Ye et al. Microbiome (2020) 8:16
https://doi.org/10.1186/s40168-020-0794-3

Microbiome

RESEARCH Open Access

# Machine learning-aided analyses of thousands of draft genomes reveal specific features of activated sludge processes

Lin Ye[1*], Ran Mei[2], Wen-Tso Liu[2], Hongqiang Ren[1] and Xu-Xiang Zhang[1*]

## Abstract

**Background:** Microorganisms in activated sludge (AS) play key roles in the wastewater treatment processes. However, their ecological behaviors and differences from microorganisms in other environments have mainly been studied using the 16S rRNA gene that may not truly represent in situ functions.

**Results:** Here, we present 2045 archaeal and bacterial metagenome-assembled genomes (MAGs) recovered from 1.35 Tb of metagenomic data generated from 114 AS samples of 23 full-scale wastewater treatment plants (WWTPs). We found that the AS MAGs have obvious plant-specific features and that few proteins are shared by different WWTPs, especially for WWTPs located in geographically distant areas. Further, we developed a novel machine learning approach that can distinguish between AS MAGs and MAGs from other environments based on the clusters of orthologous groups of proteins with an accuracy of 96%. With the aid of machine learning, we also identified some functional features (e.g., functions related to aerobic metabolism, nutrient sensing/acquisition, and biofilm formation) that are likely vital for AS bacteria to adapt themselves in wastewater treatment bioreactors.

**Conclusions:** Our work reveals that, although the bacterial species in different municipal WWTPs could be different, they may have similar deterministic functional features that allow them to adapt to the AS systems. Also, we provide valuable genome resources and a novel approach for future investigation and better understanding of the microbiome of AS and other ecosystems.

**Keywords:** Activated sludge, Metagenomics, Machine learning

## Background

Activated sludge (AS) is the largest biotechnology application in the world and is of eminent importance for the remediation of anthropogenic wastewater [1]. The pollutant removal functions of AS are achieved by microorganisms with diverse community structures, among which populations with important metabolic functions have been individually studied [2–4]. Meanwhile, AS is a unique engineered ecosystem that can be controlled by a variety of operating conditions, and its attributes make it attractive for microbial ecologists studying the behaviors of microbial community assembly [5, 6].

One major topic of AS microbiome research is investigating the core populations that are consistent occupants in a large number of AS communities and are potentially important contributors to the system performance. Such analysis has been performed using 16S rRNA gene sequencing at different scales, including one full-scale wastewater treatment plant (WWTP) in Hong Kong [7], 13 WWTPs in Denmark [8], 14 WWTPs in Asia and North America [9], and 269 WWTPs in 23 countries [1]. Core AS microbial communities were identified at both regional and global scales by counting shared species or operational taxonomic units (OTUs), implying that a small number of key microorganisms constitute an indispensable portion of the AS community regardless of geographical and operational variations. However, the 16S rRNA gene, despite a useful biomarker to explore microbial community and construct phylogeny, does not necessarily reflect microbial physiology [10]. Therefore, the in situ functions and ecological contributions of the identified core AS populations are still not clear. Moreover, vast metabolic diversity can

* Correspondence: linye@nju.edu.cn; zhangxx@nju.edu.cn
[1]State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu, China
Full list of author information is available at the end of the article

be embedded in one species or OTU, which is usually defined at 97% sequence identity or even higher levels [11]. Thus, further investigation of the AS community is warranted using more advanced approaches that could resolve metabolic potentials with higher resolution.

Metagenomics aimed at recovering population genomes and annotating genetic potentials have been applied to AS and uncovered individual microorganisms or functions that are challenging to study using other methods [12–14], demonstrating that this approach is promising for revealing greater diversity at the functional level than the analysis of 16S rRNA gene sequences. However, few efforts have been made to resolve microbial ecology, such as the core-community phenomenon in AS, using metagenomics. Furthermore, metagenomics could facilitate a comparative analysis of microbiomes of AS and other ecosystems at functional level. Microorganisms associated with freshwater systems, soil, human feces, rainwater, and stormwater have been shown to seed the activated sludge via influent sewage [15, 16]. Comparing the populations in AS and various non-AS ecosystems could provide insights into how the AS microbial community is assembled and whether the AS populations possess unique functional features that are vital to the adaption to the conditions of wastewater treatment bioreactors.

The vast diversity observed in AS and tremendous information obtained by metagenomics present new data analysis challenges. Conventional approaches mainly rely on reducing dimensionality to retrieve and visualize ecological patterns. Ordination analyses such as nonmetric multidimensional scaling and principal coordinates analysis could only present the first two or three eigenvectors that account for a limited proportion of the entire variance. Phylogenetic analysis is based on one or multiple selected conserved genes out of thousands of genes in a prokaryotic genome, which inevitably results in loss of information. In recent years, machine learning approaches have received growing attention and have been applied in genomics research [17, 18]. Unlike conventional methods, they can automatically detect patterns in data with less expert handcrafting and are therefore suitable to handle and analyze large and complex datasets such as genomic and metagenomic data [18, 19]. They can further be used to disentangle the complexity and diversity in the AS community by comparing different AS systems and comparing AS with other environments.

Here, we present 2045 high- and medium-quality bacterial and archaeal metagenome-assembled genomes (MAGs) recovered from 114 global municipal AS samples, representing one of the largest assemblies of MAGs from the municipal AS microbiome. After the recovery of the vast genomic information, we aimed to address two questions. First, is there a significant core AS community at the MAG and protein level shared by a large number of WWTPs, or are there obvious plant-specific features in the AS MAGs? Second, are the AS MAGs similar to genomes of populations from other environments, or do they have unique environment-specific traits? In addition to a novel machine learning approach, a collection of conventional methods including genome and protein comparison, phylogenetics, and ordination was applied, and their results were compared.

## Results

### 2045 MAGs were obtained from AS of different WWTPs

Approximately 1.35 Tb of metagenomic sequencing data generated from 114 AS samples of 23 municipal WWTPs located in eight countries were used to construct MAGs (Additional file 1: Figure S1, Table S1, Table S2). Among the 7548 bacterial and archaeal MAGs obtained, 2045 are estimated to have overall quality (defined as completeness − 5 × contamination) ≥ 50 [20]. The average completeness and contamination of the 2045 MAGs were 82.0% and 2.0%, respectively. Figure 1a shows that 743 of the 2045 MAGs are nearly complete (completeness ≥ 90%, average contamination 2.6%). The other two groups contain 845 (70% ≤ completeness < 90%) and 456 MAGs (50% ≤ completeness < 70%), and their average contamination values are 3.3% and 0.92%, respectively. The average contig number of these MAGs is 292, and the contig numbers have a moderate association with contamination level (Spearman's rho = 0.47, $P < 2.2e{-}16$) but not with completeness level (Spearman's rho = − 0.11, $P = 4.3e{-}08$) (Additional file 1: Figure S2). As shown in Additional file 1: Figure S2, most of the MAGs have good overall quality (high completeness and low contamination), while it was also found that some MAGs have relatively smaller contig numbers and medium-quality values (50–80%) (Additional file 1: Figure S2a), which leads to the relatively weak association between contig number and contamination level.

The 2045 MAGs were classified into 49 phyla (Fig. 1b and Additional file 1: Table S3). Among these MAGs, 21 were assigned to three archaeal phyla (*Halobacterota*, *Micrarchaeota*, and *Nanoarchaeota*). For bacteria, the phylum containing the highest number of MAGs was *Proteobacteria* (508 MAGs), followed by *Bacteroidota* (409 MAGs), *Patescibacteria* (178 MAGs), *Myxococcota* (164 MAGs), *Actinobacteriota* (161 MAGs), *Planctomycetota* (122 MAGs), *Chloroflexota* (114 MAGs), and *Acidobacteriota* (96 MAGs). The remaining MAGs were assigned to other miscellaneous bacterial phyla (Additional file 1: Table S3). To further understand the diversity among these MAGs, phylogenetic analysis was performed using the universal core gene markers predicted from each MAG [21]. Figure 1b shows that the clustering patterns in the tree are highly consistent with the taxonomy assignments, with *Proteobacteria* and *Bacteroidales* as the two most dominant clusters.

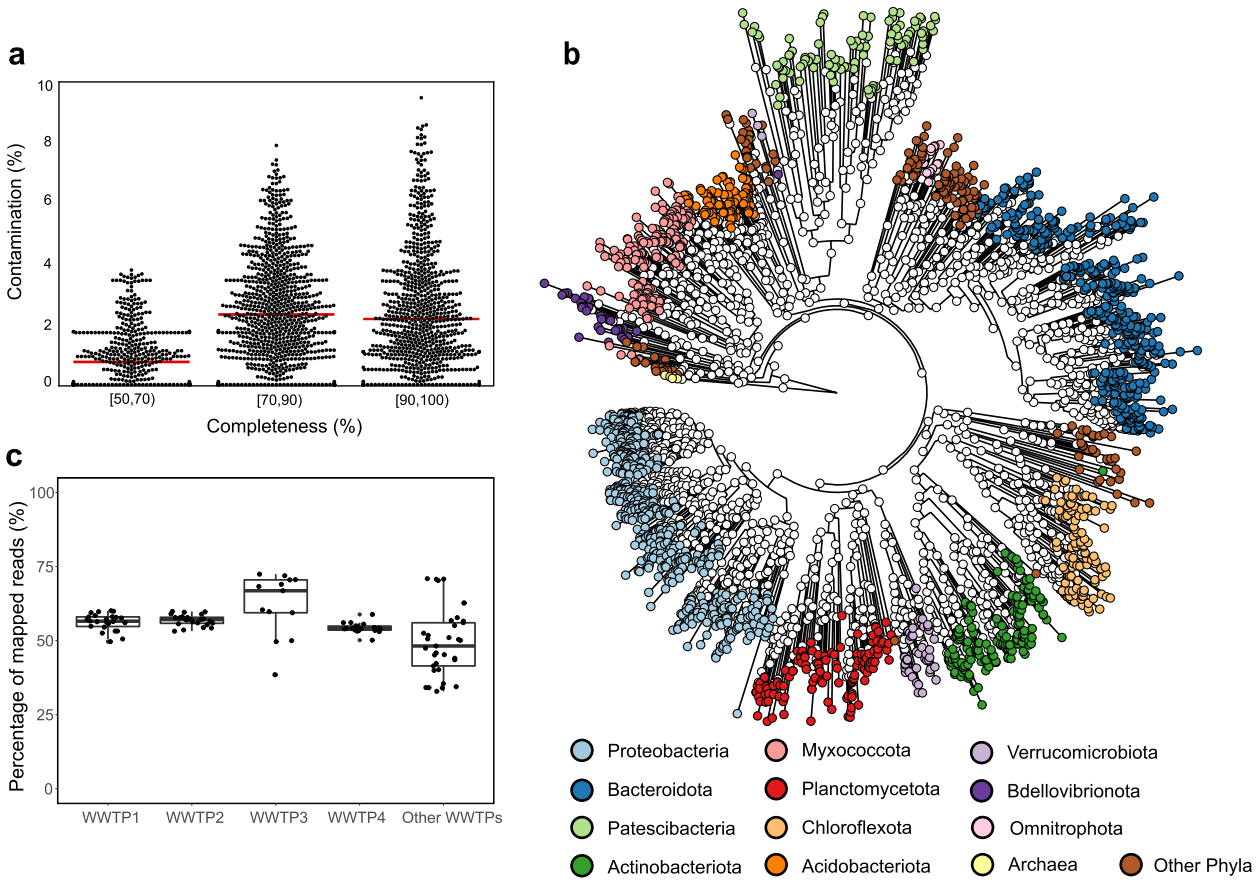


**Fig. 1** Overview of the 2045 MAGs assembled from 114 AS microbiomes. **a** Estimated completeness and contamination of the 2045 MAGs. The position of each horizontal red line refers to the mean contamination value of the corresponding group. **b** A maximum likelihood phylogenetic tree of the AS archaeal and bacterial MAGs based on universal core gene markers. The genome phylogenetic tree was generated using the universal PhyloPhlAn markers conserved across the bacterial and archaeal domains. A total of 98 MAGs with fewer than 80 universal markers were not included in this tree. The taxonomy of the MAGs was determined using GTDB-Tk, and it is shown in different colors. **c** Percentages of metagenomic sequencing reads of the different AS samples mapped to the 2045 MAGs

To estimate the representativeness of the MAGs for AS microbial genetic information, we mapped the metagenomic sequencing reads of each WWTP to the MAGs and calculated the percentage of mapped reads in each sample. As shown in Fig. 1c, 54–63% of reads (average per WWTP) of AS samples from the first four WWTPs, which have larger sequencing data volumes and significantly contribute to the AS MAG catalog, were mapped to the MAGs. For other WWTPs, the mapping ratios ranged from 34 to 72%.

### The AS MAGs show obvious plant-specific features

To evaluate the plant-specific features of the MAGs, we first analyzed the distribution of reads mapped to the MAGs obtained from different plants. As shown in Fig. 2a, most (60–87%) of the mapped metagenomic reads from each WWTP were mapped to its own MAGs. A relatively small fraction of reads in each WWTP (approximately 33% in WWTP1, 32% in WWTP2, 35% in WWTP3, and 13% in WWTP4) were mapped to MAGs from other WWTPs. MAGs of WWTP1 and WWTPs shared more mapped reads than other WWTP pairs (approximately 20% of sequencing reads of WWTP1 and WWTP2 were mapped to each other's MAGs), likely because they are located in the same city.

In addition to mapping reads to MAGs, we also computed the average nucleotide identity (ANI) values by comparing the MAGs with an all-against-all strategy. The results in Fig. 2b and Additional file 1: Figure S3 show that 214 MAG pairs have ANI > 95% between WWTP1 and WWTP2, suggesting that these 214 bacterial or archaeal species (43% MAGs in WWTP1 and 33% MAGs in WWTP2) were shared between WWTP1 and WWTP2. However, the numbers of potentially shared species between other WWTPs were relatively small. For example, no MAG pairs with ANI > 95% were observed between WWTP3 and WWTP4, and only four MAG pairs with ANI > 95% were found between WWTP1 and

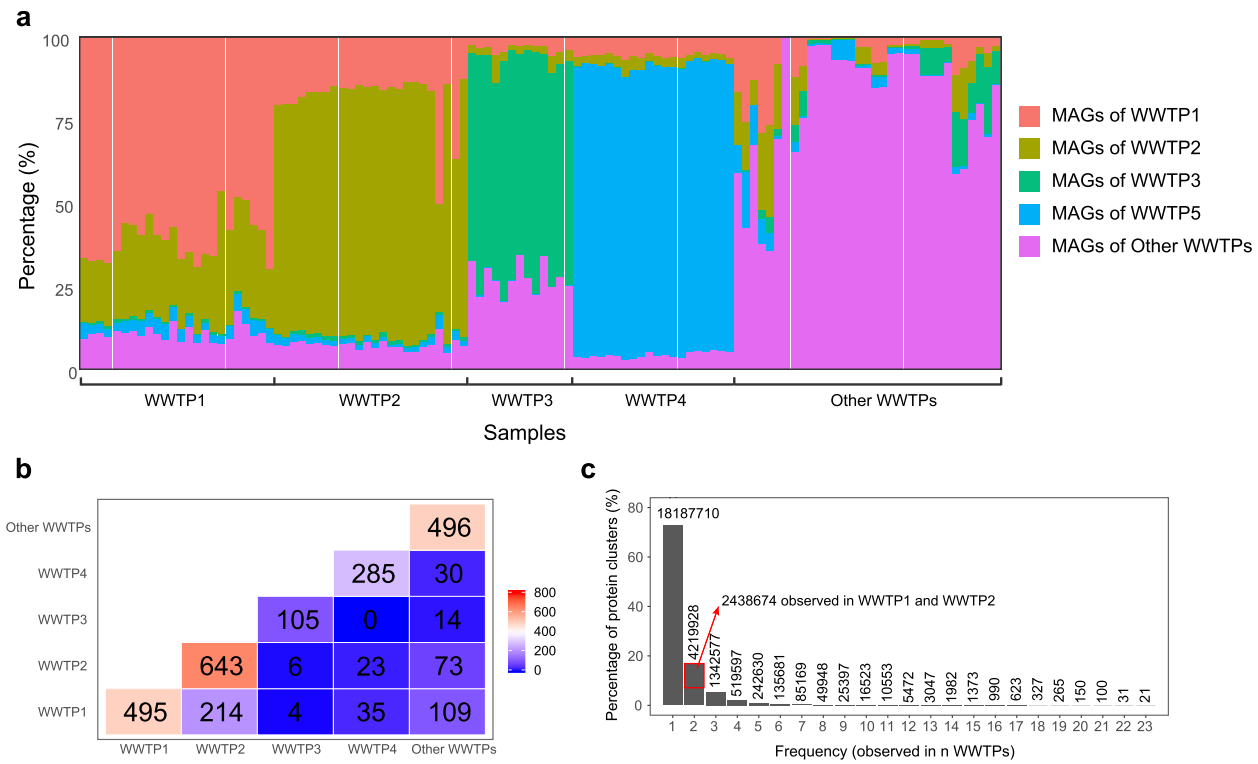


**Fig. 2** Comparison of MAGs and protein sequences in different WWTPs. **a** Relative abundance of metagenomic sequencing reads of each sample mapped to the MAGs from different WWTPs. **b** Numbers of MAG pairs with ANI > 95% between different WWTPs. The values on the diagonal also refer to the MAG number in each of the first four WWTPs and the total MAG number of other WWTPs. **c** Frequency distribution of protein clusters across WWTPs. The protein sequences predicted from all assembly contigs were clustered at an identity cutoff of 90% with CD-HIT, and then the protein clusters observed at each frequency were counted. The *y*-axis values were transformed to percentages, and the numbers on the top of bars refer to the absolute values of protein clusters observed in *n* WWTPs

WWTP3. A number of MAG pairs were also observed between WWTP1 and "other WWTPs" (109) and between WWTP2 and "other WWTPs" (73). This is probably because a large fraction (9/19) of WWTPs in "other WWTPs" are located in China and near WWTP1 and WWTP2 (Additional file 1: Table S1).

Since the MAGs represent only part (34 to 72%) of the AS microbiome according to the mapping results, we also conducted a pairwise comparison of protein sequences predicted from all assembled contigs of the first four WWTPs. Other WWTPs were not included in this comparison because of their low sequencing depths. As shown in Additional file 1: Figure S4, 62% of proteins predicted from WWTP1 are highly similar (identity > 90%) to those of WWTP2. However, only a small number of proteins predicted from WWTP3 (10–27%) and WWTP4 (7.9–28%) have highly similar hits (identity > 90%) in other WWTPs. We further identified 24,850,093 clusters (identity cutoff 90%) from the 44,212,953 protein sequences predicted from all AS samples. A frequency distribution plot (Fig. 2c) shows that 73.2% of the protein clusters were found in one WWTP, and 17.0% were found in two WWTPs. Among the protein clusters observed in two WWTPs, over half (57.8%) were shared by WWTP1 and WWTP2, which were located in the same city. Only 0.1% of total protein clusters were present in ≥ 10 WWTPs. The protein comparison results confirmed the results of read mapping and ANI calculation. It further suggested that, although a certain amount of proteins and MAGs may be shared by different WWTPs, a large proportion of bacterial populations in different WWTPs are largely different at both the DNA and protein levels, i.e., the bacterial genomes have plant-specific features.

## Phylogeny and functional features cannot well separate MAGs from AS and MAGs from other environments

In addition to comparing MAGs among different WWTPs, we also explored whether the 2024 bacterial AS MAGs obtained in this study could be distinguished from the 7164 MAGs of other non-engineered (natural and animal/human-related) environments [20]. We constructed a maximum likelihood phylogenetic tree encompassing 1000 randomly selected AS MAGs and 1000 randomly selected non-AS MAGs (Fig. 3a). The tree shows that both AS and non-AS MAGs are distributed

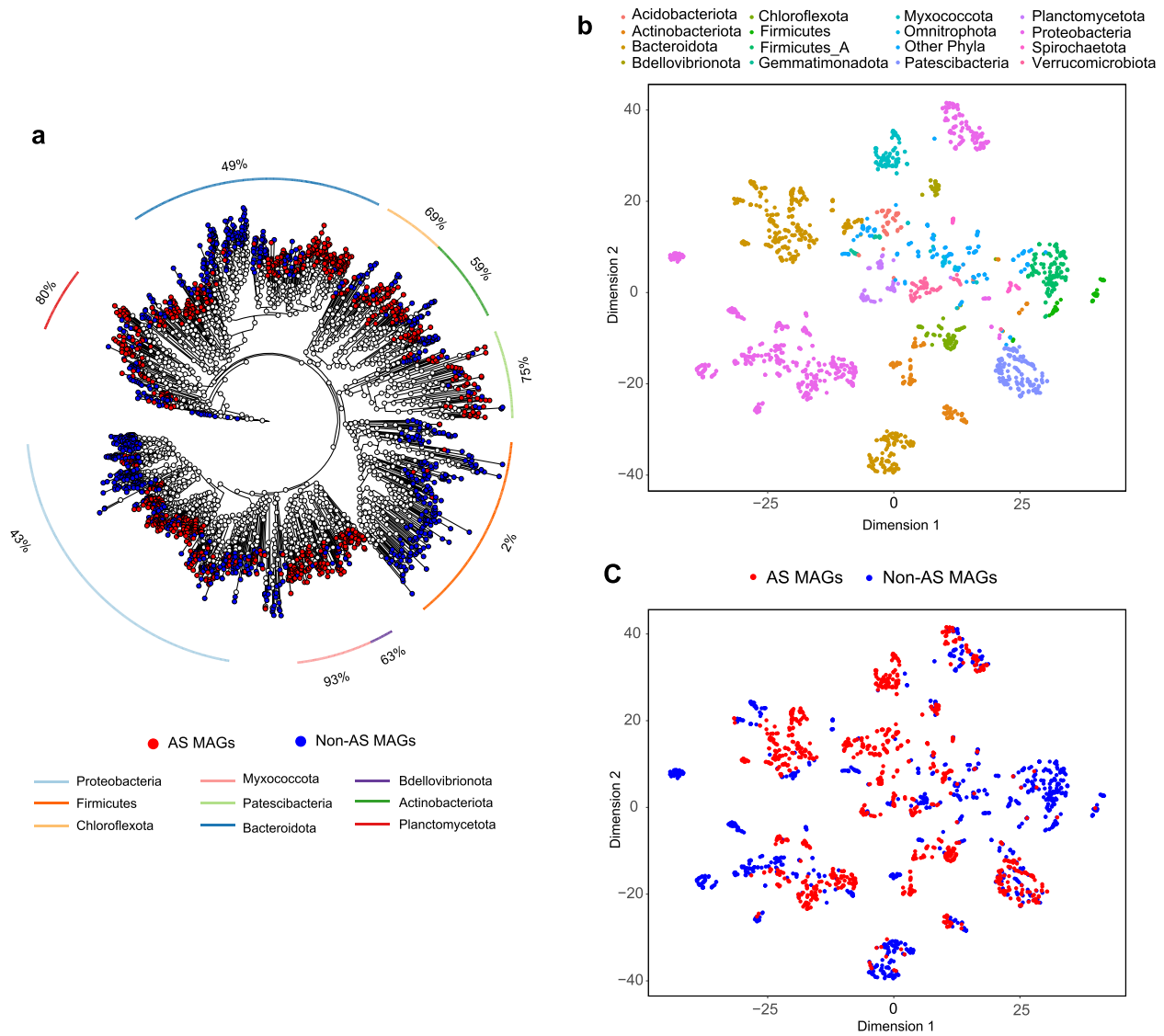


**Fig. 3** Phylogenetic and functional comparison of AS MAGs and non-AS MAGs. **a** A whole-genome maximum likelihood phylogenetic tree consisting of AS MAGs and non-AS MAGs. One thousand MAGs randomly selected from AS bacterial MAGs and 1000 MAGs randomly selected from other environments (Parks et al. [20]) were used to build this whole-genome tree with the same methods as in Fig. 1b. The outside percent value refers to the relative abundance of AS MAGs in each clade. **b** Clustering of the AS and non-AS MAGs based on the COG presence/absence matrix with the t-SNE algorithm. The 2000 MAGs in **a** were used to generate this figure. The points representing MAGs are colored according to the taxonomy of each MAG. **c** The same clustering plot as in **b**, with the red points representing AS MAGs and blue points representing non-AS MAGs

in a wide range of phyla. Non-AS MAGs were dominant in the *Firmicutes* clade (which contained only 2% AS MAGs). More AS MAGs than non-AS MAGs belonged to *Myxococcota* (93% AS MAGs) and *Planctomycetota* (80% AS MAGs). Considerable amounts of both AS and non-AS MAGs were present in most of the remaining clades. These patterns remained basically unchanged when the number of AS and non-AS MAGs used for tree construction increased. Overall, the large-scale phylogenetic analysis based on random selection shows that the AS MAGs are phylogenetically interspersed among non-AS MAGs, and no clear separation patterns were observed.

We further investigated the differences between AS and non-AS MAGs by annotating them with the database of clusters of orthologous groups of proteins (COGs). As proteins in each COG have the same domain architecture and likely have the same function [22], comparison of COG profiles may reflect the different functions encoded in the MAGs. A COG presence/absence matrix was generated for the 2024 bacterial AS MAGs and 7164 non-AS bacterial MAGs. A t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis based on the COG presence/absence matrix was able to separate MAGs associated with different phyla (Fig. 3b). However, no clear grouping patterns between AS

MAGs and non-AS MAGs (Fig. 3c) were observed, which was similar to the results of the phylogenetic tree. Most of the AS and non-AS MAGs were widely distributed and co-present in most phyla, except that few AS MAGs were observed in *Firmicutes* and some AS MAGs were separated from non-AS MAGs in the *Bacteroidota* cluster.

### A machine learning approach to distinguish between AS and non-AS MAGs based on COGs

We further explored whether machine learning can better distinguish between AS and non-AS MAGs. To do so, the COG presence/absence matrix generated from the 2024 AS and 7164 non-AS MAGs was used as an input of the random forest model (Fig. 4). After the model was constructed and trained, its accuracy and applicability were further evaluated. Both the holdout method and *k*-fold cross-validation were applied to verify the model to avoid the overfitting issue. For the holdout method, the dataset was divided into two partitions as testing (20%) and training (80%) sets. The number of trees is an important parameter affecting the accuracy of the random forest algorithm and should be tuned. As shown in Additional file 1: Figure S5, after the tree number (*n* estimators) was increased to 200, the accuracy did not increase with the number of trees, and other parameters (tree depth and max features) were also simultaneously optimized (Additional file 1: Figure S5). With optimized parameters (*n* estimators 300, tree depth 20, and max features 100), the training and testing data groups were analyzed (Fig. 5a), and the overall prediction accuracy of the random forest model achieved 96.6% (94% for AS and 97% for non-AS MAGs, Additional file 1: Table S4). Particularly, the recall (i.e., true positive rate) for non-AS MAGs was 98%, which was higher than that of the AS MAGs (91%). This result suggests that approximately 9% of AS MAGs were wrongly classified as non-AS MAGs. The F1-score, which is the harmonic average of the precision and recall, of AS and non-AS MAGs was 0.93 and 0.98, respectively. The classification accuracy obtained from 10-fold stratified cross-validation ranged from 95.0 to 95.6% (Fig. 5b), suggesting that the model is reliable and accurate, and no overfitting was observed. Receiver operating characteristic (ROC) curves also demonstrated the excellent performance (area under the ROC curve (AUC) ranged from 0.94 to 1; for the mean ROC curve, AUC = 0.98) of the random forest model (Fig. 5c).

We further investigated the quality (completeness and contamination) and phylogeny of the wrongly predicted MAGs. Figure 5d indicates that the wrongly predicted MAGs were evenly distributed among correctly predicted MAGs. There was no significant difference between the contamination values of the two groups of MAGs (*t* test, $P < 0.05$). The average contamination of the wrongly predicted MAGs (1.7%) was lower than that of the correctly predicted MAGs (2.2%), and the average completeness of the wrongly predicted MAGs (82.1%) was slightly higher than that of the correctly predicted MAGs (81.6%). This suggests that the overall quality of wrongly predicted MAGs is better than that of correctly predicted MAGs. Therefore, completeness and contamination levels may not be the major reasons leading to incorrect prediction. Phylogenetic analysis showed that erroneously predicted MAGs were distributed in various phyla, while many were associated with *Proteobacteria*, which was inherently diverse (Additional file 1: Figure S6).

### Different functional features between AS and non-AS MAGs

During the random forest model training, an importance value was assigned to each COG. The COGs with higher

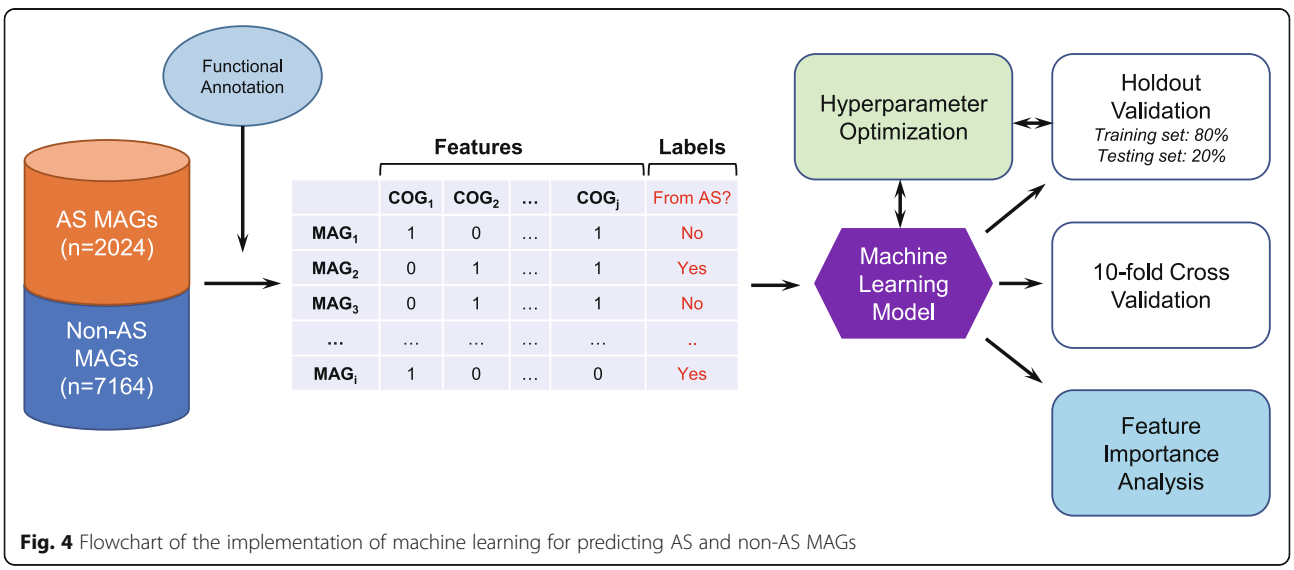


**Fig. 4** Flowchart of the implementation of machine learning for predicting AS and non-AS MAGs

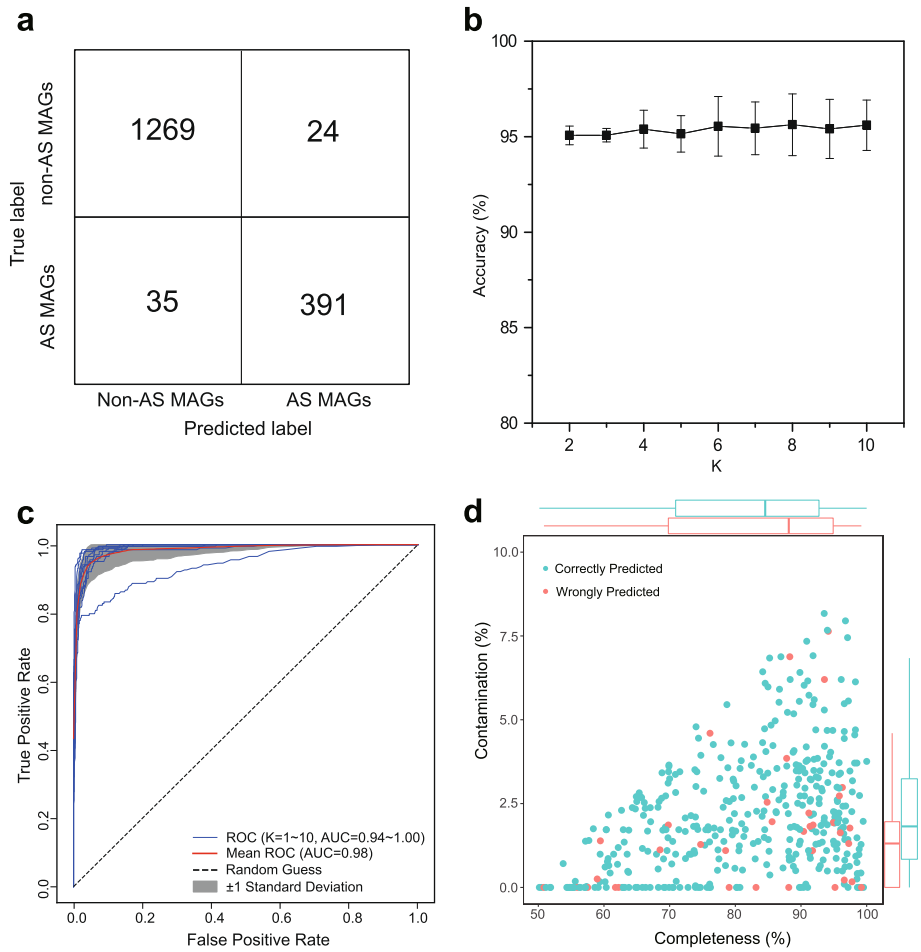


**Fig. 5** Performance of the random forest model. **a** Confusion matrix showing the performance of the random forest model on the 20% testing data group of the holdout validation. **b** Prediction accuracy of the random forest model determined based on 10-fold cross-validation. **c** ROC curves for evaluating the random forest model created from 10-fold cross-validation. **d** The completeness and contamination of correctly predicted MAGs and wrongly predicted MAGs. Boxplots along the *x*- and *y*-axes show the means and quartiles of the completeness and contamination values of correctly and wrongly predicted MAGs

importance values were more informative when the model was used to predict whether a MAG was from AS. Therefore, by analyzing the importance of each COG, the functions that differentiate the sources of MAGs can be identified. Figure 6a shows the presence/absence of the top 20 COGs based on the importance value among the MAGs (see Additional file 1: Table S5 for the importance values and descriptions). Some COGs (e.g., COG1979, 1328, 1464, 2011, and 1636) were clearly rarely present in AS MAGs. Proteins of these COGs are related to anaerobic metabolisms or functions, such as alcohol dehydrogenase and anaerobic ribonucleoside-triphosphate reductase. In contrast, several COGs (e.g., COG3324, 2114, 2107, and 3303) were more frequently observed in AS MAGs than in MAGs from other environments. Proteins of COG3324 and COG 2114 are related to sensing the nutritional contents of the surrounding media or other environmental signals [23]. Proteins of COG 3033 were annotated as tryptophanase, which catalyzes the beta-elimination reaction of L-tryptophan to yield indole, ammonium, and pyruvate, and the produced indole molecules may affect biofilm formation and multidrug exporters [24].

Many COGs besides the top 20 also contributed to the machine learning-based prediction. Among them, 148 COGs accounted for 50% of the cumulative importance, and approximately 1500 COGs were needed to reach a cumulative importance of 90% (Fig. 6b). This result indicates the highly diverse functional features of the AS microbiomes and the strong capability of the machine learning approach in capturing complex information. It also explained why the conventional phylogenetic and ordination approaches failed to separate the AS and non-AS MAGs.

## Discussion

Despite the important roles of AS microorganisms in removing various pollutants from wastewater, the microbiome

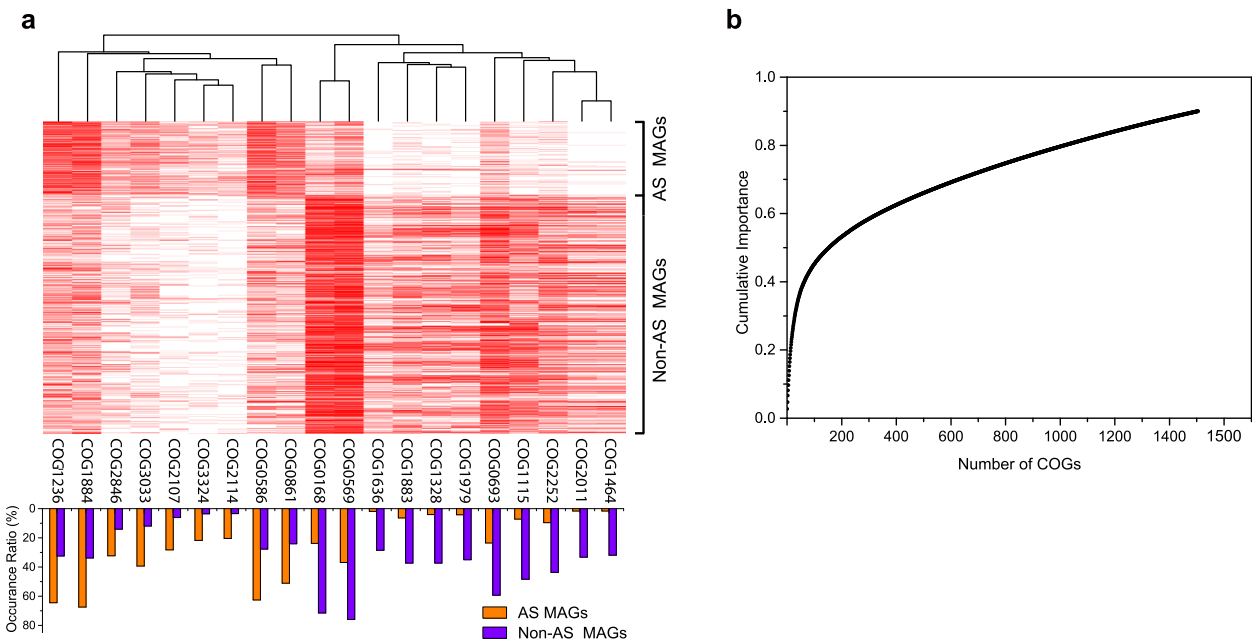


**Fig. 6** Feature importance determined by the random forest model. **a** The presence/absence of the top 20 COGs (with the highest importance values) in each MAG (heatmap). The "red" and "white" colors represent presence and absence, respectively. The bar plot shows the percentage of MAGs carrying each COG in the AS MAG group and in the non-AS MAG group. The importance values and descriptions are shown in Additional file 1: Table S5. **b** Cumulative importance values of the COGs

in AS remains largely uncharacterized. Based on metagenomic assembly and binning strategies, this study constructed an AS genome catalog consisting of 2024 bacterial and 21 archaeal MAGs recovered from 114 global municipal AS samples. This catalog likely represents the largest reported AS genome collection. Its coverage of the bacteria in AS systems is considered to be high, as up to 50–70% of the metagenomic sequencing reads could be mapped to the MAGs. Thus, this catalog could enable us to comprehensively profile the AS bacterial community structures and functions in a higher-resolution manner.

We found that the bacterial MAGs obtained from different WWTPs could be largely different according to DNA and protein comparisons, especially for WWTPs located in geographically distant areas. This suggests that AS MAGs may have plant-specific features at the genetic level and is consistent with a recent study based on 16S rRNA gene sequencing showing that municipal AS has a small, global core bacterial community [1]. Since MAGs contain much more genetic information and have more variants than 16S rRNA genes, it can be inferred that the genomes of the bacteria within the small core determined based on the 16S rRNA gene could also be largely different in different WWTPs. Therefore, the number of highly similar bacterial genomes present in different WWTPs might be very limited. The observation of small-core populations is in line with the previously reported functional redundancy in AS ecosystems [25, 26]. Although the overall functions of AS in all municipal WWTPs are carbon and nutrient removal, different operational parameters and wastewater compositions may lead to significantly different microbial communities with similar functions in different WWTPs. Moreover, we found that the similarity between MAGs of WWTP1 and WWTP2 located in the same city is higher than the similarity between MAGs of other WWTPs (Fig. 2 and Additional file 1: Figure S4). This is probably due to the similar wastewater compositions and environmental conditions in WWTP1 and WWTP2. This finding agrees with previous reports [8, 9] that regional WWTPs have more core bacteria taxa than global WWTPs [1]. Overall, the low similarity of the MAGs and proteins between different WWTPs suggests that extremely high genetic diversity is present in the AS ecosystem.

Due to the extremely high genetic complexity in AS, the phylogenetic tree and COG ordination analysis failed to distinguish between AS MAGs and non-AS MAGs. The major reason is that phylogenetic analysis and COG ordination are processes developed to reduce the dimensionality of multivariate data. For phylogenetic tree construction, only a limited number, usually a few hundreds, of genes coding universally conserved proteins are selected among 2000–3000 genes in a bacterial genome [21], leading to a concomitant loss of genetic information. Further loss occurs when the sequencing data are converted into

distances (distance-matrix methods) or likelihood estimations (maximum likelihood methods) or when singular sites are discarded (parsimony methods) [27, 28]. The ordination methods (including t-SNE) also suffer from information loss due to the dimension reduction [29]. Although dimension reduction is important in some cases to summarize significant information from redundant high-dimensional data [30], its application could miss the subtle dependencies in the datasets; for instance, the differences between AS and non-AS MAGs were not captured in this study. Here, we found that a machine learning approach (random forest model) accurately distinguished between AS MAGs and non-AS MAGs based on COG presence/absence because the random forest algorithm could take full advantages of high-dimensional data by constructing a multitude of decision trees [31].

The high prediction accuracy of machine learning also suggests that municipal WWTPs can select bacteria with specific functions. Although the bacterial species in different municipal WWTPs could be different [32], they may have similar deterministic functional traits to adapt themselves to the AS system. This idea complements the recent finding that the stochastic process is more important than deterministic factors in shaping the community assembly in AS based on 16S rRNA gene sequencing [1]. The higher resolution of genome-level analysis reveals that AS bacterial genomes have specific functional traits despite stochastic community assembly. Based on the random forest algorithm, we identified several functional features that are likely important for the bacteria in AS systems. Some features are primarily related to aerobic conditions in municipal WWTP bioreactors. Besides, we also found that COGs involved in sensing the nutritional contents or other environmental signals are important for bacteria in AS. This is probably related to the more frequent changes of loading rate and other conditions in wastewater treatment bioreactors than other natural environments (e.g., soil and sea water). Another functional feature is regulating of the biofilm formation, which also important for AS because most bacteria in AS are involved in floc (a specialized type of biofilm) formation. However, the role of many other COGs and their co-occurrence contributions to the machine learning model remain unexplained. It should also be noted that the protein functions inferred based on COG annotation may not be sufficient to reflect the detailed functional features of the AS. Future efforts are needed to investigate and confirm the functions of the proteins assigned to these COGs.

Despite the high prediction accuracy of the random forest algorithm, we also noted some false positive and false negative predictions. Further analysis shows that these erroneous results were not due to the quality (completeness and contamination) of the MAGs, suggesting that the random forest model could handle datasets with missing values (incomplete MAGs) and a certain level of noise (contaminated MAGs) well [33]. A small number of erroneous results are reasonable because AS is an open ecosystem, and extraneous microorganisms could be introduced into the AS through incoming raw sewage [8] or upstream biological processes [34]. In addition, the microorganisms in AS could also be easily spread to other environments via effluent discharge to receiving water bodies [35]. These stochastic propagation processes could not be captured by the machine learning model, and other technologies should be applied to identify these minor species.

Although high percentages of the metagenomic sequencing reads (50–75% for most samples) were included in the AS MAGs obtained in this study, a large number of bacterial genomes in the AS still remain unavailable due to the high complexity of the AS microbiome and microdiversity issues, which significantly hampers genome assembly and binning [12, 36]. Also, many MAGs may not be obtained due to the relatively low sequencing depths of some samples analyzed in this study (Additional file 1: Table S1). We anticipate that these genomes also possess functional features similar to those of the MAGs obtained in this study, and future investigations with higher sequencing depth based on long-read sequencing [37] or single-cell sequencing [38] are needed to confirm this hypothesis. In addition, although thousands of COGs were identified by the machine learning model as important functional features to distinguish between AS MAGs and non-AS MAGs, most of them could not be well annotated. Further investigation of these proteins would be beneficial to improve our understanding of the microbial ecology of AS systems and provide a theoretical foundation for optimizing AS processes. Moreover, it should be noted, like other metagenomic studies, incorrect contig assembly and false assignment of assembled contigs to MAGs [39] may also occur in the MAG catalog of this study. Therefore, caution should be taken when using this dataset in future studies and various analyses and experiments are encouraged to confirm the results.

## Conclusions

In summary, our work provides one of the largest genome resources for investigation of the AS microbiome. Based on this, we found that the AS MAGs have obvious plant-specific features and that few genomes and proteins are shared by different WWTPs, especially for WWTPs located in geographically distant areas. Despite the differences, specific functional traits of AS MAGs, including functions related to aerobic metabolism, nutrient sensing/acquisition, and biofilm formation, were identified by a machine learning approach based on the COG presence/absence matrix. These features are likely important for bacteria to adapt themselves in AS systems. By applying the machine learning approach, AS

MAGs could be differentiated from non-AS MAGs with an accuracy of 96.6%. The results demonstrated that machine learning approach could be a powerful tool for understanding the microbial ecology in different ecosystems.

## Methods

### Activated sludge sampling

In this study, 57 AS samples were collected from the aeration tanks of 11 full-scale municipal WWTPs in 8 cities of China for metagenomic sequencing (Additional file 1: Table S1). For the two WWTPs in Nanjing City, time-series sampling was conducted every month from January 2014 to December 2015, and 24 samples were obtained from each of the two WWTPs. For other WWTPs, sampling was conducted only once in each plant during the period from April 2017 to July 2017. Detailed information about the WWTPs is shown in Additional file 1: Table S1. All sludge samples were fixed in 50% (v/v) ethanol aqueous solution and transported on ice to the laboratory for DNA extraction.

### DNA extraction and metagenomic sequencing

DNA was extracted from the AS samples using the FastDNA™ SPIN Kit for Soil (MP Biomedicals, Irvine, CA, USA) following the manufacturer's protocol. The DNA concentration and quality were determined using a NanoDrop One spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and agarose gel (2%) electrophoresis. Metagenomic sequencing was conducted to obtain the entire genomic information from the sludge samples. DNA extracted from each AS sample was used for metagenomic library construction and then sequenced on an Illumina HiSeq X Ten platform (San Diego, CA, USA) with a paired-end (2 × 150) sequencing strategy. The raw metagenome reads have been deposited in the NCBI Sequence Read Archive and are available under the BioProject PRJNA556302.

### Collection of public activated sludge metagenomic data and metagenome-assembled genomes

In addition to the 57 AS metagenomes sequenced in this study, we also downloaded 57 other municipal AS metagenomic datasets reported in previous studies for assembly and genome binning. All of the datasets were generated on the Illumina HiSeq platform with paired-end sequencing strategy. The accession numbers and information of these datasets are shown in Additional file 1: Table S1, Table S2, and Fig. S1.

Moreover, a few thousands of bacterial MAGs in a previous study [20] were also used in this study. The MAGs obtained from the anaerobic digesters and laboratory-scale wastewater treatment reactors in this catalog were excluded. Because the seed sludge of these reactors is usually activated sludge, but the influent and operational conditions may be quite different from those of the typical aerobic reactors in municipal wastewater treatment plants. Therefore, their microbial communities may be quite different from those of the typical activated sludge. Finally, 7164 bacterial draft genomes recovered from the metagenomes of different environments in the previous study [20] were used to prepare the input data for the machine learning model.

### Metagenomic assembly and contig binning

The metagenomic data were trimmed and quality-filtered using Trimmomatic v 0.32 [40] with default parameters. Then, clean reads from all samples of each WWTP were assembled into contigs using MEGAHIT v1.1.1 [41] with the following parameters: --k-min 41 --min-contig-len 1000. Then, the clean reads of each sample were mapped to the assembled contigs using Bowtie2 v 2.2.9 [42]. A depth file was generated with the jgi_summarize_bam_contig_depths included in MetaBAT2 [43] based on the mapping results. Then, draft genomes were recovered based on tetranucleotide frequency and contig abundance using MetaBAT2 v 2.12.1 [43]. The quality of the recovered genome bins was assessed by using CheckM v 1.0.7 [44]. Open reading frames were predicted in the assembled scaffolds using Prodigal v 2.6.1 [45], CD-HIT v 4.7 [46] was used to group protein sequences into clusters based on sequence identity, and Diamond v0.9.24.125 [47] was used to compare the protein sequences obtained from different WWTPs.

### Merging of compatible bins and genome refining

The "merge" command of CheckM v 1.0.7 [44] was used to identify pairs of bins that could be merged according to the following criteria: (1) the completeness increased by ≥ 10% and the contamination increased by ≤ 1% when the bin pairs were merged; (2) the differences between mean GC of the bins were within 3%; (3) the mean coverage of the bins had an absolute percentage difference ≤ 25%; and (4) the bins had identical taxonomic classifications as determined by CheckM.

Genome refining was conducted with RefineM v0.0.24 [20]. Briefly, contigs with a GC or tetranucleotide distance outside the 98th percentile of the expected distributions were identified and removed. Contigs were also removed if their mean coverage had an absolute percentage difference ≥ 50% when compared with the mean coverage of the bin. The "taxon_profile" command of RefineM was used to taxonomically classify the genes constituting each bin, and contigs with divergent taxonomic classifications were removed with the "taxon_filter" command of RefineM. In addition, contigs with 16S rRNA genes that appear incongruent with the taxonomic identity of each bin were also identified and removed with RefineM. Only MAGs with an overall quality ≥ 50 (defined as completeness −5 × contamination) were used

for downstream analysis. After genome refining, the genome taxonomy was assigned using GTDB-Tk v 0.2.1 (https://github.com/Ecogenomics/GTDBTk). The ANIs between MAGs were determined using FastANI [48].

### Genome phylogenetic tree construction

The phylogenetic analyses were conducted with PhyloPhlAn [21] using the "dev" branch of the repository (https://bitbucket.org/nsegata/phylophlan/overview). The genome maximum likelihood phylogenetic tree was generated in Newick format using the 400 universal PhyloPhlAn markers conserved across the bacterial and archaeal domains with the following options: "--diversity high --accurate --min_num_markers 80." To avoid the crowd of tree branches, we used 1000 randomly selected AS MAGs and 1000 randomly selected non-AS MAGs to construct the tree. The final tree was reconstructed for visualization using GraPhlAn v1.1.3 [49].

### Functional genomic analysis

To identify protein domains in a genome, we annotated all of the MAGs using Prokka v 1.13.3 [50] with default parameters, and all protein domains were classified in different COGs. Then, a COG matrix was derived with MAGs in rows and the presence/absence of the COGs in each MAG as columns:

| | $COG_1$ | $COG_2$ | … | $COG_j$ |
|---|---|---|---|---|
| $MAG_1$ | 0 | 1 | … | 1 |
| $MAG_2$ | 1 | 0 | … | 0 |
| … | … | … | … | … |
| $MAG_i$ | 0 | 0 | … | $n_{ij}$ |

where the matrix element $n_{ij}$ equals 1 if $MAG_i$ encodes a protein ortholog belonging to $COG_j$ and equals 0 otherwise.

The COG matrix was used to perform t-SNE analysis with the Rtsne package (https://cran.r-project.org/web/packages/Rtsne) and was also used for the construction of the machine learning model.

### Development of the machine learning model

The COG matrix constructed based on the functional annotation of the MAGs obtained in the present study and the previous study [20] was used to formulate the machine learning model to distinguish bacteria from municipal AS and those from other environments. The final dataset consists of 9288 MAGs (2024 from AS and 7164 from other environments) and 2580 COGs and was used to train and test two machine learning models based on support vector machine and random forest algorithms. Random forest was chosen because it has higher accuracy than support vector machine. Moreover, the random forest algorithm is suitable for datasets with many features, especially when each of the features contributes little information [31].

The model training and evaluation were performed with scikit-learn (https://scikit-learn.org/), a Python package for machine learning. Both the holdout method and *k*-fold cross-validation were applied to verify the model. For the holdout method, the dataset was divided into two partitions as training (80%) and testing (20%) sets. The training set was used to train the model, and the unseen testing data were used to test the predictive ability. Overfitting is a common issue in machine learning that can occur in most models [51]. In this study, out-of-bag (OOB) estimates were applied to avoid overfitting. In addition, 10-fold cross-validation was conducted to verify that the model was not overfitted. The dataset was randomly partitioned into 10 mutually exclusive and approximately equal subsets, and one set was kept for testing while the others were used for training. This process was iterated with the 10 subsets. Furthermore, the COGs significantly contributing to the machine learning-based prediction were analyzed based on the feature importance provided by the random forest model.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s40168-020-0794-3.

**Additional file 1: Table S1.** Information about the WWTPs and activate sludge samples analyzed in this study. **Table S2.** Accession numbers of the metagenomic datasets used in this study. **Table S3.** Abundance of AS MAGs assigned to each phylum. **Table S4.** Prediction report of the random forest model. **Table S5.** Importance values and descriptions of the top 20 COGs identified by random forest model to differentiate the AS and non-AS MAGs. **Figure S1.** Geographical locations of the WWTPs where activated sludge samples were collected by us and other researchers. **Figure S2.** Associations between MAG completeness and number of contigs (a), and associations between MAG completeness and number of contigs (b). **Figure S3.** Venn diagram showing the shared and unique MAGs of WWTP1, WWTP2, WWTP3 and WWTP4. **Figure S4.** Profile of protein sequences identity between different WWTPs. The protein sequences predicted from all assembly contigs of each WWTP were compared each other with Diamond and then the best hits of the protein sequences were counted and summarized. **Figure S5.** Random forest parameter tuning and optimization. (a) Number of trees (n_estimators); (b) Tree depth; (c) Maximum features. **Figure S6.** Phylogeny of the erroneously predicted MAGs. The topology of this tree is exactly same with Fig. 1b. Extended lines were added to show positions of the erroneously predicted MAGs.

#### Acknowledgements

We thank Ying Yang, Feng Guo, Jizhou Duan, Xingtao Zhang, Kailong Huang, Fuzheng Zhao, Mingyuan Sun, Wenda Tao, Yongchao Xie, Peng Liu, Zhuoyan Shen, Haohao Sun, and Qingmiao Yu for assistance in the sludge sampling and sample treatment.

#### Authors' contributions

LY, ZX, and HR conceived and designed the study. LY performed metagenome assembly, genome binning, and machine learning analysis. LY and RM performed other data analysis. LY, RM, ZX, and WL wrote the manuscript. All authors have read and approved the final manuscript.

**Funding**
This study has received funding from the National Science and Technology Major Project of China (2017ZX07202003) and the National Natural Science Foundation of China (51878333, 51608256).

**Availability of data and materials**
The raw reads of AS metagenomes sequenced in this study have been deposited in the NCBI Sequence Read Archive and are available under the BioProject PRJNA556302. All sequence data analyzed in this study are available in public databases with the accession codes given in Additional file 1: Table S2. The MAGs obtained in this study are available in Figshare at https://figshare.com/projects/AS_MAGs/66554.

**Ethics approval and consent to participate**
Not applicable

**Consent for publication**
Not applicable

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu, China. [2]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

Received: 19 September 2019 Accepted: 20 January 2020
Published online: 11 February 2020

**References**

1. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, et al. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol. 2019;4:1183–95.
2. McIlroy SJ, Onetto CA, McIlroy B, Herbst F-A, Dueholm MS, Kirkegaard RH, et al. Genomic and in situ analyses reveal the Micropruina spp. as abundant fermentative glycogen accumulating organisms in enhanced biological phosphorus removal systems. Front Microbiol. 2018;9:1004.
3. Kitzinger K, Koch H, Lücker S, Sedlacek CJ, Herbold C, Schwarz J, et al. Characterization of the first "Candidatus Nitrotoga" isolate reveals metabolic versatility and separate evolution of widespread nitrite-oxidizing bacteria. MBio. 2018;9:e01186–18.
4. Guo F, Zhang T, Li B, Wang Z, Ju F, Liang Y-T. Mycobacterial species and their contribution to cholesterol degradation in wastewater treatment plants. Sci Rep. 2019;9:836.
5. Ayarza JM, Erijman L. Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. Microb Ecol. 2011;61:486–95.
6. Griffin JS, Wells GF. Regional synchrony in full-scale activated sludge bioreactors due to deterministic microbial community assembly. ISME J. 2017;11:500–11.
7. Ju F, Zhang T. Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant. ISME J. 2015;9:683–95.
8. Saunders AM, Albertsen M, Vollertsen J, Nielsen PH. The activated sludge ecosystem contains a core community of abundant organisms. ISME J. 2016;10:11–20.
9. Zhang T, Shao M-F, Ye L. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. ISME J. 2012;6:1137–47.
10. Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11:442–6.
11. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12:635.
12. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.
13. Sun H, Narihiro T, Ma X, Zhang X-X, Ren H, Ye L. Diverse aromatic-degrading bacteria present in a highly enriched autotrophic nitrifying sludge. Sci Total Environ. 2019;666:245–51.
14. Pérez MV, Guerrero LD, Orellana E, Figuerola EL, Erijman L. Time series genome-centric analysis unveils bacterial response to operational disturbance in activated sludge. mSystems. 2019;4:e00169–19.
15. McLellan S, Huse S, Mueller-Spitz S, Andreishcheva E, Sogin M. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. Environ Microbiol. 2010;12:378–92.
16. Shanks OC, Newton RJ, Kelty CA, Huse SM, Sogin ML, McLellan SL. Comparison of the microbial community structures of untreated wastewaters from different geographic locales. Appl Environ Microbiol. 2013;79:2906–13.
17. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51:12–8.
18. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
19. Liu Z, Hsiao W, Cantarel BL, Drábek EF, Fraser-Liggett C. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. Bioinformatics. 2011;27:3242–9.
20. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.
21. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 2013;4:2304.
22. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6.
23. Lory S, Wolfgang M, Lee V, Smith R. The multi-talented bacterial adenylate cyclases. Int J Med Microbiol. 2004;293:479–82.
24. Yoshida Y, Sasaki T, Ito S, Tamura H, Kunimatsu K, Kato H. Identification and molecular characterization of tryptophanase encoded by tnaA in Porphyromonas gingivalis. Microbiology. 2009;155:968–78.
25. Vuono DC, Benecke J, Henkel J, Navidi WC, Cath TY, Munakata-Marr J, et al. Disturbance and temporal partitioning of the activated sludge metacommunity. ISME J. 2015;9:425–35.
26. Valentín-Vargas A, Toro-Labrador G, Massol-Deya AA. Bacterial community dynamics in full-scale activated sludge bioreactors: operational and ecological factors driving community assembly and performance. PLoS One. 2012;7:e42524.
27. Sourdis J, Nei M. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol Biol Evol. 1988;5:298–311.
28. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17:368–76.
29. Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
30. Carreira-Perpinán MA. A review of dimension reduction techniques. In: Technical Report CS-96-09. University of Sheffield: Department of Computer Science; 1997.
31. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
32. Shchegolkova NM, Krasnov GS, Belova AA, Dmitriev AA, Kharitonov SL, Klimina KM, et al. Microbial community structure of activated sludge in treatment plants with different wastewater compositions. Front Microbiol. 2016;7:90.
33. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min. 2017;10:363–77.
34. Mei R, Kim J, Wilson FP, Bocher BT, Liu W-T. Coupling growth kinetics modeling with machine learning reveals microbial immigration impacts and identifies key environmental parameters in a biological wastewater treatment process. Microbiome. 2019;7:65.
35. Price JR, Ledford SH, Ryan MO, Toran L, Sales CM. Wastewater treatment plant effluent introduces recoverable shifts in microbial community composition in receiving streams. Sci Total Environ. 2018;613:1104–16.
36. Nelson WC, Maezato Y, Wu Y-W, Romine MF, Lindemann SR. Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia. Appl Environ Microbiol. 2016;82:255–67.
37. Loman NJ, Watson M. Successful test launch for nanopore sequencing. Nat Methods. 2015;12:303–4.
38. Woyke T, Doud DF, Schulz F. The trajectory of microbial single-cell sequencing. Nat Methods. 2017;14:1045–54.
39. Sharon I, Banfield JF. Genomes from metagenomics. Science. 2013;342:1057–8.
40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

41. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
43. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. Peer J Preprints. 2019;7:e27522v27521.
44. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
45. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
46. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
47. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
48. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
49. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. Peer J. 2015;3:e1029.
50. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
51. Domingos PM. A few useful things to know about machine learning. Commun ACM. 2012;55:78–87.

## Publisher's Note