FastAAI: Efficient Estimation of Genome Average Amino Acid Identity and Phylum-level relationships using Tetramers of Universal Proteins

Konstantinos Konstantinidis (kostas.konstantinidis@gatech.edu)
Georgia Institute of Technology
Carlos Ruiz-Perez
Georgia Institute of Technology
Kenji Gerhardt
Georgia Institute of Technology https://orcid.org/0000-0003-0644-1862
Luis Rodriguez-R
University of Innsbruck https://orcid.org/0000-0001-7603-3093
Chirag Jain
Indian Institute of Science https://orcid.org/0000-0002-4300-0794
James Tiedje
Michigan State University https://orcid.org/0000-0002-8992-6218
James Cole
Michigan State University

Article

Keywords:

Posted Date: March 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1459378/v1

FastAAI: Efficient Estimation of Genome Average Amino Acid Identity and Phylum-level relationships using Tetramers of Universal Proteins

Carlos A Ruiz-Perez[1*], Kenji Gerhardt[1*], Luis M Rodriguez-R[2,3*], Chirag Jain[4], James M. Tiedje[5], James R Cole[5], and Konstantinos T Konstantinidis[1,2,*]

[1]School of Biological Sciences and [2]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

[3]Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, Innsbruck 6020, Austria.

[4]Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, KA 560012, India

[5]Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA.

*Contributed equally

[*]To whom correspondence should be addressed.

Abstract

Estimation of whole-genome relatedness and taxonomic identification are two important bioinformatics tasks in describing environmental or clinical microbiomes. The genome-aggregate Average Nucleotide Identity (ANI) is routinely used to derive the relatedness of closely related (species level) microbial and viral genomes, but it is not appropriate for more divergent genomes. Average Amino Acid Identity (AAI) can be used in the latter cases, but no current AAI implementation can efficiently compare thousands of genomes. Here we present FastAAI, a tool that estimates whole-genome pairwise relatedness using shared tetramers of universal proteins in a matter of microseconds, providing a speedup of up to 5 orders of magnitude when compared with current methods of calculating AAI or alternative whole-genome metrics. Further, FastAAI resolves distantly related genomes related at the phylum level with comparable accuracy to the phylogeny of ribosomal rRNA genes, substantially improving on a known limitation of current AAI

implementations. Therefore, Fast AAI uniquely expands the toolbox for microbiome analysis and allows it to scale to millions of genomes.

Introduction

The diversity of microbial and viral genomes on the planet is very large, e.g., estimated at over a billion species of (just) bacteria, and most of it remains undiscovered[1]. As genome sequencing can help characterizing this diversity and has recently become routine, most microbial scientists have been overwhelmed by the amount of available data. For many researchers, a complete and thorough analysis of the available genome data is not necessary. Instead, these users would be better served with straightforward methods that can identify their unknown DNA/RNA sequence(s) and be able to discriminate between well-understood taxa and those that are potentially novel. Thus, tools that can help researchers identify the most "interesting" or relevant genomes to their unknown/query genome(s) among thousands of candidates are important. Furthermore, metagenomics, which is the sequencing of environmental DNA, allows the genetic characterization of the majority of these microbes that have resisted cultivation in the laboratory. However, current tools to analyze metagenomic data are clearly lagging behind the development of sequencing technologies (and data), and do not scale with the number of metagenome-assembled genomes (MAGs), currently in the order of hundreds of thousands, that have become available[2]. This is a major limitation for better understanding, studying, and communicating about the biodiversity of uncultivated microorganisms that run the life-sustaining biogeochemical cycles on the planet, form critical associations with their plant and animal hosts, or produce products of biotechnological value.

One commonly used pair of methods to determine the level of novelty of a newly described taxon relative to the taxa already available in reference databases are genome-aggregate average nucleotide (ANI) and amino-acid (AAI) identity. ANI represents the average nucleotide identity of all genes shared between any two genomes and offers robust resolution between strains of the same or closely related species (i.e., showing 80-100% ANI). However, ANI is not appropriate for estimating genome relatedness

(and thus, classification) among more divergent genomes because nucleotide substitutions saturate and/or sequences cannot be reliably aligned at this level[3]. Instead, the average amino acid identity (AAI) should be used for moderately divergent genomes as it effectively delineates relationships at the genus and family levels, similarly to ANI for the species level. It is also important to note that the ANI/AAI-based approach can be applied to any unknown genome sequence, even if it is incomplete, that encodes at least a few genes shared with at least a few reference genomes[4], whereas alternative approaches based on universal genes such as the small subunit ribosomal rRNA (SSU rRNA) gene are limited to sequences carrying the corresponding genes. Accordingly, the ANI/AAI approach has facilitated the analysis of an increasing number of MAG, single-cell amplified genome (SAG) or viral genome sequences and the need to perform whole-genome ANI or AAI calculations has grown exponentially in recent years.

Despite these strengths, traditional ANI/AAI calculation is based on pairwise sequence alignments, e.g., BLAST[5], and thus do not scale well with an increasing number of genomes. For instance, performing all vs. all ANI calculation among the ~13,000 reference genomes (~169 million pairwise comparisons) of all bacterial species taxonomically described to date[6] required ~1 month of processing time on a computer cluster of 500 CPUs. Performing a similar comparison among 100,000 genomes (10 billion comparisons) is expected to take more than 1 year due to the exponential increase in the number of comparisons to perform. This is prohibitively expensive, even for larger computer clusters. While a faster ANI implementation based on the numbers of k-mers shared between genomes (FastANI) has been recently described by our team[4], and has been widely used, current AAI implementations are not fast enough, especially considering that the majority of genomes in the reference database are moderately or distantly related to each other. Further, AAI does not provide robust resolution at the inter-domain and phylum levels (that is, between distantly related genomes) compared to the phylogenetic analysis of the SSU rRNA or other universal genes. Alternative approaches such as phylogenetic placement based on universal genes are accurate [7, 8], but are also computationally expensive and do not scale well with the exponentially increasing number of microbial genomes recovered from environmental samples. Thus,

there is an urgent need for a scalable tool to calculate genome relatedness, especially for deep-branching genomes with no close (e.g., same species) relatives available.

Here we present FastAAI, a bioinformatics tool aiming at solving the issues related to the speed and scalability of the AAI estimation among bacterial and archaeal genomes while providing higher resolution at the phylum and domain level. FastAAI is a standalone Python tool that leverages protein tetramer information in single-copy proteins (SCPs) "universally" present in microbial genomes to calculate genome similarities that are subsequently translated into "traditional" AAI values of relatedness. The current implementation of FastAAI can perform pairwise genome comparisons in a matter of microseconds and scales exceedingly well with a larger number of genomes compared.

Results

We focus our analysis below primarily on bacterial genomes because of their large numbers in the public datasets. FastAAI applies similarly to archaeal genomes but for genomes of other domains (e.g., eukaryotic) adjustments in the collection of universal single-copy proteins will be required for best results. For our accuracy evaluations, we primarily focus on the comparison of FastAAI to 16S rRNA gene identities since the latter gene represents the backbone of bacterial taxonomy and has been most commonly used to assess genetic relatedness among taxa/genomes. We provide computational and performance comparisons to traditional alignment-based methods such as DIAMOND[9] and EzAAI[10] as well as alternative methods such as GTDB.

FastAAI accurately estimates AAI

FastAAI is intended to replace traditional AAI in resolving relationships above the species level; at and within the species level, ANI is sensitive and efficient enough (e.g., FastANI implementation) and thus, already serves as a complementary method to AAI. Therefore, we focused our performance evaluation of FastAAI on genomes with a relatedness between 30% and 90% AAI. Using the RefSeq dataset (see

Methods), we calculated a "reference" AAI value for each pair of genomes using DIAMOND and compared predictions made with FastAAI against these values. FastAAI's estimator $\widehat{AAI}$ (see Methods for details on the calculation of $\widehat{AAI}$) achieves a highly accurate estimation of DIAMOND-based AAI between 30% and 90% $\widehat{AAI}$, with an average prediction error of only 1.3% within this range (R-squared: 0.98, RMSE: 1.3%; Figure 1). The strong linear correlation between $\widehat{AAI}$ and DIAMOND-based AAI values demonstrates that FastAAI can provide a similarly robust estimate of genome relatedness above the species level (Figure 2B and 2C). For the most distantly related genomes (e.g., related at around 30% AAI), the $\widehat{AAI}$ and DIAMOND-based AAI values deviated the most from each other, which we investigated below with comparisons against the 16S rRNA gene and showed that this was due to higher accuracy of $\widehat{AAI}$ at the domain and phylum levels.

FastAAI is faster than alternative AAI implementations

To demonstrate the gains in speed of FastAAI compared to alignment-based methods for calculating AAI, we performed an all-vs-all comparison of 100 randomly chosen RefSeq genomes (10,000 pairwise comparisons) using FastAAI, DIAMOND-based AAI, and the recently published EzAAI. EzAAI uses MMseqs2[11] as a substitute for either BLAST or DIAMOND as its alignment engine, but calculates AAI in a fundamentally similar manner; that is, finding reciprocal best-match alignments between the proteins of each pair of genomes it processes and calculating AAI based on these best matches. As DIAMOND does not independently function as a pipeline for both protein prediction and alignment in the manner that FastAAI and EzAAI do, all three tools were supplied with the same set of predicted proteins generated from the original 100 RefSeq genomes using Prodigal. Timing covered all of the remaining comparable steps of each tool, including data formatting, database creation as needed, the calculation of AAI, and writing results.

The average time per pairwise genome comparison was 9.6 seconds (S.D. = 4.16) for DIAMOND, 32.0 seconds (S.D. = 12.1) for EzAAI, and 0.031 seconds (S.D. non-applicable) for FastAAI. Although this

test already demonstrates more than two orders of magnitude speedup over both DIAMOND and EzAAI, it is in fact a nearly worst-case scenario for demonstrating the speed gains of FastAAI relative to these tools. As can be seen in Figure S1, the relative speed gain per comparison of FastAAI over DIAMOND and EzAAI increases as the datasets processed grow larger, with FastAAI becoming nearly 5 orders of magnitude faster per comparison than either DIAMOND or EzAAI in our largest test dataset (~2.29 billion pairwise comparisons) when a run started from predicted proteins.

FastAAI is faster and has a lower RAM requirement than GTDB-tk

There are alternative methods to measure relatedness among genomes and/or taxonomically classify them aside from the traditional ANI/AAI; most notably, GTDB-tk[12] uses a combination of approximate, maximum-likelihood-based phylogenetic placement and FastANI to classify query genomes. Because GTDBtk can use only its own prebuilt database, we created an equivalent FastAAI database from the same set of genomes that are used in the latest GTDB database (rel. 202, last updated Jul. 9th, 2021), totaling 45,555 bacterial and 2,339 archaeal genomes. We randomly sampled three sets of 100 genomes from these 47,894 genomes and used these sets to test both GTDBtk and FastAAI by querying each set of 100 selected genomes against the full complement of 47,894 target genomes. To ensure that the comparison between FastAAI and GTDBtk was fair, both tools were run with 10 threads, supplied with 256 GB of RAM, used their respective premade databases as search targets, and were supplied with query sets composed of genomes (not proteins or HMM results), as this is the starting point for a GTDBtk search. In addition to memory usage, we measured both the walltime time and CPU time for each tool, reflecting the user's waiting time from program start to program end and the sum of computer time across all 10 threads, respectively. On average, FastAAI completed in 9.8% of the runtime, 16.9% of the CPU time, and required only 0.4% of the RAM required by GTDBtk. All comparison results can be seen in Table 1. All times are reported as HH:MM:SS.

Table 1. FastAAI vs GTDBtk runtime and resource usage; 100 genomes vs. GTDB Release 202			
Test	**Walltime**	**CPU Time**	**Max RAM use (GB)**
FastAAI, Sample 1	00:07:46	00:46:28	0.81
GTDBtk, Sample 1	01:26:38	05:18:55	232.08
FastAAI, Sample 2	00:07:51	00:50:11	0.87
GTDBtk, Sample 2	01:17:41	04:12:42	231.67
FastAAI, Sample 3	00:08:31	00:54:35	0.98
GTDBtk, Sample 3	01:22:38	05:24:31	232.10

To further illustrate the minimal RAM required by FastAAI, we ran an additional all-vs-all tests with FastAAI on varying sizes of random samples of the GTDB genomes, ranging in size from 50 vs. 50 to the entirety of GTDB release 202 (47,894 vs 47,894 genomes). The whole-GTDB all-vs-all test was the longest-running and most memory intensive of these tests (~2.29 billion pairwise comparisons), which FastAAI was able to complete in a walltime of 1:28:27, a CPU time of 10:33:07, and using only 4.21 GB of RAM. Provided a preprocessed database (i.e., universal proteins identified and tetramers extracted), even a comparison of this enormous scale could be accomplished with FastAAI within a day, on a typical personal laptop. Performing the same task with GTDB-tk would take well in excess of 500 CPU hours and require hundreds of gigabytes of RAM that are available only on supercomputers. The results for the other tests can be seen in Table S1.

FastAAI vs. 16S rRNA gene identities and phylum-level resolution

Our comparison of DIAMOND-based AAI and 16S rRNA gene identities using the bacterial RefSeq genomes demonstrated a strong correlation (Figure S2A), with overlaps between adjacent taxonomic ranks in terms of the AAI values (e.g., about 30% of the genomes grouped at the genus level as their lowest shared rank show similar AAI values to genomes grouped at the family level), consistent with

the patterns observed in previous studies based on smaller genome datasets[13]. In these comparisons, we observed that the lowest values for 16S rRNA gene identities (70-80%) corresponded to (traditional) AAI values between 30-40%. However, the same range in AAI values (30-40%) also included genome pairs with higher 16S rRNA gene identities (80-85%), revealing substantial overlaps in AAI values (i.e., low resolution) at higher taxonomic levels, as noted also previously[13]. However, in the case of FastAAI, 16S rRNA gene identities between 70 and 80% corresponded to FastAAI's $\widehat{AAI}$ values between 32.28 and 33.57 with almost no overlap with same- vs. different-domain level comparisons (Figure S2B), revealing higher resolution of FastAAI at the phylum-domain level, comparable to that of 16S rRNA gene identities [see also Fig. S3 for discussion of a few outlier genome pairs observed.

More specifically, for 16S rRNA gene identities, we identified the inter-domain and inter-phylum valleys at 70.4 and 80.2%, respectively (Figure 2A). We observed that only 1.1% of same-domain comparisons were below 70.4%, while only 0.54% of different-domain comparisons were above this threshold, indicating limited overlap and good discrimination capacity at the domain level based on 16S rRNA gene identities. However, the same values for phylum-level comparisons revealed a lower discriminatory capacity; 34.7% of the comparisons were below the 80.2% threshold and 10.9% were above. Note that phyla are typically designated based on their branching patterns in the 16S rRNA gene phylogeny, not necessarily the 16S rRNA gene identity; hence, the results reported here do not necessarily reflect low performance of the 16S rRNA gene phylogeny approach. Our goal was to examine if convenient sequence identity thresholds for discriminating taxonomic ranks can determined that could guide future taxonomic studies and description and to evaluate FastAAI's domain- and phylum-level resolution. Compared to 16S rRNA gene identity values, we found that the thresholds discriminating domain and phylum-level comparisons using FastAAI were 35.3% and 40.5% $\widehat{AAI}$. Interestingly, we found that only 0.07% of different-domain and 0.03% of same-domain comparisons were below and above the domain-threshold (35.3%), indicating similar -if not higher- discriminatory power at this level for $\widehat{AAI}$ relative to the 16S rRNA gene. Phylum-level comparisons were nonetheless very similar to those observed for 16S rRNA with

20.4% same-phylum and 6.47% different-phylum comparisons below and above the identified threshold (40.5%), respectively. These results indicate that FastAAI offers a lower frequency of misclassifications (or higher resolving power) than (traditional) AAI and 16S rRNA gene identities at the domain and phylum level for deep-branching genomes. It should be also noted that the large number of genome pairs sharing the same phylum rank but showing identities below the discrimination threshold in all three metrics indicates that at least some of the corresponding genomes may deserve a distinct (novel) phylum designation based on genomic relatedness. We evaluated the remaining taxonomic ranks e.g., family and genus in a similar fashion, and showed that reliable FastAAI thresholds can be established for each rank, and that FastAAI's classification based on the best-matching genome largely (~95% of the cases) agreed with more sensitive tree placement classification approaches (see Supplementary Results for details).

FastAAI is robust to varying levels of genome completeness and contamination

Most studies using culture-independent genomic techniques filter recovered prokaryotic genomes according to their quality level based on estimated completeness and contamination[14]. Gene phylogenies or classification methods that rely on a single or few genes are often affected by genome incompleteness due to the absence of the required gene(s) for the analysis. In addition, contaminating gene sequences can be sources of error that can affect homology-based classification methods such as MEGAN or MyTaxa[15, 16]. In such cases, AAI is generally more robust because the number of proteins or genome fragments used in the calculations may be adequate even for relatively incomplete genomes (e.g., ~200 shared proteins are typically adequate[4]). Given that FastAAI relies on a reduced set of universal proteins, it is important to assess the consistency of the results at variable levels of genome quality. For this, we selected 500 genomes from the RefSeq dataset that were estimated to be 100% complete with 0% contaminated. We predicted the proteins for each selected genome and used FastAAI to calculate their genome relatedness against the smaller RefSeq dataset to create a baseline of reference $\widehat{AAI}$ values. Each set of proteins per genome was then randomly subsampled and *in-silico* "contaminated" to create a set of genomes with varying levels of contamination and completeness (see methods, Figure S3A). Each of these incomplete and contaminated

genomes was then searched against the RefSeq dataset and the resulting $\widehat{AAI}$ values of these searches were compared to the complete-genome results (baseline) to determine the degree of deviation from the baseline. Our results demonstrated that in the absence of contamination, FastAAI could recover almost identical $\widehat{AAI}$ values for genomes with varying levels of completeness, even down to 10% complete genomes, only deviating by, at most, 2.26% from the complete-genome-based $\widehat{AAI}$ values in the genomes of the lowest completeness (Figure S3B). This result showed that FastAAI is robust even for incomplete genomes, as long as about 8 of the universal proteins are present (roughly corresponding to 10% completeness). As expected, the contamination resulted in an increased deviation of $\widehat{AAI}$ values from the baseline. In the 100% completeness genomes, the deviation increased from 0% (100% completeness, 0% contamination) to 0.95% (100% completeness, ~37% contamination), a relatively small value considering the high level of contamination. This trend was also observed in lower completeness levels, where we observed the highest values of deviation, ranging from 2.26% to 3.98% $\widehat{AAI}$ in genomes with 10% completeness and contamination of up to 37% (a very low-quality genome; Figure S3B). Therefore, considering that proposed classification standards for the quality of assembled genomes establish a low quality-draft as a genome with <50% completeness and <10% contamination[14], we expect FastAAI to perform well, with less than 1% $\widehat{AAI}$ deviation from the true estimate, for such low quality genomes.

Discussion

This study presented FastAAI, a tool for the fast and accurate estimation of the whole-genome genetic relatedness and classification of microbial genomes that provides comparable results, if not better, with traditional methods such as AAI, and SSU rRNA gene identities, while dramatically improving in terms of speed and resolution. We also demonstrated substantial advantages over newer alternative tools for the classification of microorganisms in terms of speed and computational resources. The high accuracy and speed, especially in comparisons among deep-branching genomes (i.e., higher taxonomic levels), will be important for environmental and clinical samples that are expected to harbor substantial novel

diversity[17]. FastAAI can effectively replace the abovementioned methods for genome comparisons at the genus level and above, while reverting to the traditional AAI, or even better ANI, for closely related genomes with high FastAAI identities (species level). Thus, species-level comparisons should instead be resolved with FastANI or regular ANI comparisons, and particularly novel genomes with no apparent close relatives in the reference databases at the (same) genus level can be further processed with FastAAI and/or using phylogenetic methods when number of genomes is tractable.

FastAAI accelerates the calculation of AAI because it does not require the alignment of proteins between genomes, and instead estimates AAI between two genomes from the shared fraction of tetramers over a set of universal SCPs shared by the two genomes. This allows FastAAI to rapidly calculate AAI from massive datasets: even in all-vs-all comparisons of tens of thousands of genomes (i.e., hundreds of millions or billions of pairwise comparisons), FastAAI requires only a few hours of runtime and very modest computational resources to complete its estimation of AAI. Indeed, the overwhelming majority of runtime for FastAAI in all of our applicable tests is consumed in the preprocessing stage, i.e., protein prediction and HMM searches to identify SCPs, and the average time per comparison between a pair of genomes decreases as the number of comparisons performed increases (Figure S1). This speedup is due to the fact that even though the total number of pairwise genome comparisons increases with the square of the number of genomes, the slow pre-processing step grows linearly and only the number of fast tetramer pairwise comparisons increases with the square of the genome number. Similarly, comparison of a single novel genome to a large, preprocessed reference data set requires additional preprocessing of only the novel genome, which dominates comparison times against even our large data sets. It is important to note, however, that the preprocessing step is performed only once for each genome and can be stored in order to accelerate subsequent AAI calculations.

Notably, FastAAI retains its performance even in the presence of highly incomplete or contaminated genomes, which are commonly recovered by metagenomics surveys. Finally, from a database perspective, FastAAI retrieves the closest relative in most cases even using a highly incomplete database, which we expect that will only improve as databases become more comprehensive with genome

representatives from uncultured microorganisms. Therefore, we expect FastAAI to be a valuable tool for the scientific community aimed at the fast and accurate genome relatedness estimation and classification of microorganisms from clinical, industrial, or environmental sources.

Online Methods

FastAAI implementation

FastAAI is divided into four steps (Figure S4A). Briefly, FastAAI performs, (*i*) identification of marker proteins in each genome using a set of 122 universal single-copy proteins (SCP) represented by hidden Markov models (HMMs) (Table S2, and below), (*ii*) extraction of tetramers for each protein per genome, (*iii)* estimation of Jaccard indices based on the tetramers extracted from proteins shared by a genome pair, (*iv*) estimation of the average Jaccard index ($\bar{J}$) followed by its translation to estimated AAI values (termed $\widehat{AAI}$ to distinguish it from AAI values estimated based on alignment-based methods). The first step is performed by searching all proteins of a genome against a collection of HMMs with HMMer 3.0[18] using profile trusted cutoffs (*--cut_tc*). These models were built from a set of 122 proteins typically found in most prokaryotic genomes in a single copy[19]. The second step consists of extracting all tetramers per marker protein. Tetramers are considered as present or absent regardless of the number of times they occur in a protein sequence. In the third step, for each pair of genomes, FastAAI identifies shared marker proteins; if a protein is only present in one genome, it is excluded from the comparison. Then, for each shared marker protein in the genome pair under comparison, a Jaccard similarity value is calculated based on the shared vs. total tetramers of the corresponding protein sequences (Figure S4B). Finally, FastAAI calculates the average of the previously estimated Jaccard values for all shared proteins ($\bar{J}$), the standard deviation, and uses $\bar{J}$ values to estimate AAI values ($\widehat{AAI}$; read below for estimation formula). In cases where the inputs are DNA genome sequences, the program predicts protein sequences using Prodigal[20] and automatically selects the optimal translation table (11 or 4).

Testing Datasets

To test performance and speed, we evaluated FastAAI on three primary datasets (Table 2). The first dataset (RefSeq) comprised all complete reference genomes from the NCBI RefSeq database with available taxonomic classification as of 11-13-2019. In total, this dataset was composed of 5,000 bacterial and 328 archaeal genomes. The second dataset (TypeMat) included 10,573 genomes (10,200 *Bacteria* and 373 *Archaea*) from type material with standing in nomenclature, which are available through the MiGA webserver with release number r2019-02[6, 21]. Genomes in either dataset with completeness estimates below 50% or contamination higher than 30% based on MiGA estimates were removed from further analyses as low-quality genomes. Finally, the third dataset (labeled PhylaLite) was composed of 39 complete genomes (36 *Bacteria* and 3 *Archaea*), one genome per phylum from the TypeMat collection. In addition to these primary datasets, we included simulated datasets derived from a subset of genomes in the RefSeq dataset (n=500) to evaluate FastAAI when genomes had different levels of completeness and contamination. For this, we created *in silico* genomes with varying degrees of completeness by randomly subsampling the single-copy marker proteins of these RefSeq genomes to a given level of completeness. Then, each of the resulting genomes was modified to increase contamination by randomly adding single-copy marker proteins from distantly related organisms to achieve a given level of contamination. All metrics used, including length, N50, completeness, and contamination, are shown for all datasets in Figure S5.

Table 2. Datasets used to evaluate FastAAI

Name	No. of Genomes	Mean Completeness	Mean Contamination
TypeMat	10,573	98.34	2.42
RefSeq	5,328	98.38	1.29
PhylaLite	39	95.98	2.15

FastAAI's Method of Estimating AAI

FastAAI utilizes HMMER to identify proteins in a genome's proteome as being one of the 122 prokaryotic single-copy proteins (discussed above, Figure S4A). FastAAI considers a protein to be a particular SCP when HMMER identifies the protein's best match to be the SCP in question and the protein has no better match among the remaining SCPs. Consequently, a protein can only be classified as one SCP and each SCP can have only one protein as a representative within a single genome. No information about proteins that are not labeled as an SCP representative is retained by FastAAI; as a result, a genome can have at most 122 proteins considered in the process of estimating AAI. To estimate AAI for a pair of genomes, FastAAI determines which SCPs the two genomes share and collects the SCP representative proteins of each in matched-SCP pairs. FastAAI then calculates the Jaccard index between the tetramers of each pair and averages them to produce an initial estimator that we label as $\bar{J}$ (Figure S4B).

Comparison of FastAAI to DIAMOND-based AAI using the RefSeq dataset revealed a linear correspondence between $\bar{J}$ and AAI values (R^2: 0.96, RMSE: 2.22%; Figure S6A). However, despite the low root mean squared error (RMSE) and the high R-squared of the linear regression, AAI values predicted using the regression coefficients were consistently lower compared to the actual AAI values, especially in the inter-phylum and inter-class levels, as evidenced by the curvature over the regression line at these regions. To solve this and improve the correspondence between FastAAI's estimate and AAI values, we transformed $\bar{J}$ by applying several transformations to linearize the relationship. The linearization coefficients were optimized using a gradient descent method aiming at decreasing the RMSE, excluding values with $\bar{J} > 0.9$ (see below). Finally, to have a 1:1 correspondence, we applied a linear model on top of the initial transformation. The final transformation defined as the estimated AAI ($\widehat{AAI}$) is described by the following formula:

$$\widehat{AAI} = \mathrm{a} + \mathrm{b} * e^{(-(-\mathrm{c}*\log(\bar{J}))^{1/\mathrm{d}})}$$

$$a = -0.3087057; b = 1.810741; c = 0.2607023; d = 3.435$$

The linear regression of AAI vs. $\widehat{AAI}$ values in the 0-100% range shows a slightly lower R-squared of 0.95 and a high RMSE (22.02%; Figure S6B). This underperformance is due to data points with $\bar{J}$ values higher than 0.9 (and thus, the corresponding values of $\widehat{AAI}$) collapsing to the AAI range of ~95-100%, increasing the regression error and generating curvature in these portions of the plot (Figure S6B). The low correspondence between FastAAI's $\bar{J}$ and AAI values in the upper range is presumably due to the high sequence conservation of the universal genes (thus, low resolving power) at this (the species) level. Therefore, FastAAI summarizes these high $\bar{J}$ values as ">90% AAI" and we recommend using instead ANI (e.g., FastANI) or traditional AAI to compare genomes related at this level, as also suggested previously[4, 14]. As AAI values tend to remain greater than 30% even across domains, FastAAI is also insensitive below 30% AAI and so summarizes $\widehat{AAI}$ estimates at or below 30% as "<30% AAI."

Estimation of 16S rRNA gene identities, AAI, and GTDBtk-RED values

We used INFERNAL (INFERence of RNA ALignment) v1.1.2[22] to identify and extract 16S rRNA gene sequence identities for all genomes in the RefSeq dataset. In cases where more than one 16S rRNA gene sequence was found in a single genome, the longest sequence was selected as the representative for the genome, or if multiple sequences had the same length, the representative was randomly chosen among them. We then performed pairwise alignments and determined the identities of each pair of sequences using a custom python script (available at https://github.com/cruizperez/Bioinformatic_Tools), which uses the Needleman–Wunsch algorithm with the same parameters as implemented in EMBOSS Needle[23]. We used the aai.rb tool from the Enveomics collection[24] with DIAMOND as the searching tool to calculate pairwise AAI values. Finally, we performed classifications and extracted Relative Evolutionary Diverge (RED) values based on the Genome Taxonomy Database (GTDB) using GTDBtk with release #202 of the database.

FastAAI taxonomic classification accuracy

We evaluated the ability of FastAAI $\widehat{AAI}$ values to discriminate genomes at different taxonomic levels by using the TypeMat dataset as the ground truth and reference taxonomy. For this, we determined threshold values delimiting taxonomic ranks by finding valleys in the distribution of $\widehat{AAI}$ values between inter- vs. intra-rank comparisons for the TypeMat genome dataset. For example, genomes assigned to the same domain typically show values above 35.3% $\widehat{AAI}$ and those assigned to different domain typically have $\widehat{AAI}$ values below 35.3%. Genomes in the RefSeq dataset (and their associated taxonomy, assumed to be correct) were treated as queries, and their taxonomic classification based on the estimated $\widehat{AAI}$ values against TypeMat genomes was compared to their available taxonomy (from RefSeq). For each rank, we built a confusion matrix where we labeled the classifications obtained based on $\widehat{AAI}$ value thresholds and compared them to the available taxonomic classification of the RefSeq query genome from NCBI as follows: True positive (TP) for pairs with $\widehat{AAI}$ > taxonomic rank threshold (or just "threshold" for simplicity) and where query and reference shared the same taxon, true negative (TN) for pairs with $\widehat{AAI}$ < threshold and where the pair belonged to a different taxon in the rank, false positive (FP) for pairs with $\widehat{AAI}$ > threshold and where the pair belonged to a different taxon in the rank, and false negatives (FN) for pairs with $\widehat{AAI}$ < threshold and where the pair belonged to the same taxon. From this confusion matrix, we calculated the accuracy, recall, precision, and F1 score for each taxonomic rank as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

FastAAI availability

FastAAI is an open-source tool freely available at https://github.com/KGerhardt/FastAAI.

Acknowledgments

This work has been supported by US NSF under awards # DBI1356288, -1356380, -1759831, and -1759892.

References

1. Amann, R. and R. Rossello-Mora (2016). "After All, Only Millions?" mBio **7**(4).

2. Nayfach, S., S. Roux, R. Seshadri, D. Udwary, N. Varghese, F. Schulz, D. Wu, D. Paez-Espino, I. M. Chen, M. Huntemann, K. Palaniappan, J. Ladau, S. Mukherjee, T. B. K. Reddy, T. Nielsen, E. Kirton, J. P. Faria, J. N. Edirisinghe, C. S. Henry, S. P. Jungbluth, D. Chivian, P. Dehal, E. M. Wood-Charlson, A. P. Arkin, S. G. Tringe, A. Visel, I. M. D. Consortium, T. Woyke, N. J. Mouncey, N. N. Ivanova, N. C. Kyrpides and E. A. Eloe-Fadrosh (2021). "A genomic catalog of Earth's microbiomes." Nat Biotechnol **39**(4): 499-509.

3. Konstantinidis, K. T. and J. M. Tiedje (2007). "Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead." Curr Opin Microbiol **10**(5): 504-509.

4. Jain, C., R. L. Rodriguez, A. M. Phillippy, K. T. Konstantinidis and S. Aluru (2018). "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries." Nat Commun **9**(1): 5114.

5. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucl. Acids. Res. **25**(17): 3389-3402.

6. Rodriguez, R. L., S. Gunturu, W. T. Harvey, R. Rossello-Mora, J. M. Tiedje, J. R. Cole and K. T. Konstantinidis (2018). "The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level." Nucleic Acids Res **46**(W1): W282-W288.

7. Sunagawa, S., D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, A. Stamatakis and P. Bork (2013). "Metagenomic species profiling using universal phylogenetic marker genes." Nat Methods **10**(12): 1196-1199.

8. Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil and P. Hugenholtz (2018). "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life." Nat Biotechnol **36**(10): 996-1004.

9. Buchfink B, Xie C, Huson DH. “Fast and sensitive protein alignment using DIAMOND” (2015). Nat Methods. **12**(1):59-60.

10. Kim, D., S. Park and J. Chun (2021). "Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity." J Microbiol **59**(5): 476-480.

11. Steinegger, M. and J. Soding (2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nat Biotechnol **35**(11): 1026-1028.

12. Chaumeil, P. A., A. J. Mussig, P. Hugenholtz and D. H. Parks (2019). "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." Bioinformatics **36**(6):1925-1927.

13. Konstantinidis, K. T. and J. M. Tiedje (2005). "Towards a genome-based taxonomy for prokaryotes." J Bacteriol **187**(18): 6258-6264.

14. Bowers, R. M., N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glockner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W. T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, C. Genome Standards, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz and T. Woyke (2017). "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea." Nat Biotechnol **35**(8): 725-731.

15. Huson, D. H., S. Mitra, H. J. Ruscheweyh, N. Weber and S. C. Schuster (2011). "Integrative analysis of environmental sequences using MEGAN4." Genome Res **21**(9): 1552-1560.

16. Luo, C., R. L. Rodriguez and K. T. Konstantinidis (2014). "MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences." Nucleic Acids Res **42**(8): e73.

17. Rinke, C., P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz and T. Woyke (2013). "Insights into the phylogeny and coding potential of microbial dark matter." Nature **499**(7459): 431-437.

18. Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Comput Biol **7**(10): e1002195.

19. Lee, M. D. (2019). "GToTree: a user-friendly workflow for phylogenomics." Bioinformatics **35**(20): 4162-4164.

20. Hyatt, D., G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC Bioinformatics **11**: 119.

21. Rodriguez-R, L. M., W. T. Harvey, R. Rosselló-Mora, J. M. Tiedje, J. R. Cole and K. T. Konstantinidis (2020). Classifying prokaryotic genomes using the Microbial Genomes Atlas (MiGA) webserver. Bergey's Manual of Systematics of Archaea and Bacteria**:** 1-11.

22. Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches." Bioinformatics **29**(22): 2933-2935.

23. Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn and R. Lopez (2019). "The EMBL-EBI search and sequence analysis tools APIs in 2019." Nucleic Acids Res **47**(W1): W636-W641.

24. Rodriguez-R LM, Konstantindis K. (2016). "The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes." PeerJ Preprints **4:e1900v1**.

Main Figures

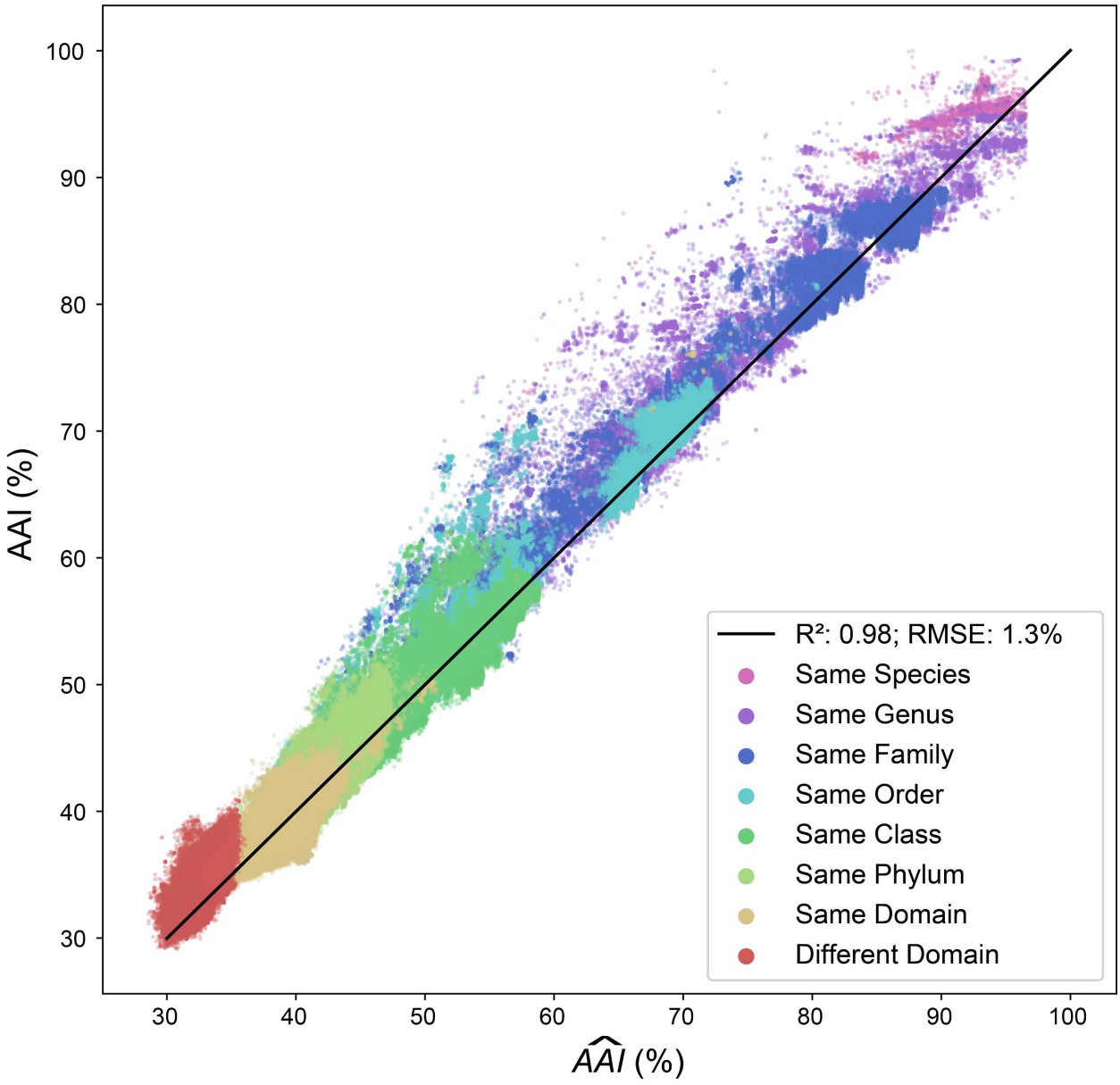
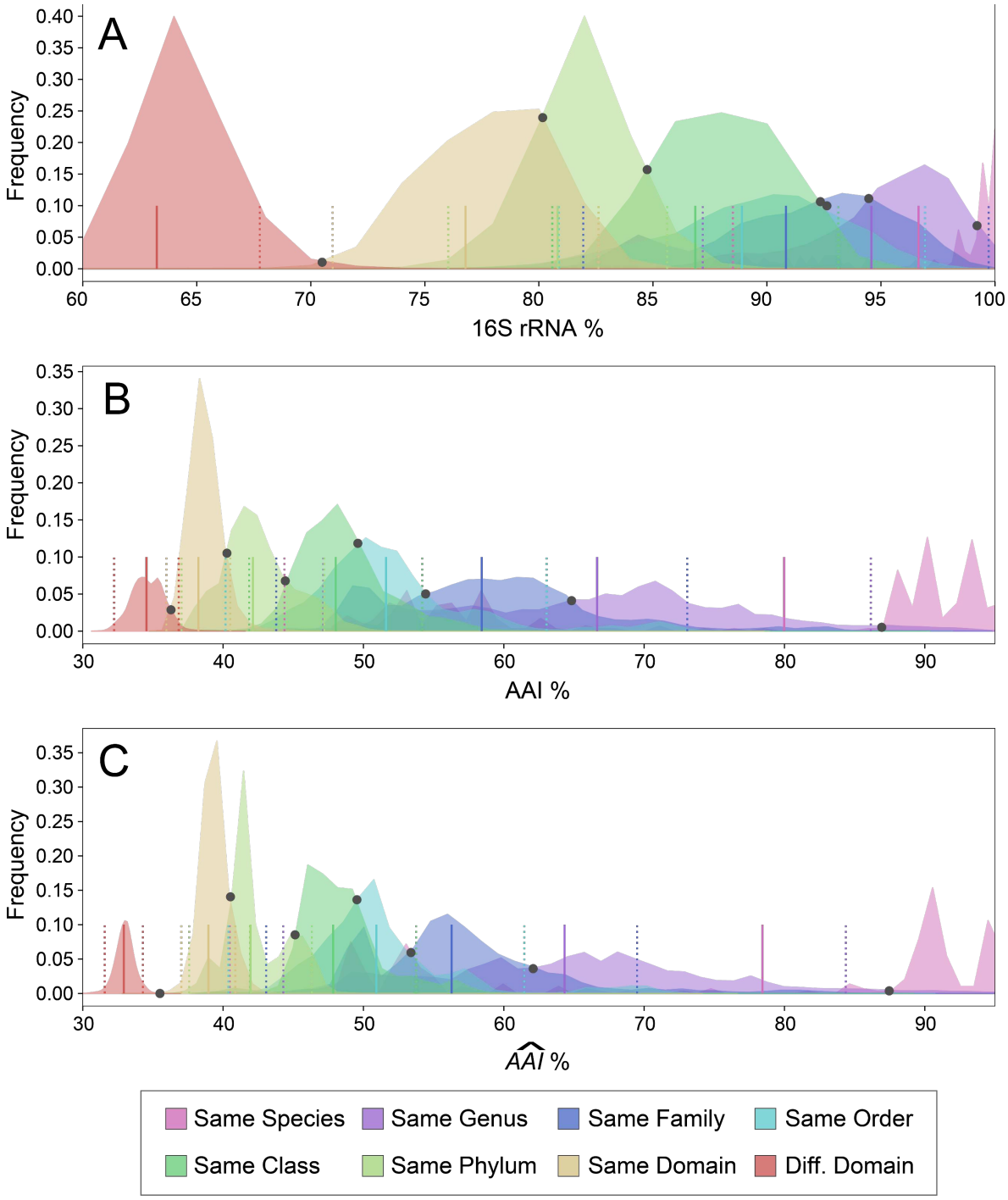


Figure 1. Identity correspondence between AAI and $\widehat{AAI}$ derived from the FastAAI value $\bar{J}$ transformation using the RefSeq dataset. Each dot represents a pairwise comparison colored by the lowest taxonomic rank (figure key) that the two genomes in the pair share. Note the linear correspondence between both values and the low error associated with the linear regression.



475

476 **Figure 2**. Distribution of 16S rRNA gene identities, AAI, and $\widehat{AAI}$ values, according to the lowest
477 taxonomic level shared by the pairs of genomes analyzed. Underlying data are based on all vs. all
478 comparisons of the TypeMat genomes. Note the large overlaps between taxonomic ranks using all
479 methods, but also the virtually no overlap between the inter- (red distribution) and intra-domain (yellow
480 distribution) groups for $\widehat{AAI}$, which is consistent to the picture based on 16S rRNA gene identities. Solid
481 lines represent the means of the distribution while dotted lines represent one standard deviation around
482 the mean.

Supplementary Files

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

- 25FastAAISupplementalText.docx