Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

ELSEVIER

Annotating unknown species of urban microorganisms on a global scale unveils novel functional diversity and local environment association

Jun Wu[a,2], David Danko[c,d,2], Ebrahim Afshinnekoo[c,d,1], Daniela Bezdan[c,d,1], Malay Bhattacharyya[e,f,1], Eduardo Castro-Nallar[g,1], Agnieszka Chmielarczyk[h,1], Nur Hazlin Hazrin-Chong[i,1], Youping Deng[j,1], Emmanuel Dias-Neto[k,1], Alina Frolova[l,1], Gabriella Mason-Buck[m,1], Gregorio Iraola[n,o,p,1], Soojin Jang[q,1], Paweł Łabaj[r,1], Patrick K. H. Lee[s,1], Marina Nieto-Caballero[t,1], Olayinka O. Osuolale[u,1], Christos A. Ouzounis[v,1], Michael H. Perlin[w,1], Bharath Prithiviraj[x,ag,1], Nicolás Rascovan[y,1], Anna Różańska[h,1], Lynn M. Schriml[z,1], Torsten Semmler[aa,1], Haruo Suzuki[ab,1], Juan A. Ugalde[ac,1], Ben Young[c,d,1], Johannes Werner[ad,1], Maria Mercedes Zambrano[ae,1], Yongxiang Zhao[af,***], Christopher Mason[c,d,**], Tieliu Shi[a,b,*], MetaSUB Consortium

[a] Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai, 200241, China
[b] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University & Capital Medical University, Beijing, 100083, China
[c] Weill Cornell Medicine, USA
[d] The Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, USA
[e] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
[f] Centre for Artificial Intelligence and Machine Learning, Indian Statistical Institute, Kolkata, India
[g] Universidad Andrés Bello, Center for Bioinformatics and Integrative Biology, Facultad de Ciencias de la Vida, Argentina
[h] Jagiellonian University, Faculty of Medicine, Department of Microbiology, Poland
[i] Department of Biological Sciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, Malaysia
[j] University of Hawaii, John A. Burns School of Mecidine, USA
[k] Medical Genomics Group, A.C. Camargo Cancer Center and LIM-27 Faculdade de Medicina, USP, São Paulo, Brazil
[l] Institute of Molecular Biology and Genetics of National Academy of Science of Ukraine, Ukraine
[m] Department of Analytical, Environmental and Forensic Sciences, King's College London, UK
[n] Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Uruguay
[o] Center for Integrative Biology, Universidad Mayor, Santiago de Chile, Chile
[p] Wellcome Sanger Institute, Hinxton, United Kingdom
[q] Institut Pasteur Korea, South Korea
[r] Małopolska Centre of Biotechnology, Jagiellonian University, Poland
[s] School of Energy and Environment, City University of Hong Kong, Hong Kong SAR, China
[t] University of Colorado at Boulder, Civil, Environmental and Architectural Department, Boulder, 80303, USA
[u] Applied Environmental Metagenomics and Infectious Diseases Research (AEMIDR), Department of Biological Sciences, Elizade University, Nigeria
[v] BCPL-CPERI, Centre for Research & Technology Hellas, Thessalonica, GR, 57001, Greece
[w] Department of Biology, Program on Disease Evolution, University of Louisville, Louisville, KY, 40292, USA
[x] Reckitt Health, Montvale, NJ, USA
[y] Aix-Marseille Université, IRD, AP-HM, IHU Méditerranée Infection, France
[z] University of Maryland School of Medicine, Institute for Genome Sciences, USA
[aa] Robert Koch Institute Berlin, Germany
[ab] Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, Japan
[ac] Millennium Initiative for Collaborative Research on Bacterial Resistance, Germany
[ad] High Performance and Cloud Computing Group, Zentrum für Datenverarbeitung (ZDV), Eberhard Karls University of Tübingen, Wächterstraße 76, 72074, Tübingen, Germany

Abbreviations:: ANI, Average Nucleotide Identity; MAG, Metagenome assembled genome; OTU, Operational taxonomic unit; BGC, Biosynthetic gene cluster.

[*] Corresponding author. Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai, 200241, China.

[**] Corresponding author. Weill Cornell Medicine, USA.

[***] Corresponding author. Biological Targeting Diagnosis and Therapy Research Center, Guangxi Medical University, Nanning, 530021, China.

E-mail address: tieliushi@yahoo.com (T. Shi).

https://doi.org/10.1016/j.envres.2021.112183
Received 16 July 2021; Received in revised form 13 September 2021; Accepted 14 September 2021
Available online 9 October 2021
0013-9351/© 2021 Published by Elsevier Inc.

Please cite this article as: M, *Environmental Research*, https://doi.org/10.1016/j.envres.2021.112183

[ae] Corporación Corpogen Research Center, Bogotá, Colombia
[af] Biological Targeting Diagnosis and Therapy Research Center, Guangxi Medical University, Nanning, 530021, China
[ag] Dept. of Biology, City University of New York, Brooklyn, 11210, NY, USA

ARTICLE INFO

Keywords:
Urban environmental microbiome
Metagenomic analysis
Novel species uncovering
Functional diversity
Biosynthetic gene clusters

ABSTRACT

In urban ecosystems, microbes play a key role in maintaining major ecological functions that directly support human health and city life. However, the knowledge about the species composition and functions involved in urban environments is still limited, which is largely due to the lack of reference genomes in metagenomic studies comprises more than half of unclassified reads. Here we uncovered 732 novel bacterial species from 4728 samples collected from various common surface with the matching materials in the mass transit system across 60 cities by the MetaSUB Consortium. The number of novel species is significantly and positively correlated with the city population, and more novel species can be identified in the skin-associated samples. The in-depth analysis of the new gene catalog showed that the functional terms have a significant geographical distinguishability. Moreover, we revealed that more biosynthetic gene clusters (BGCs) can be found in novel species. The co-occurrence relationship between BGCs and genera and the geographical specificity of BGCs can also provide us more information for the synthesis pathways of natural products. Expanded the known urban microbiome diversity and suggested additional mechanisms for taxonomic and functional characterization of the urban microbiome. Considering the great impact of urban microbiomes on human life, our study can also facilitate the microbial interaction analysis between human and urban environment.

1. Introduction

With the rapidly urbanizing world, more than half of the world's population live in urban areas and the urban microbes have had an increasingly effect on human health. For example, certain urban microbes have been implicated as the potential to increase or disrupt immunoregulation and/or exaggerate or suppress inflammation. Despite several important discoveries about taxonomic diversity in cities(Rinke et al., 2013), a large amount (~50%) of species in the urban environment are still unknown. Hence, uncovering these unknown species is essential to deeply parsing of microbial interaction between human and urban environment.

With the accelerated reduction in sequencing costs, many previously unknown species' genomes are being reconstructed. Recently, the MetaSUB Consortium was established to extend our knowledge of urban microbiomes by studying mass transit systems within multiple cities worldwide. Metagenome samples (n = 4728) were collected during 2015–2017 from 60 cities around the world (Danko et al., 2021). As expected from prior work, about 50% of the quality-checked reads in the urban samples still could not be mapped to known reference genomes (Danko et al., 2021), demonstrating the magnitude of unidentified species existing in the urban environment. To uncover these species and reveal their likely functional capabilities, metagenomic binning can be implemented to obtain genomes directly from environmental samples without prior isolation (Rinke et al., 2013; Alneberg et al., 2014; Wei et al., 2019; Cao, 2020). Metagenomic reads are assembled into contigs, and subsequently clustered into metagenome assembled genomes (MAGs) on the basis of sequence composition, depth of coverage, and taxonomic affiliations (Albertsen, 2013; Kang et al., 2019). Considering the advantage of recovering unknown genomes, a growing number of related studies have been launched to reconstruct numerous MAGs, mainly associated with the gut microbiome (Parks et al., 2018a; Nayfach et al., 2016, Nayfach et al., 2019; Pasolli, 2019bib_Nayfach_et_al_2016; Almeida, 2019). These studies have recovered thousands of previously unknown genomes which provided significant phylogenetic expansions of the tree of life (Parks et al., 2018a; Tyson, 2004; Brown et al., 2015), and spurred the development of methods for assessing the quality of recovered MAGs with regards to their estimated completeness, contamination and strain heterogeneity (Parks et al., 2015; Eren et al., 2015, Eren et al., 2021Eren et al., 2015).

Here, we uncovered 732 novel bacterial species by reconstructing 5980 MAGs using the metagenome data provided by the MetaSUB Consortium. We investigated the association between those novel species and specific geographical backgrounds, as well as their putative functional capacities. The bacterial genomes uncovered here substantially increased our knowledge of urban microbiome species diversity and will facilitate a better understanding of urban microbial biodiversity.

2. Materials and methods

2.1. Metagenomic assembly and contigs binning

The samples were collected following the Danko et al.'s protocol (Danko et al., 2021). Briefly, samples were collected from various common surfaces (e.g. seat, handrail, ticket machine, palm and floor) with matching materials in the mass transit systems of 60 cities worldwide. The metadata such as time, geolocation, and scanning barcodes were also recorded. To optimally preserve the DNA, the flocked swabs used for sample collection were preserved with a storage tube containing a buffer. The samples were stored at −80 °C before DNA extraction. DNA was prepared for Illumina sequencing using the QIAGEN Gene Reader DNA Library Prep Kit I (cat. no. 180435). AdapterRemoval (Schubert et al., 2016) (version 2.17) were used to trim adaptor sequences and to remove low quality reads with default parameters. Preprocessed sequences were aligned to human genome (hg38) using Bowtie2 (Langmead and Salzberg, 2012) (version 2.3.5). Read pairs with both ends mapped to the human genome were regarded as human reads and read pairs with only one mate mapped were discarded. Read pairs with neither mate mapped to human genome were regarded as non-human reads and used for downstream analysis.

The non-human reads were then assembled with metaSPAdes (Nurk et al., 2017) (version 3.10.1). Only contigs longer than 1000 nucleotides (nts) were considered for further processing. This resulted in 1.96e7 different contigs for a total length of 5.61e10 nt. Thereafter, these contigs were binned through MetaBAT2 (Kang et al., 2015) (version 2.12.1). Depth of coverage required for the binning was inferred by mapping the raw reads back to the corresponding contigs with Bowtie2 (Langmead and Salzberg, 2012) (version 2.3.5) with the option '–local –very-sensitive-local'. Completeness, contamination and strain heterogeneity were estimated with CheckM (Parks et al., 2015) (version 1.0.13) using the lineage_wf workflow. The QS (quality score) of each

[1] Listed alphabetically.
[2] Equal contribution.

MAG was calculated as: completeness - 5 × contamination. Using the measurement of QS, only 6107 MAGs survived with QS $\geq$ 50. On the basis of these metrics, the MAGs were classified into high-quality, medium-quality and low-quality MAGs (High quality: completeness >90%, contamination <5% and strain heterogeneity <0.5%; Medium quality: completeness >50% and contamination <5%; Low quality: Others). The GUNC was also applied as an additional check for chimerism (Orakov, 2021).

The NCBI GenBank database (Sayers, 2019) and the Integrated Gut Genomes (IGG) dataset (Nayfach et al., 2019) and The Genomes from Earth's Microbiomes (GEM) database (Nayfach, 2021) were used to assign all MAGs. From the GenBank database deposited as of July 2020, we obtained all the complete bacterial and archaea genomes (Bacterial: 19,282, Archaea: 389). The IGG database (release date: May 30, 2020) contains 23,790 representative genomes for all species. The GEM database (release date: November 30, 2020) composes of 52,515 MAGs from over 10,000 metagenomes collected from diverse microbiomes. The MinHash sketch of the reference genomes were created using Mash (Ondov, 2016) (version 2.1.1) with default parameters. Then, the Mash distance between each MAG and reference genomes was calculated. At last, the dnadiff (version 1.3) from MUMmer (Kurtz, 2004) (version 4.0.0) was further used to compare each MAG and its closest reference genome. MAGs with the fraction of the MAG aligned (aligned query, AQ) no less than 60% and whole-genome average nucleotide identity (ANI) less than 95% were regarded as assigned.

2.2. Species-level de-replication of MAGs

To explore the novel genomes, we de-replicated the MAGs using an approach similar to a previously published method (Nayfach et al., 2019; Almeida, 2019). Briefly, we first calculated the Mash distance between each pair of MAGs through creating a MinHash sketch for each MAG to perform an all-against-all comparison. Then a single-linkage clustering was built from the Mash distance and the primary clusters were identified on the basis of a Mash ANI of 0.9. The Mash ANI is short of accuracy for the incomplete genomes (Olm et al., 2017), but the extremely high computation efficiency makes it highly suitable for the primary clustering. To improve the accuracy, we performed a secondary clustering in each primary cluster with the measurement of whole-genome-based ANI (gANI) calculated using dnadiff. Genomes were clustered into OTUs using average-linkage hierarchical clustering with a gANI cut-off of 0.95 using the R (version 3.6.1) package dendextend (Galili, 2015) (version 1.12.0).

2.3. Phylogenetic and taxonomic analysis

Taxonomic annotation of each OTU was performed with GTDB-TK (Chaumeil et al., 2019) (version 0.3.2, Genome Taxonomy Database version 89) using the 'classify_wf' function with default parameters. Based on the multiple sequence alignment results generated by GTDB-Tk, the OTU-only bacterial phylogeny was built with FastTree2 (Price et al., 2010) (version 2.1.10) and visualized using Graphlan (Asnicar et al., 2015) (version 1.1.3). As the phylogenetic analysis method mentioned previously (Almeida, 2019), the phylogenetic diversity was quantified by the sum of branch lengths using phytools R package (Revell, 2012).

2.4. OTUs abundance and prevalence estimation

The relative abundance of each MAG was calculated from the alignments of the non-human reads against the assemblies of the same sample. The relative abundance in each sample was defined as the number of reads aligned to the contigs of the MAG normalized by the total number of reads in the sample. The relative abundance of an OTU in a sample was calculated as the sum of the relative abundance of MAGs in the sample belonging to the OTU. The prevalence of OTUs was determined by assessing the level of genome coverage, mean and depth evenness as the approach mentioned in a previous study (Almeida, 2019), which takes both the depth and coverage into account.

2.5. Protein sharing network-based OTU annotation

The proteins for each OTU were predicted using prodigal (version 2.6.3) (Hyatt et al., 2010) with the option "-c". Protein clusters were generated as the previous method (Bin Jang, 2019). Briefly, all-versus-all comparison was performed using Diamond with option "-e 1e-5 –sensitive". Then the protein clusters were generated by using Markov cluster algorithm (Enright et al., 2002). After the clustering, we applied the Jaccard coefficient (Real and Vargas, 1996) to measure the similarities between any two OTUs. As the MCL-based method cannot handle overlaps, the Jaccard coefficient matrix was further transformed into the topological overlap matrix using the R package WGCNA (Zhang and Horvath, 2005), which was first proposed for constructing gene co-expression networks.

2.6. Functional analysis for the OTUs

Functional analysis was performed using eggNOG-mapper v1.0.3 [38] based on the eggNOG database (Huerta-Cepas et al., 2019) (v.5.0) with options "-d bact -m diamond -override –usemem –seed_ortholog_evalue 1e-5". Brite Hierarchy form KEGG was used to screen metabolic related pathways and KEGG orthology (KO) among all the KOs annotated by eggNOG. The microbial BGCs were inferred with antiSMASH (version 4.5) (Blin et al., 2019) and the number of BGCs that matched the MIBiG repository (Kautsar et al., 2020) was determined with the option "–knowclusterblast". The co-occurrence relationship between the BGCs and genera was measured using Jaccard coefficient. The significance was calculated with hypergeometric test and adjusted using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). BGC-genus pairs with adjusted p-value less than 0.01 were regarded as significantly co-occurred. The co-occurrence network was visualized using Cytoscape (Shannon, 2003) (version 3.7.2).

3. Results

3.1. Recovering genomes from the urban microbiome

There were 4728 samples collected from various common surfaces with matching materials in the mass transit systems, such as benches (both metal and wood), subway floors, kiosks, and wall tiles of 60 cities worldwide (Fig. 1A, Supplementary Figs. S1A and B). To recover genomes, the reads were first assembled into contigs after quality control and contigs with length >1000 nt were used to generated 14,080 bins (see Methods section for details). Following the criterion defined by the minimum information about a metagenome-assembled genome, we obtained 1448 high quality MAGs (more than 90% completeness, less than 5% contamination and strain heterogeneity <0.5%) and 4532 medium quality MAGs (more than 50% completeness and less than 10% contamination) (see the Methods section for details, Fig. 1. B). In the following analyses, we focused on the 5980 MAGs that met or exceeded the medium quality. Finally, 1791 samples from 45 cities with at least one MAG were left for further analysis.

To explore how many of these MAGs belong to the annotated species, we matched them against all complete bacterial and archaeal genomes (bacterial: 19,282 and archaea: 389) available as of July 2020 from the NCBI GenBank database (Sayers, 2019). According to the previously reported genome threshold for species delineation (Jain et al., 2018; Varghese, 2015), which is at least 95% average nucleotide identity (ANI) over at least 60% of the genome, 2560 MAGs were properly matched to the bacterial kingdom. To extend the reference genome assignment, we also compared the MAGs with the public 52,515 MAGs obtained from GEM database (Nayfach, 2021), 4644 species representatives in Unified

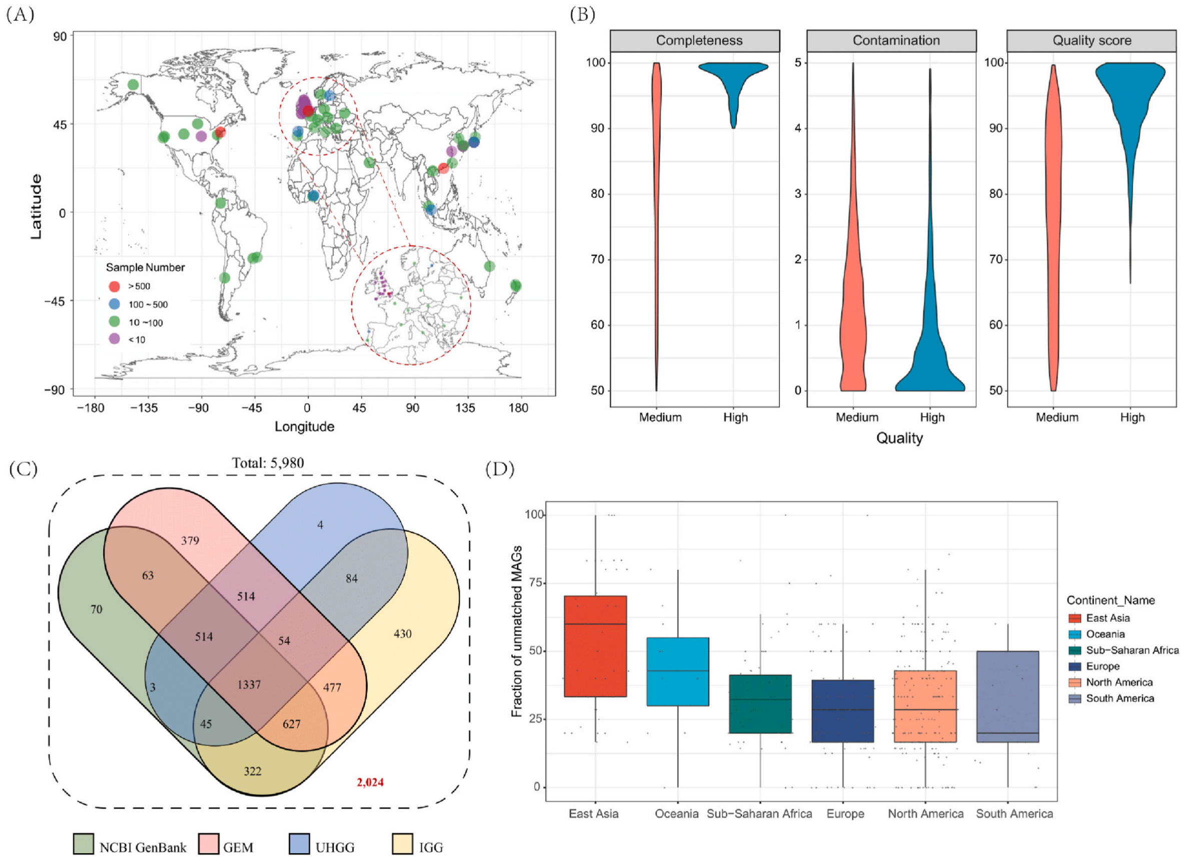

Fig. 1. Recovery of genomes from global urban microbiome and matching with the reference genomes. A: Geographical distribution of metagenomes. Number of samples were divided into four levels and colored, respectively. **B**: Quality score distribution of the MAGs that met or exceeded the medium quality. Completeness and contamination values estimated by CheckM are reported for medium- (n = 4532) and high-quality (n = 1448) MAGs. **C**: Number of the median- and high-quality MAGs matching NCBI GenBank, GEM, UHGG and IGG genomes alongside those that did not match any reference genome from either database. **D**: Fraction of MAGs that did not match any genomes from the NCBI Ref, GEM, UHGG and IGG database across different continents. Samples with total recovered MAG numbering less than 5 were excluded from the comparison.

Human Gastrointestinal Genome (UHGG) collection (Almeida, 2021) and 23,790 representative genomes for all species in the IGG dataset (Nayfach et al., 2019). Combining the alignment results with the NCBI GenBank, GEM, UHGG and IGG databases, there were still more than one-third of the MAGs novel relative to currently known genomes (Fig. 1C). To explore the association between the number of unmatched MAGs and geographic location, we compared the number of unmatched MAGs across different continents. We summarized the MAG numbers recovered in each sample, and samples with less than 5 MAGs identified were excluded from the comparison. The results showed that East Asia and Oceania sets have a larger proportion of unmatched MAGs whereas the European set possesses the smallest proportion of unmatched MAGs (Fig. 1D), indicating that these unmatched MAGs may be regional specific, and thus deeper investigation of these unmatched MAGs should be carried out.

3.2. *Taxonomic classification and geographical distribution*

Given that we identified a large number of unmatched MAGs in our datasets, we aimed to further determine their taxonomic classification to explore whether these MAGs represent novel taxa following definition of species-level OTU (Nayfach et al., 2016, 2019bib_Nayfach_et_al_2019bib_Nayfach_et_al_2016).

The 5980 MAGs were first de-replicated into estimated species-level OTUs using a two-step clustering strategy (see the Methods section for detail), yielding a total of 1304 species-level OTUs with a median quality score of 88.74 (interquartile range [IQR]: 72.59–95.57), completeness of 95.64% (IQR: 82.76%–98.79%) and contamination of 0.83% (IQR: 0.14%–1.71%). These OTUs were then classified at species-level OTUs with GTDB-tk tools (Chaumeil et al., 2019; Parks et al., 2018b). The results showed that only 2 of the 1304 OTUs were classified as archaea (*Halococcus* and *Halalkalicoccus* at genus-level, respectively). The remaining 1302 OTUs were classified as Bacteria and were shown to be consistent with previous reports from urban and rural environments (Danko et al., 2021; Liu et al., 2018; Maron et al., 2005) with the dominant assigned bacterial phyla being *Proteobacteria* (37.56%), *Actinobacteria* (27.80%), *Firmicutes* (22.04%) and *Bacteroidetes* (9.68%) (Supplementary Fig. S1C). For the following analysis, we only focused on these 1302 bacterial OTUs.

We noticed that more than half of these OTUs (56.07%) could not be classified at the species level and 6% and 0.1% could not be properly annotated at known genus and family levels respectively (Fig. 2A). Phylogenetic analysis revealed that these novel OTUs could expand the known diversity by 67% on the basis of total branch lengths (Faith, 1992), with the largest increase being within the *Proteobacteria* phylum (Fig. 2B). Several novel OTUs with high phylogenetic similarity were retrieved, belonging to *Microbacterium*, *Pseudomonas* and *Chryseobacterium*. Among these novel OTUs, the top 5 families were most represented were *Microbacteriaceae* (9.43%), *Sphingomonadaceae* (5.74%), *Xanthomonadaceae* (5.33%), *Pseudomonadaceae* (4.37%) and *Burkholderiaceae* (3.96%). Considering the amount of novel OTUs, we counted the number of both the known and unknow OTUs reconstructed in each sample. The OTU number was normalized by the sequence depth. From the results we found that the New York samples can provide

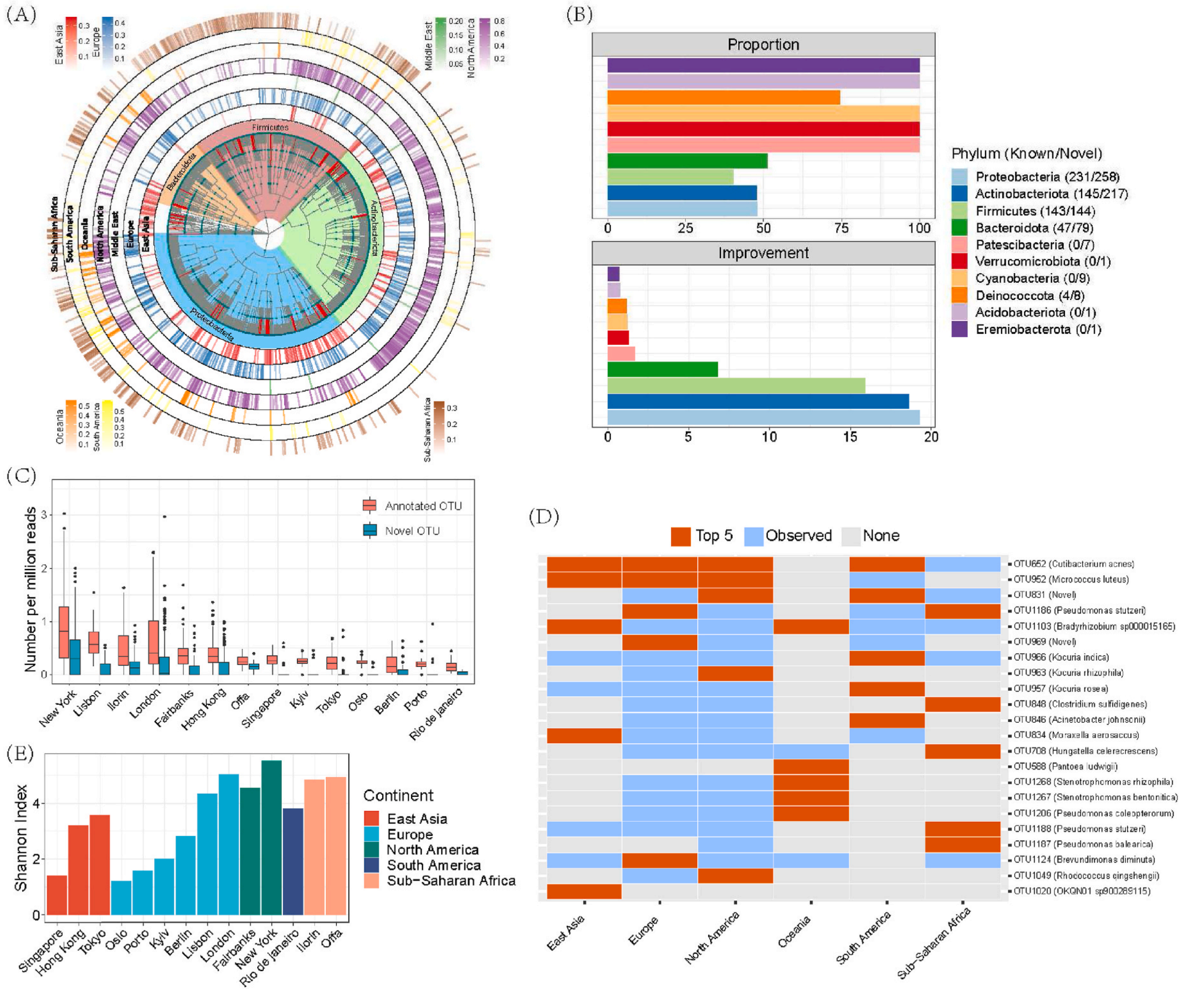

Fig. 2. Phylogeny based classification and geographical specific analysis. A: Maximum-likelihood phylogenetic tree of the 1302 recovered bacterial OTUs. The nodes and clades colored red are novel and the corresponding phylum is depicted in the first outer layer. The outermost seven layers represent the heatmap of relative abundance of OTUs in the seven continents. **B**: Level of increase in phylogenetic diversity provided by the novel OTUs, relative to the complete diversity per phylum (top) and represented as absolute total branch length (bottom). The number of classified and novel OTUs assigned to each phylum is depicted in brackets (Classified/Novel). **C**, Comparison of number of OTUs, including both the annotated and novel OTUs, reconstructed from samples obtained from different cities. **D**: The top five most frequently observed OTUs in each continent. Only continents with more than 20 OTUs are shown. **E**: The Shannon's diversity index was calculated to measure the species diversity for each city. Cities with less than 20 samples were excluded. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the most reconstructed OTUs (both annotated and novel OTUs). Although the Offa samples didn't provide too many reconstructed OTUs, but the proportion of novel OTUs in all constructed OTUs is the largest (Fig. 2C).

To determine the prevalence and diversity of the OTUs across different cities and continents, we defined the relative abundance of an OTU in a sample as the sum of the relative abundance of all MAGs belonging to this OTU. The prevalence of an OTU was estimated by considering the level of genome coverage, mean read depth, and evenness, and normalized by MAGs/OTU (see methods). To explore the geographical difference of the OTUs appearance, we calculated the frequency of OTUs observed in each continent and ranked the OTUs by descending order of their frequency (continents with less than 20 OTUs were excluded from the comparison). The five most frequently observed OTUs were listed for each continent (Fig. 2. D), and the results showed that the five most frequent OTUs in each continent mainly belong to four classes: *Actinobacteria*, *Alphaproteobacteria*, *Clostridia* and *Gammaproteobacteria*. The most frequently observed OTU (OTU652) was classified as *Cutibacterium acnes* at the species level and appeared in more than 40% of the samples. *Cutibacterium acnes* is naturally found in higher concentrations as skin flora on the chest and back, as well as in other areas with greater numbers of hair follicles (Matsen 3rdet al., 2013; Wilson, 2005). Beyond that, we found that the OTU848 and OTU708, classified as *Clostridium sulfidigenes* and *Hungatella celerecrescens* at species level, were specifically ranked at the top of the sub-Saharan African samples. In addition, two novel OTUs (OTU831 and OTU969) were ranked the highest in South America, North America and Europe, and were classified as *Psychrobacter* (OTU831) and *Kocuria* (OTU969) at the genus level. Interestingly, both two novel OTUs were almost undetectable in East Asia and Oceania (Fig. 2D). Bacteria belonging to the *Psychrobacter* genus are generally isolated from humans and can cause human infection such as endocarditis and peritonitis (Winn, 2006). *Kocuria* is a genus of gram-positive bacteria and is frequently found as normal skin flora in humans and other mammals. Sporadic reports in the literature have dealt with infections by *Kocuria* species, mostly in compromised hosts with serious underlying conditions (Savini, 2010). To investigate the species diversity for each city based on these OTUs, we then calculated the Shannon's diversity index (cities with less than 20 samples were excluded). We observed that Fairbanks and New York City (North America), London (Europe), Offa and Ilorin (Sub-Saharan Africa) were the top five locations with higher species diversity (Fig. 2E).

The above results indicated the differences of OTUs prevalence or variance among different cities or continents. To further investigate the factors that potentially associated with these differences, we attempted to explore the variation of the number of identified OTUs with respect to various factors, including city population, population density, the type of surface material that had been sampled, and coastal proximity. Our results suggested that the number of novel OTUs positively correlated with the city population and population density, supporting the hypothesis that population is likely a factor in microbial community composition (Fig. 3A and B). We also found that samples obtained from skin (left/right palm) and glass contained a higher number and more novel OTUs than from other types of surfaces (Fig. 3C). In addition, more OTUs were identified in the samples obtained from coastal cities compared to the ones from inland cities, however there was no statistically significant difference in terms of the number of novel OTUs identified (Fig. 3D).

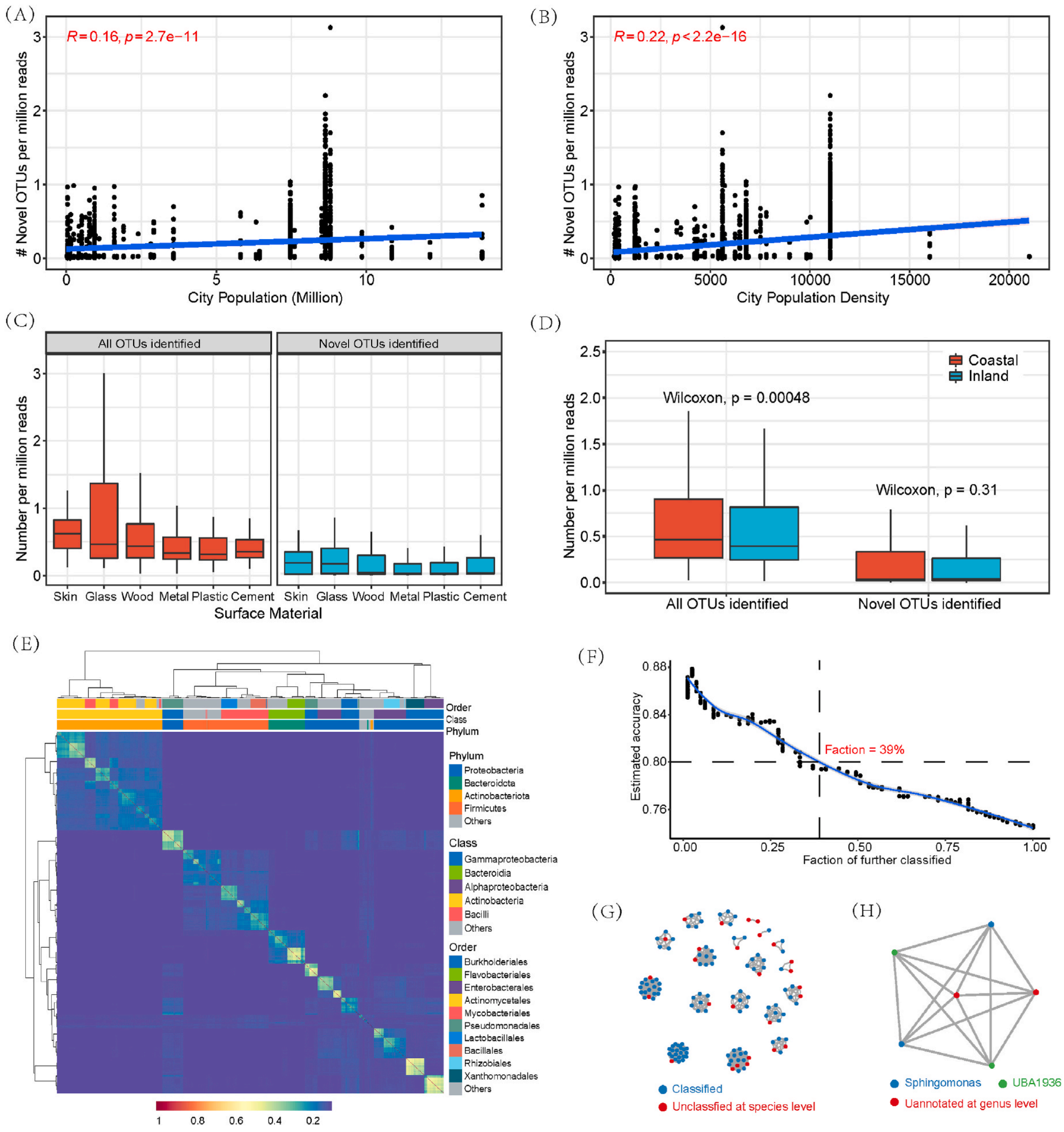

Fig. 3. Specificity analysis of novel OTUs and network-based taxonomy classification. A, Scatter plot of the city population versus the number of OTUs identified. R indicates the Pearson's coefficient. **B**, Scatter plot of the city population density versus the number of OTUs identified. **C**, Both the number of total OTUs and novel OTUs identified from samples obtained from different surface materials. Materials with less than 50 samples were excluded. **D**, Number of total OTUs and novel OTUs identified from samples obtained from coastal cities or inland cities. The Wilcox's test was used to measure the statistical significance. **E**, Heatmap depicts the Topology Overlap Matrix (TOM) among all OTUs in the analysis. The color of each cell indicates the similarity between OTUs. (blue for lower and red for higher similarity). The OTU dendrogram and module assignment are also shown along the top and colored by the phylogeny-based classification results at phylum, class and order levels respectively. **F**, Performance of the network-based classification method at genus level. Through increasing the threshold to cut the dendrogram, more OTUs can be classified along with an accuracy decrease, and 35% OTUs can be further classified with accuracy above 80%. Loess method was used to fit the model and the level of confidence interval was set as 0.95. **G**, 39% OTUs can be classified with 80% accuracy. The red dots represent the OTUs that are unclassified and blue dots represent the classified OTUs. **H**, Two novel OTUs at genus level can be further classified with 87.9% accuracy using the network-based method, and the previously unknown genus UBA1936 could also be classified as *Sphingomonas*. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.3. OTUs classification with protein sharing network

More than half of the OTUs could not be classified by phylogeny-based methods since both genome-match-based and phylogeny-based classification methods heavily depend on the information of known reference genomes. To further examine the likely biology of these novel microorganisms, we classified the novel OTUs using a protein sharing network method run only the recovered OTUs' information. The

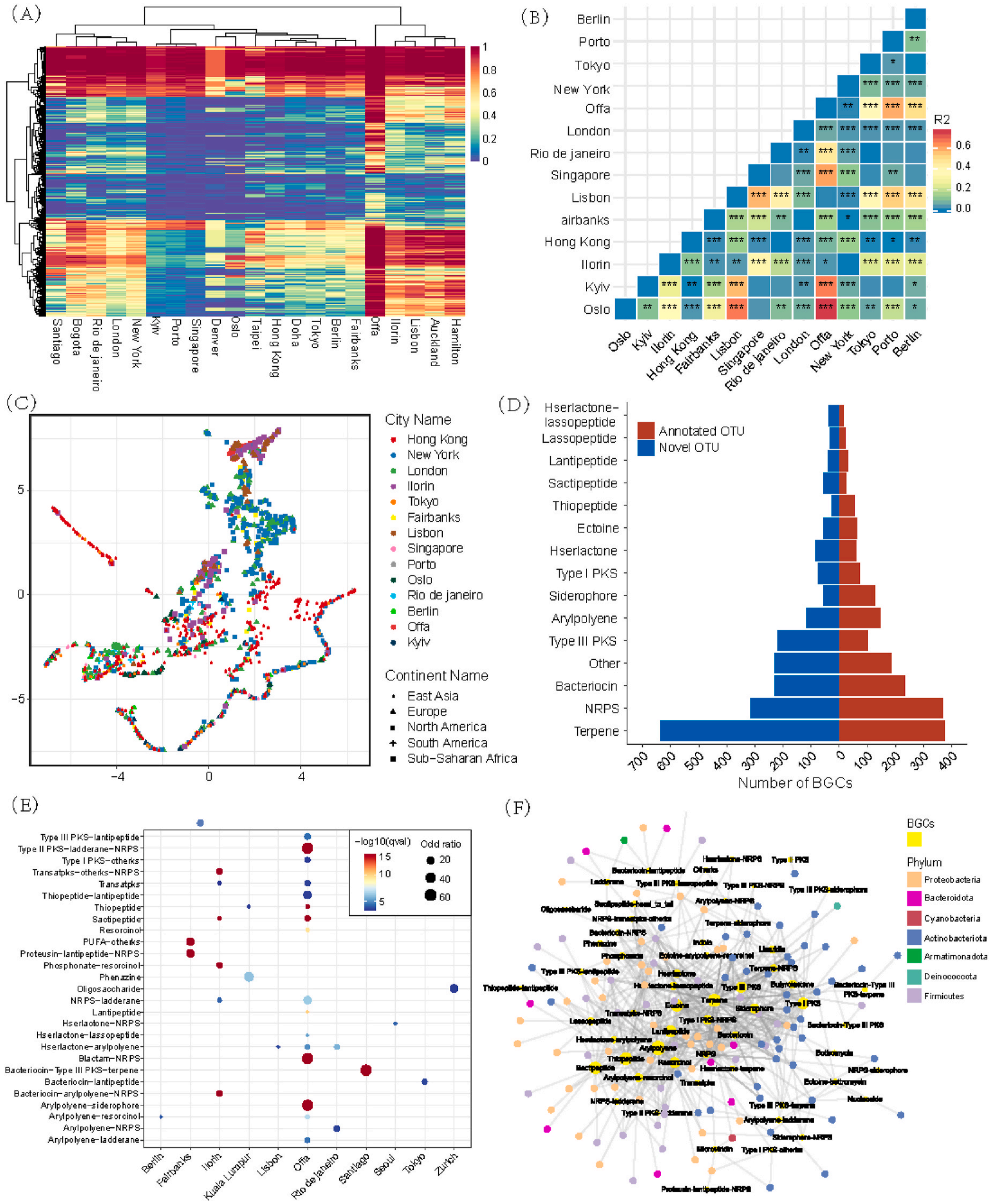

Fig. 4. Functional analyses for the uncovered OTUs. A: The heatmap depicts the observed frequency of KEGG function in each city. Blue represents lower and red higher frequency. KEGG functions absent in more than 90% samples for 90% cities are not shown. **B:** Dissimilarity among cities based on the COG function abundance. R2 is the partial R-squared. Cities with less than 20 samples were excluded. P-values are indicated as * (<0.05), ** (<0.01) or *** (<0.001). **C:** UMAP of KEGG function profiles of samples obtained from different cities. Axes are arbitrary and without meaningful scale. Cities with less than 20 samples were not shown. **D:** Number of BGCs found in all the OTUs subdivided by the BGC types. Only the fifteen most abundant categories are shown. **E:** BGCs enriched in the samples obtained from different cities. Cities with less than 20 samples were excluded. The q-value were calculated using hypergeometric test and Benjamini-Hochberg adjustment followed. **F:** Co-occurrence relationship between the BGCs and genera. The labeled nodes colored with yellow represent the BGCs and the nodes colored with other colors represented the genera. Genera nodes are colored according to the phylum annotation. The size of BGC nodes indicated the number of genera occurred together with the BGC. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

pipeline to build the protein sharing network leveraged previously proposed method (see Materials and methods) (Bin Jang, 2019). Based on the taxonomic classifications of the GTDB-tk tools, we found that TOM has a strong ability to group the OTUs at different taxonomic levels (Fig. 3E).

To evaluate the annotation performance with TOM, we calculated the module purity as a proxy accuracy metric. By choosing different thresholds, we cut the dendrogram and extracted the modules using dynamicTreeCut (Langfelder et al., 2008), and the module purity for each taxonomic level and the number of OTUs contained in the modules were calculated along with the module coverage (Supplement Fig. S2). The results showed that the TOM approach could robustly classify the OTUs at phylum, class, order and family levels, and the accuracy reached above 74% at genus level. We also noticed that even though 39% of the OTUs were unclassified at genus level with the phylogeny-based method, many could be further classified with an accuracy above 80% with the protein sharing network-based method (Fig. 3F and G). For example, two OTUs that were unclassified at genus level with the phylogeny-based method can be classified with 87.9% accuracy using the network-based method (Fig. 3H), including two genera (*Sphingomonas* and UBA1936) in the module. UBA1936 is annotated as an unclassified *Sphingomonadales* in the NCBI Taxonomy database and could also be *Sphingomonas* based on their high similarity with known *Sphingomonas* in the same module. To verify the conjecture, we inspected all the OTUs that were classified as UBA1936 at genus level and their neighbor OTUs. We found that the closest OTUs of UBA1936 were also classified as *Sphingomonas* at genus level. These results also indicated that the protein sharing network-based method can be a complimentary approach to the traditional taxonomy annotation methods.

3.4. Association between the microbial community function and the geographic location

To further uncover the functions of the 1302 species-level OTUs, we performed functional annotation for the genes predicted from these OTUs using eggNOG mapper (Huerta-Cepas, 2017, 2019). The results revealed that, 44% of predicted genes in most OTUs lack proper functional annotations (Supplementary Fig. S3A). As expected, the annotated OTUs harbor more annotated genes than novel OTUs (Supplementary Fig. S3B). To investigate the dispersion of the KO patterns across different cities, we summarized the proportion of samples in which a specific KO presented for each city (Fig. 4A). The results showed that 584 KOs presented in at least 80% of samples for all cities. We also noted the distinct KO pattern in Offa samples. To further reveal these special pathways, we extracted the KOs which presented in more than 80% Offa samples but presented in no more than one-third of samples collected from other cities. As a result, 268 genes (e.g. aacC, mecR1, and tetM) involved in 130 KOs, such as K0257 (methicillin resistance protein) and K18220 (ribosomal protection tetracycline resistance protein), were obtained (Supplementary Table S1).

We also explored the geographical distinctiveness of functional profiles (e.g. COG and KEGG) with respect to these metagenomic samples, we first extracted the COG and KEGG annotations from the eggNOG result and measured the dissimilarity through performing permutational multivariate analysis using the “adomis2” function of R package Vegans (McArdle and Anderson, 2001) with Bray-Curtis distance. The results showed that, for most cities, the functional profile was significantly distinctive. In particular, the samples obtained from Offa (Nigeria) and Oslo (Norway) showed the largest dissimilarity (ANOSIM statistic R: 0.75, p-value ≤ 0.001, Fig. 4B, Fig. S4A). Temperature between these two cities might be one of the factors which contributes to this significant functional dissimilarity. The geographical distinctiveness of KEGG function abundance was also tested using permutational multivariate analysis, and similar results were observed (Supplement Figs. S4B and C). In contrast, we observed that the KEGG function profiles of the samples obtained from Europe and North America were almost inseparable. However, a small group of samples obtained from Hong Kong showed significantly specific KEGG function profiles (Fig. 4C). Furthermore, we also estimated the dissimilarity of the samples obtained from different surface materials and the results showed that the skin samples retained the highest specificity for the KEGG function followed by the cement samples (Fig. S4D).

Moreover, we surveyed the presence of BGCs within each OTU using antiSMASH v4.5 (Blin et al., 2019). We detected 4407 BGCs, of which about half encode for Terpenes, Nonribosomal peptides (NRPS), and Bacteriocins. In addition, less than 30% of the detected BGCs had a positive match in the Minimum Information about a Biosynthetic Gene (MIBiG) cluster database (Kautsar et al., 2020), which was similar to the percentage reported for the human gut samples (Almeida, 2019). Of note, the BGCs found in the novel OTUs are significantly more than those found in annotated OTUs (Fig. 4D), further proving the valuableness of the deeper mining of these novel species. The number of BGCs encoding for Terpene or Type III Polyketide synthase (PKS) detected in the novel OTUs was nearly twice as high as those in the annotated OTUs. Terpene synthases are widely distributed in bacteria and have been reported to carry potent antimicrobial activities (Yamada, 2015; Mahizan et al., 2019). The type III PKS-derived polyketides have the potential to serve as a variety of purposes, such as pharmaceuticals, nutraceuticals, or plastic/fuel precursors (Palmer and Alper, 2019). To further reveal the geographical specific BGCs, we evaluated the enrichment of identified BGCs in different cities. The results showed that the Offa samples harbored the most specific BGCs including a series of antibiotic biosynthetic clusters (hybrid clusters), such as β-lactams-NRPS, thiopeptide, and phenazine (Fig. 4E). Beta-lactams are the most widely used class of antibiotics that have specificity for bacteria, while thiopeptides belong to a growing class of sulfur-rich, highly modified heterocyclic peptide antibiotics. Phenazines are commonly considered to be antibiotics, but they can also participate in environmental redox reactions, especially with iron. These results also revealed the different biosynthesis characteristics for different cities.

We also investigated the association between the BGCs and genera, and found that the top five BGCs associated with the most genera occurring together are arylpolyene, terpene, thiopeptide, ectoine, and sactipeptide (Fig. 4F). The co-occurring genera for these top 5 BGCs are also shown (Fig. S5). For example, arylpolyene, which is a highly abundant class of bacterial natural products and functionally related to antioxidative carotenoids (Schoner, 2016), significantly co-occurred together with *Brevundimonas*, *Pseudomonas,* and *Pantoea*. The terpenes significantly associated with *Pseudomonas*, *Corynebacterium,* and *Chryseobacterium*, and are significantly the largest class of natural products, with roles in mediating antagonistic and beneficial interactions among organisms (Gershenzon and Dudareva, 2007; Li and Yin, 2019). Based on these results, we also found that the genus *Pseudomonas* harbored the most co-occurring BGCs (37.9%) followed by *Bacillus* (29.3%) and smaller sets of other genera.

4. Discussion

With the rapid progress of urbanization, the urban environment is playing an increasing role in the human-microbe interactions. More and more studies have revealed that the environmental microbiome have a deep impact on human health. Previous studies have noted that about 50% of high-quality (>Q30) reads in urban samples could not be mapped to known reference genomes (Danko et al., 2021). Although the absolute amount of the taxonomic novelty is limited (n = 732 new species), the new species and OTUs still provide new insights into the composition of urban microbiome and the relationship between the urban, environmental microbiome and human activities.

Although we recovered thousands of MAGs, half of which are previously unknown, it is clear that the metagenome-based methods still suffer the limitation of low recovery rates for the recovery of less abundant species. We also observed that the MAGs recovery rate in the urban environmental microbial samples data is much smaller as compared to human gut samples (average 3.8 bins per sample for urban environment vs. 20.6 bins per sample for human gut (Almeida, 2019)), probably due to the higher microbial diversity in the urban environment as compared to the human gut, reinforcing that a much higher sequencing depth (>50M reads/sample vs. ~10) would likely be needed for the urban environmental microbiome analysis.

However, the number of novel species uncovered is significantly and positively correlated with the number of sequenced samples, indicating that an increased number of sample collections, as well as sequencing depth, could improve genome recovery. The larger number of novel OTUs observed in the cities with higher population density, as well as in skin samples (relative to the other surfaces and sample types) indicates that human activities may significantly influence the microbial components, and that host-environment associations should examined in detail in environmental microbiome analysis. Despite the coastal cities tend to have bigger population and population density, we did observe significant difference between the coastal and inland cities in terms of the number of novel OTUS. It may attribute to other environmental factors (e.g. temperature, rainfall and humidity) and a deeper exploration should be considered in the future.

In addition, we also show that the protein function profile has a strong ability to group the OTUs at different taxonomic levels, giving support to the concept of "phylogenetic inertia" that suggests that more closely related species will be more functionally similar whereas distantly related species will be less functionally similar (Dreiss et al., 2015). This result indicated that the function profile based taxonomic annotation method can be used as supplement to the traditional, phylogeny-based methods. Similar to the genus/species composition, the microbial function profile can also possess the geospatial stratification. The BGC results revealed that the number of BGCs detected in the novel OTUs was nearly twice as high as in the annotated OTUs, often are coding for terpene, which has been reported to carry potent antimicrobial activity, indicating that many natural compounds with potential antimicrobial activity are yet to be identified in the urban environmental microbiome. The geographical specificity of BGCs suggested that environmental microbes in different cities may possess their specific biosynthetic products. The co-occurrence between the BGCs and genera shown in this study can further guide the discovery of new and natural antimicrobials from environmental microbes.

Therefore, a comprehensive collection of bacterial genomes may allow a detailed microbial ecosystem description, providing reliable genome recovery from MAGs, as well a deeper understanding of functional associations. The novel bacterial species uncovered from urban environments can help guide improved city and biodiversity mapping, and serve as a metric that quantifies the changes of the microbiota in urban ecological systems.

5. Conclusions

In this study, we used a range of computational tools to recover thousands of species-level OTUs, using the first global urban metagenome data, and observed that half of these OTUs could not be classified at species level. We revealed the composition and functional dispersion of these novel OTUs, which were also associated with the population density, across different cities and countries. Moreover, we also retrieved the BGCs from these novel OTUs and the co-occurrence relationship between BGCs and genera was also analyzed. Considering the close relationship between the urban environment and public health, these uncovered novel OTUs can provide new insight into the hygienic environment monitoring. Our results help to fill the gap of unknown species with respect to the urban environmental microbiome and can provide more comprehensive information for investigating the unexplored biodiversity in other biomes.

Declarations

Ethic approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The OTU annotation, KEGG prevalence, OTU-BGCs occurrence, plasmid, rRNA, tRNA and viral associated data as well the code used to generate data and figures is available at https://github.com/Junwu302/Urban-MAGs. Metadata and the sequencing reads can be founded in the publicly available repository on Pangea (https://pangea.gimmebio.com/).

Funding

This work was supported by National Natural Science Foundation of China grants (Nos. 31801118, 31671377), Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01, 20692191500), Beihang University & Capital Medical University Plan (BHME-201904), The Special Fund of the Pediatric Medical Coordinated Development Center of Beijing Hospitals Authority (No. XTCX201809) and the Open Research Fund (KLATASDS2105) of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU.

Authors' contributions

J. Wu, C. Mason and T.L. Shi conceived the study. J. Wu wrote the manuscript and performed the binning and downstream bioinformatics analyses. T.L. Shi and C. Mason finalized the manuscript. D. Danko developed the assembly pipeline. C. Ouzounis, E. Disas-Neto, J. Wernr, M. M. Zambrano, Y. Osuolale, D. Bertrand and E. Elhaik revised the manuscript and contributed to the interpretation of the data. All authors read and approved the final manuscript.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this manuscript.

Acknowledgements

We thank all the members in the MetaSUB consortium for generating the raw data used in this study. We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, the Scientific Computing Unit (SCU), XSEDE Supercomputing Resources, as well as funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, the National Institutes of Health (R01AI151059), the NSF (1840275), and the Alfred P. Sloan Foundation (G-2015-13964).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2021.112183.

Contributing members of the MetaSUB consortium

Marcos Abraao, Muhammad Afaq, Ireen Alam, Gabriela E

Albuquerque, Kalyn Ali, Lucia E Alvarado-Arnez, Sarh Aly, Jennifer Amachee, Maria G. Amorim, Majelia Ampadu, Nala An, Núria Andreu Somavilla, Michael Angelov, Verónica Antelo, Catharine Aquino, Mayra Arauco Livia, Luiza F Araujo, Jenny Arevalo, Lucia Elena Alvarado Arnez, Fernanda Arredondo, Matthew Arthur, Sadaf Ayaz, Silva Baburyan, Abd-Manaaf Bakere, Katrin Bakhl, Thais F. Bartelli, Kevin Becher, Joseph Benson, Denis Bertrand, Silvia Beurmann, Christina Black, Brittany Blyther, Bazartseren Boldgiv, Gabriela P Branco, Christian Brion, Paulina Buczansla, Catherine M Burke, Irvind Buttar, Jalia Bynoe, Sven Bönigk, Kari O Bøifot, Hiram Caballero, Alessandra Carbone, Anais Cardenas, Ana V Castro, Ana Valeria B Castro, Astred Castro, Simone Cawthorne, Jonathan Cedillo, Salama Chaker, Allison Chan, Anastasia I Chasapi, Gregory Chem, Jenn-Wei Chen, Michelle Chen, Xiaoqing Chen, Ariel Chernomoretz, Daisy Cheung, Diana Chicas, Hira Choudhry, Carl Chrispin, Kianna Ciaramella, Jake Cohen, David A Coil, Colleen Conger, Ana F. Costa, Delisia Cuebas, Aaron E Darling, Pujita Das, Lucinda B Davenport, Laurent David, Gargi Dayama, Paola F De Sessions, Chris K Deng, Monika Devi, Felipe S Dezem, Sonia Dorado, LaShonda Dorsey, Steven Du, Alexandra Dutan, Naya Eady, Stephen Eduard Boja Ruiz, Jonathan A Eisen, Miar Elaskandrany, Lennard Epping, Juan P Escalera-Antezana, Iqra Faiz, Luice Fan, Nadine Farhat, Kelly French, Skye Felice, Laís Pereira Ferreira, Gabriel Figueroa, Denisse Flores, Marcos AS Fonseca, Jonathan Foox, Aaishah Francis, Pablo Fresia, Jacob Friedman, Jaime J Fuentes, Josephine Galipon, Laura Garcia, Annie Geiger, Samuel M Gerner, Dao Phuong Giang, Matías Giménez, Donato Giovannelli, Dedan Githae, Samantha Goldman, Gaston H Gonnet, Juana Gonzalez, Irene González Navarrete, Tranette Gregory, Felix Hartkopf, Arya Hawkins-Zafarnia, Nur Hazlin Hazrin-Chong, Tamera Henry, Samuel Hernandez, David Hess-Homeier, Yui Him Lo, Lauren E Hittle, Nghiem Xuan Hoan, Irene Hoxie, Elizabeth Humphries, Shaikh B Iqbal, Riham Islam, Sharah Islam, Takayuki Ito, Tomislav Ivankovic, Sarah Jackson, JoAnn Jacobs, Esmeralda Jiminez, Ayantu Jinfessa, Takema Kajita, Amrit Kaur, Fernanda de Souza Gomes Kehdy, Vedbar S Khadka, Shaira Khan, Michelle Ki, Gina Kim, Hyung Jun Kim, Sangwan Kim, Ryan J King, Kaymisha Knights, Ellen Koag, Nadezhda Kobko-Litskevitch, Giuseppe KoLoMonaco, Michael Kozhar, Nanami Kubota, Sheelta S Kumar, Lawrence Kwong, Rachel Kwong, Ingrid Lafontaine, Manolo Laiola, Isha Lamba, Hyunjung Lee, Lucy Lee, Yunmi Lee, Emily Leong, Marcus H Y Leung, Chenhao Li, Weijun Liang, Moses Lin, Yan Ling Wong, Priscilla Lisboa, Anna Litskevitch, Tracy Liu, Sonia Losim, Jennifer Lu, Simona Lysakova, Gustavo Adolfo Malca Salas, Denisse Maldonado, Krizzy Mallari, Tathiane M Malta, Maliha Mamun, Yuk Man Tang, Sonia Marinovic, BrunnaMarques, NicoleMathews, YuriMatsuzaki, MadelynMay, EliasMcComb, AdiellMelamed, Wayne Menary, Ambar Mendez, Katterinne N Mendez, Irene Meng, Ajay Menon, Mark Menor, Nancy Merino, Cem Meydan, Karishma Miah, Tanja Miketic, Eric Minwei Liu, Wilson Miranda, Athena Mitsios, Natasha Mohan, Mohammed Mohsin, Karobi Moitra, Laura Molina, Eftar Moniruzzaman, Sookwon Moon, Isabelle de Oliveira Moraes, Maritza S Mosella, Maritza S Mosella, Josef W Moser, Christopher Mozsary, Amanda L Muehlbauer, Oasima Muner, Muntaha Munia, Naimah Munim, Tatjana Mustac, Kaung Myat San, Areeg Naeem, Mayuko Nakagawa, Masaki Nasu, Bryan Nazario, Narasimha Rao Nedunuri, Aida Nesimi, Aida Nesimi, Gloria Nguyen, Hosna Noorzi, Avigdor Nosrati, Houtan Noushmehr, Diana N. Nunes, Kathryn O'Brien, Niamh B O'Hara, Gabriella Oken, Rantimi A Olawoyin, Kiara Olmeda, Itunu A Oluwadare, Tolulope Oluwadare, Jenessa Orpilla, Jacqueline Orrego, Melissa Ortega, Princess Osma, Israel O Osuolale, Oluwatosin M Osuolale, Rachid Ounit, Christos A Ouzounis, Subhamitra Pakrashi, Rachel Paras, Andrea Patrignani, Ante Peros, Sabrina Persaud, Anisia Peters, Robert A Petit III, Adam Phillips, Lisbeth Pineda, Alketa Plaku, Alma Plaku, Brianna Pompa-Hogan, Max Priestman, Bharath Prithiviraj, Sambhawa Priya, Phanthira Pugdeethosal, Benjamin Pulatov, Angelika Pupiec, Tao Qing, Saher Rahiel, Savlatjon Rahmatulloev, Kannan Rajendran, Aneisa Ramcharan, Adan Ramirez-Rojas, Shahryar Rana, Prashanthi Ratnanandan, Timothy D Read, Hugues Richard, Alexis Rivera, Michelle Rivera, Alessandro Robertiello, Courtney Robinson, Anyelic Rosario, Kaitlan Russell, Timothy Ryan Donahoe, Krista Ryon, Thais S Sabedot, Thais S Sabedot, Mahfuza Sabina, Cecilia Salazar, Jorge Sanchez, Ryan Sankar, Paulo Thiago de Souza Santos, Zulena Saravi, Thomas Saw Aung, Thomas Saw Aung, Nowshin Sayara, Steffen Schaaf, Anna-Lena M Schinke, Ralph Schlapbach, Jason R Schriml, Felipe Segato, Marianna S. Serpa, Heba Shaaban, Maheen Shakil, Hyenah Shim, Yuh Shiwa Shaleni K Singh, Eunice So, Camila Souza, Jason Sperry, Kiyoshi Suganuma, Hamood Suliman, Jill Sullivan, Jill Sullivan, Fumie Takahara, Isabella K Takenaka, Anyi Tang, Emilio Tarcitano, Mahdi Taye, Alexis Terrero, Andrew M Thomas, Sade Thomas, Masaru Tomita, Xinzhao Tong, Jennifer M Tran, Catalina Truong, Stefan I Tsonev, Kazutoshi Tsuda, Michelle Tuz, Carmen Urgiles, Brandon Valentine, Hitler Francois Vasquez Arevalo, Valeria Ventorino, Patricia Vera-Wolf, Sierra Vincent, Renee Vivancos-Koopman, Andrew Wan, Cindy Wang, Samuel Weekes, Xiao Wen Cai, Johannes Werner, David Westfall, Lothar H Wieler, Michelle Williams, Silver A Wolf, Brian Wong, Tyler Wong, Hyun Woo Joo, Rasheena Amy Zhang, Shu Zhang, Yang Zhang, Yuanting Zheng.

References

Albertsen, M., et al., 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. 31 (6), 533–538.

Almeida, A., et al., 2019. A new genomic blueprint of the human gut microbiota. Nature 568 (7753), 499–504.

Almeida, A., et al., 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. 39 (1), 105–114.

Alneberg, J., Bjarnason, B., de Bruijn, I., et al., 2014. Binning metagenomic contigs by coverage and composition. Nat. Methods 11 (11), 1144–1146.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., Segata, N., 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 3, e1029.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300.

Bin Jang, H., et al., 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. 37 (6), 632–639.

Blin, K., Shaw, S., Steinke, K., et al., 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 47 (W1), W81–W87.

Brown, C.T., Hug, L., Thomas, B., et al., 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523 (7559), 208–211.

Cao, J., et al., 2020. Diversity and abundance of resistome in rhizosphere soil, 63. Science China-Life Sciences, pp. 1946–1949.

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

Danko, D., Bezdan, D., Afshin, EE, et al., 2021. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell 184 (13), 3376–3393.e17.

Dreiss, L.M., Burgio, K.R., Cisneros, L.M., et al., 2015. Taxonomic, functional, and phylogenetic dimensions of rodent biodiversity along an extensive tropical elevational gradient. Ecography 38 (9), 876–888.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30 (7), 1575–1584.

Eren, A.M., Esen, Ö.C., Quince, C., et al., 2015. Anvi'o: an advanced analysis and visualization platformfor 'omics data. Peerj 3, e1319.

Eren, A.M., Kiefl, E., Shaiber, A, et al., 2021. Community-led, integrated, reproducible multi-omics with anvi'o. Nature Microbiology 6 (1), 3–6.

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61 (1), 1–10.

Galili, T., 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31 (22), 3718–3720.

Gershenzon, J., Dudareva, N., 2007. The function of terpene natural products in the natural world. Nat. Chem. Biol. 3 (7), 408–414.

Huerta-Cepas, J., et al., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol. Biol. Evol. 34 (8), 2115–2122.

Huerta-Cepas, J., et al., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47, D309–D314.

Hyatt, D., Chen, G.L., LoCascio, P.F., et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. 11 (1), 1–11.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9 (1), 1–8.

Kang, D.W.D., Froula, J., Egan, R., Wang, Z., 2015. MetaBAT, An efficient tool for accurately reconstructing single genomes from complex microbial communities. Peerj 3, e1165.

Kang, D.W.D., Li, F., Kirton, E., et al., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. Peerj 7, e7359.

Kautsar, S.A., et al., 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 48, D454–D458.

Kurtz, S., et al., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5 (2), 1–9.

Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24 (5), 719–720.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359.

Li, W., Yin, W.B, et al., 2019. Genetic mining of the "dark matter" in fungal natural products. Science China Life Sciences 62, 1250–1252.

Liu, H., Zhang, X., Zhang, H., et al., 2018. Effect of of air pollution on the total bacteria and pathogenic bacteria in different sizes of particulate matter. Environ. Pollut. 233, 483–493.

Mahizan, N.A., Yang, S.K., Moo, C.L., et al., 2019. Terpene derivatives as a potential agent against antimicrobial resistance (AMR) pathogens. Molecules 24 (14), 2631.

Maron, P.A., et al., 2005. Assessing genetic structure and diversity of airborne bacterial communities by DNA fingerprinting and 16S rDNA clone library. Atmos. Environ. 39, 3687–3695.

Matsen, F.A., 3rd, et al., 2013. Origin of propionibacterium in surgical wounds and evidence-based approach for culturing propionibacterium from surgical sites. J Bone Joint Surg Am 95, e1811–e1817.

McArdle, B.H., Anderson, M.J., 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82 (1), 290–297.

Nayfach, S., et al., 2021. A genomic catalog of Earth's microbiomes. Nat. Biotechnol. 39 (4), 499–509.

Nayfach, S., Rodriguez-Mueller, B., Garud, N., Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 26 (11), 1612–1625.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., Kyrpides, N.C., 2019. New insights from uncultivated genomes of the global human gut microbiome. Nature 568 (7753), 505–510.

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27 (5), 824–834.

Olm, M.R., Brown, C.T., Brooks, B., Banfield, J.F., 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11 (12), 2864–2868.

Ondov, B.D., et al., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology 17, 1–14.

Orakov, A., et al., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. 22 (1), 1–19.

Palmer, C.M., Alper, H.S., 2019. Expanding the chemical palette of industrial microbes: metabolic engineering for type III PKS-derived polyketides. Biotechnol. J. 14 (1), e1700463.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25 (7), 1043–1055.

Parks, D H, Rinke, C, Chuvochina, M, et al., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life[J]. Nat. Microbiol. 2 (11), 1533–1542.

Parks, D H, Chuvochina, M, Waite, D W, et al., 2018b. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life[J]. Nat. Biotechnol. 36 (10), 996–1004.

Pasolli, E., et al., 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from human microbiome metagenomes spanning age, geography, and lifestyle. Cell 176 (3), 649–662.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS One 5 (3), e9490.

Real, R., Vargas, J.M., 1996. The probabilistic basis of Jaccard's index of similarity. Syst. Biol. 45 (3), 380–385.

Revell, L.J., et al., 2012. Phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3 (2), 217–223.

Rinke, C., Schwientek, P., Sczyrba, A., et al., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499 (7459), 431–437.

Savini, V., et al., 2010. Drug sensitivity and clinical impact of members of the genus Kocuria. J. Med. Microbiol. 59 (12), 1395–1402.

Sayers, E.W., et al., 2019. Database resources of the national center for biotechnology information. Nucleic Acids Res. 47 (suppl_1), D23–D28.

Schoner, T.A., et al., 2016. Aryl polyenes, a highly abundant class of bacterial natural products, are functionally related to antioxidative carotenoids. Chembiochem 17 (3), 247–253.

Schubert, M., Lindgreen, S., Orlando, L., 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes 9 (1), 1–7.

Shannon, P., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.

Wilson, M., 2005. Microbial inhabitants of humans: their ecology and role in health and disease. Cambridge University Press.

Tyson, G.W., et al., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428 (6978), 37–43.

Varghese, N.J., et al., 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 43 (14), 6761–6771.

Wei, F., Wu, Q., Hu, Y., et al., 2019. Conservation metagenomics: a new branch of conservation biology. Sci. China Life Sci. 62, 168–178.

Winn, W.C., 2006. Koneman's Color Atlas and Textbook of Diagnostic Microbiology. Lippincott williams & wilkins.

Yamada, Y., et al., 2015. Terpene synthases are widely distributed in bacteria. Proceedings of the National Academy of Sciences 112 (3), 857–862.

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4 (1).