# Global within-species phylogenetics of sewage microbes suggest that local adaptation shapes geographical bacterial clustering

**Marie Jespersen**
Technical University of Denmark
**Patrick Munk**
Technical University of Denmark https://orcid.org/0000-0001-8813-4019
**Joachim Johansen**
University of Copenhagen https://orcid.org/0000-0001-7052-1870
**Rolf Kaas**
Technical University of Denmark https://orcid.org/0000-0002-5050-8668
**Henry Webel**
University of Copenhagen
**Håkan Vigre**
Technical University of Denmark
**Henrik Nielsen**
Clinical Microbiomics A/S https://orcid.org/0000-0003-2281-5713
**Simon Rasmussen**
University of Copenhagen https://orcid.org/0000-0001-6323-9041
**Frank Aarestrup** (fmaa@food.dtu.dk)
Professor https://orcid.org/0000-0002-7116-2723

**Article**

**Keywords:**

**Posted Date:** February 22nd, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1321305/v1

# Global within-species phylogenetics of sewage microbes suggest that local adaptation shapes geographical bacterial clustering


## Authors

Marie Louise Jespersen[1,2], Patrick Munk[1], Joachim Johansen[2], Håkan Vigre[1], Rolf Sommer Kaas[1], Henry Webel[2], Henrik Bjørn Nielsen[3], Simon Rasmussen[2], Frank M. Aarestrup[1].

## Affiliations

[1]National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark

[2] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N, Denmark

[3] Clinical-Microbiomics A/S, Copenhagen, Denmark

## Author List Footnotes

Correspondence: Simon Rasmussen (simon.rasmussen@cpr.ku.dk) and Frank M. Aarestrup (fmaa@food.dtu.dk)

**Abstract**

Most investigations of geographical differences within microbial species are limited to focusing on a single species. Here, we investigate the global differences for multiple bacterial species by using a dataset of 757 metagenomics sewage samples from 101 different countries worldwide. The within-species variations were identified by performing unsupervised genome reconstructions that reduce database and mapping biases, and the analyses were further expanded by using gene focused approaches. Applying these methods, we recovered 3,353 near complete (NC) metagenome assembled genomes (MAGs) encompassing 1,439 different MAG species and found that within-species genomic variation was often coherent with regional separation. Additionally, we found that the variation of organelle genes correlated less with geography compared to metabolic and membrane associated genes, suggesting that the global differences of these species are caused by regional environmental selection rather than limitations on dissemination. From the combination of the large and globally distributed dataset with the in-depth analysis methods, we present the most comprehensive investigation of global within-species phylogeny from metagenomics data to date.

## Introduction

Sewage samples have proven useful for surveillance of antimicrobial resistance (AMR)[1,2] and infectious diseases, e.g. poliovirus, norovirus, and rotavirus[3,4]. Very recently, sewage samples have been used in the surveillance of the Covid-19 pandemic[5–7]. In supplement to such surveillance activities, understanding of the microbial community residing in sewage is important, because sewage has been suggested to comprise a reservoir of AMR and at the same time, provide an environment for potential genetic transfer between the bacteria in the community[8]. Several studies have examined the bacterial composition of sewage samples by 16S rRNA investigations[9–11] or mapping to reference databases[12]. However, such investigations are limited to species previously identified and furthermore, it can be difficult to distinguish closely related species from 16S rRNA analysis, thus, species living in sewage that are closely related to species from the human gut could be confused. Within the last decade, investigations of microbiomes in human hosts, soil, plants, and more have found that differences in bacterial communities correlates with geography[11,13–16]. Among bacterial species isolated from clinical infections such as *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Escherichia coli*, within-species diversity correlating with geography has been observed in multiple studies[17–19]. Some of these geographical differences could be a result of local environmental selection but may also be due to the effect of dispersal limitations on local prevalence. These findings challenge the long-standing, Baas Becking ecological hypothesis that *"Everything is everywhere,* but *the environment selects"*[20].

In the attempt to disentangle the effect of environmental selection and/or dispersal limitations, researchers have studied the within-species diversity not only from isolates, but also in metagenomics data. Correlations between diversity within species and geography have been identified in bacterial species, such as *Eubacterium rectale* and *Candidatus pelagibacter*, from marine metagenomes[21] and

human gut microbiomes[22]. Another study did not identify significant geographical differences within subspecies of for instance *Bacteroides vulgatus* and *Alistipes putredinis* from the human gut[23]. However, the subjects included in that study were limited to North Americans and Europeans. It has often been difficult to obtain comparable samples in a standardized way across large geographies. The Global Sewage dataset, containing samples from 101 different countries, serves as an ideal candidate for a broader investigation of regional within-species diversity. A phylogenetic analysis of 79 samples focusing only on reference mapping to known bacterial species has previously been performed[12] and found geographical clustering for environmental and human commensal bacteria.

In this study, we aim to investigate the microbial community of sewage by constructing metagenome assembled genomes (MAGs) and determine the within-species phylogeny on a global scale, using 757 sewage samples from 241 sites and a total of 101 different countries. With these phylogenies, we increase the depth of the analysis by comparing geographical clustering between different genes when stratified by the cellular localization of the encoded proteins. Both in terms of sample size and global reach this study is the most comprehensive investigation of within species diversity among sewage species to date. From our analysis, we identified 3,353 near complete (NC) MAGs from 1,439 different MAG species and found that variation within species correlated with geographical separation. Furthermore, we found that genes associated with organelles displayed on average 10% less geographical variation compared to other groups of genes, suggesting that the geographical clustering is primarily due to environmental selection. Thus, we confirm the fundamental microbial ecology doctrine that microbes are globally dispersed but selected by the environment.

## Results

### Predominant bacteria in sewage do likely not originate from the human gut

To identify bacterial genomes from sewage across the world, we used a combination of two different metagenomics genome binners (VAMB[24] and MetaBAT2[25]). From 757 samples across 101 different countries (**Figure 1a** and **Supplementary Figure 1a**), we were able to create 3,353 NC MAGs assigned to a total of 1,439 different MAG species. Of the MAGs we detected, 3,301 were annotated to bacteria and 52 to archaea. The taxonomic distribution of the identified MAGs comprised 37 phyla, 75 classes, 151 orders, 259 families, 419 genera, and 215 species. However, we could only annotate 699 MAGs (20.8%) at species level, leaving 2,654 unknowns. Likewise, there were unannotated MAGs at genus (29%), family (6%), and order (2%) level (see **Supplementary Data 1** for complete taxonomic annotations). The identified MAG species captured a wide range of taxonomy from the known microbial tree of life (**Figure 1b**) and geographical origin of the sample did not correlate with phylogeny ($R^2$=0.01) (**Figure 1c**). Thus, we do not expect that MAG reconstruction was confounded by sample origin.

As the sewage was collected from urban areas, we were interested in knowing how large a fraction of the MAGs that could be associated with the human gut microbiome. The 3,353 NC MAGs we identified were less than the number of NC MAGS (5,036) found from binning of 1,000 human faecal samples[24], however, these gut MAGs represented a lower number of different MAG species (645), suggesting that the binning of sewage metagenomes is more complicated than binning of human gut samples. Additionally, we found that only 1.2% of all the identified NC MAGs could be annotated to the human gut microbiome, similar to a previous study based on mapping of reads from a subset of the Global Sewage samples, where 3.7% of the reads were found to be associated with the human microbiome[2]. In contrast to this, other studies using 16S rRNA marker genes have found a higher

proportion (15% and 4.3-28.7%)[10,11]. The difference in these results could be due to the limitations of each of the different methods for bacterial identification. Even though mapping of reads is restricted to the contents of reference databases, this method has the advantage of needing less coverage for a read to map, compared to the high read coverage that is necessary for a successful *de novo* assembly, which is a prerequisite to genome binning. Due to this high coverage need, genome binning has limitations in detecting low-abundance species. Thus, there are advantages and disadvantages for all methods and genome binning can be used to detect prevalent, including novel, bacterial genomes.

To further investigate differences in the bacterial composition, we compared the overall ratio of phyla between the seven World Health Organization (WHO) regions and a pool of human gut samples (**Figure 2a and Supplementary Figure 1b**). The phyla compositions of the identified MAGs were similar between all sewage samples regardless of geographical origin. We found a high proportion of Firmicutes and Proteobacteria in all regions, which is coherent with 16S rRNA analyses of influent sewage in Hong Kong and USA[9,11]. These results also agree with the taxonomic annotation of metagenomics reads from influent sewage in Portugal[26]. However, the phylum ratio from 1,000 diverse human gut samples was significantly different from the sewage phylum ratio (two-sample Kolmogorov-Smirnov, $P < 1e\text{-}15$), again suggesting that the MAGs identified by binning were not primarily the fraction of bacteria in the sewage originating from the human gut. To investigate the bacterial composition of the sewage further, we used the abundance of the MAG species across all samples. To identify abundances, we mapped the reads from the sewage samples and a selection of human gut samples to a combined catalogue of the sewage and human MAG species. When comparing the abundances of different bacterial phyla, we again saw similar distributions across the different regions (**Figure 2b**). Additionally, these results suggested that the bacteria found in the

human samples were a smaller fraction of the ones found in sewage, and that the human samples were less diverse than the sewage samples. This could be due to the high intra-sample diversity, as a result of the many gut and other microbiomes mixed together in the sewage samples, which is reflected by the higher Shannon diversity index of the sewage samples (~4 for sewage samples vs. 2.5 for human gut samples, **Figure 2c**). Furthermore, when mapping reads to all the MAG species identified in either human gut or sewage samples, the MAG species captured around 70% of the diversity of the human samples, whereas only 41% of the sewage reads were mapped. Collectively, these results suggest that the binning of bacterial MAGs from the species-rich sewage is more complex compared to faecal samples, probably due to the increased alpha diversity and strain mixing in the sewage samples. Additionally, the most abundant species in the sewage samples were probably the species capable of surviving in this environment, which was not necessarily the ones originating from the human gut. This difference in species composition is supported by the lower number of genes involved in oxidative phosphorylation found in the most prevalent MAG species from human gut samples as compared to sewage samples (**Figure 2d**).

**Sewage bacteria vary according to geography**

To infer the phylogeny and identify geographical clustering within single species, we identified MAG species that were present across multiple samples. From the 1,439 different MAG species, only 41 contained MAGs from ten or more different samples. For each of these 41 species, we identified orthologous genes and found between 1,437 and 4,967 different orthologous genes for each species. We inferred a maximum likelihood tree for each gene and created an overall species tree for each MAG species using ASTRAL. For instance, Cluster 5, which we annotated as a species in the *Brahymonas* genus, contained between 1 and 12 MAGs from each of the geographical regions and

were based on 2,298 orthologous gene trees (**Figure 3a**). To evaluate geographical clustering, we performed a PERMANOVA test on the distance matrices from the ASTRAL trees with more than one genome from at least two regions, leading to a total of 33 tested species trees (**Supplementary Figure 2**). Out of these, 12 clustered significantly according to the regional origin of the samples the MAGs were identified in (**Figure 3b**). Geographical clustering was identified as the $R^2$ value from the PERMANOVA test, which describes how much of the variation in the data that can be explained from the regional origin of the samples. Other metagenomic studies of sewage and human gut samples have also shown that some species vary according to geographical dispersion, while others do not[12,22,24]. One study of the human gut microbiome found a higher number of species within the phylum Firmicutes, where variation was geographically separated[22]. However, we did not find any phyla enriched in geographical clustering (**Supplementary Figure 3**). On average 56% of the variation in the 12 significant trees could be explained by geography. This is much higher than the average of 19% variation in the significant trees found by a similar study of the human gut microbiome[24], suggesting that more of the genomic variation found in environmental bacteria correlates with geography than the variation found in gut microbes. In four additional trees (C26, C28, C34, and C38), we found a high degree of geographical clustering (>50%), but this clustering was not significant, potentially due to a lower number of MAGs in these trees. The geographical clustering suggests that either dispersal limitations or local selection of these species cause within-species variations from adjacent areas to be more similar.

**Genes associated with organelles contribute less to geographical variation**

To investigate whether the observed geographical clustering was the result of dispersal limitations or local selection, we examined whether the clustering varied between different groups of genes. We hypothesised that proteins on the surface of the bacteria would be subject to selective pressure from

the environment, because of their interaction with the surroundings, whereas the intracellular proteins, not directly interacting with the environment, would vary as a consequence of genetic drift. To test this hypothesis, we divided the orthologous genes into groups based on the GO term cellular component annotation and selected genes associated with organelles and membranes for comparison. The genes associated with organelles were mostly coding for proteins that were a part of the ribosome (85%) and to a lesser extend proteins bound to the chromosome (9%), acting as part of the flagellum (5%), or polyhedral organelles (1%). For many of the orthologous genes (up to 91%), it was not possible to annotate them to any cellular component GO term. To include these unannotated genes in the analysis, we grouped them with gene groups other than membrane and organelles. When investigating the biological process GO term annotation of this gene group, the largest fraction (on average 45%) of genes were annotated to be part of a metabolic pathway and we therefore considered this group to represent metabolic genes.

In soil bacteria[27] and throughout a wide variety of different prokaryotes[28] ribosomal genes have been found to have lower nucleotide diversity. In this study as well, the genes in the organelle group varied less than genes in the other two gene groups for most (94%) of the investigated species (**Supplementary Figure 4**). However, we tested whether different amounts of variation within the gene sequences affected the $R^2$ values and found that this was not the case (**Supplementary Table 1**). Thus, differences in $R^2$ values found between gene groups was not a result of a reduced variation of the organelle genes. In addition to this, we tested for the effect of different gene lengths, as well as the number of MAGs and the distribution of different regions in a tree. For one MAG species (C26) we found that the $R^2$ values correlated with the number of MAGs in the tree and regional entropy, therefore, this cluster was excluded from further analyses. For the remaining 32 MAG species, none of these factors correlated with the $R^2$ values, meaning that they were not relevant to include in subsequent statistical testing.

With a Kruskal-Wallis Rank-Sum test on $R^2$ values from the three gene groups (membrane, organelle, and metabolic), we found nine MAG species with significantly ($P < 0.05$) different geographical clustering between the groups after Benjamini Hochberg (BH) correction of p-values. One of the MAG species with significant differences were Cluster 5, *Brachymonas*, in which the trees of genes associated with organelles had significantly lower $R^2$ values, than trees of genes from the other two groups (**Figure 3c)**. The lower $R^2$ values in organelle gene trees were also found for the eight other significant MAG species (**Figure 3d and Supplementary Figure 5**)(two-sided Wilcoxon Rank Sum test, $P < 0.05$**, Supplementary Data 2**). Moreover, six of the nine significant MAG species were also significantly clustered according to geography in the ASTRAL species tree and the three remaining clusters (C34, C37, and C38) also showed a high degree of geographical clustering but were not significant. Likewise, we saw lower $R^2$ in organelle genes for the majority (16 of 23) of the remaining MAG species without significant difference. To support these findings, we calculated the dN/dS ratio for the nine MAG species with significant differences between organelle and membrane $R^2$ values. For all the nine MAG species, the dN/dS ratios were larger for membrane genes than organelle genes, suggesting more positive selection of the membrane genes (**Supplementary Figure 6)**. Furthermore, this difference was significant for seven MAG species (C3, C5, C13, C14, C34, C37, and C38).

Collectively, our results show similar levels of geographical clustering in the membrane genes and metabolic genes, whereas the genes associated with the organelles displayed significantly less geographical clustering. We expect that the organelle genes would have followed a similar evolutionary trajectory as the metabolic and surface genes if the clustering was a result of dispersal limitations. Our results thus suggest that the geographical clustering is primarily due to regional selection and not dispersal limitations.

## Discussion

Here, we present a comprehensive investigation of how sewage bacteria are dispersed globally. Our results showed that the phylum composition of the bacteria identified by metagenomic binning in untreated sewage, from the inlet to the wastewater treatment plants, were similar between all regions of the world. However, we also found that in general these bacteria showed a degree of geographical clustering for the within-species diversity. Interestingly, the genomic variation in organelle genes showed less regional clustering than genes involved in metabolic and membrane functions, suggesting that the clustering observed is primarily due to selection rather than dispersal limitations. Thus, the bacteria residing in sewage can spread globally, but are under evolutionary pressure to adjust to the different environments across the world. This selection pressure combined with the co-existence of multiple bacterial species and the presence of antimicrobial resistance genes (ARGs) and antimicrobial drugs in sewage create a high probability for transferral of ARGs between bacteria[8]. To prevent global transmission of these genes, it is important to better understand how sewage bacteria are globally disseminated.

The WHO has a goal of delaying the dissemination and emergence of AMR through monitoring[29]. We have previously suggested that sewage sampling is a desirable strategy for such surveillance activities[30]. Metagenomics binning can be used to identify novel, bacterial genomes that are not present in reference databases. Here, we found that binning of shotgun sequences could identify a fraction of the bacteria residing in the sewage samples and that the bacteria originating from the human gut were a small subgroup of the microbiome found in these samples. This is a reminder that when using sewage samples for surveillance activities, the detection method should be selected carefully. If monitoring human-derived pathogens in sewage, one needs to consider that DNA from these organisms is a very small fraction of the total pool of microbial DNA. Mapping to reference genomes can be useful for the monitoring of the global spread or the local levels of a particular

species, like we have seen for the covid-19 pandemic[7]. However, metagenomic assembly and binning could be used to identify potential candidates for surveillance and possibly to clarify the genomic context in which ARGs are emerged and disseminated. Thus, there is more potential for new and important discoveries using metagenomics binning of sewage samples.

In conclusion, this study shows a clear geographical phylogenetic clustering of bacterial species and underpin the importance of including samples from the entire world if global conclusions must be made.

## Methods

### Global sewage dataset

Samples were collected and handled as part of the Global Sewage project[2,31]. In brief, untreated sewage samples were collected before the inlet to the wastewater treatment plant at sample sites. DNA was extracted and fragmented from the untreated sewage and libraries were sequenced using Illumina paired-end sequencing. 757 samples from 241 sites spanning 101 different countries across the world were included in this project. A complete list of samples included can be found in **Supplementary Data 3**.

### Genome binning

Quality trimming of sequencing reads was performed as described in Hendriksen *et al.* 2019[2]. We assembled forward, reverse, and singleton reads with metaSpades (v3.13)[32] using kmer sizes between 27-127 bp with an interval of 20 bp. Scaffolds above 1,000 bp were saved for further analyses. For binning with MetaBAT2 (v2.10.2)[25], we filtered contigs to a minimum size of 1,500 bp and performed single-sample binning with MetaBAT2 using default settings. For binning with VAMB (v3.0.1)[24], we combined contigs >2,000 bp from all samples into one catalogue. From a pilot run with VAMB[24] binning, we found no increase in the number of NC MAGs when using contigs >1,500bp, therefore we chose a minimum contig size of 2,000bp to reduce mapping time. We mapped reads from each sample to the contig catalogue using Minimap2 (v2.6)[33]. Afterwards, jgi_summarize_bam_contig_depths from MetaBAT (v2.10.2)[25] was used to calculate abundances of contigs in each sample. The output abundances were combined into a matrix, normalized using vambtools[24], and used as input to VAMB. We calculated and normalized Tetra Nucleotide Frequencies (TNFs) using vambtools as well. From the contig catalogue, we obtained contig names and lengths and used them as input to VAMB along with the normalized TNFs. We ran VAMB using

the memory mapping mode available at the github repository. VAMB was run using a GPU with a mini-batch size of 256 and a network of 48 latent and 1,024 hidden neurons.

We assessed the quality of all MAGs with a size above 1 Mbp using CheckM (v1.1.3) lineage_wf[34]. We defined the quality of MAGs as in Almeida *et al*. 2019[35] where NC MAGs were defined as >0.9 completeness and <0.05 contamination. We used NC MAGs in our further analysis. To group similar MAGs (likely to be different variations of the same bacterial species) into clusters, we used dRep compare (v2.2.3)[36] with a mash threshold of 90% and an Average Nucleotide Identity (ANI) threshold of 95%. dRep dereplication was run with the same thresholds and the resulting score was used to select between MAGs from the same sample within one cluster, to avoid redundancy if a bin was identified both by MetaBAT and VAMB. Additionally, the dereplication results were used to select the best MAG species representative for each cluster of MAGs. We assessed the taxonomy of all NC MAGs with GTDB-Tk classify_wf (v0.3.2)[37].

For comparison of the taxonomic distribution of sewage MAGs to human gut species taxonomy, the MAGs created in Nissen *et al* 2021[24] were used. Difference in phylum distributions between human and sewage samples were tested for significance with a two-sample Kolmogorov-Smirnov test in R. Additionally, these human gut MAGs were dereplicated like the MAGs from the sewage data and used together with reads from 15 randomly selected human samples (no infant or diseased hosts) from the dataset from Almeida *et al.* 2019[35] for abundance comparisons. The abundances of MAG species were obtained using CoverM (v 0.6.1)[38], by mapping to the best representative MAG species genomes and filtering based on the expected/covered ratio of a genome (>=0.5). The expected coverage of a genome was calculated as described in Rasmussen *et al.* 2015[39]. The fraction of sewage MAGs likely to originate from the human gut microbiome was found by using MASH (v2.0)[40] to map to the Unified Human Gastrointestinal Genome (UHGG) catalogue[41] and identified as within a mash distance of 0.05 to any genomes in this catalogue.

## Phylogeny

We reconstructed phylogenetic trees containing all the identified MAGs using the marker gene set from GTDB-Tk. One tree was created including the GTDB-Tk reference species and another without these. Both trees were inferred with FastTree (v2.1.11)[42] and rooted on a *Thermotogae*, because this is the bacterial phylum most closely related to Archaea[43]. We visualized the trees using iTol (v1.0)[44]. For the 41 dRep clusters spanning ten or more samples, a separate tree of the MAGs belonging to each cluster was inferred. For this, we used Prodigal (v2.6.3)[45] protein predictions from GTDB-Tk as input to Sonicparanoid (v1.3.4)[46], to identify orthologous genes, using the fast mode. To align DNA sequences of all identified orthologous, we used MAFFT (v7.453)[47]. Samples were excluded from the alignments if they had more than one copy of an orthologous gene, to avoid uncertainty of which gene was used to infer phylogeny. We used TrimAl (v1.4)[48] to convert alignments to phylip format, prior to building a separate phylogenetic tree for each gene using IQ-TREE (v1.6.8)[49] with automatic model selection. Trees were created if a gene was observed in at least three samples, which is the lowest possible number of samples that a tree can be inferred from with IQ-TREE. To infer the overall species tree phylogeny, we used all the gene trees from a specific MAG species as input to ASTRAL (v5.7.4)[50]. In this tree, IQ-TREE was used to correct branch lengths with the ASTRAL tree as constrained tree input. We used the ggtree (v2.0.4) package in R[51] for visualization of species trees.

## Functional annotation

To assign functional annotation to the genes, we used the Prodigal protein predictions from GTDB-Tk as input to InterProScan (v5.36-75.0)[52]. From the InterProScan output, we then extracted the GO-term annotation and used the GO.db-package (v2.1)[53] in R to get the annotations within the Cellular Component category. We grouped the genes into the top-level annotations within this category, to

get overall groupings for comparisons between gene groups. We selected the groups membrane and organelle for further analysis, and the remaining genes were combined into one collapsed group. The genes involved in oxygen tolerance were likewise identified from the InterProScan output, by identifying the genes involved in the KEGG pathway map00190, oxidative phosphorylation.

**dN/dS calculation**

Codeml (paml (v.4.9j)[54]) was used for genewise dN/dS calculation. Genes with genetic variation between samples were identified with the snppos_analyzer from CSI phylogeny (v1.4)[55] and only these genes were input to codeml. Furthermore, to make sure that gene alignments were in frame, only alignments starting with a start codon (ATG, TTG, or GTG) were included. The phylip format gene alignments from the phylogeny reconstruction were converted to fasta files using TrimAl (v1.4)[48] and stop codons were removed prior to dN/dS calculation. Along with the alignments, the gene tree files from IQ-TREE (v1.6.8)[49] were used as input to codeml paml (v.4.9j)[54]. One dN/dS ratio per gene was calculated with codeml by setting the model option to 0 and the seqtype option to codons. Additionally, optimization was performed one branch at the time (method: 1) and ambiguous sites were removed from the calculation (cleandata: 1), otherwise default settings were used.

**Statistical testing**

To determine the amount of geographical variation for both GTDB-TK tree, species trees, and gene trees, we used the adonis2 function from the vegan package in R (v2.5-6)[56] to perform a Permutational multivariate analysis of variance (PERMANOVA) according to geography. Prior to this testing, multiple MAGs from the same city within one tree were limited to one representative MAG based on the dRep score. In addition to this, a MAG was removed from the tree if it was the only representative of a region in this tree. To adjust for multiple testing, we corrected the p-values from these tests using

348 Benjamini & Hochberg[57]. For some short genes with low variance, the $R^2$-values outputted from the
349 PERMANOVA test were negative (**Supplementary Figure 7**), these were excluded from the
350 analysis. It is possible to get a negative R2 value when the fitted model is worse than a horizontal
351 line.
352 To identify species with any significant differences in $R^2$- or dN/dS-values between gene groups, we
353 grouped the values of the gene trees according to the Cellular Component annotations and used a
354 Kruskal-Wallis test on the values from the different groups. Afterwards, we applied a Wilcoxon Rank
355 Sum Test on the groups from MAG species displaying significance ($P < 0.05$) from the Kruskal-
356 Wallis test, to identify which of the three groups that were differing from each other.
357 To support the comparisons of $R^2$ values between gene groups, we investigated if different gene
358 qualities could bias the $R^2$ value. This was done by calculating the Pearson Correlation Coefficient
359 (PCC) for $R^2$ values according to the number of samples in the tree, gene variation, gene length, and
360 regional entropy. Number of MAGs was counted in a tree after removing duplicate city and single
361 region samples, as it was done prior to the PERMANOVA. Gene variation was obtained as the mean
362 fraction of varying sites across all pairwise sequences (mean pi). Gene lengths were identified as the
363 number of positions in the fasta output from Sonicparanoid. Regional entropy was calculated as:

364 $$-\sum_{i=1}^{n} \ln\left(p_i{}^{p_i}\right)$$

365 where $n$ was the total number of regions in the tree, and $p_i$ was the proportion of samples belonging
366 to a specific region. We performed these calculations on all gene trees tested with PERMANOVA.
367

## References


1. Karkman, A., Berglund, F., Flach, C.-F., Kristiansson, E. & Larsson, D. G. J. Predicting clinical resistance prevalence using sewage metagenomic data. *Commun Biol* **3**, 711 (2020).
2. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (03 08, 2019).
3. Deshpande, J. M., Shetty, S. J. & Siddiqui, Z. A. Environmental surveillance system to track wild poliovirus transmission. *Appl. Environ. Microbiol.* **69**, 2919–2927 (2003).
4. Santiso-Bellón, C. *et al.* Epidemiological Surveillance of Norovirus and Rotavirus in Sewage (2016–2017) in Valencia (Spain). *Microorganisms* **8**, 458 (2020).
5. Randazzo, W., Cuevas-Ferrando, E., Sanjuán, R., Domingo-Calap, P. & Sánchez, G. Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. *Int. J. Hyg. Environ. Health* **230**, 113621 (2020).
6. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environ. Sci. Technol. Lett.* **7**, 511–516 (2020).
7. Izquierdo-Lara, R. *et al.* Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg. Infect. Dis.* **27**, 1405–1415 (2021).
8. Fouz, N. *et al.* The Contribution of Wastewater to the Transmission of Antimicrobial Resistance in the Environment: Implications of Mass Gathering Settings. *Trop Med Infect Dis* **5**, (2020).
9. Cai, L., Ju, F. & Zhang, T. Tracking human sewage microbiome in a municipal wastewater treatment plant. *Appl. Microbiol. Biotechnol.* **98**, 3317–3326 (2014).

10. Newton, R. J. *et al.* Sewage reflects the microbiomes of human populations. *MBio* **6**, e02574 (2015).
11. Shanks, O. C. *et al.* Comparison of the microbial community structures of untreated wastewaters from different geographic locales. *Appl. Environ. Microbiol.* **79**, 2906–2913 (2013).
12. Ahrenfeldt, J. *et al.* Metaphylogenetic analysis of global sewage reveals that bacterial strains associated with human disease show less degree of geographic clustering. *Sci. Rep.* **10**, 3033 (2020).
13. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
14. Griffiths, S. M. *et al.* Host genetics and geography influence microbiome composition in the sponge Ircinia campana. *J. Anim. Ecol.* **88**, 1684–1695 (2019).
15. Coller, E. *et al.* Microbiome of vineyard soils is shaped by geography and management. *Microbiome* **7**, 140 (2019).
16. Greenlon, A. *et al.* Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15200–15209 (2019).
17. Nicolas-Chanoine, M.-H. *et al.* Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *J. Antimicrob. Chemother.* **61**, 273–281 (2008).
18. Rasigade, J.-P. *et al.* Global distribution and evolution of Panton-Valentine leukocidin-positive methicillin-susceptible Staphylococcus aureus, 1981-2007. *J. Infect. Dis.* **201**, 1589–1597 (2010).
19. Gladstone, R. A. *et al.* Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates. *Microb*

*Genom* **6**, (2020).

20. O'Malley, M. A. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **39**, 314–325 (9/2008).
21. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* vol. 26 1612–1625 (2016).
22. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
23. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
24. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00777-4.
25. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
26. Lira, F., Vaz-Moreira, I., Tamames, J., Manaia, C. M. & Martínez, J. L. Metagenomic analysis of an urban resistome before and after wastewater treatment. *Sci. Rep.* **10**, 8174 (2020).
27. Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K. & Banfield, J. F. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *The ISME Journal* vol. 14 1834–1846 (2020).
28. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).
29. World Health Organization. *Monitoring and evaluation of the global action plan on*

*antimicrobial resistance*. (2019).

30. Aarestrup, F. M. & Woolhouse, M. E. J. Using sewage for surveillance of antimicrobial resistance. *Science* **367**, 630–632 (2020).
31. Patrick Munk, Christian Brinch, Frederik Duus Møller, Thomas N. Petersen, Rene S. Hendriksen, Anne Mette Seyfarth, Jette S. Kjeldgaard, Christina Aaby Svendsen, Bram van Bunnik, Fanny Berglund, [Global Sewage Surveillance Consortium], D. G. Joakim Larsson, Marion Koopmans, Mark Woolhouse, Frank M. Aarestrup. Global sewage metagenomics provides unparalleled insight into spatial, taxonomic, and genomic evolution of antimicrobial resistance.
32. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (05 2017).
33. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
34. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
35. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (04 2019).
36. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: A tool for fast and accurate genome de-replication that enables tracking of microbial genotypes and improved genome recovery from metagenomes. *Cold Spring Harbor Laboratory* 108142 (2017) doi:10.1101/108142.
37. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2019)

464 doi:10.1093/bioinformatics/btz848.
465 38. Woodcroft, B. CoverM. https://github.com/wwood/CoverM.
466 39. Rasmussen, S. *et al.* Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell*
467 **163**, 571–582 (2015).
468 40. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash.
469 *Genome Biol.* **17**, 132 (2016).
470 41. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
471 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
472 42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood
473 Trees for Large Alignments. *PLoS ONE* vol. 5 e9490 (2010).
474 43. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between
475 domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
476 44. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new
477 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
478 45. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
479 identification. *BMC Bioinformatics* **11**, 119 (2010).
480 46. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference.
481 *Bioinformatics* **35**, 149–151 (2019).
482 47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
483 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
484 48. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
485 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
486 (2009).
487 49. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective

488 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**,
489 268–274 (2015).
490 50. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree
491 reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
492 51. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Ggtree : An r package for visualization
493 and annotation of phylogenetic trees with their covariates and other associated data. *Methods*
494 *Ecol. Evol.* **8**, 28–36 (2017).
495 52. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*
496 **30**, 1236–1240 (2014).
497 53. Carlson, M., Falcon, S., Pages, H. & Li, N. GO. db: A set of annotation maps describing the
498 entire Gene Ontology. *R package version* **3**, 10–18129 (2017).
499 54. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–
500 1591 (2007).
501 55. Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the problem of
502 comparing whole bacterial genomes across different sequencing platforms. *PLoS One* **9**,
503 e104984 (2014).
504 56. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan
505 McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H.
506 Stevens, Eduard Szoecs, Helene Wagner. vegan: Community Ecology Package. R package
507 version 2.5-6. (2019).
508 57. Hochberg, B. Y. A. Controlling the False Discovery Rate: A Practical and Powerful Approach
509 to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, (1995).
510

**Acknowledgements**

We thank Anders Gorm Pedersen for a fruitful discussion on the phylogenetic analysis. This work was supported by The Novo Nordisk Foundation (NNF16OC0021856: Global Surveillance of Antimicrobial Resistance). S.R., J.J., and H.W. was supported by the Novo Nordisk Foundation (grant NNF14CC0001).

**Author contributions**

F.M.A. and S.R. conceived the idea and guided the analyses. M.L.J performed the analysis. P.M. performed the metagenomics binning with MetaBAT2. P.M., J.J., H.V., H.W., R.S.K., H.B.N., S.R., and F.M.A. provided guidance for the analysis and input for interpretation of results. M.L.J., S.R., and F.M.A. drafted the paper with contributions from all co-authors. All authors have read and accepted the final version of the manuscript.

**Data availability**

The raw reads are available in the European Nucleotide Archive (ENA) under the accession numbers: PRJEB40798, PRJEB40816, PRJEB40815, PRJEB27621, and ERP015409.

**Code availability**

The code used in this paper is available on GitHub at https://github.com/marieljespersen/Sewage_MAG_phylogeny

**Competing interests**

H.B.N. is employed at Clinical-Microbiomics A/S. The additional authors declare no competing interests.

## Figures

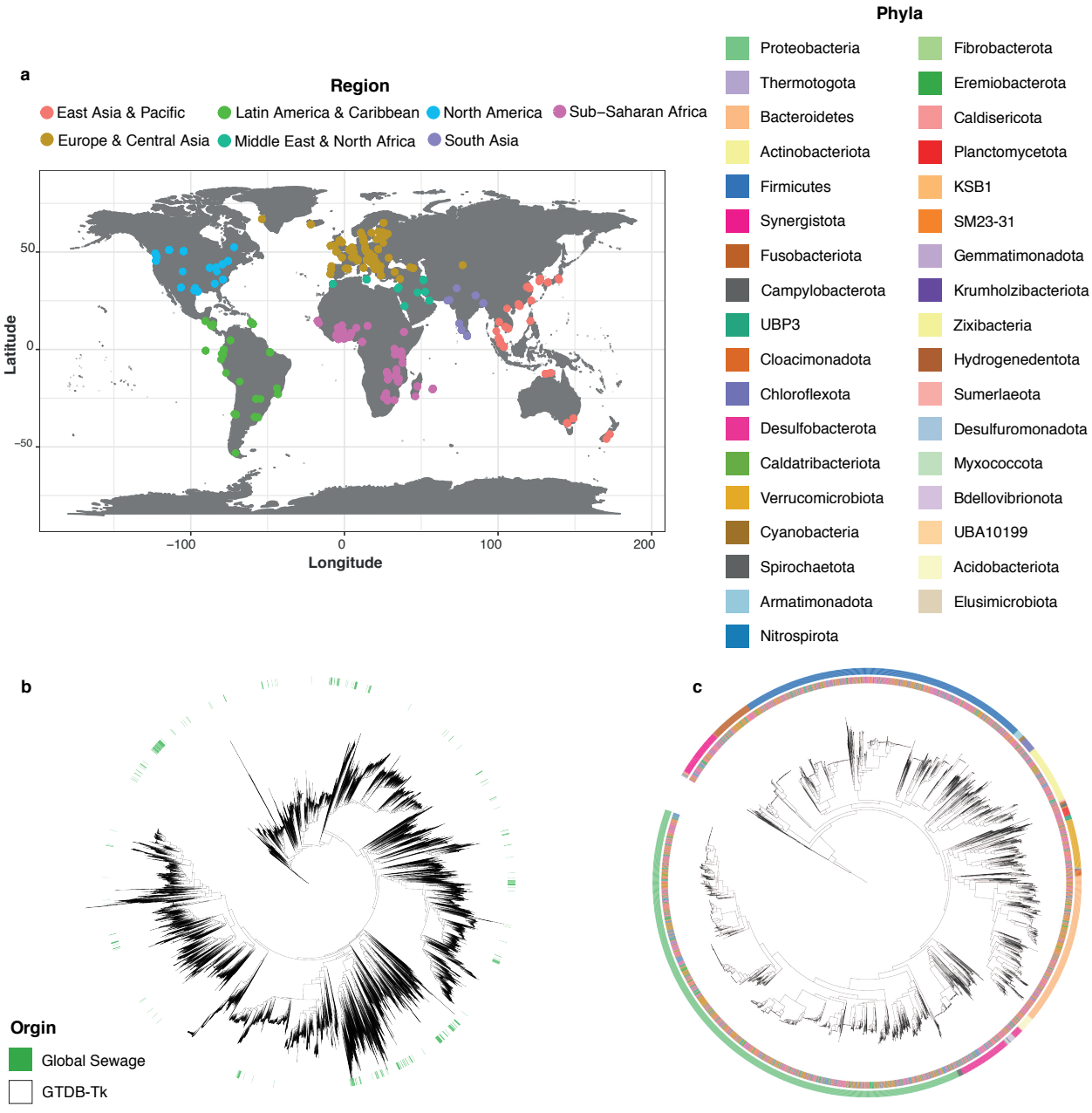
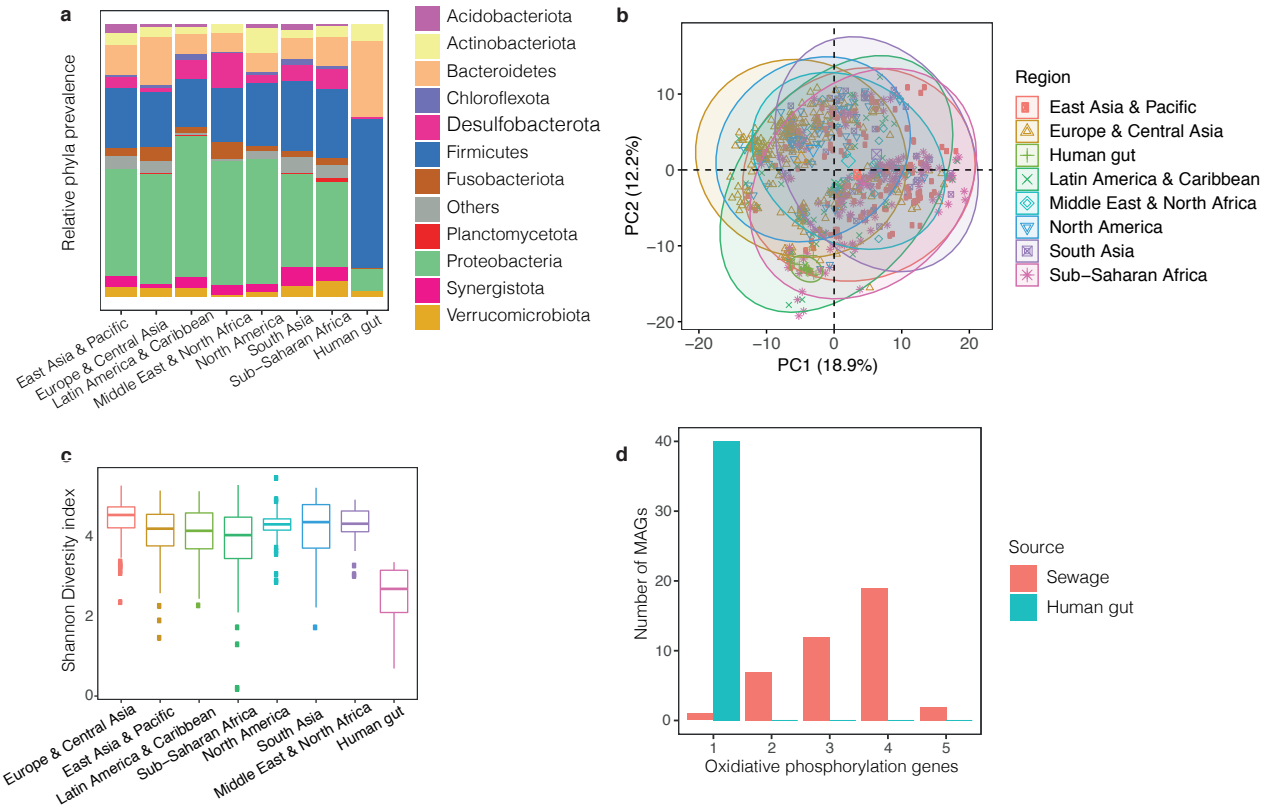



**Figure 1. Distribution of samples and MAGS. a)** World map of sampling sites. 757 sewage samples were collected from 241 different sampling sites spanning 101 different countries. Sampling sites are highlighted and coloured according to the regional grouping from the World Health Organization (WHO). Sampling times can be found in **Supplementary Figure 1a. b)** Maximum likelihood (ML) tree of marker gene, amino acid alignments for all bacterial MAGs and bacterial genomes included

542 in GTDB-Tk. The MAGs identified from sewage are scattered throughout the tree of known bacterial
543 species. **c)** ML tree of marker gene, amino acid alignments for all bacterial MAGs identified in this
544 study. The identified MAGs are clustered according to phyla rather than geographical origin. Inner
545 band is showing the geographical origin of samples according to WHO region, colours follow legend
546 in b. Outer band is showing the phyla of MAGs, coloured according to the legend on top.
547




548

549 **Figure 2. Bacterial composition of sewage. a)** Relative frequency of the 11 most prevalent phyla
550 between the different regions. The ratio plotted is the taxonomy of the combined pool of MAGs from
551 all samples of a certain region. In this plot, only the 11 phyla found in more than ten samples across
552 regions were plotted, while the remaining MAGs were grouped into one combined category (others).
553 A plot of all phyla identified can be found in **Supplementary Figure 1b.** The phyla ratio between
554 sewage samples of all regions were similar, but the phyla distribution in the human gut samples were
555 different from these. **b)** PCA plot of clr transformed read count abundance of the combined data set
556 of 2,084 MAG species obtained in this work and in Nissen *et al.* 2021[24]. The bacterial abundances in
557 sewage samples across all regions were similar, however, the bacteria in the human gut samples
558 consists of only a small fraction of the variety of bacteria found in sewage. **c)** Shannon Diversity
559 index comparison between sewage samples from different regions and human gut samples. The
560 Shannon diversity index was calculated from Transcripts Per Kilobase Million (TPM) using the vegan

561 package in R[56]. The alpha diversity is similar between sewage samples from all regions, but the
562 diversity in the human gut samples is lower. **d)** Frequency of genes involved in the oxidative
563 phosphorylation pathway. These genes were identified from the best representative genome of the 40
564 most prevalent MAG species from the Global Sewage data set and from the MAG species obtained
565 in Nissen *et al.* 2021[24]. Colours according to sample origin of the MAG species genome. Only one
566 gene involved in oxidative phosphorylation is found in all MAG species from sewage, whereas up to
567 five genes from this pathway is found in the human gut MAG species.
568

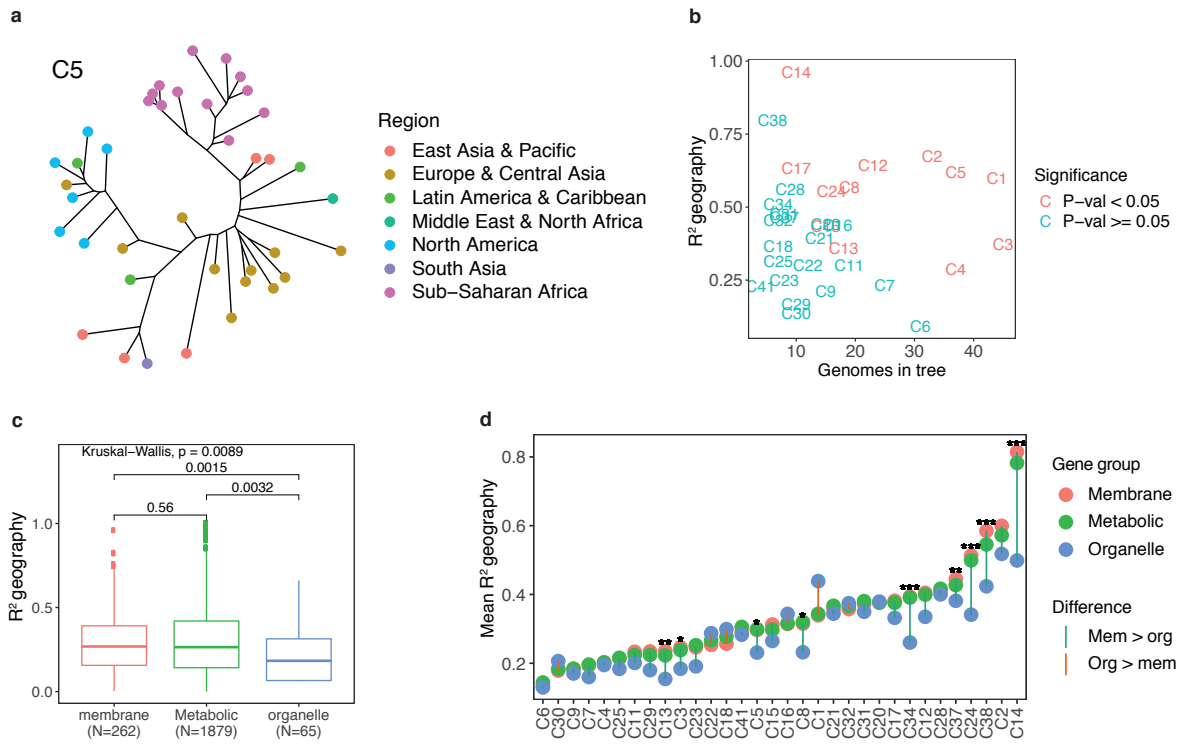



**Figure 3. Geographical clustering. a)** ASTRAL species tree of *Brachymonas*, Cluster 5. This tree was created from 2,298 orthologous gene trees. The tips in the tree are coloured according to WHO regions. PERMANOVA testing of this tree gave an $R^2$ of 0.62 and a p-value < 0.001. **b)** Geographical clustering of 33 MAG species. Results from PERMANOVA testing on ASTRAL species trees; the $R^2$ values are plotted against the number of MAGs in the tested tree. MAGs that were the only representative of a region in a tree were excluded from the tree prior to PERMANOVA testing. Additionally, trees could only be tested if MAGs from at least two different regions were present in the tree. Points were coloured according to significance level of the PERMANOVA p-value after Benjamini-Hochberg (BH) correction. In 12 ASTRAL species trees the MAGs were significantly clustered according to regional origin. **c)** Geographical clustering in different cellular component groups of C5. The geographical R2 values of *Brachymonas* at the y axis, Cluster 5 divided by the different gene groups (membrane, metabolic, and organelle) at the x axis. We tested for significant

differences between the groups with a Kruskal-Wallis Rank Sum Test (P=0.032), and afterwards for significance levels between gene groups with a Pairwise Wilcoxon Rank Sum test (P=0.008). P-values from both tests were BH corrected. **d)** Geographical clustering in different cellular component groups of all tested MAG species. Mean $R^2$ values from PERMANOVA testing on gene trees in each cellular component group plotted for all 32 tested MAG species. Differences in geographical $R^2$ values between groups, where tested like described previously with a Kruskal-Wallis Rank Sum Test and subsequent Pairwise Wilcoxon Rank Sum test. The significance level highlighted is of the p-value between genes from organelles and membrane from the Wilcoxon test adjusted by BH correction. We found the $R^2$ values of the organelle genes to be lower than the ones from the membrane genes in 25 MAG species, of which nine were significantly lower. *: P<0.05, **: P<0.01, ***: P<0.001.

# Supplementary Files

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

- SupplementaryMaterial.pdf
- SData1.xlsx
- SData2.xlsx
- SData3.xlsx