AAI-generated trees and of trees generated using 16s rRNA results in a number of examples in which the tree generated using 16s rRNA does not provide accurate separation between strains. For example, the *Escherichia*/*Shigella* tree generated based on AAI (Fig. 4a) well separates the *E. coli O157*:*H7* strains and *E. coli O55*:*H7* strains, which were mixed in the tree generated from 16s rRNA identities (Fig. 4b). In addition, the 16s tree failed to show an immediate common ancestor for *Shigella dysenteriae* and *Escherichia coli O157*:*H7* [50], while this is clearly seen in the AAI tree. In contrast, the trees generated based on several marker genes were able to capture these separations (Additional file 1: Figure S3). Similar results can also be seen for the *Streptococcus* and *Bacillus* lineages (Additional file 1: Figures S4 and S5).

To quantify the usefulness of the marker genes in reconstructing similar relationships between strains to those observed using genome-wide information, we computed the split distance between each marker gene tree and the AAI tree (Additional file 1: Table S3, "Methods" section). For all three lineages, many marker genes outperformed the 16s rRNA gene. In fact, out of 74 marker genes (including 16s rRNA), the 16s rRNA gene was only the 48th, 64th, and 31st best-performing gene in reconstructing the AAI tree of *Escherichia*/*Shigella*, *Bacillus*, and *Streptococcus*, respectively. The top 10 marker genes that best reconstructed each lineage are listed in Table 1. The genes that were most useful for reconstruction of the AAI-based tree vary between lineages, suggesting that marker genes need to be individually selected for each specific lineage of interest.

Marker genes that are least conserved in their sequences within a lineage can be used to generate trees most similar to that inferred using AAI within that lineage

The above finding that different marker genes seem to be most useful for reconstructing the AAI-based tree of different bacterial lineages raises the question of how to choose the correct marker genes for examining a lineage

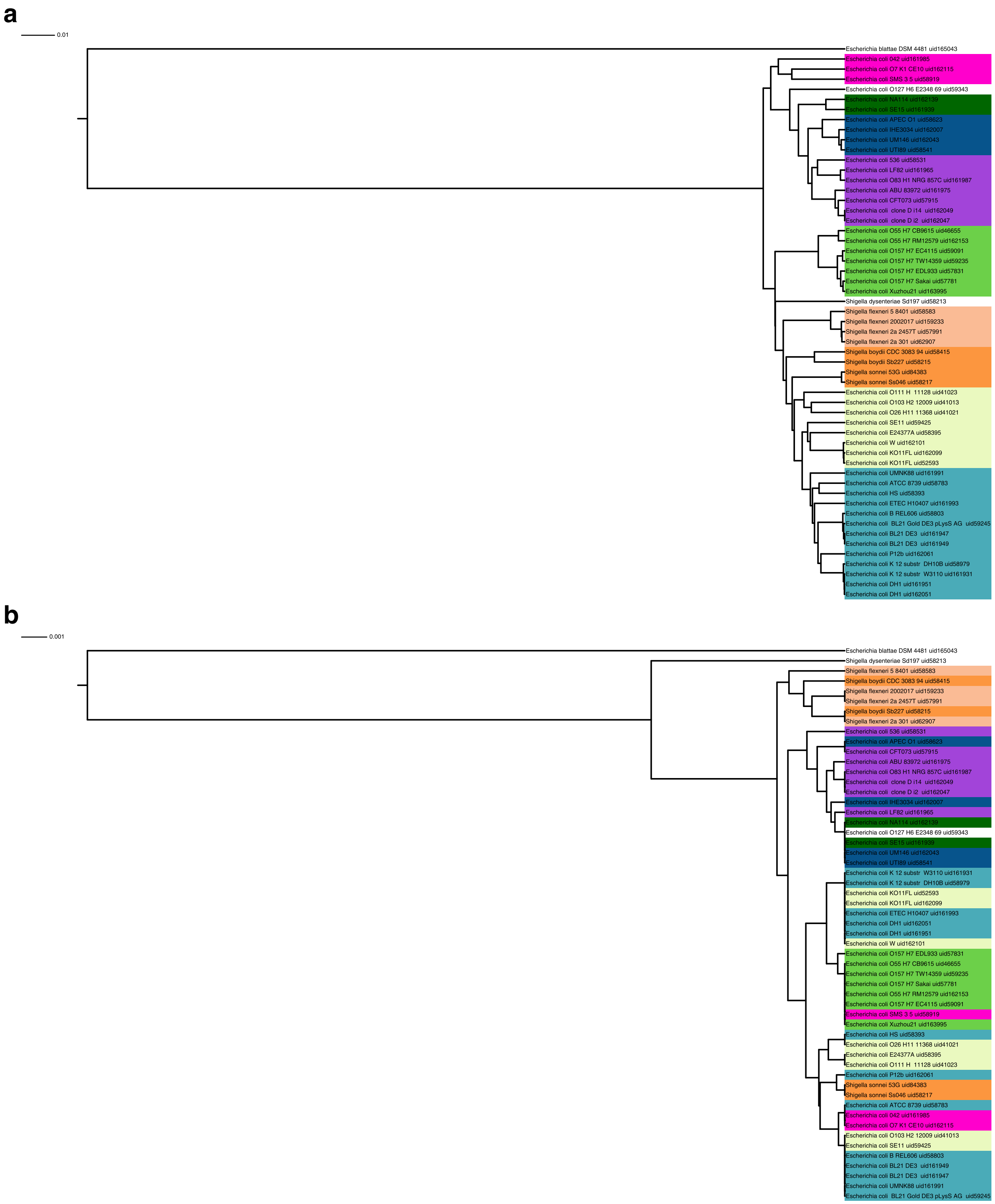
a
0.01
Escherichia blattae DSM 4481 uid165043
Escherichia coli 042 uid161985
Escherichia coli O7 K1 CE10 uid162115
Escherichia coli SMS 3 5 uid58919
Escherichia coli O127 H6 E2348 69 uid59343
Escherichia coli NA114 uid162139
Escherichia coli SE15 uid161939
Escherichia coli APEC O1 uid58623
Escherichia coli IHE3034 uid162007
Escherichia coli UM146 uid162043
Escherichia coli UTI89 uid58541
Escherichia coli 536 uid58531
Escherichia coli LF82 uid161965
Escherichia coli O83 H1 NRG 857C uid161987
Escherichia coli ABU 83972 uid161975
Escherichia coli CFT073 uid57915
Escherichia coli clone D i14 uid162049
Escherichia coli clone D i2 uid162047
Escherichia coli O55 H7 CB9615 uid46655
Escherichia coli O55 H7 RM12579 uid162153
Escherichia coli O157 H7 EC4115 uid59091
Escherichia coli O157 H7 TW14359 uid59235
Escherichia coli O157 H7 EDL933 uid57831
Escherichia coli O157 H7 Sakai uid57781
Escherichia coli Xuzhou21 uid163995
Shigella dysenteriae Sd197 uid58213
Shigella flexneri 5 8401 uid58583
Shigella flexneri 2002017 uid159233
Shigella flexneri 2a 2457T uid57991
Shigella flexneri 2a 301 uid62907
Shigella boydii CDC 3083 94 uid58415
Shigella boydii Sb227 uid58215
Shigella sonnei 53G uid84383
Shigella sonnei Ss046 uid58217
Escherichia coli O111 H 11128 uid41023
Escherichia coli O103 H2 12009 uid41013
Escherichia coli O26 H11 11368 uid41021
Escherichia coli SE11 uid59425
Escherichia coli E24377A uid58395
Escherichia coli W uid162101
Escherichia coli KO11FL uid162099
Escherichia coli KO11FL uid52593
Escherichia coli UMNK88 uid161991
Escherichia coli ATCC 8739 uid58783
Escherichia coli HS uid58393
Escherichia coli ETEC H10407 uid161993
Escherichia coli B REL606 uid58803
Escherichia coli BL21 Gold DE3 pLysS AG uid59245
Escherichia coli BL21 DE3 uid161947
Escherichia coli BL21 DE3 uid161949
Escherichia coli P12b uid162061
Escherichia coli K 12 substr DH10B uid58979
Escherichia coli K 12 substr W3110 uid161931
Escherichia coli DH1 uid161951
Escherichia coli DH1 uid162051
b
0.001
Escherichia blattae DSM 4481 uid165043
Shigella dysenteriae Sd197 uid58213
Shigella flexneri 5 8401 uid58583
Shigella boydii CDC 3083 94 uid58415
Shigella flexneri 2002017 uid159233
Shigella flexneri 2a 2457T uid57991
Shigella boydii Sb227 uid58215
Shigella flexneri 2a 301 uid62907
Escherichia coli 536 uid58531
Escherichia coli APEC O1 uid58623
Escherichia coli CFT073 uid57915
Escherichia coli ABU 83972 uid161975
Escherichia coli O83 H1 NRG 857C uid161987
Escherichia coli clone D i14 uid162049
Escherichia coli clone D i2 uid162047
Escherichia coli IHE3034 uid162007
Escherichia coli LF82 uid161965
Escherichia coli NA114 uid162139
Escherichia coli O127 H6 E2348 69 uid59343
Escherichia coli SE15 uid161939
Escherichia coli UM146 uid162043
Escherichia coli UTI89 uid58541
Escherichia coli K 12 substr W3110 uid161931
Escherichia coli K 12 substr DH10B uid58979
Escherichia coli KO11FL uid52593
Escherichia coli KO11FL uid162099
Escherichia coli ETEC H10407 uid161993
Escherichia coli DH1 uid162051
Escherichia coli DH1 uid161951
Escherichia coli W uid162101
Escherichia coli O157 H7 EDL933 uid57831
Escherichia coli O55 H7 CB9615 uid46655
Escherichia coli O157 H7 TW14359 uid59235
Escherichia coli O157 H7 Sakai uid57781
Escherichia coli O55 H7 RM12579 uid162153
Escherichia coli O157 H7 EC4115 uid59091
Escherichia coli SMS 3 5 uid58919
Escherichia coli Xuzhou21 uid163995
Escherichia coli HS uid58393
Escherichia coli O26 H11 11368 uid41021
Escherichia coli E24377A uid58395
Escherichia coli O111 H 11128 uid41023
Escherichia coli P12b uid162061
Shigella sonnei 53G uid84383
Shigella sonnei Ss046 uid58217
Escherichia coli ATCC 8739 uid58783
Escherichia coli 042 uid161985
Escherichia coli O7 K1 CE10 uid162115
Escherichia coli O103 H2 12009 uid41013
Escherichia coli SE11 uid59425
Escherichia coli B REL606 uid58803
Escherichia coli BL21 DE3 uid161949
Escherichia coli BL21 DE3 uid161947
Escherichia coli UMNK88 uid161991
Escherichia coli BL21 Gold DE3 pLysS AG uid59245

Fig. 4 Trees of the *Escherichia/Shigella* lineage generated **a** based on AAI and **b** based on 16s rRNA percent identity. The latter fails to separate some major divisions of this lineage and poorly reconstitutes the genome-wise similarity relationship

Table 1 The identity of the top 10 best performing markers differs between the three examined lineages

Escherichia/Shigella	*Streptococcus*	*Bacillus*
argS	*nrdA*	*cdsA*
glnS	*serS*	*glyA*
ileS	*dnaN*	*hisS*
valS	*cpsG*	*leuS*
infB	*fusA*	*pheT*
fusA	*proS*	*tmk*
topA	*pyrG*	*infB*
Ffh	*tyrS*	*ksgA*
proS	*ftsY*	*pheS*
Tmk	*glnS*	*topA*

of interest. We hypothesized that genes that are less conserved in their sequences within a lineage should be more capable of capturing the evolutionary differences between its members and therefore should allow for the reconstruction of better trees. To test this hypothesis, we ranked the different marker genes according to their percent identities within the three examined lineages ("Methods" section). Genes with higher ranks (closer to 74) have lower percent identities (are less conserved in their sequences). At the same time, genes with lower ranks (closer to 1) have higher levels of sequence conservation. As is shown in Fig. 5, the rank of genes strongly correlates with their AAI-based tree reconstruction performance, for the three lineages examined (p values $\ll 0.001$ according to the Spearman correlation test). This means that marker genes that are the least conserved in their sequences within each lineage can be used to better reconstruct the AAI-based tree for the given lineage.

An example of this trend can be seen when considering the marker gene with the lowest levels of sequence conservation for the *Bacillus* lineage, *coaE*. The correlation observed between AAI and the percent identity of *coaE* is much better than that observed between AAI and the percent identity of the 16s rRNA gene (Fig. 6a). Furthermore, *coaE* reconstructs a tree that is much more similar to the one obtained using AAI (Fig. 6b). Specifically, when we used the 16s rRNA gene to reconstruct the *Bacillus* tree, we could not distinguish between closely related genomes. However, when we used *coaE*, such closely related genomes could be clustered as they were in the AAI-derived tree (Fig. 6b).

It was previously suggested that *Bacillus anthracis* (the causative agent of anthrax), *Bacillus cereus*, and *Bacillus thuringiensis* are genetically one species in spite of their widely different phenotypes and pathogenic effects. In agreement with previous studies into *Bacillus* [51], the trees we reconstruct for *Bacillus* using AAI or *coaE* both show that *Bacillus* has two major clades, with *B. anthracis* forming a monophyletic branch in one of the clades (Fig. 6b). In contrast, the 16s rRNA reconstructed tree does not capture the two-clade division and also does not capture the monophyletic branch of *B. anthracis*. This is noteworthy because it demonstrates that by using marker genes, individually chosen to best separate particular lineages of interest, we may in some cases be able to separate pathogens (e.g., anthrax) from less pathogenic members of their lineage.

Combined, our results suggest that the relative levels of sequence conservation of marker genes provide a useful metric for the selection of the best markers to use to infer degrees of genome-wide similarity within specific lineages of interest.

Conclusions

The 16s rRNA gene was used in many microbiome studies in which it was demonstrated that fluctuations in microbiome structure correlate with ailments such as obesity [52, 53], caries and periodontal disease [54, 55], gastrointestinal disease [56, 57], and urinary tract infections [58]. At the same time, as we show here and others have shown before us, 16s rRNA is not very useful when it comes to inferring the degree of genome-wide similarity of closely related prokaryotic strains. Understanding whether and how fluctuations in specific species abundances contribute to the observed correlations may thus require separation of strains at higher resolution than possible using 16s rRNA. For example, in microbiome studies that aim to separate certain members from a lineage such as the pathogenic ones, or revisit a classification scheme, a marker that can better distinguish between closely related genomes may be required.

Here, we present an analysis quantifying the extent to which 16s rRNA similarity predicts genome-wide similarity between both closely and distantly related prokaryotes. We showed that 16s rRNA is one of the best marker genes for inferring genome-wide gene sequence similarity (as estimated using AAI) for distantly related genomes. At the same time, many marker genes can be more useful than 16s rRNA, when it comes to estimating genome-wide levels of gene sequence similarity, for closely related prokaryotes. Finally, we show that the marker genes that are least conserved in their sequences within a lineage of interest are the most useful in inferring trees that resemble those constructed based on genome-wide similarity information (as measured using AAI).

We show that less sequence-conserved markers are less useful than 16s rRNA and other more sequence-conserved markers for inferring genome-wide levels of gene sequence similarity between distantly related prokaryotes. At the same time, less sequence-conserved markers are more useful for inferring similarity of closely related prokaryotes. Why is this the case? Less conserved genes are

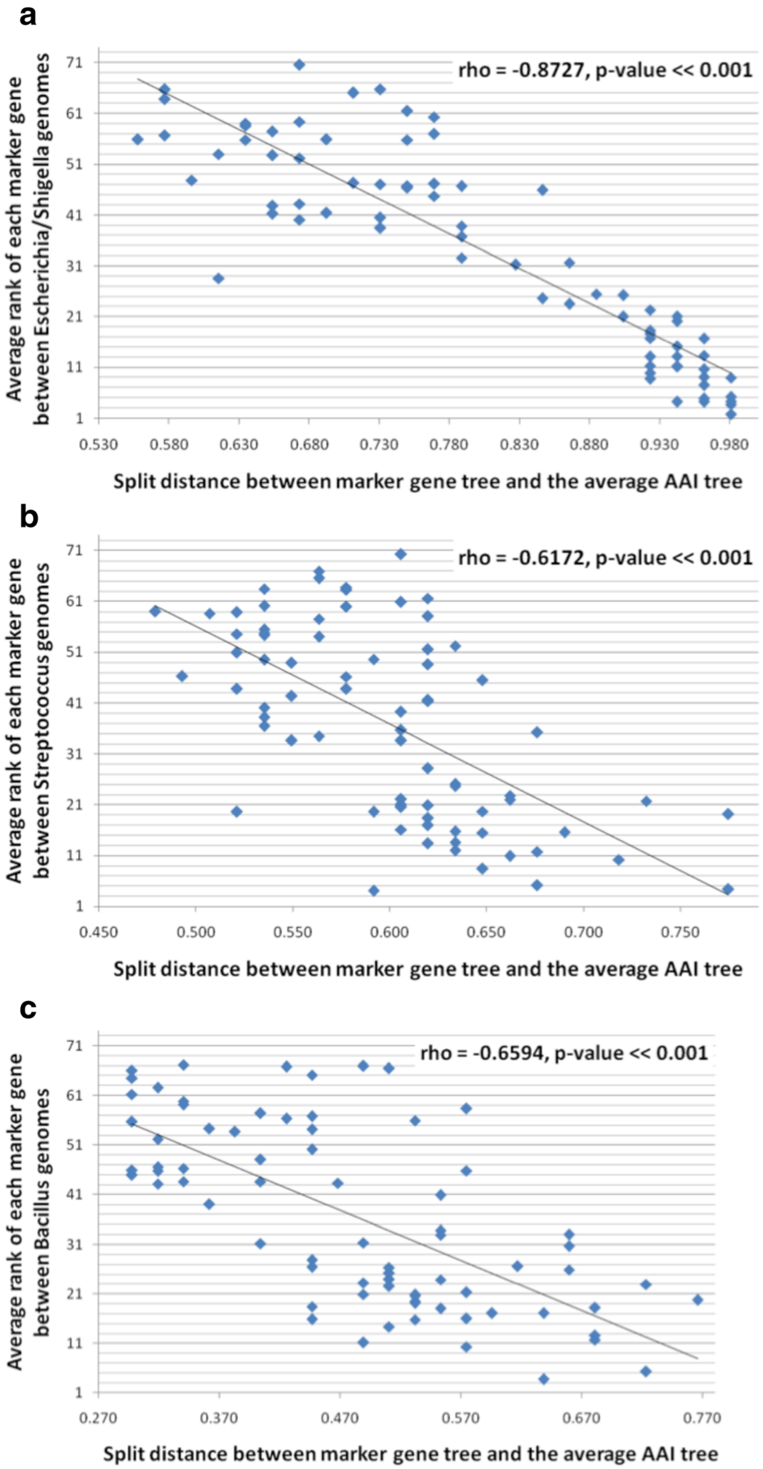

Fig. 5 Trees generated using less sequence-conserved marker genes better reconstitute the tree generated using AAI, for three lineages: **a** *Escherichia/Shigella*, **b** *Streptococcus*, and **c** *Bacillus*. The percent identity rank of each gene represents its relative sequence conservation within the lineage. The split distance represents the resemblance between the tree generated using each marker gene and the tree generated using AAI. Spearman correlation coefficients and *p* values are shown, and a *linear trend line* is drawn that minimizes the squared error

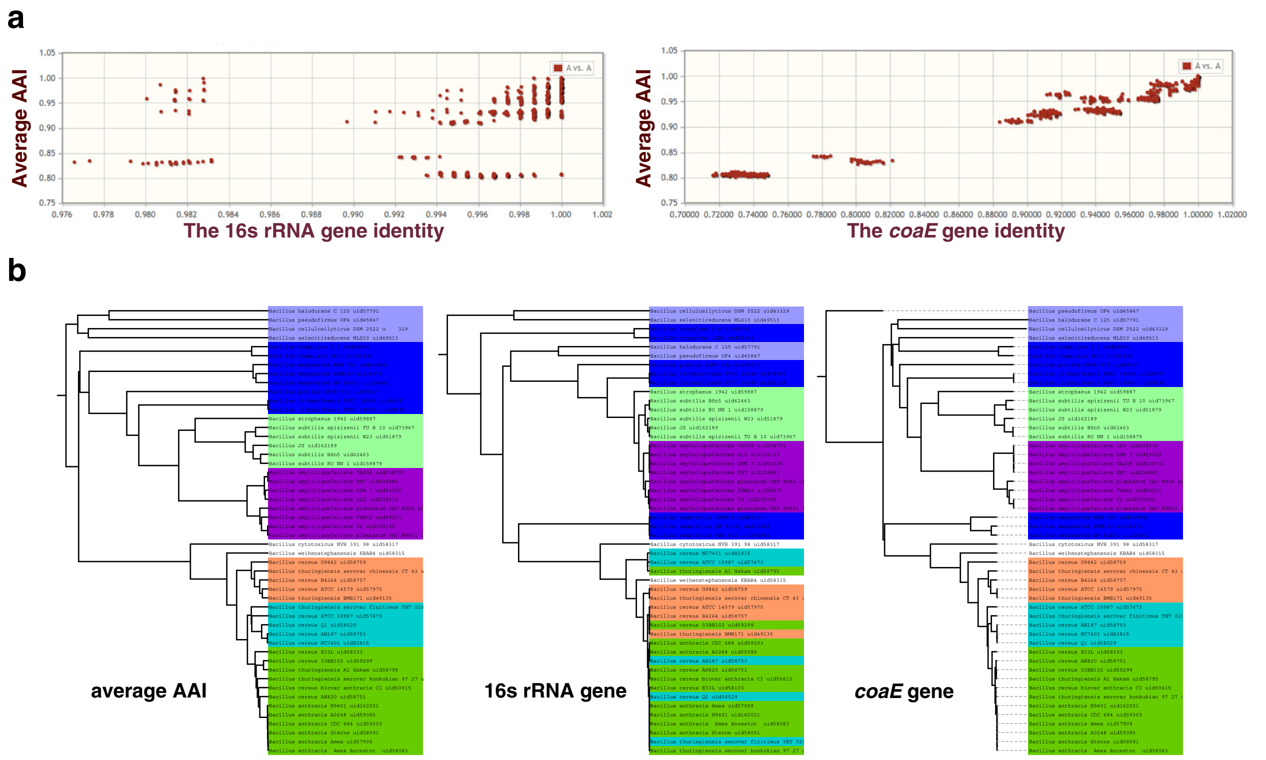

Fig. 6 Comparison of the 16s rRNA gene and the *coaE* gene within the *Bacillus* lineage, for **a** the gene identity of 16s rRNA or *coaE* vs. the average AAI and **b** tree reconstructed from 16s rRNA or *coaE* gene in comparison with the tree reconstructed from AAI. As the least sequence-conserved marker gene within the *Bacillus* lineage, *coaE* correlates well with the AAI in percent identities and is able to reconstruct the tree resembling that reconstructed from AAI

ones that have, per definition, acquired more changes in their sequences. These changes can include indels and multiple hits at the same exact sites, which will make alignments and subsequent quantification of levels of similarity less reliable at larger evolutionary distances. For such distantly related genomes, levels of similarity between more sequence-conserved markers, such as 16s rRNA, will result in better inference of genome-wide levels of similarity. At the same time, the fact that less sequence-conserved markers accumulate more substitutions gives them more power to detect differences between closely related strains. This makes them more useful for predicting genome-wide similarity at short distances.

Prokaryotic genomes tend to undergo substantial horizontal gene transfer (HGT). Such HGT can greatly dilute any signal pertaining to the relationship between strains and lead to false inferences. This is true no matter what method of inference is used, be it 16s rRNA-based, based on other markers, or even based on whole microbiome or whole genome data (such as AAI). While whole genome (AAI)-based classification can therefore also be quite noisy, it nevertheless provides us with the best possible inference of the extent of similarity between genomes, since they utilize data from all available loci, rather than from a specific gene alone. Using data from entire genomes is not, however, always feasible, especially not in microbiome studies in which most bacteria cannot be cultured. Shotgun metagenome studies allow one to combine information from large number of loci, but in a much more complex way, since rarely will multiple loci be sequenced from exactly the same strain of prokaryote. Additionally, such studies are often beyond the budget of many researchers. This is why 16s rRNA has become so useful for microbiome studies, and this is also why other marker genes may be useful. However, a concern was raised that markers other than 16s rRNA may be less useful, because they may undergo more HGT than 16s rRNA [28]. Here we show that even under the potential impact of HGT, some less sequence-conserved markers outperform 16s rRNA, in predicting genome-wide levels of gene similarity (as estimated using genome-wide AAI data), for closely related bacterial lineages. This in turn may suggest that as a group, these markers do not undergo more HGT than 16s rRNA within the lineages in which they outperform 16s rRNA.

We find that markers that provide the best resolution, for a specific lineage, are those that are least conserved in their sequences within that lineage. Therefore, less sequence-conserved markers can be chosen for specific lineages of interest that would provide more accurate reconstruction of the diversity in ecosystems and allow for higher resolution microbiome studies. In order to facilitate easy marker choice, we have built the POGO database (Database of Pairwise-comparisons Of Genomes and conserved Orthologous genes) [59]. POGO-DB allows users to rank markers according to their relative levels of sequence conservation within a lineage of interest and also allows users to compare the percent identity of different markers to AAI for the same lineage. Therefore, users can use POGO-DB to verify whether the 16s rRNA gene has good correlation to the AAI within their lineages of interest or to select the best alternative markers for the lineages in which they are interested, in an informed manner.

An advantage of using 16s rRNA has been that its conservation is such that universal primers can be designed to amplify its sequence across all prokaryotes. This is likely not possible for most other marker genes and especially not for the ones that are least conserved in their sequences. However, as we demonstrate, the less sequence-conserved marker genes are most useful for inferring finer grained strain differences for particular lineages of interest. Therefore, in order to utilize these markers, where they are most useful, it is only necessary to design primers that can recognize areas of the marker genes' sequences that are conserved within a lineage of interest. To demonstrate the feasibility of designing such primers, we designed 10 primer sets for the 10 marker genes that are least conserved in their sequences in the *Escherichia/Shigella* lineage (Additional file 1: Table S4). These primer sets were designed to allow for the amplification of these 10 markers across all sequenced *Escherichia* and *Shigella* strains. For 7 of the 10 markers, the designed primers enable the amplification of over 70 % of the full gene length. In order to assist researchers in designing lineage-specific primers for markers of interest, we have added a feature to POGO-DB that allows users to download the sequences of a marker gene from a list of specified genomes. These sequences can then be aligned to identify regions that are conserved within the lineage of interest and that can be used to design primers that will amplify the gene within that lineage.

Availability of data and materials

The data set supporting the results of this article is available in the POGO database, http://pogo.ece.drexel.edu.

Additional file

Additional supplementary information may be found in the online version of this article.

Additional file 1: Figure S1. The number of genomes in which each marker gene is identified. Out of the 79 potential marker genes, 73 are present in at least 90 % of the genomes. **Figure S2.** Spearman's correlation between each marker gene and the average AAI for all complete genomes. Genes are ordered in the same way as in Fig. 3. **Figure S3.** Trees generated based on AAI and on percent identities of each marker gene (including 16s rRNA), for the *Escherichia/Shigella* clade. **Figure S4.** Trees generated based on AAI and on percent identities of each marker gene (including 16s rRNA), for the *Streptococcus* clade. **Figure S5.** Trees generated based on AAI and on percent identities of each marker gene (including 16s rRNA), for the *Bacillus* clade. **Table S1.** List of 79 potential marker genes surveyed, out of which 73 were found to be present in at least 90 % of the genomes. **Table S2.** Alternative names of 79 potential marker genes surveyed. **Table S3.** Split distances between UPGMA tree generated using AAI and that generated using the percent identities of each marker gene, shown in correspondence with the average percent identity ranks of the marker genes. **Table S4.** Designed primers for each of the 10 genes that were least conserved in their sequences in the *Escherichia/Shigella* lineage. (ZIP 3355 kb)

Abbreviations

AAI: amino acid identity; HGT: horizontal gene transfer; UPGMA: Unweighted Pair Group Method with Arithmetic Mean.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YL carried out all analyses described in the manuscript and co-wrote the manuscript. GR and RH designed and coordinated the study and co-wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Work reported here was run on hardware supported by Drexel's University Research Computing Facility. This work initiated with funds provided by the Louis and Bessie Stein Family Fellowship to support exchanges between Drexel and Israeli Universities. The work performed by the Drexel Ecological and Evolutionary Signal-processing and Informatics (EESI) lab was also supported in part by a National Science Foundation (NSF) CAREER award number #0845827, NSF award number #1120622, and Department of Energy (DOE) Office of Science (BER) award DE-SC0004335. RH was supported by an ERC FP7 CIG grant (No. 321780), by a BSF grant (No. 2013463), by a Yigal Allon Fellowship awarded by the Israeli Council for Higher Education, and by the Robert J. Shillman Career Advancement Chair. Research by RH is carried out in the Rachel & Menachem Mendelovitch Evolutionary Process of Mutation & Natural Selection Research Laboratory.

Author details

[1]School of Biomedical Engineering, Science and Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. [2]Ecological and Evolutionary Signal-processing and Informatics Laboratory, Electrical & Computer Engineering Department, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. [3]Rachel & Menachem Mendelovitch Evolutionary Processes of Mutation & Natural Selection Research Laboratory, Department of Genetics and Developmental Biology, the Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, 31096 Haifa, Israel.

Received: 20 August 2015 Accepted: 22 March 2016
Published online: 03 May 2016

References

1. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004;38:525–52. doi:10.1146/annurev.genet.38.072902.091216.

2. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12:69. doi:10.1186/S12915-014-0069-1.
3. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, et al. The Human Microbiome Project: a community resource for the healthy human microbiome. Plos Biol. 2012;10(8):e1001377. doi:10.1371/Journal.Pbio.1001377.
4. Shade A, Caporaso JG, Handelsman J, Knight R, Fierer N. A meta-analysis of changes in bacterial and archaeal communities with time. ISME J. 2013;7(8):1493–506. doi:10.1038/ismej.2013.54.
5. Kittelmann S, Seedorf H, Walters WA, Clemente JC, Knight R, Gordon JI, et al. Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. PLoS One. 2013;8(2):e47879. doi:10.1371/journal.pone.0047879.
6. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16s rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010;76(12):3886–97. doi:10.1128/Aem.02953-09.
7. Tian RM, Cai L, Zhang WP, Cao HL, Qian PY. Rare events of intra-genus and intra-species horizontal transfer of the 16s rRNA gene. Genome biology and Evolution. 2015. doi:10.1093/gbe/evv143.
8. Janda JM, Abbott SL. 16s rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761–4. doi:10.1128/Jcm.01228-07.
9. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37:D141–5. doi:10.1093/Nar/Gkn879.
10. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72. doi:10.1128/Aem.03006-05.
11. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590–6. doi:10.1093/nar/gks1219.
12. Vos M, Quince C, Pijl AS, de Hollander M, Kowalchuk GA. A comparison of rpoB and 16s rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS One. 2012;7(2):e30600. doi:10.1371/journal.pone.0030600.
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. doi:10.1371/journal.pone.0057923.
14. Ong SH, Kukkillaya VU, Wilm A, Lay C, Ho EXP, Low L, et al. Species identification and profiling of complex microbial communities using shotgun Illumina sequencing of 16s rRNA amplicon sequences. PLoS One. 2013;8(4):e60811. doi:10.1371/journal.pone.0060811.
15. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG et al. Oligotyping: differentiating between closely related microbial taxa using 16s rRNA gene data. Methods Ecol Evol. 2013;4(12). doi:10.1111/2041-210X.12114.
16. Zaneveld JR, Lozupone C, Gordon JI, Knight R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res. 2010;38(12):3869–79. doi:10.1093/Nar/Gkq066.
17. Kitahara K, Miyazaki K. Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16s rRNA. Mob Genet Elements. 2013;3(1):e24210. doi:10.4161/mge.24210.
18. Achenbach LA, Carey J, Madigan MT. Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. Appl Environ Microbiol. 2001;67(7):2922–6. doi:10.1128/AEM.67.7.2922-2926.2001.
19. Walsh DA, Bapteste E, Kamekura M, Doolittle WF. Evolution of the RNA polymerase B′ subunit gene (rpoB′) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. Mol Biol Evol. 2004;21(12):2340–51. doi:10.1093/Molbev/Msh248.
20. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. Use of 16s rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007;73(1):278–88. doi:10.1128/Aem.01177-06.
21. Torriani S, Felis GE, Dellaglio F. Differentiation of Lactobacillus plantarum, L. pentosus, and L. paraplantarum by recA gene sequence analysis and multiplex PCR assay with recA gene-derived primers. Appl Environ Microbiol. 2001;67(8):3450–4. doi:10.1128/AEM.67.8.3450-3454.2001.
22. Verbekea TJ, Sparling R, Hill JE, Links MG, Levin D, Dumonceaux TJ. Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (cpn60 UT): application to Thermoanaerobacter species. Syst Appl Microbiol. 2011;34:171–9. doi:10.1016/j.syapm.2010.11.019.
23. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008;9(10):R151. doi:10.1186/Gb-2008-9-10-R151.
24. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12 Suppl 2:S4. doi:10.1186/1471-2164-12-S2-S4.
25. Segata N, Huttenhower C. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS One. 2011;6(9):e24704. doi:10.1371/journal.pone.0024704.
26. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10(12):1196–9. doi:10.1038/nmeth.2693.
27. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50. doi:10.1038/nature11711.
28. Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006;311:1283–7. doi:10.1126/science.1123061.
29. Zeigler DR. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol. 2003;53:1893–900. doi:10.1099/ijs.0.02713-0.
30. Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol. 2005;1:106–12. doi:10.1371/journal.pcbi.0010024.
31. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6:805–14. doi:10.1038/nrg1709.
32. Bragg L, Tyson GW. Metagenomics using next-generation sequencing. Methods Mol Biol. 2014;1096:183–201. doi:10.1007/978-1-62703-712-9_15.
33. Koslicki D, Foucart S, Rosen G. WGSQuikr: fast whole-genome shotgun metagenomic classification. PLoS One. 2014;9:e91784. doi:10.1371/journal.pone.0091784.
34. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4. doi:10.1038/nmeth.2066.
35. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243. doi:10.7717/peerj.243.
36. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28:1033–4. doi:10.1093/bioinformatics/bts079.
37. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 2011;12:R44. doi:10.1186/gb-2011-12-5-r44.
38. Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, et al. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput Biol. 2011;7:e1001061. doi:10.1371/journal.pcbi.1001061.
39. Konstantinidis KT, Ramette A, Tiedje JM. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl Environ Microbiol. 2006;72:7286–93. doi:10.1128/AEM.01398-06.
40. 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. doi:10.1186/1471-2105-4-41.
41. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.
42. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2005;33(Database issue):D34–8. doi:10.1093/nar/gki063.
43. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc. 2007;2007:pdb top17. doi:10.1101/pdb.top17.
44. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
45. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005;187(18):6258–64. doi:10.1128/JB.187.18.6258-6264.2005.
46. Puigbo P, Garcia-Vallve S, McInerney JO. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics. 2007;23(12):1556–8. doi:10.1093/bioinformatics/btm135.
47. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R. Using QIIME to analyze 16s rRNA gene sequences from microbial communities. Curr Protoc Microbiol. 2012;Chapter 1:Unit 1E 5. doi:10.1002/9780471729259.mc01e05s27.
48. Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM. cpnDB: a chaperonin sequence database. Genome Res. 2004;14(8):1669–75. doi:10.1101/gr.2649204.

49. Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE. The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. PLoS One. 2012;7(11):e49755. doi:10.1371/journal.pone.0049755.
50. Hershberg R, Tang H, Petrov DA. Reduced selection leads to accelerated gene loss in Shigella. Genome Biol. 2007;8(8):R164. doi:10.1186/Gb-2007-8-8-R164.
51. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, et al. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis—one species on the basis of genetic evidence. Appl Environ Microbiol. 2000;66(6):2627–30.
52. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology—human gut microbes associated with obesity. Nature. 2006;444(7122):1022–3. doi:10.1038/Nature4441022a.
53. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4. doi:10.1038/nature07540.
54. Yang F, Zeng XW, Ning K, Liu KL, Lo CC, Wang W, et al. Saliva microbiomes distinguish caries-active from healthy human populations. ISME J. 2012;6(1):1–10. doi:10.1038/Ismej.2011.71.
55. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, et al. Deep sequencing of the oral microbiome reveals signatures of periodontal disease. PLoS One. 2012;7(6):e37919. doi:10.1371/journal.pone.0037919.
56. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65. doi:10.1038/nature08821.
57. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174–80. doi:10.1038/nature09944.
58. Siddiqui H, Nederbragt AJ, Lagesen K, Jeansson SL, Jakobsen KS. Assessing diversity of the female urine microbiota by high throughput sequencing of 16s rDNA amplicons. BMC Microbiol. 2011;11:244. doi:10.1186/1471-2180-11-244.
59. Lan YM, Morrison JC, Hershberg R, Rosen GL. POGO-DB-a database of pairwise-comparisons of genomes and conserved orthologous genes. Nucleic Acids Res. 2014;42(D1):D625–32. doi:10.1093/Nar/Gkt1094.