bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Global Genetic Cartography of Urban Metagenomes and Anti-Microbial Resistance

David Danko[1,2,°], Daniela Bezdan[1,2,°], Ebrahim Afshinnekoo[1,2,†], Sofia Ahsanuddin[3,†], Chandrima Bhattacharya[1,2,†], Daniel J Butler[1,2,†], Kern Rei Chng[4,†], Francesca De Filippis[5,†], Jochen Hecht[6,†], Andre Kahles[7,†], Mikhail Karasikov[7,†], Nikos C Kyrpides[8,†], Marcus H Y Leung[9,†], Dmitry Meleshko[1,2,†], Harun Mustafa[7,†], Beth Mutai[10,6,†], Russell Y Neches[8,†], Amanda Ng[4,†], Marina Nieto-Caballero[11,†], Olga Nikolayeva[12,†] , Tatyana Nikolayeva[12,†], Eileen Png[4,†], Jorge L Sanchez[13,†], Heba Shaaban[1,2,†], Maria A Sierra[1,2,†], Xinzhao Tong[9,†], Ben Young[1,2,†], Josue Alicea [1,2,‡], Malay Bhattacharyya[14,‡], Ran Blekhman[15,‡], Eduardo Castro-Nallar[16,‡], Ana M Cañas[13,‡], Aspassia D Chatziefthimiou[17,‡], Robert W Crawford[18,‡], Youping Deng[19,‡], Christelle Desnues[20,‡], Emmanuel Dias-Neto[21,‡], Daisy Donnellan[13,‡], Marius Dybwad[22,‡], Eran Elhaik[23,‡], Danilo Ercolini[5,‡], Alina Frolova[24,‡], Alexandra B Graf[25,‡], David C Green[26,‡], Iman Hajirasouliha[1,2,‡], Mark Hernandez[11,‡], Gregorio Iraola[27,‡], Soojin Jang[28,‡], Frank J Kelly[26,‡], Kaymisha Knights[13,‡], Paweł P Łabaj[29,‡], Patrick K H Lee[9,‡], Per Ljungdahl[30,‡], Abigail Lyons[13,‡], Gabriella Mason-Buck[31,‡], Ken McGrath[32,‡], Emmanuel F Mongodin[33,‡], Milton Ozorio Moraes[34,‡], Niranjan Nagarajan[4,‡], Houtan Noushmehr[35,‡], Manuela Oliveira[36,‡], Stephan Ossowski[37,‡], Olayinka O Osuolale[38,‡], Orhan Özcan[39,‡], David Paez-Espino[8,‡], Nicolas Rascovan[40,‡], Hugues Richard[41,‡], Gunnar Rätsch[7,‡], Lynn M Schriml[33,‡], Torsten Semmler[42,‡], Osman U Sezerman[39,‡], Leming Shi[43,44,‡], Le Huu Song[45,‡], Haruo Suzuki[46,‡], Denise Syndercombe Court[31,‡], Dominique Thomas[13,‡], Scott W Tighe[47,‡], Klas I Udekwu[30,‡], Juan A Ugalde[48,‡], Brandon Valentine[13,‡], Dimitar I Vassilev[49,‡], Elena Vayndorf[50,‡], Thirumalaisamy P Velavan[51,‡], María M Zambrano[52,‡], Jifeng Zhu[13,‡], Sibo Zhu[43,44,‡], Christopher E Mason[1,2,^], The International MetaSUB Consortium*

[°] Equal Contribution
[†] Researcher, Listed Alphabetically
[‡] Principal Investigator, Listed Alphabetically
* Complete List Attached

1 Weill Cornell Medicine 1300 York Ave., New York, NY 10065
2 The Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine 1305 York Ave., Y13-05, New York, NY 10065
3 Icahn School of Medicine at Mount Sinai Icahn School of Medicine at Mount Sinai, New York, New York 10029
4 Genome Institute of Singapore 60 Biopolis Street, Genome, Singapore 138672
5 Department of Agricultural Sciences, Division of Microbiology, University of Naples Federico II nan
6 Centre for Genomic Regulation Dr. Aiguader 88, 08003, The Barcelona Institute of Science and Technology, Barcelona, Spain
7 ETH Zurich, Department of Compter Science, Biomedical Informatics Group ETH Zurich, Universitätstrasse 6, 8092 Zurich
8 Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.
9 School of Energy and Environment, City University of Hong Kong, Hong Kong SAR, China School of Energy and Environment, City University of Hong Kong, Tat Chee Avenue, Kowloon
10 Kenya Medical Research Institute US Army medical Research Directrorate, Kenya.
11 University of Colorado at Boulder Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, CO
12 ETH Zurich, Functional Genomics Center Zurich Functional Genomics Center Zurich, ETH Zurich, University of Zurich, Winterthurerstrasse 190, 8057 Zurich
13 Weill Cornell Medicine Clinical and Translational Science Center 407 East 61st Street, 2nd Floor, RR-C1 New York, NY 10065
14 Machine Intelligence Unit, Indian Statistical Institute, Kolkata Machine Intelligence Unit, Indian Statistical Institute, Kolkata, 203 B. T. Road, Kolkata
15 University of Minnesota University of Minnesota, MN, USA
16 Universidad Andrés Bello, Center for Bioinformatics and Integrative Biology, Facultad de Ciencias de la Vida Avenida República 330, Santiago, Chile
17 Weill Cornell Medicine Qatar Education City Al Luqta St, Ar Rayyan, Qatar
18 California State University Sacramento Tschannen Science Complex 3000, 6000 J Street, Sacramento, CA, 95819
19 University of Hawaii John A. Burns School of Mecidine University of Hawaii John A. Burns School of Mecidine, 651 Ilalo Street, Honolulu, HI 96822
20 Aix-Marseille Université, Mediterranean Institute of Oceanology, Université de Toulon, CNRS, IRD, UM 110 163 Avenue de Luminy, Case 901, Bâtiment OCEANOMED-Méditerranée,13288 Marseille, Cedex 09, France
21 A.C.Camargo Cancer Center Laboratory of Medical Genomics, AC Camargo Cancer Center, Rua Taguá 440, São Paulo
22 Norwegian Defence Research Establishment FFI, Kjeller, Norway Norwegian Defence Research Establishment FFI, PO Box 25, NO-2027 Kjeller, Norway
23 Department of Animal & Plant Sciences, University of Sheffield Department of Animal & Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK
24 Institute of Molecular Biology and Genetics of National Academy of Science of Ukraine Institute of Molecular Biology and

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genetics of National Academy of Science of Ukraine, 150 Zabolotnogo str., Kyiv 03143, Ukraine
25 University of Applied Sciences Vienna FH Campus Wien, University of Applied Sciences, Vienna
26 Department of Analytical, Environmental and Forensic Sciences MRC Centre for Environment & Health, King
27 Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Uruguay Institut Pasteur Montevideo, Mataojo 2020, Montevideo 11400, Uruguay
28 Institut Pasteur Korea Institut Pasteur Korea, 16, Daewangpangyo-ro 712 beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, 13488
29 Małopolska Centre of Biotechnology, Jagiellonian University Małopolska Centre of Biotechnology, Jagiellonian University, Gronostajowa 7A, 30-387 Kraków
30 Stockholm University Frescativägen, 114 19 Stockholm, Sweden
31 Department of Analytical, Environmental and Forensic Sciences King's Forsenics, King's College London
32 Microba 388 Queen St, Brisbane, 4000
33 University of Maryland School of Medicine, Institute for Genome Sciences University of Maryland School of Medicine, Institute for Genome Sciences
34 Fundação Oswaldo Cruz Laboratório de Hanseníase, Fundação Oswaldo Cruz FIOCRUZ, Rio de Janeiro, Rio de Janeiro, Brasil
35 University of São Paulo, Ribeirão Preto Medical School University of São Paulo, Ribeirão Preto Medical School
36 Instituto de Patologia e Imunologia Molecular da Universidade do Porto Instituto de Patologia e Imunologia Molecular da Universidade do Porto
37 Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. 2 Institute of Medical Genetics and Applied Genomics, University of Tübingen, Calwerstrasse 7, 72076 Tübingen, Germany.
38 Applied Environmental Metagenomics and Infectious Diseases Research, Department of Biological Sciences, Elizade University 002, Wuraola-Adeojo Avenue, Ilara-Mokin, Ondo State
39 Acibadem Mehmet Ali Aydınlar University Acibadem Mehmet Ali Aydınlar University
40 Aix-Marseille Université, IRD, AP-HM, IHU Méditerranée Infection 19-21 Boulevard Jean Moulin, 13005 Marseille, France
41 Sorbonne University, Faculty of science, Institute of Biology Paris-Seine, Laboratory of Computational and Quantitative Biology 4 Place Jussieu, 75005 Paris, France
42 Robert Koch Institute Berlin Robert Koch Institute, Nordufer 20, 13353 Berlin
43 State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University Shanghai, China, 200438
44 Department of Epidemiology, School of Public Health, Fudan University Fudan University, Shanghai, China, 200032
45 Institute of Tropical Medicine, Vietnamese-German Center of Excellence 108 Military Central Hospital, N 1, Tran Hung Dao Str, Hai Ba Trung Dist
rict, 10000, Hanoi
46 Keio University Endo 5322, Fujisawa, Kanagawa, 252-0882
47 University of Vermont -Extreme Microbio 149 Beaumont ave Univ of Vermont Cancer Center HSRF 303 Burlington, VT 05405
48 Millennium Initiative for Collaborative Research on Bacterial Resistance Millennium Initiative for Collaborative Research on Bacterial Santiago, Chile
49 Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski" 5 James Bourchier Blvd., Sofia 1164, Bulgaria
50 Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK
51 Institute of Tropical Medicine, Univeristätsklinikum Tübingen, Wilhelmstrasse 27, 72074,Tübingen
52 Corporación Corpogen Corporación Corpogen, Carrera 4 #20-41, Bogotá, DC, 110311, Colombia

^To whom correspondence should be addressed: Christopher E. Mason (chm2042@med.cornell.edu).

Abstract

Although studies have shown that urban environments and mass-transit systems have distinct genetic profiles, there are no systematic studies of these dense, human/microbial ecosystems around the world. To address this gap in knowledge, we created a global metagenomic and antimicrobial resistance (AMR) atlas of urban mass transit systems from 58 cities, spanning 3,741 samples and 4,424 taxonomically-defined microorganisms collected for from 2015-2017. The map provides annotated, geospatial details about microbial strains, functional genetics, antimicrobial resistance, and novel genetic elements, including 10,928 novel predicted viral species. Urban microbiomes often resemble human commensal microbiomes from the skin and airways, but also contain a consistent “core” of 61 species which are predominantly not human commensal species. Conversely, samples may be accurately (91.4%) classified to their city-of-origin using a linear support vector machine over taxa. These data also show that AMR density across cities varies by several orders of magnitude, including many AMRs present on plasmids with specific cosmopolitan distributions. Together, these results constitute a high-resolution global metagenomic atlas, which enables the discovery of new genetic components of the built human environment, highlights potential forensic applications, and provides an essential first draft of the global AMR burden of the world’s cities.

Keywords: Built Environment, metagenome, global health, antimicrobial resistance, AMR markers

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The high-density urban environment has historically been home to only a fraction of all people, with the majority living in rural areas or small villages. Since 2014, the situation has reversed, with an estimated 55% of the world's population now living in urban areas (United Nations, 2018). Since the introduction of germ theory and John Snow's work on cholera, it has been clear that people in cities interact with microbes in ways that may be markedly different than in rural areas (Neiderud, 2015). Microbes in the built environment have been implicated as a possible source of contagion (Cooley *et al*., 1998) and other syndromes, like allergies, are associated with increasing urbanization (Nicolaou *et al*., 2005). These data indicate that cities in general have an impact on human health, although the mechanisms underlying impact are broadly variable and often little understood. Indeed, our understanding of microbial dynamics in the urban environment outside of pandemics is limited (Gilbert and Stephens, 2018).

Technological advances in next-generation sequencing (NGS) and metagenomics have created an unprecedented opportunity for rapid, global studies of microorganisms and their hosts, providing researchers, clinicians, and policymakers with a more comprehensive view of the functional dynamics of microorganisms in a city. The NGS methods enable culture-independent sampling of the microorganisms, with the potential for both taxonomic and functional annotation. This is particularly important for surveillance of microorganims as they acquire antimicrobial resistance (AMR) (Fresia *et al*., 2019). Metagenomic methods enable nearly real-time monitoring of organisms, AMR genes, and pathogens as they emerge within a given geographical location, and also have the potential to reveal hidden microbial reservoirs and detect microbial transmission routes as they spread around the world (Zhu *et al*., 2017). A molecular map of urban environments will enable significant new research on the impact of urban microbiomes on human health.

Urban transit systems - such as subways and buses - are a daily contact interface for the billions of people who live in cities. Notably, urban travelers bring their commensal microorganisms with them as they travel and come into contact with organisms and mobile elements present in the environment, including AMR markers. The urban microbiome, however, has only been profiled with comprehensive metagenomic methods in a select few cities (Afshinnekoo *et al*., 2015; Hsu *et al*., 2016; Kang *et al*., 2018; MetaSUB International Consortium, 2016; Mason *et al*., 2016) on a limited number of occasions. This leaves a gap in scientific knowledge about a microbial ecosystem that affects the majority of the global population. Human commensal microbiomes have been found to vary widely based on culture and geography and geographically-constrained studies, while informative, may prove to miss similar differences (Brito *et al*., 2016). Moreover, data on urban microbes and AMR genes are urgently needed in developing nations, where antimicrobial drug consumption is expected to rise by 67% by 2030 (United Nations, 2016; Van Boeckel *et al*., 2015), both from changes in consumer demand for livestock products and an expanding use of antimicrobials - both of which can alter AMR profiles of these cities.

The International Metagenomics and Metadesign of Subways and Urban Biomes (MetaSUB) Consortium was launched in 2015 to address this gap in knowledge on the density, types, and dynamics of urban metagenomes and AMR profiles. Since then, we have developed standardized collection and sequencing protocols to process 3,741 samples across 58 cities worldwide (**Table 1, S1**). Sampling took place at three major time points: a pilot study in 2015-16 and two global city sampling days (gCSD) in mid 2016 and mid 2017, with most samples taken on June 21st. Each sample was sequenced with an average of 6M 125bp paired-end reads using Illumina NGS sequencers (see Methods). To create a consistent analysis of our large dataset, we generated an open-source analysis pipeline (MetaSUB Core Analysis Pipeline, CAP), which includes a integrated set of peer-reviewed, metagenomic tools for taxonomic identification, k-mer analysis, AMR gene prediction, functional profiling, annotation of particular microbial species and geospatial mapping.

This first global metagenomic study of urban microbiomes reveals a consistent "core" urban microbiome across all cities, but also distinct geographic variation, which may reflect epidemiological variation and that also enables new forensic, source-tracking capabilities. More importantly, our data demonstrate that a significant fraction of the urban microbiome remains to be characterized. Though 1,000 samples are sufficient to discover roughly 80% of the observed taxa and AMR markers, we continue to observe novel taxa and genes at an ongoing discovery rate of approximately one new species and one new AMR marker for every 10 samples. Notably, this genetic variation is affected by various environmental factors (e.g., climate, surface type, latitude, etc.) and samples show greater diversity near the equator. Moreover, sequences associated with AMR markers are widespread, though not necessarily abundant, and show geographic specificity. Here, we present the results of our global analyses and a set of tools developed to access and analyze this extensive atlas, including: two interactive map-based visualizations for samples (metasub.org) and AMRs (resistanceopen.org), an indexed search tool over raw sequence data

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(dnaloc.ethz.ch), a Git repository for all analytical pipelines and figures, and application programming interfaces (APIs) for computationally accessing results (github.com/metasub/metasub_utils).

Results

We first investigated the distribution of microbial species across the global urban environment. Specifically, we asked whether the urban environment represents a singular type of microbial ecosystem or a set of related, but distinct, communities, especially in terms of biodiversity. We observed a bimodal distribution of taxa prevalence across our dataset (**Figure 1**), which we used to define two separate sets of taxa based on the inflection points of the distribution: the putative "sub-core" set of urban microbial species that are consistently observed (>70% of samples) and the less common "peripheral" (<25% of samples) species. We also define a set of true "core" taxa which occur in almost all samples (>95%) from the global dataset (**Table 1**).

Table 1: Sample Counts, The number of samples collected from each region.

Region	Pilot	CSD16	CSD17	Other Studies	Total
East Asia	22	25	1244	0	1291
Europe	131	235	845	71	1282
Middle East	0	66	14	0	80
North America	28	221	158	184	591
Oceania	0	31	32	0	63
South America	20	43	29	34	126
Sub Saharan Africa	0	92	191	0	283
Background Control	0	18	49	0	67
Lab Negative Control	9	0	24	7	40
Positive Control	9	1	33	7	50

Applying these thresholds, we identified 1,145 microbial species (**Figure 1B**) that make up the sub-core urban microbiome with 61 species in the "core" microbiome (**Figure 1A**). Core and sub-core taxa classifications were further evaluated for sequence complexity and genome coverage on a subset of samples. Of the 1,206 taxa with prevalence greater than 70%, 69 were flagged as being low-quality classifications, based on genome coverage and marker abundance (see methods). The sub-core microbiome was principally bacterial, with just one eukaryotic taxa identified (*S. cerevisiae*). The three most common bacterial phyla across the world's cities ordered by the number of species observed were Proteobacteria, Actinobacteria, and Firmicutes. To test for possible geographic bias in our data we normalized the prevalence for each taxa by the median prevalence within each city. The two normalization methods broadly agreed (**Figure 1B**).

This trend can also be observed in the global co-occurrence map of microbial species (**Figure 1C**) which shows a single, strongly co-occurring group and a minor tail of co-occurrence but no secondary cluster. Despite their global prevalence, the core taxa are not uniformly abundant across all cities. Many species exhibited a high standard deviation and kurtosis (calculated using Fisher's definition and normal kurtosis of 0) than other species (**Figure 1D**). Furthermore, some species show distinctly high mean abundance, often higher than the core species, but more heterogeneous global prevalence. For example, *Salmonella enterica* is identified in less than half of all samples, but is the 12th most abundant species based on the fraction of mapped DNA. The most relatively abundant microbial species was *Cutibacterium acnes* (**Figure 1D**), which had a comparatively stable distribution of abundance across all samples. *C. acnes* is known as a prominent member of the human skin microbiome. To test for any biases arising from uneven geographic sampling, we measured the relative abundance of each taxon by calculating the fraction of reads classified to each particular taxon, and compared the raw distribution of abundance to the distribution of median abundance within each city (**Figure 1D**); the two measures closely aligned. Also, an examination of the positive and negative controls indicates that these results are not likely due to contamination or batch effect (**Figure S1**). In total, we observed 61 core taxa (>95%), 1,145 sub-core taxa (70-95%) 2,466 peripheral taxa (<25%), and 4,424 taxa across all samples.

The taxonomic classification used (KrakenUniq) revealed that much of the diversity in our data represents sequences of novel or unknown provenance. Indeed, the average fraction of DNA unclassified at the species level across all of our samples is approximately 50% (**Figure 1E**), mirroring results seen by previous urban microbiome works (Afshinnekoo *et al*., 2015; Hsu *et al*., 2016). As humans are a major part of the urban environment, the DNA in our samples could be expected to resemble commensal human

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

microbiomes. To investigate this, we compared non-human DNA fragments from our samples to a randomized set of 50 samples from 5 commensal microbiome sites in the Human Microbiome Project (HMP) (Consortium *et al*., 2012) (stool, skin, airway, gastrointestinal, urogenital). We used MASH to perform a k-mer based comparison of our samples vs. the HMP samples, which showed more similarity between MetaSUB samples and those from skin and airway sites (**Figure 1F**, **Figure S2**). A similar comparison based on taxonomic profiles, rather than k-mers, revealed greater similarity between MetaSUB samples and the skin and airway commensal microbiomes (**Figure S2**) than other body sites.

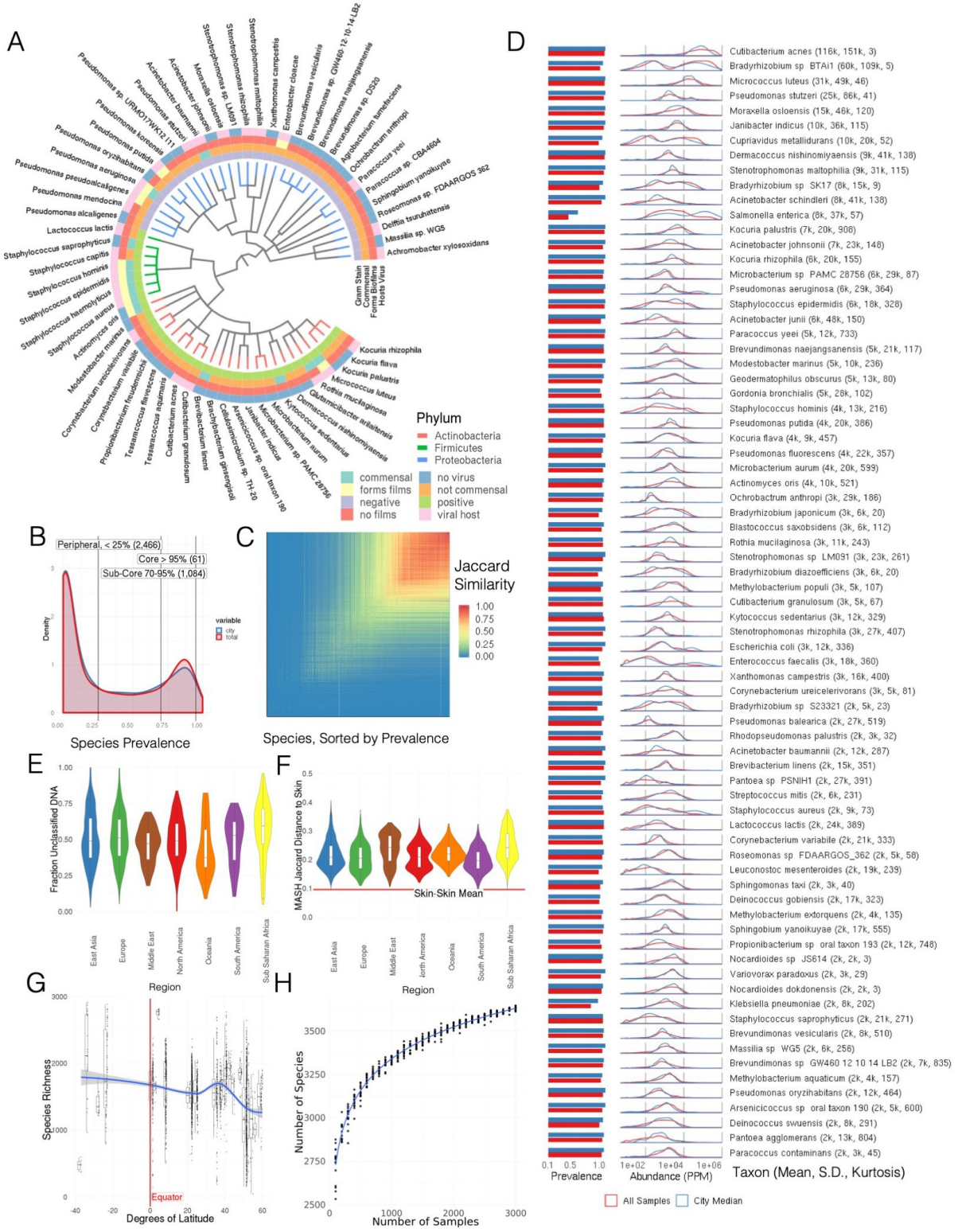

Figure 1: The core and global metagenome *A) Taxonomic tree showing 61 core taxa, colored by phylum and annotated according to gram stain, ability to form biofilms, predicted association with a virus, and whether the*

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bacteria is a human commensal species B) Distribution of species prevalence from all samples and normalized by cities. Vertical lines show defined group cutoffs. C) Co-occurrence of species within samples. Co-occurrence is measured by Jaccard Index, species are sorted by prevalence. D) prevalence and distribution of relative abundances of the 75 most abundant taxa. Mean relative abundance, standard deviation, and kurtosis of the abundance distribution are shown. E) Fraction of DNA in each sample which was not classified, by continent. F) MASH (k-mer based) similarity between MetaSUB samples and HMP skin microbiome samples, by continent. G) Species Richness (Alpha Diversity) vs. Latitude. H) Rarefaction analysis showing the number of species detected in randomly-chosen sets of samples.

Given that a large fraction of DNA in our samples could not be classified, it is possible that the unclassified DNA in our samples is from novel taxa which are not human commensals, such as those from soil. To test this, we processed 28 metagenomic soil samples (Bahram *et al*., 2018) using the same pipeline, and then compared these soil samples to MetaSUB samples using MASH. Our samples were very dissimilar from the soil samples (**Figure S2**), even in comparison to human skin microbiomes. This suggests that the unclassified DNA may represent heretofore uncharacterized taxa that are not known commensals being shed into the environment. To estimate the number of likely unobserved taxa, we performed a rarefaction analysis on the taxa that were identified. By estimating the number of taxa identified for different numbers of samples, we see a diminishing trend (**Figure 1H**), but one that is not yet approaching the asymptote. Indeed, the rarefaction curve did not reach a plateau and, even after including all samples, it still shows an expected marginal discovery rate of roughly one new species for every 10 samples added to the study. Thus, we estimate that most (~80%) of the classifiable taxa in a the urban microbiome could be identified with 1,000 samples. However, this new diversity is likely not evenly distributed, with the bulk of new discoveries expected to be made from equatorial regions and cities in the Southern hemisphere.

To examine this quantitatively, we next investigated the impact of geography on sample composition. In ecology, an increasing distance from the equator is associated with a decrease in taxonomic diversity (O'Hara *et al*., 2017). Indeed, the MetaSUB dataset recapitulates this result and identifies a significant, albeit small, decrease in taxonomic diversity as a function of absolute latitude ($p < 2e^{-16}$, $R^2 = 0.06915$); samples are estimated to lose 6.9 species for each degree of latitude away from the equator (**Figure 1G**). The effect of latitude on species diversity is not purely monotonic, since several cities have higher species diversity then their latitude would predict. This is expected, as latitude is only a rough predictor of a city's climate. While this is an observation consistent with ecological theory, the samples are heavily skewed by the location of the target cities, as well as the prevalence of those cities in specific latitude zones of the northern hemisphere.

Global Diversity Varies by Geography

Despite the core urban microbiome present in almost all samples, there was also clear subsets of variation in both taxonomy and localization. We calculated the Jaccard distance between samples measured by the presence and absence of species, and then performed a dimensionality reduction of the data using UMAP (Uniform Manifold Approximation and Projection, McInnes *et al*. (2018)) for visualization (**Figure 2A**). Jaccard distance was correlated with distance based on Jensen-Shannon Divergence, which accounts for relative abundance, and k-mer distance calculated by MASH, which is based on the k-mer distribution in a sample (**Figure S3A, B, C**). In principle, Jaccard distance can be influenced by read depth, however the total number of species identified stabilized at 100,000 reads (**Figure S4**) compared to an average of 6.01M reads per sample. Samples collected from Europe and East Asia were distinct, but the separation between other continents was less clear. A similar trend was found in an analogous analysis based on functional pathways rather than a taxonomy (**Figure S5**), which indicates geographic stratification of the metagenomes at both the functional and taxonomic levels. These findings confirm and extend earlier analyses performed on a fraction of the MetaSUB data which were run as a part of CAMDA Challenges in years 2017 and 2018 (Harris *et al*., 2018; Walker and Datta, 2019).

We next quantified the degree to which metadata covariates influence the taxonomic composition of the samples using MAVRIC, a statistical tool to estimate the sources of variation in a count-based dataset (Moskowitz and Greenleaf, 2018). We identified several covariates which influenced the taxonomic composition of our samples, including: continent, elevation above sea-level, average temperature, surface material, and proximity to the coast. The most important factor, which could explain 11% of the variation in isolation, was the continent from which a sample was taken. The other four factors ranged from explaining 2% to 7% of the possible variation in taxonomy in isolation (**Table S2**). We note that many of the factors were confounded with one another so together explain less diversity than their sum. Surprisingly, the population density of the sampled city appeared to have no discernible effect on taxonomic variation, indicating diversity is driven more by non-human factors. This is consistent with our

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

observation that the urban microbiome is not merely derivative of the human skin microbiome.

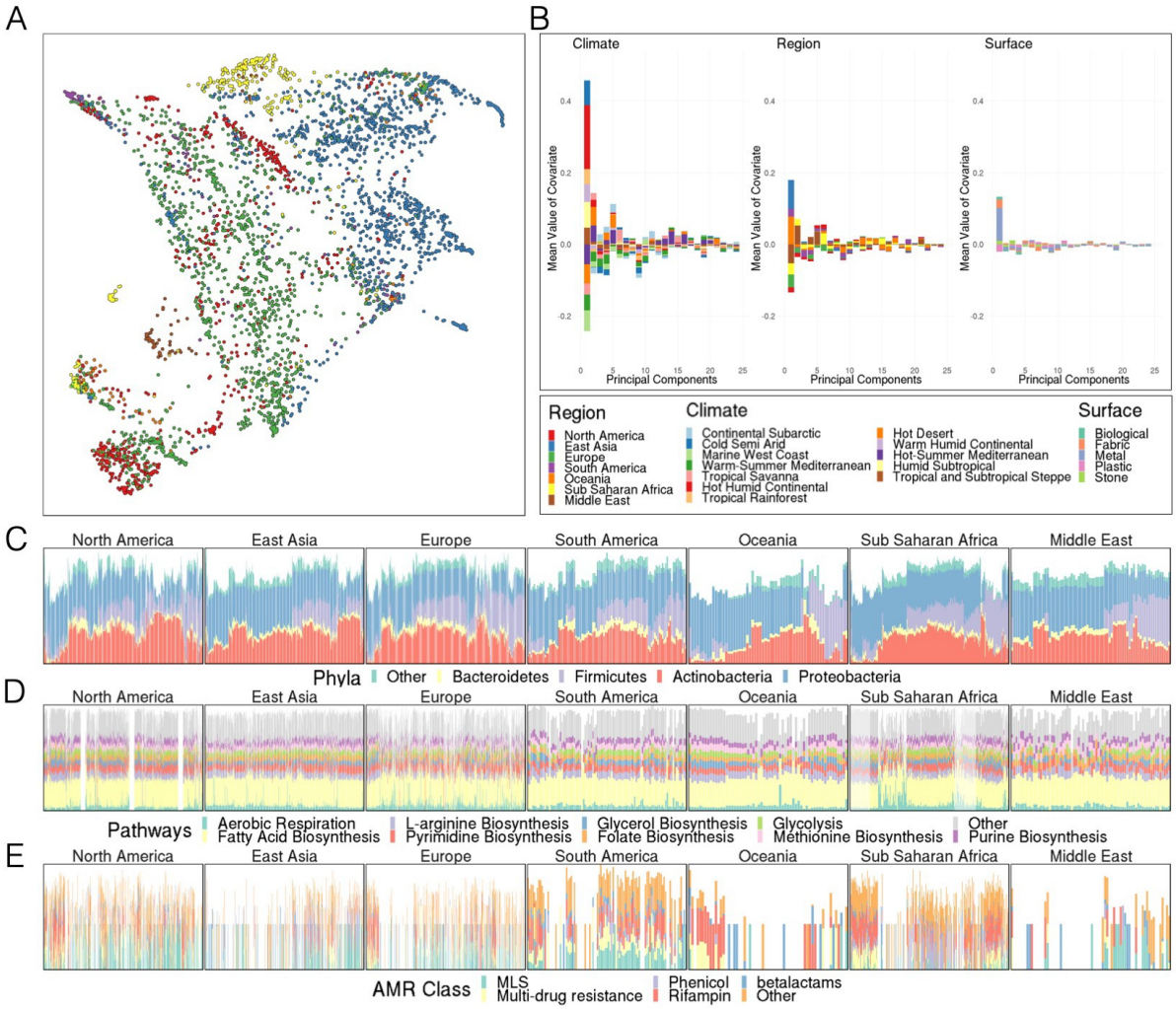

Figure 2: Functional, Taxonomnic, and AMR differences at global scale. *A) UMAP of taxonomic profiles based on Jaccard distance between samples. Colored by the region of origin for each sample. Axes are arbitrary and without meaningful scale. B) Association of the first 25 principal components of sample taxonomy with climate, continent, and surface material. C) Distribution of major phyla, sorted by hierarchical clustering of all samples and grouped by continent. D) Distribution of high-level groups of functional pathways, using the same order as taxa (C). E) Distribution of AMR genes by drug class, using the same order as taxa (C). Note that MLS is macrolide-lincosamide-streptogramin.*

To quantify how the principle covariates, climate, continent, and surface material impacted the taxonomic composition of samples, we also performed a Principal Component Analysis (PCA) on our taxonomic data normalized by proportion and identified principal components (PCs) which were strongly associated with a metadata covariate in a positive or negative direction (PCs were centered so an average direction indicates an association). We found that the first two PCs (representing 28.0% and 15.7% of the variance of the original data, respectively) associated strongly with the city climate while continent and surface material associate less strongly (**Figure 2B**).

Next, we tested whether geographic proximity (in km) of samples to one another had any effect on the variation, since samples taken from nearby locations could be expected to more closely resemble one another. Indeed, for samples taken in the same city, the average JSD (Jensen-Shannon distance) was weakly predictive of the taxonomic distance between samples, with every increase of 1km in distance between two samples representing an increase of 0.056% in divergence ($p < 2e^{-16}$, R2 = 0.01073, **Figure S2**). This suggests a "neighborhood effect" for sample similarity analogous to the effect described by Meyer *et al*. (2018), albeit a very minor one. To reduce bias that could be introduced by samples taken from precisely the same object we excluded all pairs of samples within 1km of one another.

At a global level, we examined the prevalence and abundance of taxa and their functional profiles between

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cities and continents. These data showed a fairly stable phyla distribution across samples, but the relative abundance of these taxa is unstable (**Figure 2C**) with some continental trends. In contrast to taxonomic variation, functional pathways were much more stable across continents, showing relatively little variation in the abundance of MetaCyc categories annotated by HUMAnN2 (**Figure 2D**). This pattern may also be due to the more limited range of pathway classes and their essential role in cellular function, in contrast to the much more wide-ranging taxonomic distributions examined across metagenomes. Classes of antimicrobial resistance were observed to vary by continent as well. Many samples had no AMR genes identified but clusters of AMR classes were observed to occur in groups of taxonomically similar samples (**Figure 2E**).

We then quantified the relative variation of taxonomic and functional profiles by comparing the distribution of pairwise distances in taxonomic and functional profiles. Both profiles were equivalently normalized to give the probability of encountering a particular taxon or pathway. Taxonomic profiles have a mean pairwise Jensen-Shannon Divergence (JSD) of 0.61 while pathways have a mean JSD of 0.099. The distributions of distances are significantly different (Welch's t-test, unequal variances, $p < 2e^{-16}$), which is consistent with observations from the Human Microbiome Project, where metabolic function varied less than taxonomic composition (HMP Consortium *et al*., 2012; Lloyd-Price *et al*., 2017) within samples from a given body site.

Diversity and Novel Sequences

We next examined the uniqueness of metagenomic sequences from the various MetaSUB cities, including those which do not map to any known species. To facilitate characterization of novel sequences, we created a searchable web interface (GeoDNA, **Figure 3A**) to enable mapping of raw sequences against our dataset. Users can submit sequences that are processed against a k-mer graph-based representation of our data (see methods). Query sequences are mapped to samples, metadata, and a set of likely sample hits, with a tunable sequence similarity (e.g. 80% or 90%) that is returned to the user. This allows researchers a new capacity to map the diversity of this dataset and rapidly identify the range of various genetic sequences around the world, including from novel genetic elements, which showed geographical and continental specificity (**Figure 3A**).

Since novel sequences and various taxonomic groups showed differentially abundance between cities, we next examined how predictive these sequences could be for tracing a samples' origin. To quantify the strength of the geographic separation, we trained a Linear Support Vector Machine (SVM) classifier to predict the city-of-origin from binary taxonomic data (present or absent) reduced to 1,000 dimensions by principle component analysis. We trained the model on 90% of the samples in our dataset and evaluated its performance on the remaining 10%. This classifier achieved 91.4% precision and 91.4% recall on the held-out test data, and the results on test data were not dominated by cities with many samples (**Figure S6**). We also added gaussian noise to each variable in a test dataset (mean of 0, standard deviation proportional to the mean of the variable) and performed repeated classification testing. The classifier was robust to noise until the standard deviation of that noise exceeded the mean of the variable, at which point performance of the classifier rapidly deteriorated (**Figure S6**). The successful geographic classification of samples demonstrates distinct city-specific trends in the detected taxa, that may enable future forensic biogeographical capacities.

While the taxonomic distributions are geographically distinct for our sampled cities (**Figure S7**), predicting the geographic origin of samples from cities not represented in the training dataset would likely be harder. This application is relevant for forensics. As a preliminary investigation into the predictive forensic capacity of our dataset, we also trained classifiers to predict various metadata features on held-out cities. Models were trained as above, except that all samples from a given city were held out, and PCA was used to reduce the data to just 100 dimensions. After training the model we classified samples from the held-out city for the appropriate metadata features (see methods). Mean accuracy by city for features varied (**Table S3**), from 38% for climate classification to 65% for city population. Different held-out cities were more predictable than others (**Figure S6**), in particular features for a number of British cities were roughly similar and could be predicted effectively. Predicting different metadata features enables forensic classification of samples by assigning smaples to several broadly limiting categories, thus reducing a sample's potential provenance.

However, the utility of geospatial reduction is variable, since city-specific taxa are not uniformly distributed (**Figure 3B**). To quantify this, we developed a score to reflect how endemic a given taxon is within a city, which reflects upon the forensic usefulness of a taxon. The Endemicity score (ES) is defined as ES = P (City Taxa) | P (Taxa City), with the probabilities both given by empirically observed fractions. This score is designed to simultaneously reflect the chance that a taxon could identify a given city and that that taxon could be found within the given city. A high endemicity score for a taxon in a given city could

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be evidence of the evolutionary advantage that the taxon has adapted to for a particular environment. However, neutral evolution of microbes within a particular niche is also possible and the ES alone does not distinguish between these two hypotheses. While the ES only considers taxa which are found in a city, a forensic classifier could also take advantage of the absence of taxa for a similar metric. The ES shows a clear inflection of values at 0.000025 (**Figure 3B**, inset), below that level we filter scores. Some cities, like Offa (Nigeria), host many endemic taxa while others, like Zurich (Switzerland), host fewer endemic species. Since some cities from well sampled continents (e.g., Lisbon, Hong Kong) also host many endemic species, these data suggest that ES may indicate interchangeability or local pockets of microbiome variation for some locations.

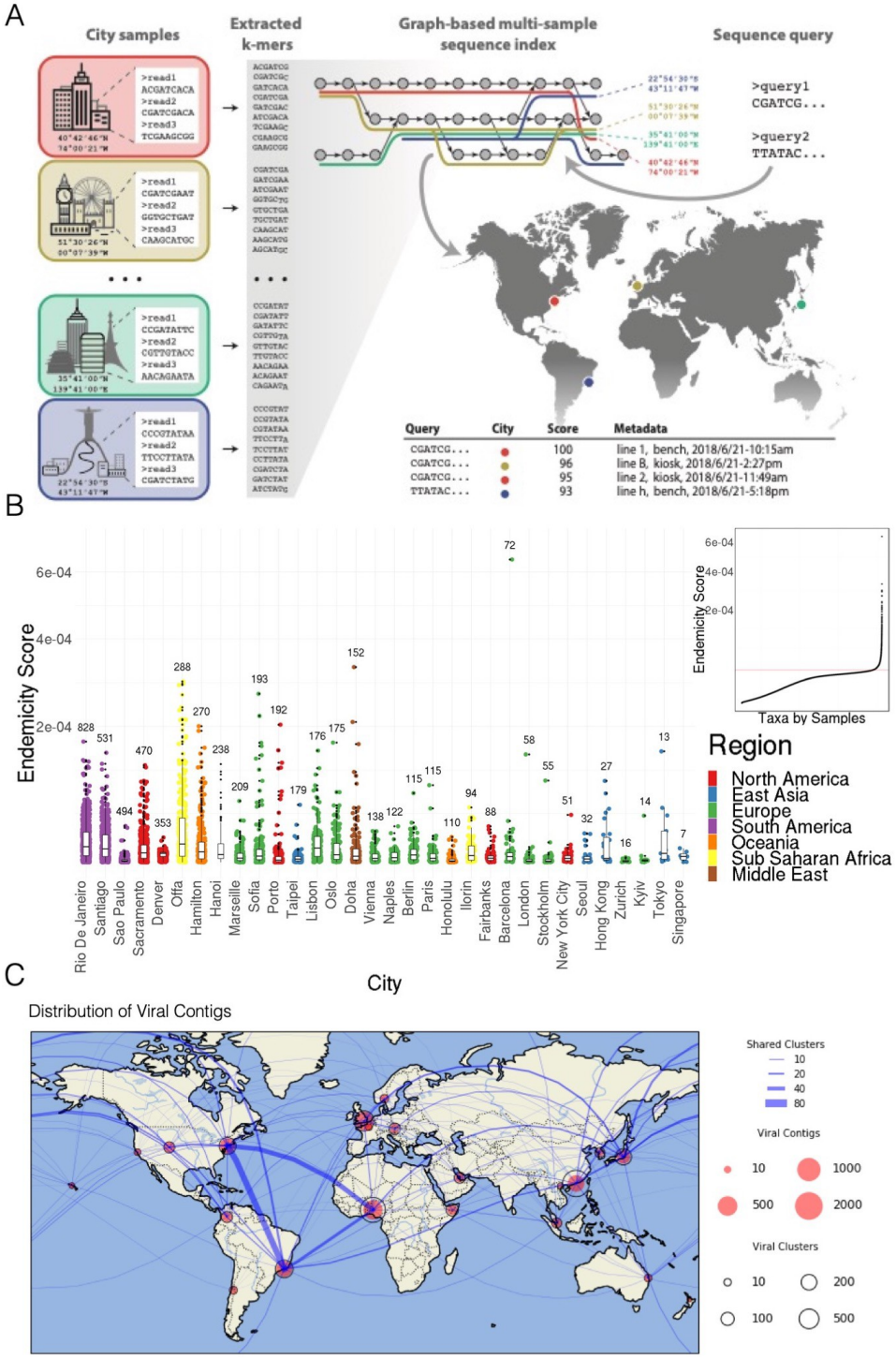

Figure 3: Quantifying known and novel urban diversity. *A) Schematic of GeoDNA representation generation – Raw sequences of individual samples for all cities are transformed into lists of unique k-mers (left). After filtration, the k-mers are assembled into a graph index database. Each k-mer is then associated with its respective city label and other informative metadata, such as geo-location and sampling*

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

information (top middle). Arbitrary input sequences (top right) can then be efficiently queried against the index, returning a ranked list of matching paths in the graph together with metadata and a score indicating the percentage of k-mer identity (bottom right). The geo- information of each sample is used to highlight the locations of samples that contain sequences identical or close to the queried sequence (middle right). B) Distributions of Endemicity scores within cities C) The planetary distribution of MetaSUB viral clusters. Solid red circles indicate the number of viral contigs recovered in each region (Section2.5). Black open circles indicate the number of viral species recovered in each region. Blue lines indicate the number of viral clusters that are shared between regions, thicker lines indicating more viral clusters in common.

We then investigated our *de novo* assembled contigs from MetaSPAdes relative to the Joint Genome Institute (JGI) known virus and retroviruses (VR) datasets and annotation pipeline. Our samples identified 16,584 predicted uncultivated viral genomes (UViGs), including 2,009 clusters containing a total of 6,979 UViGs and 9,605 singleton UViGs, for a total of 11,614 predicted viral species. We compared the predicted species to known viral sequences in the IMG/VR system, which contains viral genomes from all NCBI isolates, a curated set of prophages, and 730k UViGs from other studies. Of the 11,614 species discovered in our data, 94.1% did not match any viral sequence in IMG/VR (Paez-Espino et al., 2019) at the species level, indicating that urban microbiomes contain significant diversity not observed in other environments. Next, we attempted to identify possible hosts for our predicted UViGs. For the 686 species with similar sequences in IMG/VR, we projected known host information onto 2,064 MetaSUB UViGs. Additionally, we used CRISPR-Cas spacer matches in the IMG/M system to assign possible hosts to a further 1,915 predicted viral species. Finally, we used a JGI database of 20 million metagenome derived CRISPR spacers to provide further rough taxonomic assignments. We note that virus-host associations were found for 41% of species in the core microbiome (**Figure 1A, Supplementary Data Packet**).

Finally, these predicted viral species from samples collected within 10, 100 and 1000 kilometers of one another were agglomerated to examine their planetary distribution at different scales (**Figure 3C**). At any scale, most viral clusters appear to be weakly cosmopolitan; the majority of their members were found at or near one location, with a few exceptions. These cosmopolitan viral clusters also exhibit an observational effect; regions in which more samples were recovered share more viral clusters (notably, the northeast United States, Lagos and the southern Brazil).

Antimicrobial Resistance (AMR) Markers Form Distinct Clusters

Quantification of antimicrobial diversity and AMRs are key components of global antibiotic stewardship. Yet, predicting antibiotic resistance from genetic sequences alone is challenging, and detection accuracy depends on the class of antibiotics (i.e., some AMR genes are associated to main metabolic pathways while others are uniquely used to metabolize antibiotics). As a first step towards a global survey of antibiotic resistance in urban environments, we mapped reads to known antibiotic resistance genes, using the MegaRES ontology and alignment software (**Figure 4**). We first performed a rarefaction analysis on the set of all resistance genes in the dataset, across the "panresistome" (**Figure 4A**). Similar to the rate of detected species, the panresistome also shows an open slope with an expected rate of discovery of one new AMR gene per 10 samples. Given that AMR gene databases are rapidly expanding and that single AMR markers were not commonly found in many samples, it is likely that future analyses will identify many more resistance genes in this data. Additionally, AMR genes show a "neighborhood" effect within samples that are geographically proximal, analogous to the effect seen for taxonomic composition (**Figure 4B**). Excluding samples where no AMR genes were detected, the Jaccard distance between sets of AMR genes increases with distance for pairs of samples in the same city; as with taxonomic composition, the overall effect is weak and noisy but significant.

We then quantified the mapped AMR relative abundances using reads/kilobase/million mapped reads (RPKM) for 20 classes of antibiotic resistance genes detected in our samples (**Figure 4C, D**). 2,210 samples had some sequence which was identified as belonging to an AMR gene, but no consistent core set of genes was identified. The most common classes of antibiotic resistance genes were for macrolides, lincosamides, streptogamines (MLS), and betalactams. Despite being relatively common, antibiotic resistance genes were universally in low abundance compared to functional genes, with measured values for resistance classes typically ranging from 0.1 – 1 RPKM compared to values of 10 - 100 RPKM for typical housekeeping genes. In spite of the low abundance of the marker genes themselves, some samples contained sequences from hundreds of distinct AMR genes, where others were only a few. Moreover, clusters of high AMR diversity were not evenly distributed across the world (**Figure 4E**), with some cities showing far more resistance genes on average (15-20X) than others (e.g. Offa), while other cities had bimodal distributions (e.g. Seoul), indicating significant regional variation.

We next identified AMR genes that were found on assembled contigs. Contigs were identified as either representing bacterial plasmids or chromosomes and were given taxonomic assignments, and then annotated

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with the number of AMR genes identified. Prominent AMR genes identified on plasmid contigs formed co-occurring clusters (**Figure 5A**). Specifically, we identified thirteen clusters of at least two AMR genes that consistently occurred on the same contigs and were found on at least 1,000 contigs. All of these genes were found to have cosmopolitan ranges, but were not necessarily found in equal abundances across regions. A<R genes in clusters were of many different types but there were notable clusters for vancomycin resistance, tetracycline resistance, and Multi-drug efflux complexes. Some clusters were dominated by a single type of resistance gene while others had genes with multiple sources.

Contigs with at least one AMR gene from both chromosomes and plasmids had a mode of three AMR genes. Chromosomal contigs had more AMR genes than plasmid contigs on average but this does not account for the larger size of chromosomal contigs. Three taxa groups, Acinetobacteria, Bacteroidetes, and Tenericutes, had AMR genes predominantly on chromosol contigs. One taxa group, Deinococcus-Thermus, had more AMR genes identified on plasmid contigs than chromosomes. AMR genes from three taxa groups, Cyanobacteria, Deinococcus-Thermus, and Tenericutes, had geographically constrained ranges while AMRs from other taxa had cosmopolitan ranges.

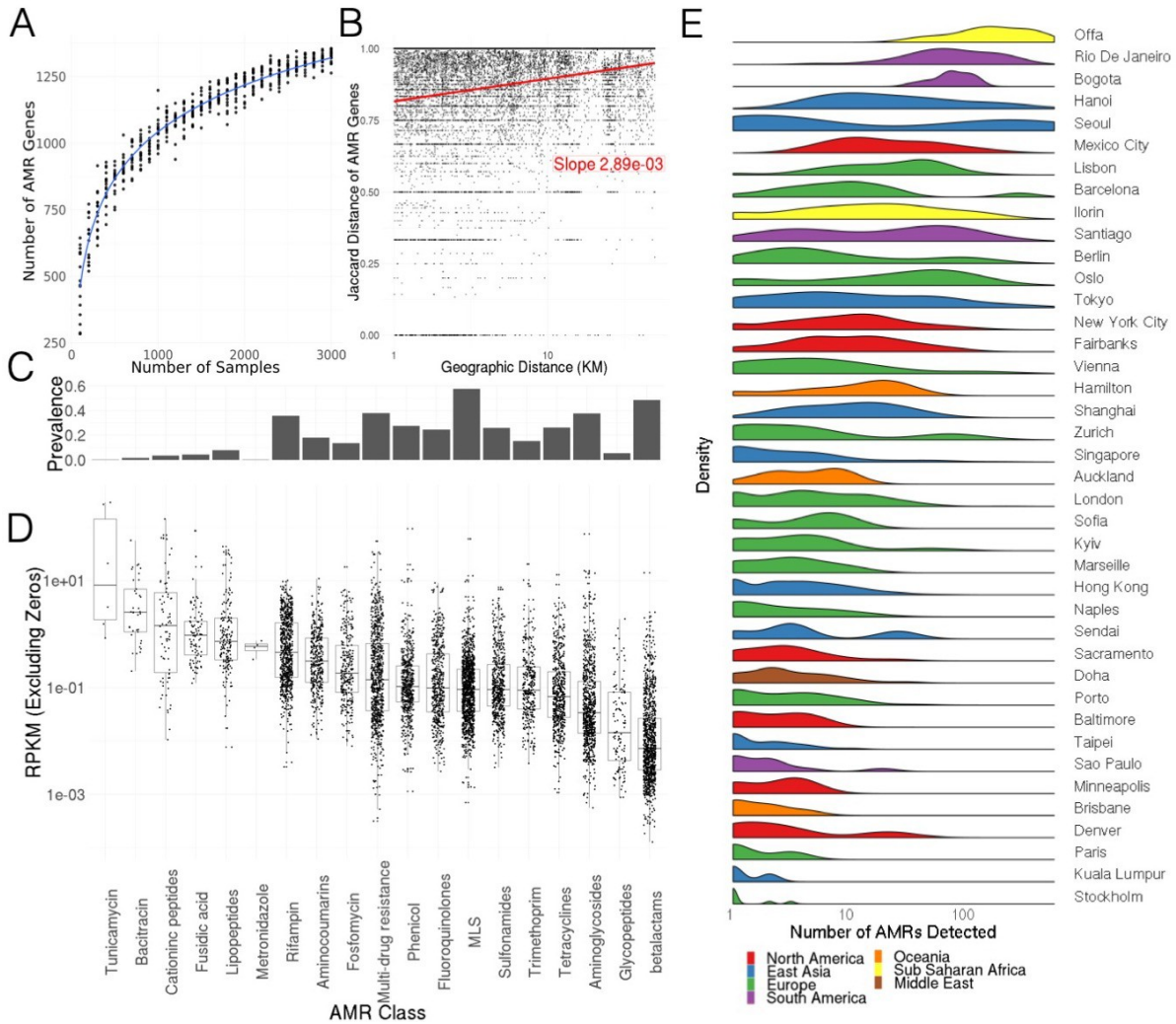

Figure 4: AMR Orthologs are common but varied in distribution. A) Rarefaction analysis of AMR genes over samples B) Co-occurrence of AMR genes decreases with distance, nearby samples are more likely to be similar C) Prevalence of AMR genes with resistance to particular drug classes D) Abundance of AMR gene classes when detected, by drug class E) Number of detected AMR genes by city.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

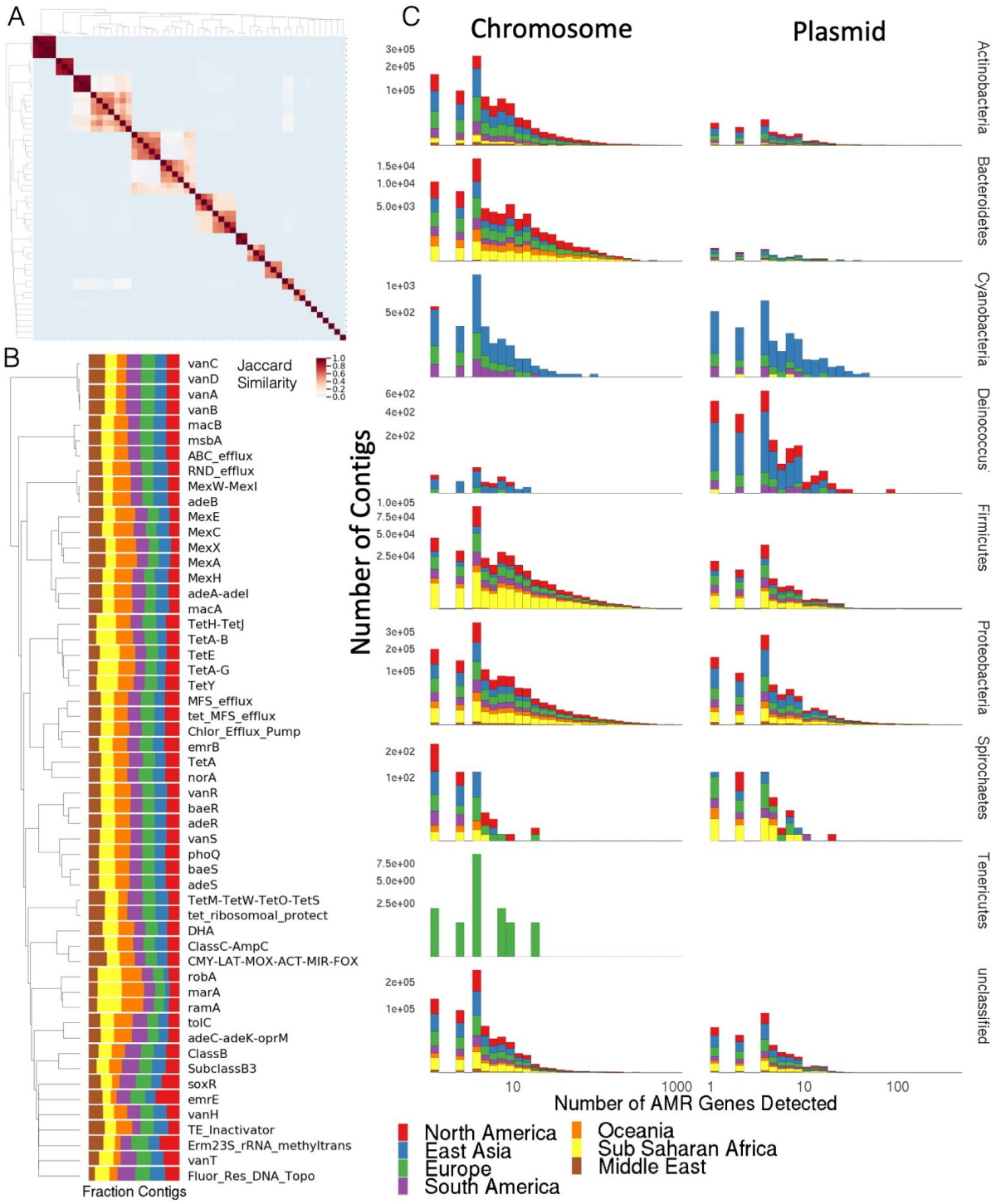

Figure 5: AMRs Form Distinct Clusters A) Co-occurrence of prominent AMR genes found on plasmids. Heatmap of the Jaccard index for pairs of genes according to the contigs where they are identified. Only genes which were identified on 1,000 contigs are more are shown. B) Geographic distribution of AMR genes in (A).
c) The number of AMR genes identified on assembled contigs based on the type of contig and taxonomic classification.

Discussion

Here, we have generated a comprehensive global, urban metagenomic atlas and identified a core microbiome present across 58 cities, consisting of a wide range of mostly bacterial and fungal species. We also identified species that are geographically constrained, and also showed that these can be used to infer the likely city of origin. Many of these species are associated with commensal microbiomes from human skin and airways, but we observed that urban microbiomes are nevertheless distinct from both human and soil microbiomes. Notably,

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

no species from the Bacteroidetes, a prominent group of human commensal organisms (Eckburg *et al*., 2005; Qin *et al*., 2010), was identified in the core urban microbiome. Notably, no archaea or viruses were identified in the group of sub-core microorganisms, although this analysis did not include viruses newly discovered in this study. For viruses in particular, this may be affected by the sampling or DNA extraction methods used, by limitations in sequencing depth, or by missing annotations in the reference databases used for taxonomic classification, which is principally problematic with phages. It is worth noting that potentially prevalent RNA viruses are omitted with our DNA-based sampling.

Unique taxonomic composition and association with covariates specific to the urban environment suggest that urban microbiomes should be treated as ecologically distinct from both surrounding soil microbiomes and human commensal microbiomes. Though these microbiomes undoubtedly interact with the urban environment, they nonetheless represent distinct ecological niches with different genetic profiles. While our metadata covariates were associated with the principal variation in our samples, they do not explain a large proportion of the observed variance. It remains to be determined whether variation is essentially a stochastic process or if a deeper analysis of our covariates proves more fruitful. We have observed that less important principal components (roughly PCs 10-100) are generally less associated with metadata covariates but that PCs 1-3 do not adequately describe the data alone. This is a pattern that was observed in the human microbiome project as well, where minor PCs (such as our Figure 2B) were required to separate samples from closely related body sites.

Much of the urban microbiome likely represents novel diversity, as our samples contain a significant proportion of unclassified DNA. More sensitive methodologies only marginally increase the proportion of DNA that can be classified (see methods), and the sequencing error rate is <1%. As such, we consider the un-mapped DNA likely to be components of truly novel organisms and viruses, as evidenced by comparison to the JGI/VR databases. Indeed, the fraction of predicted viral sequences which belonged to previously unobserved taxa was particularly high in our study (94.1%). However, taxonomic associations of these viruses to observed microbial hosts suggests these results are not spurious. This rate of discovery may prove prescient for novel taxa in other domains, and novel discovery of taxa may help to reduce the large fraction of DNA which cannot currently be classified.

Many of the identified taxa are frequently implicated as infectious agents in a clinical setting including specific Staphylococcus, Streptococcus, Corynebacterium, Klebsiella and Enterobacter species. However, there is no clear indication that the species identified in the urban environment are pathogenic and further in-depth study is necessary to determine the clinical impact of urban microbiomes. This includes microbial culture studies, specifically searching for virulence factors and performing strain-level characterization. Seasonal variation also remains open to study as the majority of the samples collected here were from two summer global City Sampling Days (June 21, 2016 and 2017). Further studies, some generating novel data, will need to explore whether the core microbiome shifts over the course of the year, with particular interest in the role of the microbiome in flu transmission (Cáliz *et al*., 2018; Korownyk *et al*., 2018).

As metagenomics and next-generation sequencing becomes more available for clinical (Wilson *et al*., 2019) and municipal use (Hendriksen et al., 2019), it is essential to contextualize the AMR markers or presence of new species and strains within a global and longitudinal context. The most common AMR genes were found for two classes of antibiotics: MLS (macrolide-lincosamide-streptogramin) and beta-lactams. MLS represents three groups of antibiotics with a mechanism of action of inhibiting bacterial protein synthesis. Macrolides, with strong Gram-positive and limited Gram-negative coverage, are prevalently used to treat upper respiratory, skin, soft tissue and sexually transmitted infections, amongst others. Beta-lactam antibiotics are a major class of antibiotics including penicillins, cephalosporins, monobactams, carbapenems and carbacephems, all of which are widely used to treat infections. Antimicrobial resistance has surged due to the selection pressure of widespread use of antibiotics and is now a global health issue plaguing communities and hospitals worldwide. Antimicrobial resistance genes are thought to spread from a variety of sources including hospitals, agriculture and water (Bougnom and Piddock, 2017; Klein *et al*., 2018). The antimicrobial classes particularly impacted by resistance include beta-lactamases, gylcopeptides and fluoroquinolones, all of which we found antimicrobial resistance genes for across our samples, notably with uneven distribution of AMR genes across cities. This could be the result of some of combination of different levels of antibiotic use, differences in the urban geography between cities, population density, presence of untreated wastewater, or reflect the background microbiomes in different places in the world. Techniques to estimate antibiotic resistance from sequencing data remain an area of intense research, as certain classes of AMR gene (ie. fluoroquinolones) are sensitive to small mutations and it is possible that our methods represent a broader metric of resistance. Further research is needed to fully explore AMR genes in the urban environment, including culture studies, which directly measure the phenotype of resistance.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In summary, this study presents a first molecular atlas of urban and mass-transit metagenomics from across the world. By facilitating large scale epidemiological comparisons, it represents a first critical step towards quantifying the clinical role of environmental microbiomes and provides requisite data for tracking changes in ecology or virulence. Previous studies have already demonstrated a role for precision clinical metagenomics in managing infectious disease and global health (Afshinnekoo *et al*., 2017; Gardy and Loman, 2018; Ladner *et al*., 2019). As an AMR and genetic metagenomic atlas, this data has the potential to aid physicians, public health officers, government officials, and others in diagnosis, clinical decision making and policy within their communities.

Data Availability

Reads with the human DNA filtered and low quality bases removed are available for download from Wasabi (an Amazon S3 clone) with individual URLs located here: https: //github.com/MetaSUB/metasub_utils. In addition to raw reads higher level results (e.g. taxonomic profiles, functional pathways, etc.) are available in the MetaSUB data packet also available for download from Wasabi. For instructional purposes, we also provide a simplified data packet for teaching which includes balanced numbers of samples from each city and completely filled metadata tables.

Jupyter notebooks used to generate all the figures and statistics in this study can be found at https://www.github.com/MetaSUB/main_ paper_figures. Interactive data visualizations are available on https://www.metagenscope.com and GeoDNA, an interface to search query DNA sequences against MetaSUB samples, is available at (dnaloc.ethz.ch/). MetaSUB metadata is available in the data-packet or may be downloaded from https://github.com/ MetaSUB/MetaSUB-metadata. Programs used for analysis of data may be found at https://github.com/ MetaSUB/MetaSUB_CAP and https://github.com/dcdanko/capalyzer. Additional tools and resources are described here https://github.com/MetaSUB/ bioinformatics_management.

The MetaSUB dataset is built and organized for full accessibility to other researchers. This is consistent with the concept of Open Science. Specifically, we built our study with the FAIR principles in mind: Findable, Accessible, Interoperable and Reusable. To make our study reproducible, we released an open source version-controlled pipeline, called the MetaSUB Core Analysis Pipeline (CAP). The CAP is intended to improve the reproducibility of our findings by making it easy to apply a number of analyses consistently to a large dataset. This pipeline includes all steps from extracting data from raw sequence data to producing refined results like taxonomic and functional profiles. The CAP itself is principally composed of other open peer-reviewed scientific tools, with some bridging and parsing scripts. Every tool in the CAP is open-source with a permissive license. The CAP is available as a docker container for easier installation in some instances and all databases used in the CAP are available for public download. The CAP is versioned and includes all necessary databases allowing researchers to replicate results.

To make our results more reproducible and accessible, we have developed a program to condense the outputs of the Core Analysis Pipeline into a condensed data-packet. This data packet contains results as a series of Tidy-style data tables with descriptions. The advantage of this setup is that result tables for an entire dataset can be parsed with a single command in most high-level analysis languages like Python and R. This package also contains Python utilities for parsing and analyzing data packets which streamlines most of the boilerplate tasks of data analysis. All development of the CAP and data packet builder (Capalyzer) package is open-source and permissively licensed.

Acknowledgements

We thank GitHub for providing private repositories to the MetaSUB consortium at no cost. We thank XSEDE and Philip Blood for their support of this project, which enabled all the *de novo* assemblies.
We would like to thank the Epigenomics and Genomics Core Facilities at Weill Cornell Medicine, funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, Igor Tulchinsky, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), the National Institutes of Health (R01ES021006, R25EB020393, 1R21AI129851, 1R01MH117406), TRISH (NNX16AO69A:0107, NNX16AO69A:0061), the NSF (1840275), the Bill and Melinda Gates Foundation (OPP1151054) and the Alfred P. Sloan Foundation (G-2015-13964), Swiss National Science Foundation grant #407540_167331 “Scalable Genome Graph Data Structures for Metagenomics and Genome Annotation” as part of Swiss National Research Programme (NRP) 75 “Big Data.”
Discovery of novel viral sequences was work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract number DE-AC02-05CH11231 and used

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

resources of the National Energy Research Scientific Computing Center, supported by the Office of Science of the US Department of Energy.
MetaSUB Sweden was supported by Stockholm Health Authority (Region Stockholm) grant SLL 20160933 awarded to KIU.
MetaSUB Seoul was supported by the Institut Pasteur Korea (2015MetaSUB) and a National Re- search Foundation of Korea (NRF) grant (NRF-2014K1A4A7A01074645, 2017M3A9G6068246).
Metasub Chile was supported by funding from CONICYT Fondecyt Iniciación grant 11140666 and 11160905, as well as funding from the Millennium Science Initiative of the Ministry of Economy, Devel- opment and Tourism, Government of Chile.
MetaSUB Japan was supported by research funds from Keio University, the Yamagata prefectural government and the City of Tsuruoka.
MetaSUB Austria and Ukraine acknowledge the bilateral AT-UA collaboration fund (WTZ:UA 02/2019; Ministry of Education and Science of Ukraine, UA:M/84-2019).
MetaSUB Ukraine was supported by research funds from Kyiv Academic Univeristy, Ministry of Education and Science of Ukraine grant 0118U100290. MetaSUB Ukraine would like to express gratitude to Kyiv Metro for the support of sampling days.
MetaSUB Barcelona was supported by the Spanish Ministry of Economy and Competitiveness, 'Cen- tro de Excelencia Severo Ochoa 2013-2017, the CERCA Programme / Generalitat de Catalunya, the "la Caixa" Foundation, the CRG-Novartis-Africa mobility programme 2016 and TMB Director Eladio De Miguel Sainz
Work in Colombia was partially funded by Colciencias (project No. 639677758300)
Sampling was carried out in compliance with regulations and permissions from local authorities (Azienda Napoletana Mobilità s.p.a. in Naples, Italy; Régie des Transports Métropolitains in Marseille, France; Transmilenio and ANLA permit 1484 in Bogotá, Colombia; Nigerian Railway Corporation (NRC) [Ilorin and Offa Branch] and Kwara Express Transport).
Since 2017 MetaSUB has partnered with the Critical Assessment of Massive Data Analysis program (CAMDA) camda.info, a full conference track at the Intelligent Systems for Molecular Biology (ISMB) Conference, for two city-prediction challenges, which addressed the problem of geographically locating samples with Open Science tools. This Open Science approach has generated multiple interesting results and concepts relating to urban microbiomics, resulting in several related publications biologydirect.biomedcentral.com/articles/collections/camdaproc and we thank them for their support as well.
Finally, we thank Mark Van Oene and other members of Illumina, Inc., who helped make the study possible.

We thank the many sampling volunteers who were trained and who made this study possible, including Sara Abdul Majid, Natasha Ab- dullah, Ait-hamlat Adel, Nayra Aguilar Rojas, Affifah Saadah Ahmad Kassim, Faisal S Al-Quaddoomi, Gabriela E Albuquerque, Alex Alexiev, Maria G Amorim, Muhammad Al-Fath Amran, Watson Andrew, Harilanto Andrianjakarivony, Álvaro Aranguren, Carme Arnan, Freddy Asenjo, Juliette Auvinet, Nuria Aventin, Thais F Bartelli, Erdenetsetseg Batdelger, François Baudon, Carla Bello, Médine Benchouaia, Hannah Benisty, Anne-Sophie Benoiston, Diego Benítez, Juliana Bernardes, Tristan Bitard-Feildel, Lucie Bittner, Guillaume Blanc, Julia Boeri, Kevin Bolzli, Alexia Bordigoni, Ciro Borrelli, Sonia Bouchard, Jean-Pierre Bouly, Alessandra Breschi, Alan Briones, Aszia Burrell, Alina Butova, Dayana Calderon, Angela Cantillo, Miguel Carbajo, Katerine Carrillo, Laurie Casalot, Sofia Castro, Jasna Chalangal, Starr Chatziefthimiou, Francisco Chavez, Allaeddine Chettouh, Erika Cifuentes, Sylvie Collin, Romain Conte, Flavia Corsi, Cecilia N Cossio, Ana F Costa, Bruno D'Alessandro, Ophélie Da Silva, Katherine E Dahlhausen, Natalie R Davidson, Eleonora De Lazzari, Stéphane Delmas, Chloé Dequeker, Alexandre Desert, Valeriia Dotsenko, Cassie L Ettinger, Emile Faure, Fazlina Fauzi, Aubin Fleiss, Aubin Fleiss, Juan Carlos Forero, Mathilde Garcia, Catalina García, Sonia L Ghose, Liliana Godoy, Andrea Gon- zalez, Camila Gonzalez-Poblete, Charlotte Greselle, Sophie Guasco, Nika Gurianova, Sebastien Halary, Eric Helfrich, Aliaksei Holik, Chiaki Homma, Michael Huber, Stephanie Hyland, Andrea Hässig, Roland Häusler, Nathalie Hüsser, Badamnyambuu Iderzorig, Mizuki Igarashi, Shino Ishikawa, Sakura Ishizuka, Kohei Ito, Sota Ito, Tomoki Iwashiro, Marisano James, Marianne Jaubert, Marie-Laure Jerier, Guill- laume Jospin, Nao Kato, Inderjit Kaur, Akash Keluth Chavan, Mahshid Khavari, Maryna Korshevniuk, Jonas Krebs, Andrii Kuklin, Antonietta La Storia, Juliana Lago, Elodie Laine, Olha Lakhneko, Gerardo de Lamotte, Romain Lannes, First Name Middle Initial Last Name, Madeline Leahy, Vincent Lemaire, Dagmara Lewandowska, Manon Loubens, Olexandr Lykhenko, Salah Mahmoud, Natalka Makogon, Dim- itri Manoir, German Marchandon, Natalia Marciniak, Vincent Matthys, Arif Asyraf Md Supie, Irène Mauricette Mendy, Roy Meoded, Mathilde Mignotte, Ryusei Miura, Kunihiko Miyake, Maria D Moc- cia, Mauricio Moldes, Jennifer Molinet, Orgil-Erdene Molomjamts, Mario Moreno, Maureen Muscat, Cristina Muñoz, Francesca Nadalin, Dorottya Nagy-Szakal, Ashanti Narce, Hiba Naveed, Thomas Neff, Wan Chiew Ng, Elsy Ngwa, Agier Nicolas, Pierre Nicolas, Abdollahi Nika, Diana N Nunes, Javier Quilez Oliete, Nils Ordioni, Mitsuki Ota, Francesco Oteri, Yuya Oto, Coral Pardo-Este, Young-Ja Park, Jananan Pathmanathan, Manuel Perez, Melissa P Pizzi, María Gabriela Portilla, Leonardo Posada, Catherine E. Pugh,

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kyrylo Pyrshev, Sreya Ray Chaudhuri, Hubert Rehrauer, Renee Richer, Paula Rodríguez, Paul Roldán, Sandra Roth, Maria Ruiz, Mariia Rybak, Ikuto Saito, Yoshitaka Saito, Khaliun Sanchir, Kai Sasaki, Kaisei Sato, Masaki Sato, Ryo Sato, Seisuke Sato, Yuma Sato, Oli Schacher, Christian Schori, Felipe Sepulveda, Marianna S Serpa, Juan C Severyn, Sarah Shalaby, Hikaru Shirahata, Jordana M Silva, Gwenola Simon, Kasia Sluzek, Rebecca Smith, Yuya Sonohara, Nicolas Sprinsky, Stefan G Stark, Chisato Suzuki, Sora Takagi, Kou Takahashi, Naoya Takahashi, Tomoki Takeda, Soma Tanaka, Emilio Tarcitano, Andrea Tassinari, Eunice Thambiraja, Antonin Thiébaut, Antonin Thiébaut, Takumi Togashi, Yuto To- gashi, Anna Tomaselli, Itsuki Tomita, Nora C Toussaint, Takafumi Tsurumaki, Yelyzaveta Tymoshenko, Mariko Usui, Sophie Vacant, Laura E Vann, Jhovana L Velasco Flores, Fabienne Velter, Riccardo Vice- domini, Tomoro Warashina, Ayuki Watanabe, Tina Wunderlin, Olena Yemets, Tetiana Yeskova, Shusei Yoshikawa, Stas Zubenko.

Materials and Methods

Metadata Collection and Cleaning

Metadata from individual cities was collected from a standardized form and set of fields using the Kobo Toolbox MetaSUB App and website. The principle fields collected were the location of sampling, the material being sampled, the type of object being sampled, the elevation above or below ground, and the station or line where the sample was collected. However, several cities were unable to use the provided apps for various reasons and submitted their metadata as separate spreadsheets, which were uploaded later. Additionally, certain metadata features, such as those related to sequencing and quality control, were added after initial sample collection.

To collate various metadata sources, we built a publicly available program which assembled a large master spreadsheet with consistent sample unique ID (UUIDs). After assembling the originally collected data attributes we added normalized attributes based on the original metadata to account for surface material, control status, and features of individual cities. A full description of ontologies used is provided as part of the collating program.

Sample Collection and Preparation

In this study we have benchmarked various types of swabs and DNA preservative tubes, including Copan Liquid Amies Elution Swab (ESwab,Copan Diagnostics, Cat.:480C, 2016) referred to as 'copan swab' and Isohelix Swabs (Mini-Swab, Isohelix Cat.:MS-02, 2017) referred to as 'isohelix swabs', which were combined with 2D Thermo Scientific^{TM} Matrix^{TM} storage tubes (3741-WP1D-BR/Matrix 1.0 ml/EA) referred to as 'matrix tube', which have been prefilled with the preservative liquid Zymo DNA/RNA Shield^{TM} (R1100-250) referred to as 'Zymo shield.' Copan swabs do not contain a preservative and need to be kept on ice after sampling until DNA extraction to prevent DNA degradation. Samples collected with Copan swabs have therefore been transported on dry ice and stored at -80C. Isohelix swabs have been stored in matrix tubes containing 400μl Zymo shield preservative. Matrix tubes have been transported at RT and stored at -80C until DNA extraction. All the Isohelix swabs were moistened by submerging the swab for a few seconds in preservative media, and all surfaces in the subway were swabbed for 3 min during the two global city sampling days in 2016 and 2017.

In-Lab controls

As positive lab controls, we used 30μl of the ZymoBiOMICS Microbial Community standard (Catalog #D6300), which we added to an empty sterile cup, followed by swabbing with Copan Liquid Amies Elution Swab (ESwab, Copan Diagnostics, Cat.:480C) for 1.5min / 3 minutes. As negative (background) lab control we used 50μl of the final resuspension buffer (MoBio PowerSoil R DNA Isolation Kit, Cat.:12888- 100), which we have added to an empty sterile cup followed by swabbing for 3 min (Fig.S1)! Furthermore, the working space has been swabbed for 1.5 min / 3 min before and after treatment with 10% bleach (**Figure S2**) to test for background contamination rates. To identify the background levels of biological material in the air at sample areas, a Copan swab has been held for 1.5 min - 3 min in the air. To estimate the source and amount of contamination in commercial swab and tube products used for MetaSUB, we tested all consumables in triplicates in the sterilized hood (UV light and 10% bleach wiped with ethanol, **Figure S2**).

DNA Extraction from Isohelix swabs using ZymoBiomics 96 MagBead

The Isohelix swab head and the entire 400 μl of DNA/RNA Shield-solubilized sample were transferred into ZR BashingBead Lysis Tubes (0.1 & 0.5 mm) (Cat# S6012-50) to which an additional 600 μl of DNA/RNA Shield was added. Mechanical lysis using bead beating was performed on a maximum of 18 samples simultaneously using the Scientific Industries Vortex-Genie 2 with Horizontal-(24) Microtube Adapter (Cat # SI-0236 and SI-H524) at maximum power for 40 minutes. The resulting lysate (400 μl) was transferred to NuncTM 96-Well

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Polypropylene DeepWell Storage Plates (Cat # 278743), followed by DNA extraction using the ZymoBIOMICS 96 MagBead Kit (Lysis Tubes) (Catalog # D4308) on the Hamilton Star according to manufacturer instructions.

DNA extraction from Copan swabs using MoBio PowerSoil DNA

Droplets in the Copan Liquid Amies Elution Swab tube (ESwab, Copan Diagnostics, Cat.:480C (http://goo.gl/8a9uCP)) were spun down at 300rpm/1min. Next, the swab pad was transferred to a Mo-Bio PowerSoil R DNA vial containing beads using sterile scissors, which we sterilized by flaming with 100% ethanol. The remaining 400-500µl Copan Amines liquid has been transferred into an Eppendorf tube and centrifuged at full speed to collect bacteria and debris in a pellet. The pellet was finally transferred to the same MoBio PowerSoil R DNA vial also containing the corresponding swab pad. Mo- Bio PowerSoil R DNA Isolation Kit, Cat.:12888-100 (https://goo.gl/65rcn2) was used according to manufacturer's instructions except for the following modifications: Both swab and pellet have been re-suspended with 135µl C1 buffer (MoBio PowerSoil R DNA). Sample homogenization was performed using either TissueLyser II (Qiagen) with 2 cycles of 3 minutes at 30Hz (https://goo.gl/hBg8Lb), or using the Vortex-Genie 2 (Vortex Catalog #13000-V1-24) adaptor and vortex at maximum speed for 10 minutes. The eluted samples have been additionally purified and concentrated by Beckmann Coulter Agencourt AMPure XP (Cat.:A63881) purification (1.8X) and eluted into 12µl - 50µl elution buffer. Subsequently, DNA was quantified using Qubit R Assay (Catalog #Q32854).

Quality Control

Sequencing quality

We measured sequencing quality based on 5 metrics: number of reads obtained from a sample, GC content, Shannon's entropy of *k*-mers, post PCR Qubit score, and recorded DNA concentration before PCR. The number of reads in each sample was counted both before and after quality control, we used the number of reads after quality control for our results though the difference was slight. GC content was estimated from 100,000 reads in each sample after low quality DNA and human reads had been removed. Shannon's entropy of *k*-mers was estimated from 10,000 reads taken from each sample.

Strain Contamination

We used BLASTn to align nucleotide assemblies from case samples to control samples. We used a threshold of 8,000 base pairs and 99.99% identity as a minimum to consider two sequences homologous. This threshold was chosen to be sensitive without solely capturing conserved regions. We identified all connected groups of homologous sequences and found approximate taxonomic identifications by aligning contigs to NCBI-NT using BLASTn searching for 90% nucleotide identity over half the length of the longest contig in each group.

K-Mer Based Analyses

We generated 31-mer profiles for raw reads using Jellyfish. All k-mers that occurred at least twice in a given sample were retained. We also generated MASH sketches from the non-human reads of each sample with 10 million unique minimizers per sketch. We calculated the Shannon's entropy of k-mers by sampling 31-mers from a uniform 10,000 reads per sample. Shannon's entropy of taxonomic profiles was calculated using the CAPalyzer package.

Sequence Preprocessing

Sequence data were processed with AdapterRemoval (v2.17, Schubert *et al*., 2016) to remove low quality reads and reads with ambiguous bases. Subsequently reads were aligned to the human genome (hg38, including alternate contigs) using Bowtie2 (v2.3.0, fast preset, Langmead and Steven L Salzberg, 2013). Read pairs where both ends mapped to the human genome were separated from read pairs where neither mate mapped. Read pairs where only one mate mapped were discarded. Hereafter, we refer to the read sets as human reads and non-human reads.

Computational Analysis

Taxonomic Analysis

We generated taxonomic profiles by processing non-human reads with KrakenUniq (v0.3.2 Breitwieser *et al*. (2018)) using a database based on all draft and reference genomes in RefSeq Microbial (bacteria, fungi, virus, and archaea) ca. March 2017. KrakenUniq was selected because it was highly performant, as it has been demonstrated to be comparable or having higher sensitivity than the best tools identified in a recent benchmarking study (McIntyre *et al*., 2017) on the same comparative dataset. KrakenUniq reports the number of unique marker k-mers assigned to each taxon, as well as the total number of reads, the fraction of available marker k-mers found, and the mean copy number of those k-mers. We found that requiring more k-mers to identify a species resulted in a roughly linear decrease in the total number of species identified without a plateau or any other clear point to set a threshold (**Figure S2A**).

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

At a minimum we required three reads assigned to a taxa with 64 unique marker k-mers. This setting captures a group of taxa with low abundance but reasonable (10-20%) coverage of the k-mers in their marker set (**Figure S2C**). However, this also allows for a number of taxa with very high (105) duplication of the identified marker k-mers and very few k-mers per read, which is unlikely (**Figure S2D**). Thus, we further filtered these taxa by applying a filter where the number of reads not exceed 10 times the number of unique k-mers, unless the set of unique k-mers was saturated (> 90% completeness). We include a full list of all taxonomic calls from all samples including diagnostic values for each call. We do not attempt to classify reads below the species level in this study.

We further evaluated prominent taxonomic classifications for sequence complexity and genome coverage. For each microbe evaluated, we calculated two indices: the average topological entropy of reads assigned to the microbe and the Gini-coefficient of read positions on the microbial genome. For brevity we refer to these as mean sequence entropy (MSE) and coverage equality (CE). The formula for topological entropy of a DNA sequence is described by Koslicki (2011). Values close to 0 correspond to low-complexity sequences and values near 1 are high complexity. In this work we use a word size of 3 with an overall sequence length of 64 since this readily fits into our reads. To find the MSE of a microbial classification we take the arithmetic mean of the topological entropy of all reads that map to a given microbial genome in a sample. Reads mapping to a microbial genome are assigned to a contiguous 10kbp bin and the Gini-coefficient of all bins is calculated. Like MSE, the Gini-coefficient is bounded in [0, 1]. Lower values indicate greater inequality, very low values indicate that a taxon may be misidentified from conserved and near conserved regions. We downloaded one representative genome per species evaluated and mapped all reads from samples to using Bowtie2 (sensitive-local preset). Indices were processed from alignments using a custom script. Species classifications with an average MSE less than 0.75 or CE less than 0.1 were flagged as low quality.

To determine relative abundance of taxa, we rarefied samples to 100,000 classified reads, computed the proportion of reads assigned to each taxon, and took the distribution of values from all samples. This was the minimum number of reads sufficient to maintain taxonomic richness (**Figure S2B**). We chose sub-sampling (rarefaction) based on the study by Weiss *et al*. (2017), showing that sub-sampling effectively estimates relative abundance. Note that we use the term prevalence to describe the fraction of samples where a given taxon is found at any abundance and we use the term relative abundance to describe the fraction of DNA in a sample from a given taxon.

MASH Analysis

We compared our samples to metagenomic samples from the Human Microbiome Project and a metagenomic study of European soil samples using MASH (Ondov et al., 2016), a fast k-mer based comparison tool. We built MASH sketches from all samples with 10 million unique k-mers to ensure a sensitive and accurate comparison. We used MASH's built-in Jaccard distance function to generate distances between our samples and HMP samples. We then took the distribution of distances to each particular human commensal community as a proxy for the similarity of our samples to a given human body site.

We also compared our samples to HMP and soil samples using taxonomic profiles generated by MetaPhlAn v2.0 (Segata *et al*., 2012). We generated taxonomic profiles from non-human reads using MetaPhlAn v2.0 and found the cosine similarity between all pairs of samples. We used the Microbe Directory (Shaaban *et al*., 2018) to annotate taxonomic calls. The Microbe Directory is a hand curated, machine readable, database of functional annotations for 5,000 microbial species.

Functional Analysis

We analyzed the metabolic functions in each of our samples by processing non-human reads with HUMAnN2 (Franzosa *et al*., 2018). We aligned all reads to UniRef90 using DIAMOND (Buchfink *et al*., 2014) and used HUMAnN2 to produce estimate of pathway abundance and completeness. We filtered all pathways that were less than 50% covered in a given sample but otherwise took the reported pathway abundance as is after relative abundance normalization (using HUMAnN2's built-in script). High level categories of functional pathways were found by grouping positively correlated pathways and manually annotating resulting clusters.

Assembly and Plasmid Annotations

All samples were assembled using metaSPAdes (v3.8.1 Nurk *et al*. (2017)) with default settings. Assembled scaffolds of at least 1,500bp of length were annotated using PlasFlow (v1.1 Krawczyk *et al*. (2018)) using default settings. PlasFlow predicts whether a contigs is likely from a chromosome or plasmid and gives a rough taxonomic annotation.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Analysis of Antimicrobial Resistance Genes

We generated profiles of antimicrobial resistance genes using MegaRes (v1.0.1, Lakin et al. (2017)). To generate profiles from MegaRes, we mapped non-human reads to the MegaRes database using Bowtie2 (v2.3.0, very-sensitive presets, Langmead and Salzberg, 2013). Subsequently, alignments were analyzed using ResistomeAnalyzer (commit 15a52dd github.com/cdeanj/resistomeanalyzer) and normalized by total reads per sample and gene length to give RPKMs. MegaRes includes an ontology grouping resistance genes into gene classes, AMR mechanisms, and gene groups. To reduce spurious mapping from gene homology we used BLASTn to align all MegaRes AMR genes against themselves. We considered any connected group of genes with an average nucleotide identity of 80% across 50% of the gene length as a set of potentially confounded genes. We collapsed all such groups into a single metagene with the mean abundance of all constituent genes. Before clustering genes, we removed all genes which were annotated as requiring SNP verification to predict resistance.

In addition to MegaRes, we mapped non-human reads from all samples to the amino acid gene sequences in the Comprehensive Antibiotic Resistance Database (McArthur *et al*., 2013) using DIAMOND. While we do not use this analysis explicitly in this study, we provide the results as a data table. Assembled contigs were annotated for AMR genes using metaProdigal (Hyatt *et al*., 2010), HMMER3 (Eddy, 2011), and ResFam (Gibson *et al*., 2015) as described by Rahman et al. (2018). All predicted gene annotations with an e-value higher than 10^{-10} were discarded.

Beta Diversity

Intersample (beta) diversity was measured by using Jaccard distances. We note that Jaccard distances do not use relative abundance information. Matrices of Jaccard distances were produced using built in SciPy functions treating all elements greater than 0 as present. Hierarchical clustering (average linkage) was performed on the matrix of Jaccard distances using SciPy (https://www.scipy.org/). Dimensionality reduction of taxonomic and functional profiles was performed using UMAP (McInnes *et al*., 2018) on the matrix of Jaccard distances with 100 neighbors (UMAP-learn package, random seed of 42). We did not use Principal Component Analysis as a preprocessing step before UMAP as is sometimes done for high dimensional data.

Alpha Diversity

Intrasample (alpha) diversity was measured by using Species Richness and Shannon's Entropy. We took species richness as the total number of detected species in a sample after rarefaction to 1 million reads. Shannon's entropy is robust to sample read depth and accounts for the relative size of each group in diversity estimation. Shannon's entropy is typically defined as $H = \sum a_i log_2 a_i$, where a_i is the relative abundance of taxon *i* in the sample. For alpha diversity based on k-mers or pathways, we substituted the relative abundance of a species for the relative abundance of the relevant type of object.

GeoDNA Sequence Search

For building the sequence graph index, each sample was processed with KMC (version 3) to convert the reads in FASTA format into lists of *k*-mer counts, using different values of *k* ranging from 13 to 19 in increments of 2. All *k*-mers that contained the character "N" or occurred in a sample less than twice were removed. For each value of *k*, we built a separate index, consisting of a labeled *de Bruijn* graph, using an implicit representation of the complete graph and a compressed label representation based on Multiary Binary Relation Wavelet Trees (Multi-BRWT). To build the index, for each sample the KMC *k*-mer count lists were transformed into *de Bruijn* graphs, from which path covers in the form of contig sets were extracted and stored as intermediate FASTA files. The contig sets of each sample were then transformed into annotation columns (one column per sample) by mapping them onto an implicit complete *de Bruijn* graph of order *k*. All annotation columns were then merged into a joint annotation matrix and transformed into Multi-BRWT format. Finally, the topology of the Multi-BRWT representation was optimized by relaxing its internal tree arity constraints to allow for a maximum arity of 40. The site for searching is publicly available (dnaloc.ethz.ch).

Viral Discovery

We followed the protocol described by Paez-Espino *et al*. (2017). Briefly, we used an expanded and curated set of viral protein families (VPFs) as bait in combination with recommended filtering steps to identify 16,584 UViGs directly from all MetaSUB metagenomic assemblies greater than 5kb. Then, the UViGs were clustered with the content of the IMG/VR system (a total of over 730k viral sequences including isolate viruses, prophages, and UViGs from all kind of habitats). The clustering step relied on a sequence-based classification framework (based on 95% sequence identity across 85% of the shortest sequence length) followed by the Markov clustering (mcl). This approach yielded 2,009 viral clusters (ranging from 2-611 members) and 9,605 singletons (or viral clusters of 1 member), sequences that failed to cluster with any sequence from the dataset or

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the references from IMG/VR, resulting in a total of 11,614 vOTUs. We define viral species from vOTUs as sequences sharing at least 95% identity over 85% of their length. Out of this total MetaSUB viral diversity, only 686 vOTUs clustered with any known viral sequence in IMG/VR.

Identifying Host-Virus Interactions

We used two computational methods to reveal putative host-virus connections (Paez-Espino *et al*., 2016). First, for the 686 vOTUs that clustered with viral sequences from the IMG/VR system, we projected the known host information to all the members of the group (total of 2,064 MetaSUB UViGs). Second, we used bacterial/archaeal CRISPR-Cas spacer matches (from the IMG/M 1.1 million isolate spacer database) to the UViGs (allowing only for 1 SNP over the whole spacer length) to assigned a host to 1,915 MetaSUB vOTUs. Additionally, we also used a database of over 20 million CRISPR-Cas spacers identified from metagenomic contigs from the IMG/M system with taxonomy assigned. Since some of these spacers may derive from short contigs these results should be interpreted with caution.

Unmapped DNA is not similar to any known sequence

A large proportion of the reads in our samples were not mapped to any reference sequences. To estimate the proportion of reads which could be assigned with more permissive metrics, we took 10k read subsets from the reads of 152 samples (post-human removal) and mapped the subsets to the NCBI-NT database with BLASTn. On average, 46.3% of non-human reads could be mapped to some known reference sequence (>90% average BLAST nucleotide identity roughly equivalent to the rank of Genus or Family) compared to an average 41.3% of reads mapped using our approach with KrakenUniq. We note that our approach to estimate the fraction of reads that could be classified using BLASTn does not account for hits to low quality taxa which would ultimately be discarded in our pipeline regardless.

Sequencing quality scores show expected trends

We measured sequencing quality based on 5 metrics: number of reads obtained from a sample, GC content (taken after removing human reads), Shannon's entropy of k-mers (from 10,000 reads sampled from each sample), post PCR Qubit score, and recorded DNA concentration before PCR. We observed good separation of negative and positive controls based on both PCR Qubit and k-mer entropy (**Figure S7**). Distributions of DNA concentration and the number of reads were as expected. GC content was broadly distributed for negative controls while positive controls were tightly clustered, expected since positive controls have a consistent taxonomic profile. Comparing the number of reads before and after quality control did not reveal any major outliers.

Testing for batch effects

A major concern for this study was that our results could be conflated with batch effect. For practical purposes, most of the flowcells used to sequence our samples contained samples from a randomly selected region. The median flowcell used in our study contained samples from just 3 cities and 2 continents. However, two flowcells covered 18 cities from 5 or 6 continents respectively. When samples from these flowcells were plotted using UMAP, the major global trends we described were recapitulated (**Figure S8A**). Further, when plotting samples by PCR qubit and k-mer entropy (the two metrics that most reliably separated our positive and negative controls) and overlaying the flowcell used to sequence each sample, only one outlier flowcell was identified, which was the used to sequence a large number of background control samples (**Figure S8B**). Plots of the number of reads against city of origin and surface material (Figures **S8C & D**) showed a stable distribution of reads across cities. Analogous plots of PCR Qubit scores were less stable than the number of reads, but showed a clear drop for control samples (**Figures S8E & F**). These results indicate that batch effects are likely to be minimal.

Strain contamination is rare or absent

Despite good separation of positive and negative controls, we identified several species in our negative controls which were also identified as prominent taxa in the dataset as a whole. Our dilemma was that a microbial species that is common in the urban environment might also reasonably be expected to be common in the lab environment. In general, negative controls had lower k-mer complexity, fewer reads, and lower post-PCR Qubit scores than case samples, and no major flowcell-specific species were observed. Similarly, positive control samples were not heavily contaminated. These results suggest samples are high quality but do not systematically exclude the possibility of contamination.

Previous studies have reported that microbial species whose relative abundance is negatively correlated with DNA concentration may be contaminants. We observed a number of species that were negatively correlated with DNA concentration (**Figure S6A**), but this distribution followed the same shape as a null distribution of

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

uniformly randomly generated relative abundances (**Figure S6B**). We also plotted correlation with DNA concentration against each species mean relative abundance across the entire dataset (**Figure S6C**). Species that were negatively correlated with DNA concentration were clearly more abundant than uncorrelated species, this suggests that there may be a jackpot effect for prominent species in samples with lower concentrations of DNA, but is not generally consistent with contamination.

Finally, we compared assemblies from negative controls to assemblies from our case samples searching for regions of high similarity that could be from the same microbial strain. We reasoned that uncontaminated samples may contain the same species as negative controls but were less likely to contain identical strains. Only 137 case samples were observed to have any sequence with high similarity to an assembled sequence from a negative control (8,000 base pairs minimum of 99.99% identity). The identified sequences were principally from Bradyrhizobium and Cutibacterium. Since these genera are core taxa (observed in nearly every sample) and high similarity was only identified in a few samples, we elected not to remove species from these genera from case samples.

K-Mer based metrics correlate with taxonomic metrics

We found clear correlations between three pairwise distance metrics (**Figures S3A, B, C**): k-mer based Jaccard distance (MASH), taxonomic Jaccard distance, and taxonomic Jensen-Shannon divergence. This suggests that taxonomic variation reflects meaningful variation in the underlying sequence in a sample. We also compared alpha diversity metrics (**Figure S3D**): Shannon entropy of k-mers, and Shannon entropy of taxonomic profiles. These metrics for pairwise distances were correlated, though some noise was present (**Figure S3**).

References

Afshinnekoo, E., Chou, C., Alexander, N., Ahsanuddin, S., Schuetz, A. N., and Mason, C. E. (2017). Precision metagenomics: Rapid metagenomic analyses for infectious disease diagnostics and public health surveillance. *Journal of Biomolecular Techniques*, 28(1):40–45.

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J. M., Reeves, D., Gandara, J., Chhangawala, S., Ahsanuddin, S., Simmons, A., Nessel, T., Sundaresh, B., Pereira, E., Jorgensen, E., Kolokotronis, S. O., Kirchberger, N., Garcia, I., Gandara, D., Dhanraj, S., Nawrin, T., Saletore, Y., Alexander, N., Vijay, P., Hénaff, E. M., Zumbo, P., Walsh, M., O'Mullan, G. D., Tighe, S., Dudley, J. T., Dunaif, A., Ennis, S., O'Halloran, E., Magalhaes, T. R., Boone, B., Jones, A. L., Muth, T. R., Paolantonio, K. S., Alter, E., Schadt, E. E., Garbarino, J., Prill, R. J., Carlton, J. M., Levy, S., and Mason, C. E. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Systems*, 1(1):72–87.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Altschul et al.. 1990. Basic Local Alignment Search Tool.pdf.

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., and Bork, P. (2018). Structure and function of the global topsoil microbiome.

Bougnom, B. P. and Piddock, L. J. (2017). Wastewater for Urban Agriculture: A Significant Factor in Dissemination of Antibiotic Resistance.

Breitwieser, F. P., Baker, D. N., and Salzberg, S. L. (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome biology*, 19(1):198.

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., Birren, B. W., Xavier, R. J., Blainey, P. C., Singh, A. K., Gevers, D., and Alm, E. J. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND.
Cáliz, J., Triadó-Margarit, X., Camarero, L., and Casamayor, E. O. (2018). A long-term survey unveils strong seasonal patterns in the airborne microbiome coupled to general and regional atmospheric circulations. *Proceedings of the National Academy of Sciences*, 115(48):12229–12234.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Consortium, T. H. M. P., Human, T., Project, M., Consortium, T. H. M. P., Human, T., and Project, M. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14.

Cooley, J. D., Wong, W. C., Jumper, C. A., and Straus, D. C. (1998). Correlation between the prevalence of certain fungi and sick building syndrome. *Occupational and Environmental Medicine*, 55(9):579–584.

Eckburg, P. B., Mian, M. F., Surette, M. G., Bienenstock, J., Forsythe, P., and Sargent, M. (2005). Diversity of the Human Intestinal Microbial Flora. *Science*, 308(5728):1635–1638.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10).

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., and Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 15(11):962–968.

Fresia, P., Antelo, V., Salazar, C., Giménez, M., D'Alessandro, B., Afshinnekoo, E., Mason, C., Gonnet,

G. H., and Iraola, G. (2019). Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. *Microbiome*, 7(1).

Gardy, J. L. and Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveil- lance system.

Gibson, M. K., Forsberg, K. J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME Journal*, 9(1):207–216.

Gilbert, J. A. and Stephens, B. (2018). Microbiology of the built environment.

Harris ZN, Dhungel E, Mosior M, Ahn T. Massive metagenomic data analysis using abundance-based machine learning. Biol Direct. 2019; 14: 12.

Hendriksen, R. S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S. K., Kjeldgaard, J., Kaas, R. S., Clausen, P. T. L. C., Vogt, J. K., Leekitcharoenphon, P., van de Schans, M. G. M., Zuidema, T., de Roda Husman, A. M., Rasmussen, S., Petersen, B., Bego, A., Rees, C., Cassar, S., Coventry, K., Collignon, P., Allerberger, F., Rahube, T. O., Oliveira, G., Ivanov, I., Vuthy, Y., Sopheak, T., Yost, C. K., Ke, C., Zheng, H., Baisheng, L., Jiao, X., Donado-Godoy, P., Coulibaly, K. J., Jergović, M., Hrenovic, J., Karpíšková, R., Villacis, J. E., Legesse, M., Eguale, T., Heikinheimo, A., Malania, L., Nitsche, A., Brinkmann, A., Saba, K. S., Kocsis, B., Solymosi, N., Thorsteinsdottir, T. R., Hatha, A. M., Alebouyeh, M., Morris D., Cormican, M., O'Connor, L., Moran-Gilad, J., Alba, P., Battisti, A., Shakenova, Z., Kiiyukia, C., Ng'eno, E., Raka, L., Avsejenko, J., B¯erzin, š, A., Bartkevics, V., Penny, C., Rajandas, H., Parimannan, S., Haber, M. V., Pal, P., Jeunen, G.-J., Gemmell, N., Fashae, K., Holmstad, R., Hasan, R., Shakoor, S., Rojas, M. L. Z., Wasyl, D., Bosevska, G., Kochubovski, M., Radu, C., Gassama, A., Radosavljevic, V., Wuertz, S., Zuniga-Montanez, R., Tay, M. Y. F., Gavačová, D., Pastuchova, K., Truska, P., Trkov, M., Esterhuyse, K., Keddy, K., Cerdà-Cuéllar, M., Pathirage, S., Norrgren, L., Örn, S., Larsson, D. G. J., Heijden, T. V. d., Kumburu, H. H., Sanneh, B., Bidjada, P., Njanpop-Lafourcade, B.-M., Nikiema-Pessinaba, S. C., Levent, B., Meschke, J. S., Beck, N. K., Van, C. D., Phuc, N. D., Tran, D. M. N., Kwenda, G., Tabo, D.-a., Wester, A. L., Cuadros-Orellana, S., Amid, C., Cochrane, G., Sicheritz-Ponten, T., Schmitt, H., Alvarez, J. R. M., Aidara-Kane, A., Pamp, S. J., Lund, O., Hald, T., Woolhouse, M., Koopmans, M. P., Vigre, H., Petersen, T. N., Aarestrup, F. M., and project consortium, T. G. S. S. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications*, 10(1):1124.

Hsu, T., Joice, R., Vallarino, J., Abu-Ali, G., Hartmann, E. M., Shafquat, A., DuLong, C., Baranowski, C., Gevers, D., Green, J. L., Morgan, X. C., Spengler, J. D., and Huttenhower, C. (2016). Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. *mSystems*, 1(3):e00018–16.

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

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kang, K., Ni, Y., Li, J., Imamovic, L., Sarkar, C., Kobler, M. D., Heshiki, Y., Zheng, T., Kumari, S., Wong, J. C. Y., Archna, A., Wong, C. W. M., Dingle, C., Denizen, S., Baker, D. M., Sommer, M. O. A., Webster, C. J., and Panagiotou, G. (2018). The Environmental Exposures and Inner- and Intercity Traffic Flows of the Metro System May Contribute to the Skin Microbiome and Resistome. *Cell Reports*, 24(5):1190–1202.e5.

Klein, E. Y., Van Boeckel, T. P., Martinez, E. M., Pant, S., Gandra, S., Levin, S. A., Goossens, H., and Laxminarayan, R. (2018). Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proceedings of the National Academy of Sciences*, 115(15):E3463–E3470.

Korownyk, C., Liu, F., and Garrison, S. (2018). Population level evidence for seasonality of the human microbiome. *Chronobiology International*, 35(4):573–577.

Koslicki, D. (2011). Topological entropy of DNA sequences. *Bioinformatics*, 27(8):1061–1067.

Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6):e35–e35.

Ladner, J. T., Grubaugh, N. D., Pybus, O. G., and Andersen, K. G. (2019). Precision epidemiology for infectious disease control.

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., Rovira, P., Abdo, Z., Jones, K. L., Ruiz, J., Belk, K. E., Morley, P. S., and Boucher, C. (2017). MEGARes: An antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Research*, 45(D1):D574–D580.

Langmead and Steven L Salzberg (2013). Bowtie2. *Nature methods*, 9(4):357–359.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O., and Huttenhower, C. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, 550(7674):61–66.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7):3348–3357.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

McIntyre, A. B., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., and Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1).

MetaSUB International Consortium. Mason, C., Afshinnekoo, E., Ahsannudin, S., Ghedin, E., Read, T., Fraser, C., Dudley, J., Hernandez, M., Bowler, C., Stolovitzky, G., Chernonetz, A., Gray, A., Darling, A., Burke, C., ?abaj, P. P., Graf, A., Noushmehr, H., Moraes, S., Dias-Neto, E., Ugalde, J., Guo, Y., Zhou, Y., Xie, Z., Zheng, D., Zhou, H., Shi, L., Zhu, S., Tang, A., Ivankovi?, T., Siam, R., Rascovan, N., Richard, H., Lafontaine, I., Baron, C., Nedunuri, N., Prithiviraj, B., Hyat, S., Mehr, S., Banihashemi, K., Segata, N., Suzuki, H., Alpuche Aranda, C. M., Martinez, J., Christopher Dada, A., Osuolale, O., Oguntoyinbo, F., Dybwad, M., Oliveira, M., Fernandes, A., Oliveira, M., Fernandes, A., Chatziefthimiou, A. D., Chaker, S., Alexeev, D., Chuvelev, D., Kurilshikov, A., Schuster, S., Siwo, G. H., Jang, S., Seo, S. C., Hwang, S. H., Ossowski, S., Bezdan, D., Udekwu, K., Udekwu, K., Lungj- dahl, P. O., Nikolayeva, O., Sezerman, U., Kelly, F., Metrustry, S., Elhaik, E., Gonnet, G., Schriml, L., Mongodin, E., Huttenhower, C., Gilbert, J., Hernandez, M., Vayndorf, E., Blaser, M., Schadt, E., Eisen, J., Beitel, C., Hirschberg, D., Schriml, L., and Mongodin, E. (2016). The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome*, 4(1):24.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Meyer, K. M., Memiaghe, H., Korte, L., Kenfack, D., Alonso, A., and Bohannan, B. J. (2018). Why do microbes exhibit weak biogeographic patterns? *ISME Journal*, 12(6):1404–1413.

Moskowitz, D. M. and Greenleaf, W. J. (2018). Nonparametric analysis of contributions to variance in genomics and epigenomics data. *bioRxiv*.

Neiderud, C. J. (2015). How urbanization affects the epidemiology of emerging infectious diseases. *African Journal of Disability*, 5(1).

Nicolaou, N., Siddique, N., and Custovic, A. (2005). Allergic disease in urban and rural populations: Increasing prevalence with increasing urbanization. *Allergy: European Journal of Allergy and Clinical Immunology*, 60(11):1357–1360.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, 27(5):824–834.
\
O'Hara, N. B., Reed, H. J., Afshinnekoo, E., Harvin, D., Caplan, N., Rosen, G., Frye, B., Woloszynek, S., Ounit, R., Levy, S., Butler, E., and Mason, C. E. (2017). Metagenomic characterization of ambulances across the USA. *Microbiome*, 5(1):125.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, 17(1):132.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, 536(7617):425–430.

Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N., and Kyrpides, N. C. (2017). Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nature Protocols*, 12(8):1673– 1682.

Paez-Espino, D., Roux, S., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Huntemann, M., Reddy, T. B., Pons, J. C., Llabrés, M., Eloe-Fadrosh, E. A., Ivanova, N. N., and Kyrpides, N. C. (2019). IMG/VR v.2.0: An integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research*, 47(D1):D678–D686.

Qin, J., Li, R., Raes, J., and Arumugam, M. (2010). A human gut microbial gene catalogue established by metagenomic sequencing: Commentary.

Rahman, S. F., Olm, M. R., Morowitz, M. J., and Banfield, J. F. (2018). Machine Learning Leverag- ing Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome. *mSystems*, 3(1).

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1).

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811.

Shaaban, H., Westfall, D. A., Mohammad, R., Danko, D., Bezdan, D., Afshinnekoo, E., Segata, N., and Mason, C. E. (2018). The Microbe Directory: An annotated, searchable inventory of microbes' characteristics. *Gates Open Research*, 2:3.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vazquez-Baeza, Y., Gonzalez, A., Morton, J. T., Mirarab, S., Xu, Z. Z., Jiang, L., Haroon, M. F., Kanbar, J., Zhu, Q., Song, S. J., Kosciolek, T., Bokulich, N. A., Lefler, J., Brislawn, C. J., Humphrey, G., Owens, S. M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J. A., Clauset, A., Stevens,
R. L., Shade, A., Pollard, K. S., Goodwin, K. D., Jansson, J. K., Gilbert, J. A., and Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463.
United Nations (2016). Political declaration of the high-level meeting of the General Assembly on antimicrobial

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

resistance. Technical report.

United Nations (2018). World Urbanization Prospects: The 2018 Revision. Key facts. Technical report.

Van Boeckel, T. P., Brower, C., Gilbert, M., Grenfell, B. T., Levin, S. A., Robinson, T. P., Teillant, A., and Laxminarayan, R. (2015). Global trends in antimicrobial use in food animals. *Proceedings of the National Academy of Sciences of the United States of America*, 112(18):5649–54.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.

Walker AR, Datta S. Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data. Biol Direct. 2019; 14: 11.

Wilson, M. R., Sample, H. A., Zorn, K. C., Arevalo, S., Yu, G., Neuhaus, J., Federman, S., Stryke, D., Briggs, B., Langelier, C., Berger, A., Douglas, V., Josephson, S. A., Chow, F. C., Fulton, B. D., DeRisi, J. L., Gelfand, J. M., Naccache, S. N., Bender, J., Dien Bard, J., Murkey, J., Carlson, M., Vespa, P. M., Vijayan, T., Allyn, P. R., Campeau, S., Humphries, R. M., Klausner, J. D., Ganzon, C. D., Memar, F., Ocampo, N. A., Zimmermann, L. L., Cohen, S. H., Polage, C. R., DeBiasi, R. L., Haller, B., Dallas, R., Maron, G., Hayden, R., Messacar, K., Dominguez, S. R., Miller, S., and Chiu, C. Y. (2019). Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *New England Journal of Medicine*, 380(24):2327–2340.

Zhu, Y.-G., Gillings, M., Simonet, P., Stekel, D., Banwart, S., and Penuelas, J. (2017). Microbial mass movements. *Science*, 357(6356):1099–1100.

Contributing Members of the MetaSUB Consortium

Marcos Abraao, Muhammad Afaq, Ireen Alam, Kalyn Ali, Lucia E Alvarado-Arnez, Sarh Aly, Jennifer Amachee, 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, Kevin Becher, Joseph Benson, Denis Bertrand, Silvia Beurmann, Christina Black, Brittany Blyther, Bazartseren Boldgiv, Gabriela P Branco, Christian Brion, John Brownstein, 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, David A Coil, Colleen Conger, 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 Ep- ping, 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 Fran- cis, 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,

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hyunjung Lee, Lucy Lee, Yunmi Lee, Emily Leong, Shawn Levy, 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, Tathi- ane M Malta, Tathiane M Malta, Maliha Mamun, Yuk Man Tang, Sonia Marinovic, Brunna Marques, Nicole Mathews, Yuri Matsuzaki, Madelyn May, Elias McComb, Adiell Melamed, 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, 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, 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, Alessan- dro 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 Sa- yara, Steffen Schaaf, Anna-Lena M Schinke, Ralph Schlapbach, Jason R Schriml, Felipe Segato, 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, 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, Car- men 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 Wright, Ryota Yamanaka, Jingcheng Yang, Hirokazu Yano, George C Yeh, Tsoi Ying Lai, Laraib Zafar, Amy Zhang, Shu Zhang, Yang Zhang, Yuanting Zheng.

Conflicts of Interest

CEM, RO, and NO are board members and equity stakeholders in Biotia, Inc., although no aspects of this company reflect a conflict for purposes of this study.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Materials

Table S1: Sample Counts

Region	project / city	Pilot	CSD16	CSD17	Other	Total
Control	Background Control	0	18	49	0	67
	Lab Negative Control	9	0	24	7	40
	Positive Control	9	1	33	7	50
East Asia	Continent Total	22	25	1244	0	1291
	Hanoi	0	0	16	0	16
	Hong Kong	0	0	713	0	713
	Kuala Lumpur	0	0	14	0	14
	Sendai	0	0	16	0	16
	Seoul	12	0	78	0	90
	Shanghai	10	0	0	0	10
	Singapore	0	0	192	0	192
	Taipei	0	0	94	0	94
	Tokyo	0	25	112	0	137
	Yamaguchi	0	0	9	0	9
Europe	Continent Total	131	235	845	71	1282
	Barcelona	7	96	0	0	103
	Belfast	0	0	0	5	5
	Berlin	0	55	0	0	55
	Birmingham	0	0	0	7	7
	Bradford	0	0	0	4	4
	Bury	0	0	0	6	6
	Eastbourne	0	0	0	6	6
	Eden	0	0	0	5	5
	Edinborough	0	0	0	6	6
	Islington	0	0	0	5	5
	Jaywick	0	0	0	6	6
	Kensington	0	0	0	5	5
	Kyiv	0	0	96	0	96
	Lands End	0	0	0	5	5
	Lisbon	0	84	0	0	84
	London	0	0	535	0	535
	Marseille	0	0	16	0	16
	Naples	0	0	16	0	16
	Newcastle	0	0	0	5	5
	Oslo	12	0	16	0	28
	Paris	0	0	16	0	16
	Porto	112	0	0	0	112
	Sofia	0	0	16	0	16
	Stockholm	0	0	64	0	64
	Swansea	0	0	0	6	6
	Vienna	0	0	16	0	16
	Zurich	0	0	54	0	54
Middle East	Continent Total	0	66	14	0	80
	Doha	0	66	14	0	80

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S1: Sample Counts Cont.

continent	project city	Pilot	CSD16	CSD17	Other	Total
North America	Continent Total	28	221	158	184	591
	Baltimore	0	0	14	0	14
	Denver	0	23	16	0	39
	Fairbanks	0	100	0	0	100
	Mexico City	10	0	0	0	10
	Minneapolis	0	0	16	0	16
	New York City	0	82	96	184	362
	Sacramento	18	16	0	0	34
	San Francisco	0	0	16	0	16
Oceania	Continent Total	0	31	32	0	63
	Auckland	0	15	0	0	15
	Brisbane	0	0	16	0	16
	Hamilton	0	16	0	0	16
	Honolulu	0	0	16	0	16
South America	Continent Total	20	43	29	34	126
	Bogota	0	17	0	0	17
	Montevideo	20	0	0	0	20
	Rio De Janeiro	0	0	0	34	34
	Santiago	0	26	0	0	26
	Sao Paulo	0	0	29	0	29
Sub Saharan Africa	Continent Total	0	92	191	0	283
	Ilorin	0	66	191	0	257
	Offa	0	26	0	0	26

Table S2: Covariate Variance.
The sample variance that can be explained by each factor, in isolation.

Factor	Variance Explained
City Population Density	0%
City Ave June Temp	4%
City Elevation	4%
Coastal City	2%
Surface Material	3%
Koppen Climate Classification	3%
Setting	6%
Above/Below Ground	7%
Continent	11%

Table S3: Classification Accuracy for Different Metadata Features. Mean classification accuracy for a classifier trained to predict a given feature using all samples except those from a given city and tested on the held out city. Reported number is the mean of the mean accuracy in each city.

Feature	Accuracy
city_ave_june_temp_c	0.552829
city_elevation	0.576737
city_koppen_climate	0.385825
city_latitude	0.515680
city_longitude	0.522026
city_population_density	0.532092
city_total_population	0.658395
coastal_city	0.601269
continent	0.585694

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

surface_ontology_coarse	0.622925
surface_ontology_fine	0.374484

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

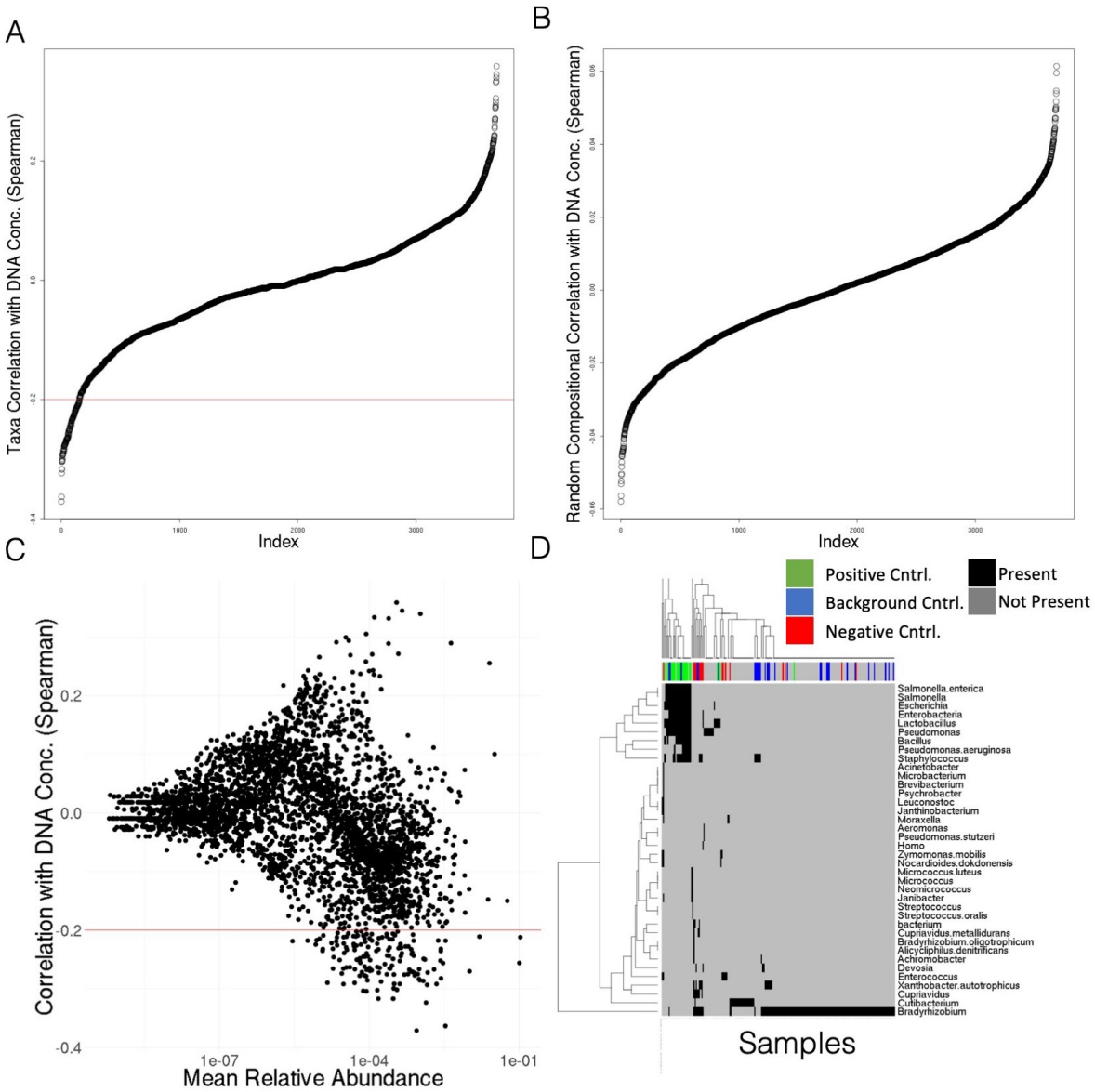

Figure S1: Sample and Taxonomic Distributions. A) Correlation of taxonomic (species) relative abundances with DNA concentration B) Correlation of randomly generated compositional vectors with DNA concentration. Note the same shape but lower magnitude C) Correlation of taxa with DNA Concentration vs the mean relative abundance of that taxa D) Presence (black) absence (grey) heatmap of taxa found in controls and other samples. Colored bar at top, red are negative controls, blue are background, green are positive. Case samples with homology are grey. Case samples without homology to control sequences are not shown.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

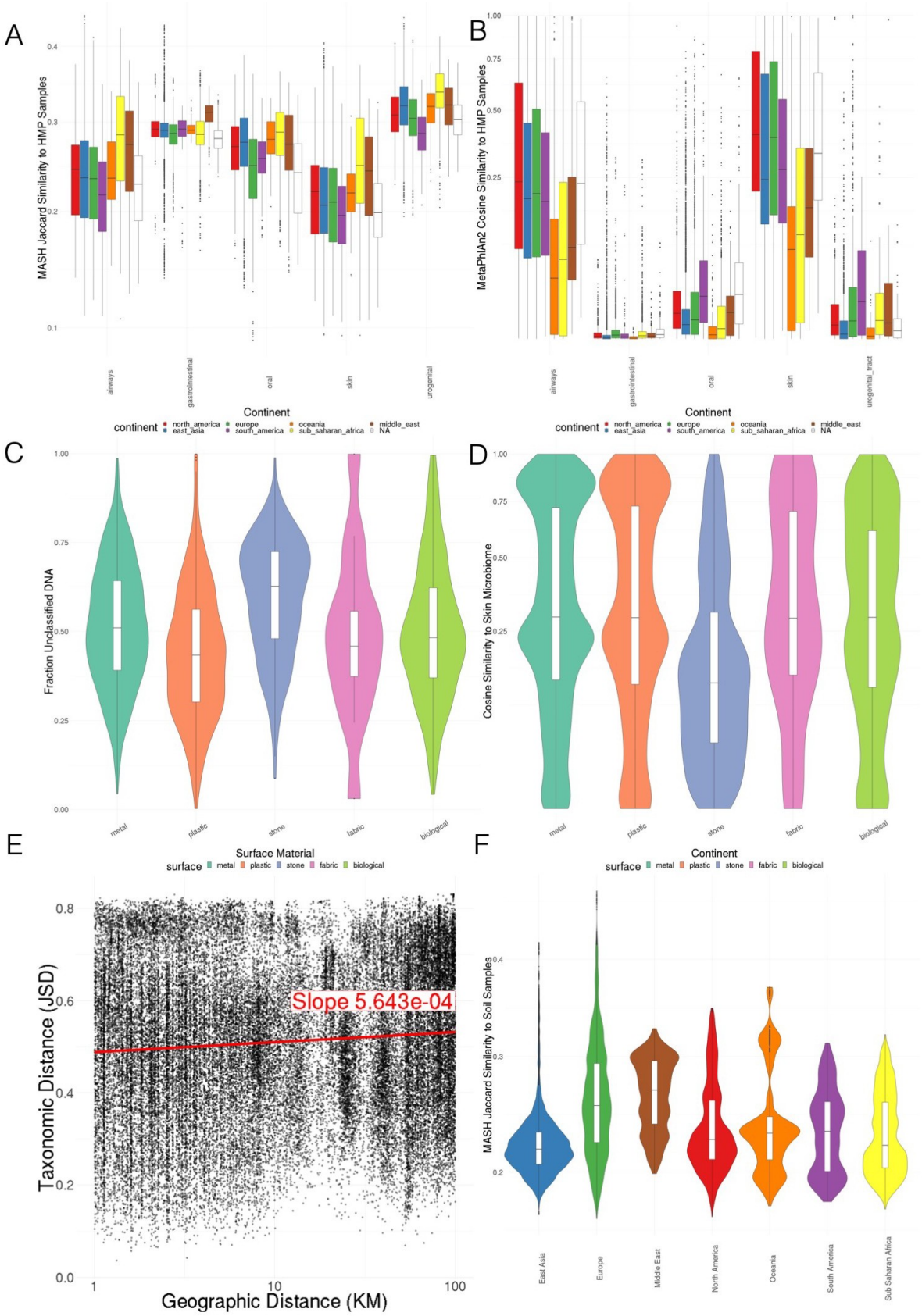

Figure S2: A) MASH *k*-mer Jaccard similarity to representative HMP samples, colored by continent B) MetaPhlAn v2.0 cosine similarity to representative HMP samples, colored by continent C) Fraction unclassified DNA by surface material D) Cosine similarity to MetaPhlAn v2.0 skin microbiome profile by surface E) Jensen-Shannon distance between pairs of taxonomic profiles vs Geographic Distance F) MASH *k*-mer Jaccard similarity to representative soil samples, colored by continent.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

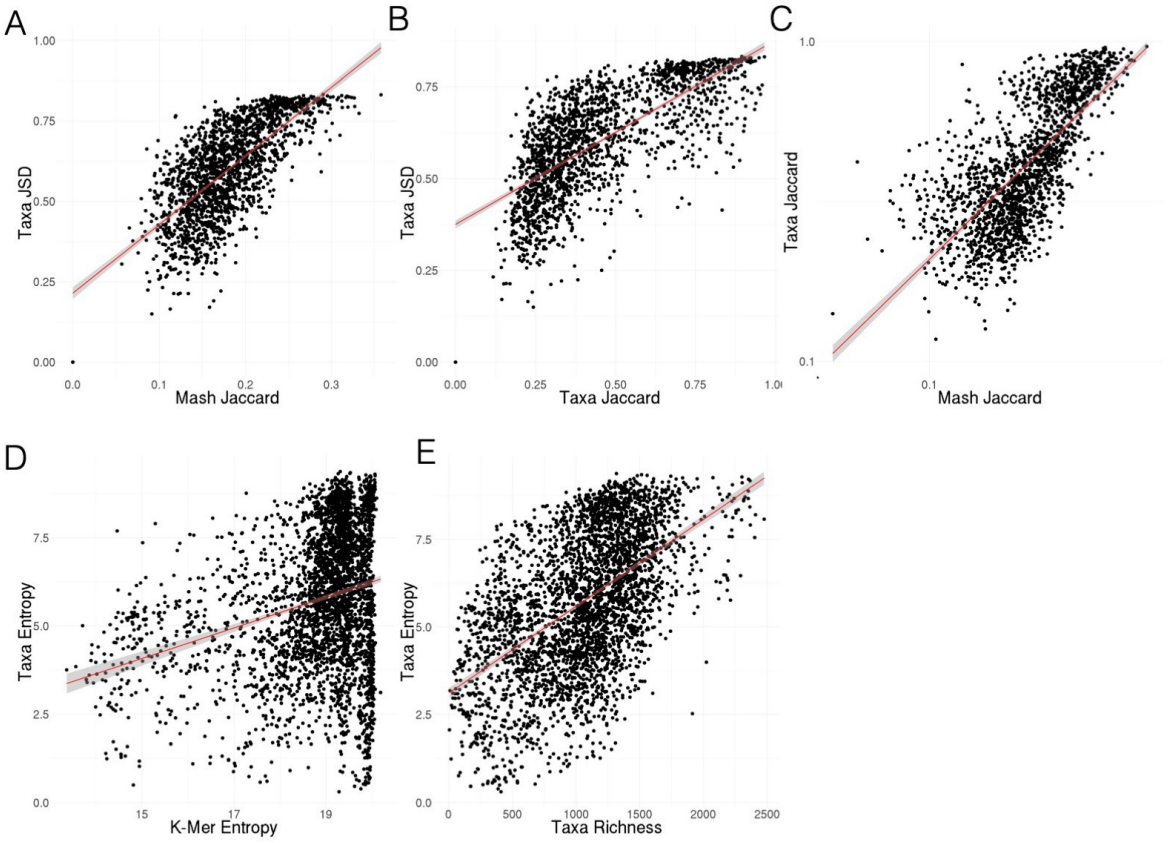

Figure S3: K-mer Diveristy of Samples. A) Jensen-Shannon Divergence of taxonomic profiles vs MASH Jaccard distance of *k*-mers B) Jensen-Shannon Divergence of taxonomic profiles vs Jaccard distance of taxonomic profiles. C) Jaccard distance of taxonomic profilesvs MASH Jaccard distance of *k*-mers D) Shannon's Entropy of taxonomic profiles vs Shannon's Entropy of *k*-mers E) Taxonomic richness (number of species) vs Shannon's Entropy of taxonomic profiles.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

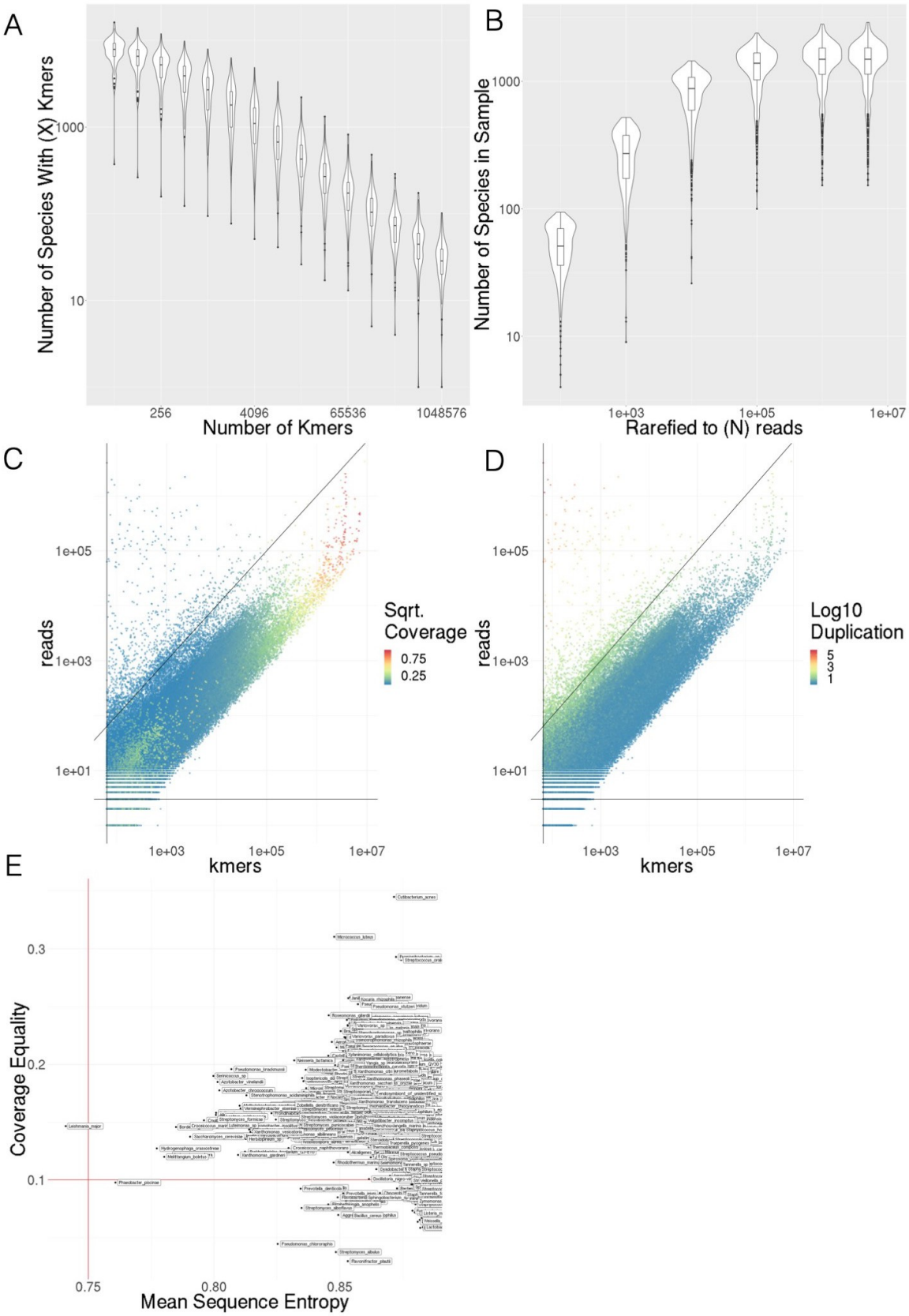

Figure S4: Sequencing Quality Control Metrics. A) Number of species detected as k-mer threshold increases for 100 randomly selected samples B) Number of species detected as number of sub-sampled reads increase C) k-mer counts compared to number of reads for species level annotations in 100 randomly selected samples, colored by coverage of marker k-mer set D) k-mer counts compared to number of reads for species level annotations in 100 randomly selected samples, colored by average duplication of k-mers E) Comparison of Mean Sequence Entropy and Coverage Equality for core and sub-core taxa. Thresholds are shown by red lines.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

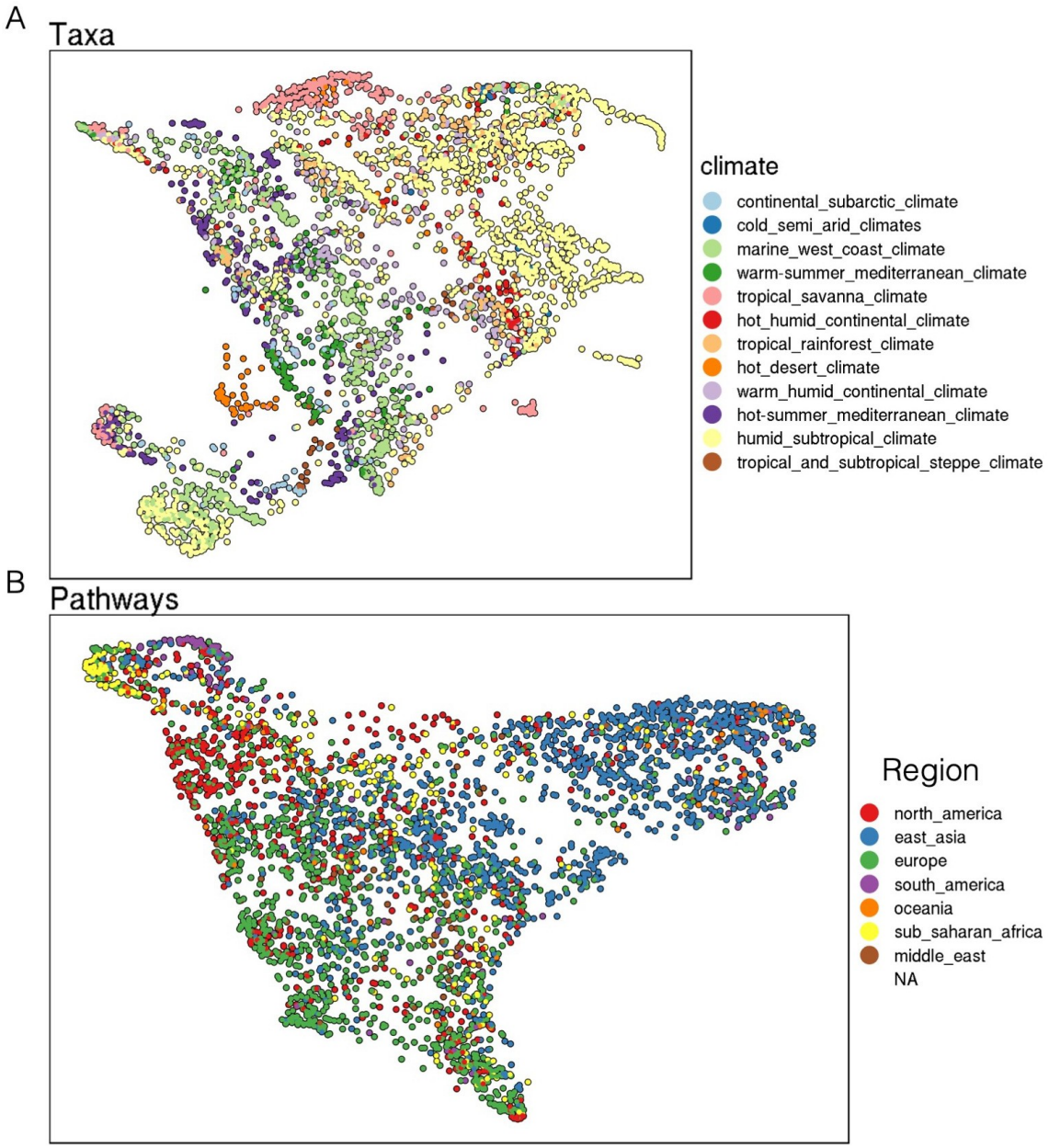

Figure S5: Climate and Geographical Clustering. A) UMAP of taxonomic profiles colored by climate classification B) UMAP of functional pathways colored by continent

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

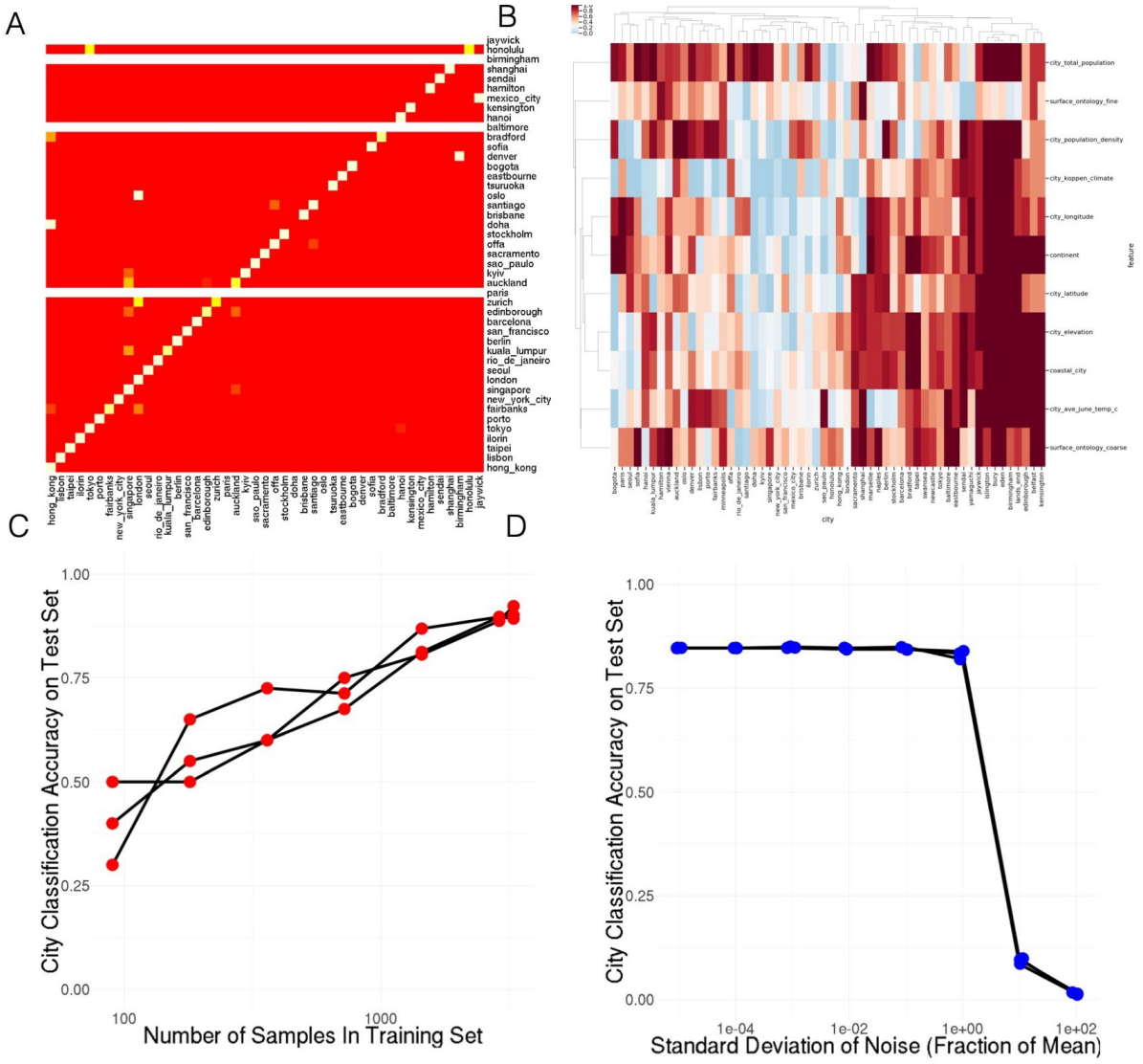

Figure S6: Predictive Capacity Testing. A) Confusion matrix of SVM classifier on test data. White lines are cities which were not in test set B) Heatmap showing classification accuracies of classifiers trained to predict a given feature using all samples except those from a given city and tested on the held out city C) Performance of Linear SVM Classifier with Increasing Data D) Effect of noise on a Linear SVM trained to predict a sample's city from its taxonomic profile. Noise is scaled to the magnitude of each taxa within the profile.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

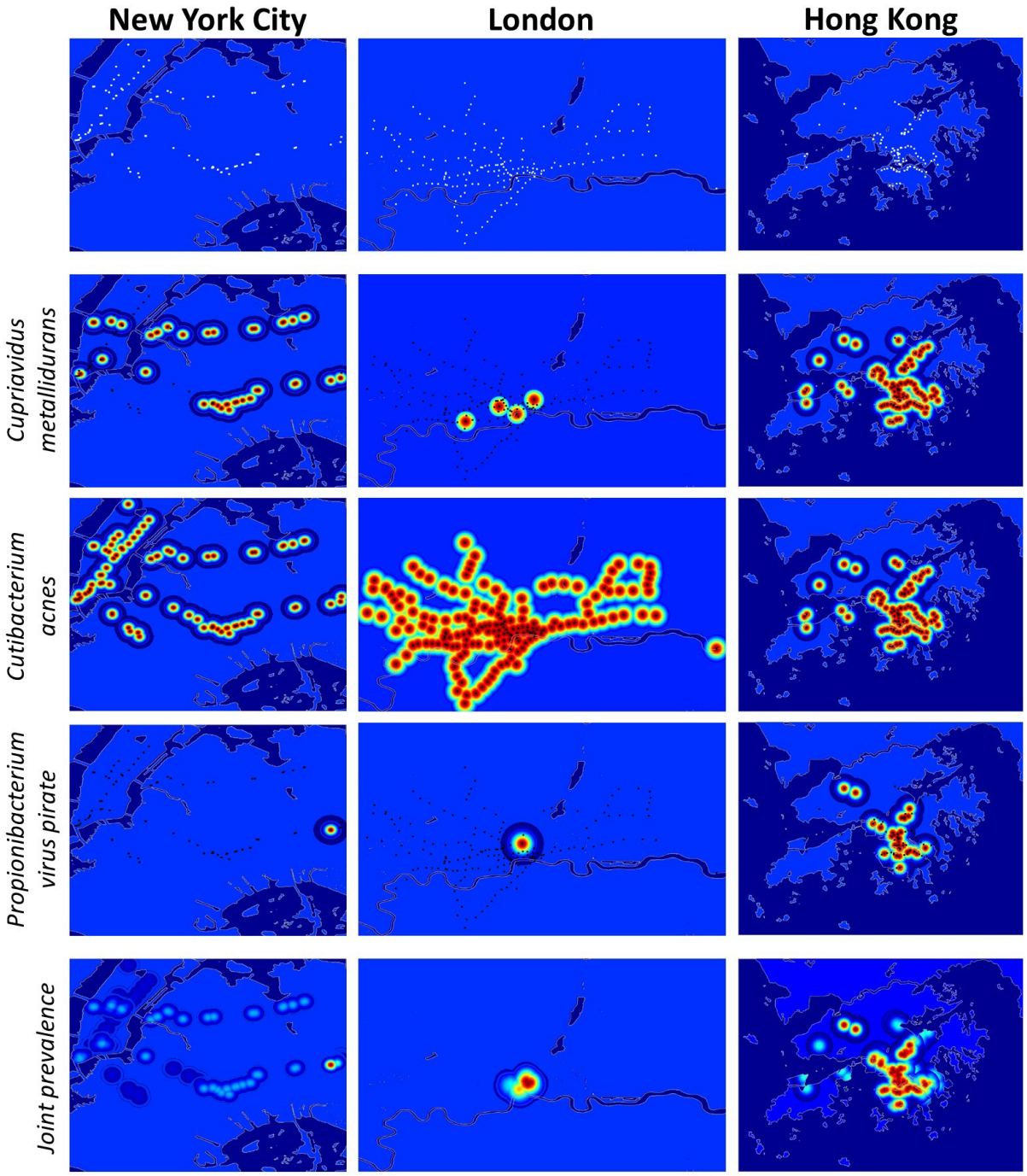

Figure S7: Example Geographic taxonomic Distributions. Distributions of taxa were estimated by fitting Gaussian distributions to sampling locations where the taxa was found with standard deviations based on the geographic distance between observations. Top Row) Sampling sites in three major cities Rows 2-4) Estimated distribution of different example species in major cities Row 5) Estimated distribution of three species together in major cities.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

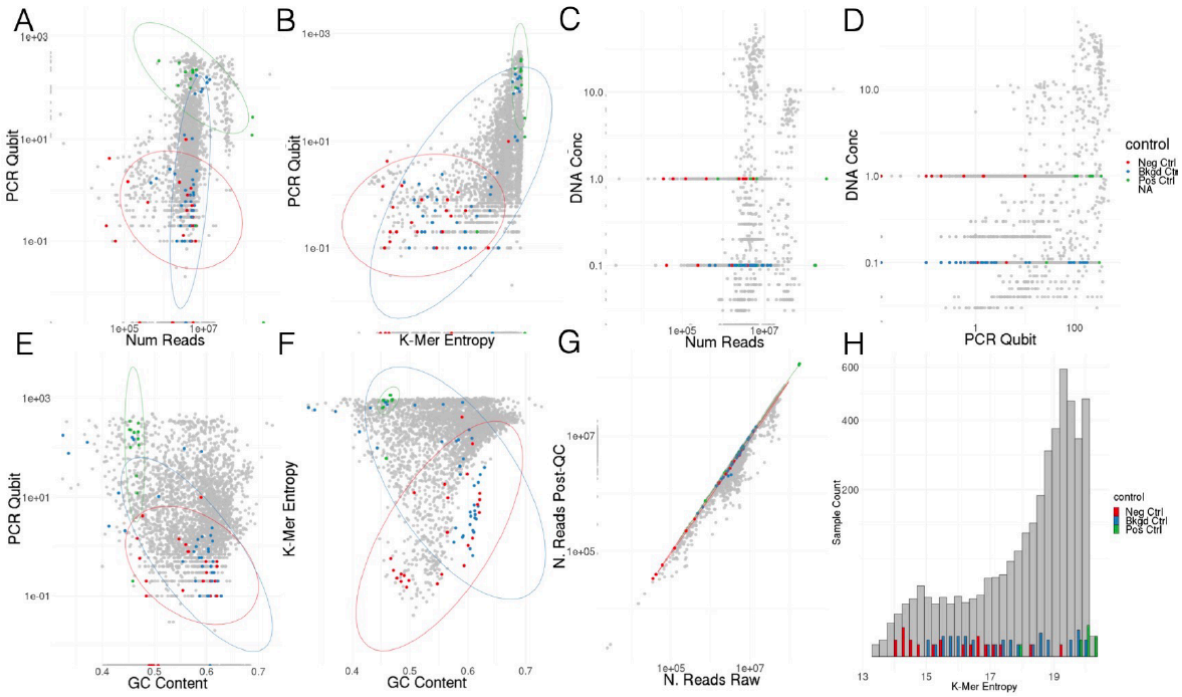

Figure S8: Comparisons of different sequencing quality control metrics with controls marked. A-F) Comparisons of the raw reads, PCR Qubit scores, manually recorded DNA concentrations, k-mer Shannon entropy, and GC fraction of quality controlled reads G) Comparison of read counts before and after quality control but before human reads were removed H) Histogram showing the number of samples with different k-mer entropies.

bioRxiv preprint doi: https://doi.org/10.1101/724526; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

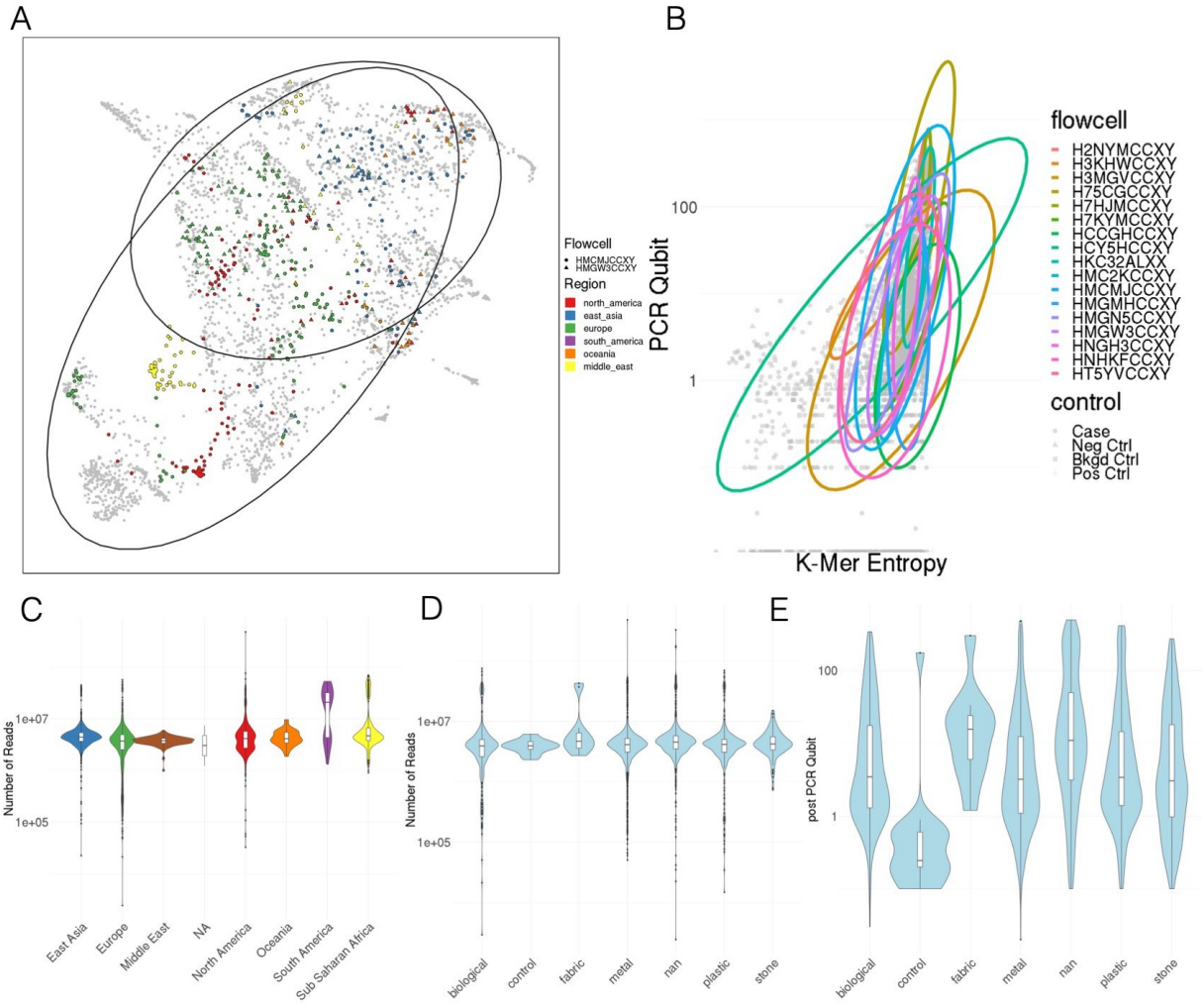

Figure S9: UMAP profiles. A) UMAP of taxonomic profiles from geographically diverse flowcells B) Flowcells vs quality control metrics C) Number of reads by region D) number of reads by surface material E) PCR Qubit by surface material.