


OPEN

Metaphylogenetic analysis of global sewage reveals that bacterial strains associated with human disease show less degree of geographic clustering

Johanne Ahrenfeldt, Madina Waisi, Isabella C. Loft, Philip T. L. C. Clausen, Rosa Allesøe, Judit Szarvas, Rene S. Hendriksen, Frank M. Aarestrup & Ole Lund *

Knowledge about the difference in the global distribution of pathogens and non-pathogens is limited. Here, we investigate it using a multi-sample metagenomics phylogeny approach based on short-read metagenomic sequencing of sewage from 79 sites around the world. For each metagenomic sample, bacterial template genomes were identified in a non-redundant database of whole genome sequences. Reads were mapped to the templates identified in each sample. Phylogenetic trees were constructed for each template identified in multiple samples. The countries from which the samples were taken were grouped according to different definitions of world regions. For each tree, the tendency for regional clustering was determined. Phylogenetic trees representing 95 unique bacterial templates were created covering 4 to 71 samples. Varying degrees of regional clustering could be observed. The clustering was most pronounced for environmental bacterial species and human commensals, and less for colonizing opportunistic pathogens, opportunistic pathogens and pathogens. No pattern of significant difference in clustering between any of the organism classifications and country groupings according to income were observed. Our study suggests that while the same bacterial species might be found globally, there is a geographical regional selection or barrier to spread for individual clones of environmental and human commensal bacteria, whereas this is to a lesser degree the case for strains and clones of human pathogens and opportunistic pathogens.

One of the basic dogma in microbiology has for almost a century been that we for microorganisms consider that “everything is everywhere but the environment selects”^{1,2}.

A large number of papers about the global transmission events of bacterial clones have been published, including descriptions of emergence and spread of specific clones of *Vibrio cholera*, MRSA, *Escherichia coli*, *Clostridium difficile*^{3–6} and many other bacterial pathogens. The main focus has been on pathogenic clones and virtually nothing is known about the global phylogeny of commensal species and clones.

The gut microbiota has so far mainly been studied in relation to diet, use of medication and diseases^{7,8}, mostly within countries⁹ and in some studies between countries^{10,11}. These studies have looked at the species or genera composition of the microbiota and the interaction between species, while virtually no details on within species phylogeny have been investigated.

The same has been the case for environmental bacteria; there are numerous projects, which have sequenced the metagenome of different niches¹², but not with much focus on the importance of geographical locations or within species phylogeny.

Almost all studies into the within species phylogeny of bacterial species have been conducted using whole genome sequencing of single cultivated isolates. A recent metagenomic study where DNA was isolated both directly from faeces and from isolates cultured from the faeces, demonstrated that most pairs of isolates and metagenomic samples were adjacent to each other in a phylogenetic tree¹³.

DTU Food, Technical University of Denmark, Kongens Lyngby, Denmark, 2800, Denmark. *email: olund@food.dtu.dk

Obtaining biological samples from global sources can be logistically difficult and further complicated by ethical constraints if the samples originate from humans. Recently, we have collected untreated human urban sewage and then conducted metagenomic sequencing as a proof of concept in establishing a global monitoring of antimicrobial resistance. Untreated, as described in the original study, means that it has not run through a sewage treatment plant, but the samples were taken from the inlet into the plants or in countries/regions where sewage is not treated, it was taken from where the sewage runs into the environment. 79 locations in 60 countries were sampled (see Supplementary Table S1 for a complete list of countries and regions)¹⁴.

The advantage of metagenomics is the feasibility to detect all genes related to all living organisms present in the sample analysed. Thus, the metagenomics approach has the potential to determine phylogenies among all bacterial species present in urban sewage and enable the study of real-time occurrences and transmissions across all bacterial species to detect changes attributed to climate, trade and travel.

Here, we present a bioinformatic pipeline able to construct meta-phylogenetic trees based on multi-metagenomic samples. The reference genomes representing the bacteria in urban sewage samples were determined and the reads were mapped to the identified bacterial reference genomes. The genetic single nucleotide polymorphism (SNP) distance between each sample containing genetic material similar to a given reference sequence was calculated and trees inferred from the distance matrices. It was investigated if different groups of bacteria showed different clustering patterns.

Results

Generation of bacterial reference genome database. A database of bacterial species was generated from NCBI (see methods), in which redundancies, very similar bacterial reference genomes, were removed by homology reduction using the Hobohm 1 algorithm¹⁵. Out of 6,510 complete bacterial genomes downloaded from NCBI, 3,721 were left after the homology reduction. The taxonomic composition of the database includes 34 phyla, 64 orders, 138 classes, 301 families, 843 genera, 2,319 species and 3,721 unique bacterial templates. The most abundant genera are *Burkholderia*, *Pseudomonas*, *Bacillus* and *Vibrio*. 467 of the genera contained only a single unique bacterial template. Out of the 2,319 species the most abundant were *Helicobacter pylori*, *Salmonella enterica* and *Escherichia coli* (see Supplementary Table S2).

Bacterial template identification. A total of 15,513 hits to bacterial templates were identified in the 79 sewage samples (step A. All steps shown in the methods section); these belonged to 996 unique bacterial reference templates. In the pre-processing (step B) all hits with a depth below 1 were discarded, leaving 11,691, belonging to 834 unique bacterial templates. Mapping was done on these 11,691 bacterial templates to create 11,691 consensus sequences (step C). After the distance calculation pre-processing (step D), where only consensus sequences with less than 40% unknown bases were kept, there was 1,504 left. These belonged to 115 unique bacterial templates. The distance calculation (step E) was run for each of these 115 unique bacterial templates, resulting in 115 distance matrices. The phylogeny was inferred (step F) on all the distance matrices where the bacterial template was observed in more than three samples, resulting in a total of 95 phylogenetic trees. An overview of the pipeline results after each step can be seen in Supplementary Fig. S2. Supplementary Table S3 provides an overview of the taxonomic composition of the remaining bacterial templates after the various steps in the pipeline.

The fraction of bacterial reference template hits for each genus in each region is shown in Fig. 1. A total of 279 different genera were identified with a depth above one and have been plotted and sorted according to the total abundance. Fifty-five genera as well as an unclassified group are all present in all regions, and 64 genera were only identified in one region. The most abundant genera and present in all regions are *Acinetobacter*, *Pseudomonas*, *Streptococcus*, *Acidovorax*, *Enterobacter*, *Bifidobacterium*, *Escherichia*, *Klebsiella*, and *Lactobacillus*.

To provide an overview of the species that are the most common in the urban sewage samples, a count matrix with the presence/absence of each species in each sample was created. The sample counts were aggregated, and the 80 most frequently occurring species are shown in a heatmap (Fig. 2). The urban sewage samples had a high occurrence of the species *Escherichia coli*, *Streptococcus suis* and *Pseudomonas fluorescens*.

Phylogenetic trees. Of the original 996 unique bacterial templates identified in the pre-processing step, B, the final output of the MetaPhylogeny pipeline, applied to the 79 urban sewage samples, was 95 phylogenetic trees. (Supplementary Dataset 1). In Fig. 3 the presence/absence of the samples in each tree is shown together with the total number of taxa in each tree and the total number of occurrences per sample.

Distance-based clustering. The distance matrices for each template were tested for clustering by comparing distances within country-groupings (intra-regional distances) to distances between country-groupings (inter-regional distances). This was conducted for all distance matrices containing more than one region and at least four samples by a modified Welch *t*-test, providing a *p*-value and an intra/inter-regional distance ratio. The countries were grouped according to World Bank regions (WB-R), World Bank income level (WB-IL), WHO regions (WHO-R) and WHO health impact (WHO-HI).

The ratios for each organism were grouped afterwards according to organism classification, annotated by the two schemes EID2 plus (EID2p) and Five class classification (5CC) (Fig. 4). The lower the ratio, the higher degree an organism will cluster according to the specified country grouping. For each regional grouping, the difference in clustering according to the intra/inter-regional distance ratio was tested using Wilcoxon rank sum test in R¹⁶ comparing all organism classifications. The results for EID2p are shown in Supplementary Table S4, and for 5CC in Supplementary Table S5. The significant differences can be found as stars above the boxplot in Fig. 4.

Figure 4 shows that when using the EID2p classification of the bacteria into commensal, environmental and pathogen, the distribution of ratios in both the environmental and commensal bacteria are more clustered than the pathogens, when using the WB-R (a) and WHO-R (c) grouping of countries but not when using the WB-IL

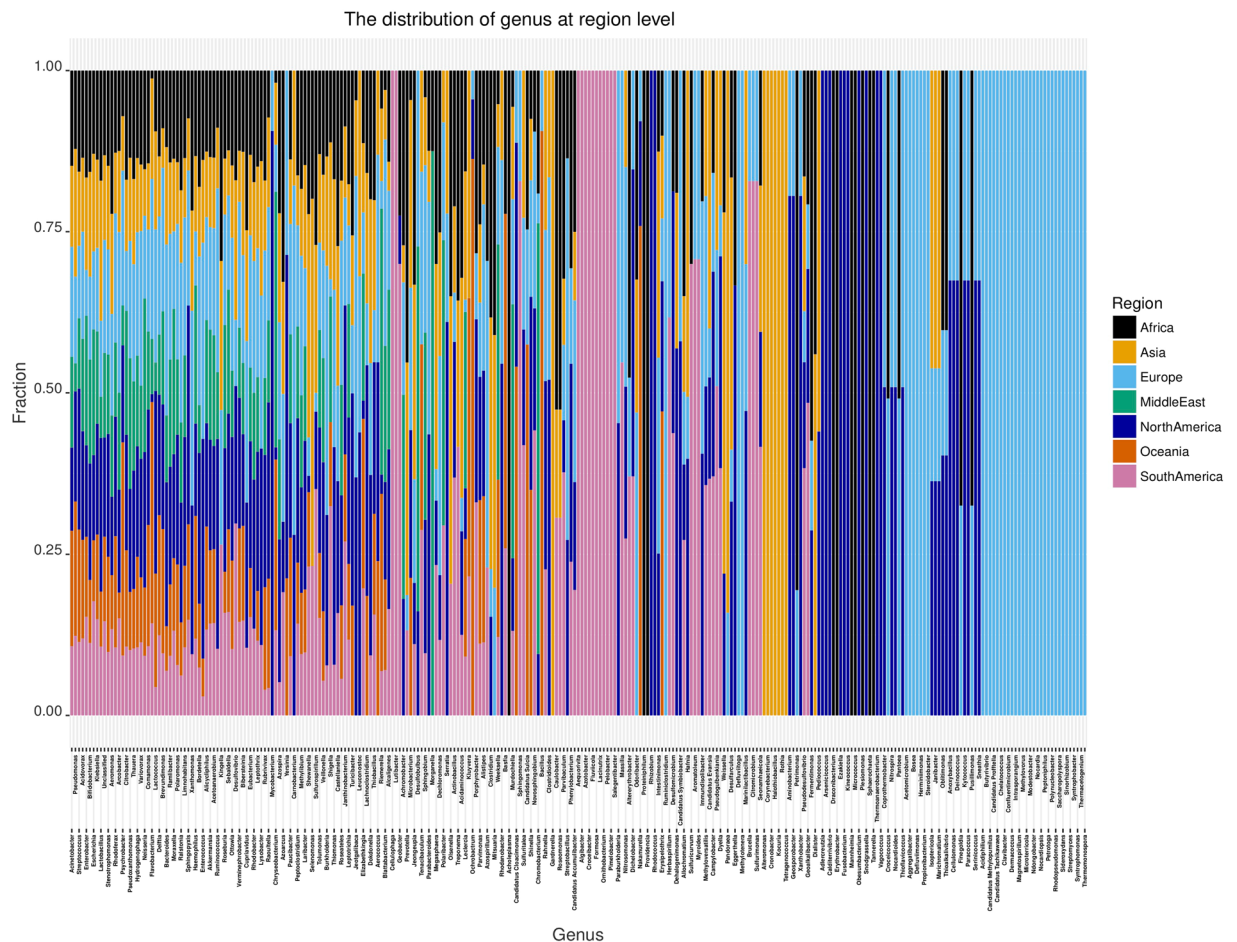


Figure 1. The distribution of identified genera in the different regions. The data have been standardized according to the number of samples in each region and the abundance for each genus has been calculated. The genera are sorted by abundance from high to low.

(b) and WHO-HI (d) clustering. For the EID2p based classification, there is a significant difference in the clustering of pathogens to environmental organisms and to commensal organisms for both WB-R (environmental p-value: 0.039, commensal p-value: 0.014) and WHO-R (environmental p-value: 0.039, commensal p-value: 0.0039), where the pathogenic bacteria has the least degree of regional clustering, and environmental and commensal the most. There is a significant difference in the clustering of commensal compared to both pathogens and environmental organism in the WHO-HI grouping (pathogen p-value: 0.022, environmental p-value: 0.029) where the commensal group is more regionally clustered and the others. No significant differences were observed for the WB-IL country grouping.

When the 5CC classification of bacteria into five categories was used (Fig. 4, bottom row (e-h)), there is no clear pattern of significantly different clustering, except that for both WB-R and WHO-R, the commensal group cluster significantly different from the opportunistic pathogen group (WB-R p-value: 0.026, WHO-R p-value: 0.028), where the commensal is the more regionally clustered group. When looking at the median ratio in Fig. 4, bottom row (e-h), although the difference between the groups is not significant, the tendency is the commensals and environmental have the lowest ratio, i.e. highest degree of clustering, and the three pathogen groups have a higher ratio and thereby less regional clustering. This goes for all but the WB-IL grouping.

Countries grouped together consistently according to regions, with a few exceptions. These countries can be seen in the Supplementary Tables S6 and S7. It is noteworthy to mention that all of Latin and North America are grouped together according to WHO, while the World Bank divides the region into two. Similarly, WHO groups Israel with Europe, while World Bank groups Israel with the Middle-Eastern countries. In addition, Pakistan changes from South Asia in WB-R to Eastern Mediterranean in WHO-R and Malta changes from Middle East & North Africa in WB-R to Europe in WHO-R.

To illustrate the levels of regional clustering for the classes from 5CC for the WB-R data, the phylogenetic trees representing the strain from the commensal, environmental and pathogenic classes, with the p-values (from the distance-based Welch *t*-test) closest to the median is shown in Fig. 5a–c. If the median was between two trees, the tree with the most samples was chosen. The median p-values are: commensal 0.0346 (Fig. 5a), environmental 0.0303 (Fig. 5b) and pathogen 0.605 (Fig. 5c).

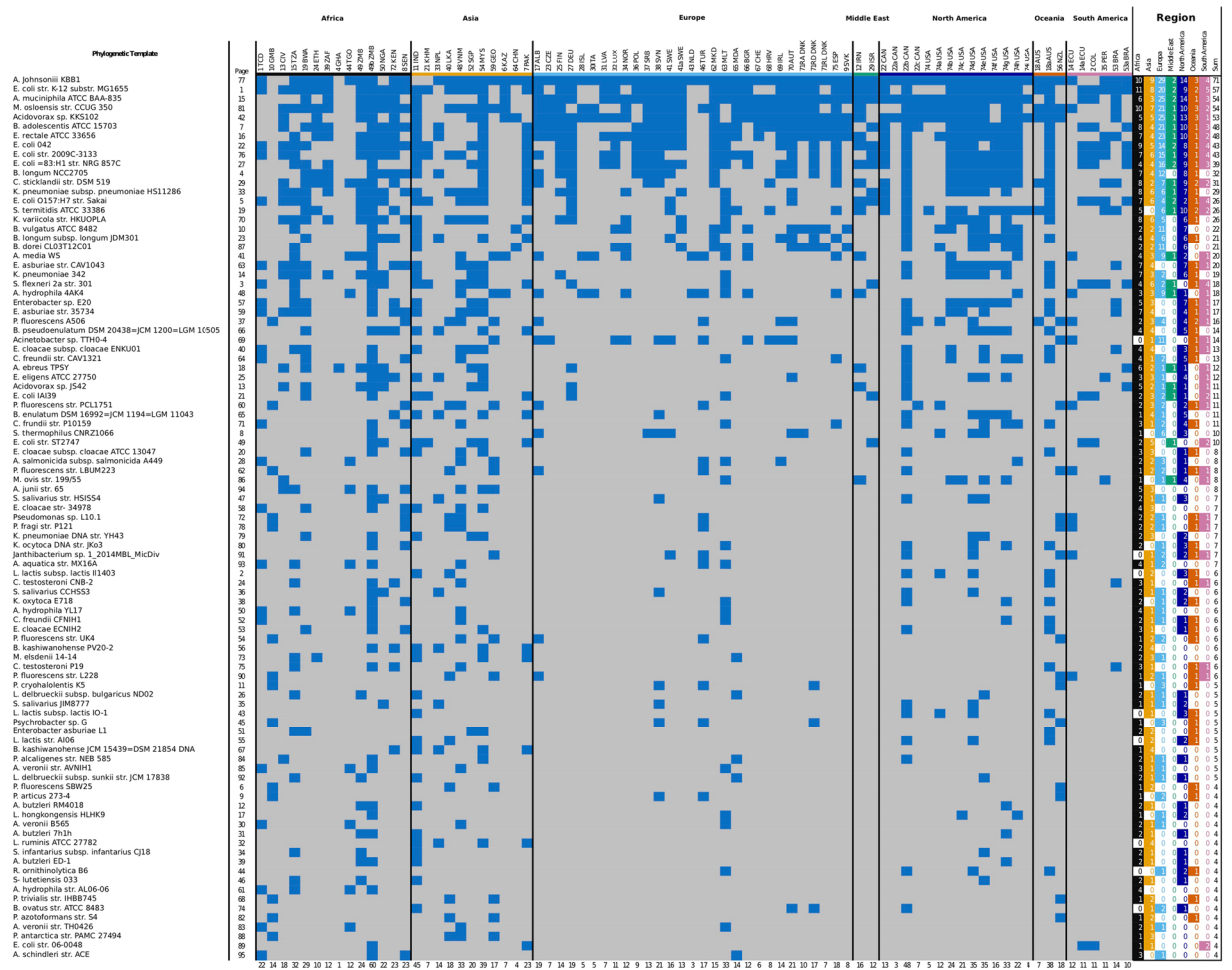


Figure 3. A presence/absence map of the samples in the 95 phylogenetic trees. Blue is presence, grey is absence. The samples are sorted by region according to continent and the trees by number of taxa. Page refers to the page number in the collated pdf file of all the phylogenetic trees. Each region has a column, if a region is represented in a phylogeny, it is marked by a coloured field. The number of samples per tree is summed in the last column, and the number of phylogenies that each sample appear in is summed in the last row.

this is a consistent pattern or potential bias in the study. We have identified at least two biases that can affect this conclusion. Firstly, in-sufficient sequencing depth in our study can affect the ability to detect rare species. Secondly, as we only have 79 samples across the world, we cannot be certain that we sequence everything.

In the mapping step of this pipeline, all of the raw reads from a sample are mapped to each of the identified template genomes for this sample, to create the consensus sequences, which we use for distance calculation. This induces the risk that reads are mapped incorrectly, i.e. for areas on the template genomes where there is a close resemblance to other genomes, we risk mapping wrong reads to that part and hereby create a false “core” genome. This could affect the distance calculation, as the areas where wrong reads are mapped could be filled with either wrong bases or will be blurred by the noise from erroneously aligned bases and therefore marked as ambiguous. However, due to the strictness of our thresholds for mapping the reads, we believe that if any of these scenarios are true, it is the latter; that some mutations may not be found due to noise in the consensus sequence. The mapping algorithm requires an ungapped alignment with a score of 50, with a match score of 1 and penalty for mismatch at -3. Our base-calling method only calls the base if the most frequently found base at the given position is found at a significance level of 0.001²². Furthermore, this method of creating a consensus sequence from metagenomic raw reads was tested by Joensen *et al.* 2017, in a study where both metagenomic raw reads and raw reads from individual isolated colonies from the same faecal samples were mapped to the same reference genome and the phylogeny was inferred by the use of NDtree. They found that in most cases, the metagenomic and the isolated sequences from the same samples were placed together in the tree¹³.

When we create the database of template sequences, used both in identifying the species in each sample and later to map the raw reads to, we only use the NCBI RefSeq²³ genomes. This can and does lead to a bias in the species we find, as we can only look for the well-studied organisms that are in this curated database. But we found it to be of greater importance to use well-described genomes, and perhaps identify fewer species, than to use genomes where we are not as certain of their origin and how well assembled they are.

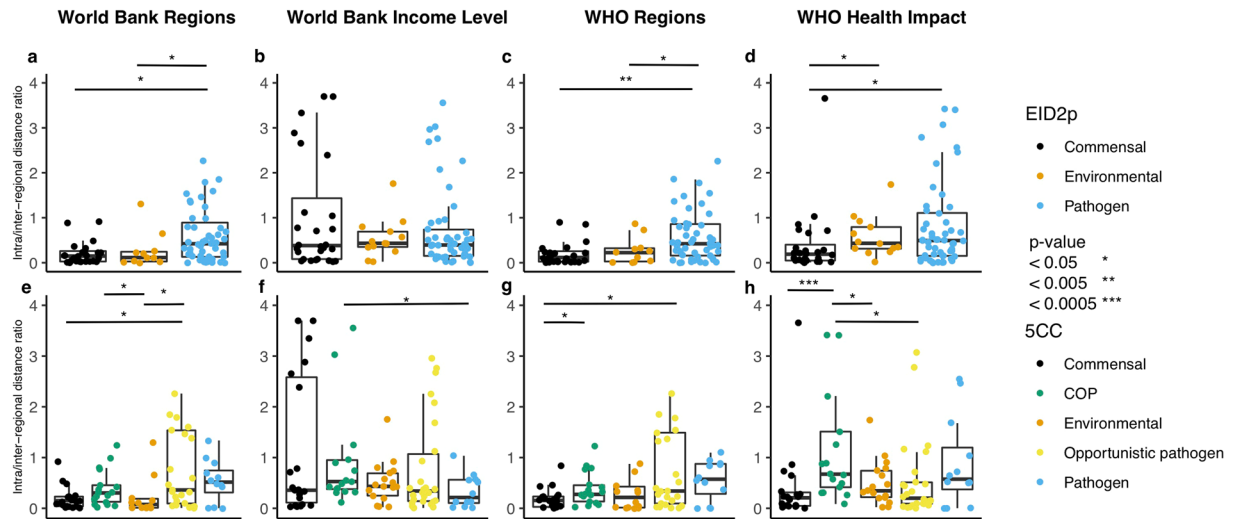


Figure 4. Boxplot showing the overall distribution of intra/inter-regional distance ratios for each organism group, with dots representing a single template organism. The top row (a–d) shows organism classification according to EID2p. The bottom row (e–h) shows organism classification according to 5CC. Significantly different clustering is marked with asterisk(s). Ratios above 4 not shown, the numbers are WB-R 4, WB-IL 1, WHO-R 4 and WHO-HI 1. COP, colonizing opportunistic pathogen.

In the step prior to the distance calculation (step D), we discard consensus sequences with more than 40% unknown bases. This gave us 1,504 consensus sequences. We initially set the thresholds at both 20% and 30%, which resulted in 518 and 1,023 consensus sequences, respectively. We deemed both numbers to be too small to give us enough trees. When we looked at the percentages of unknown bases over all the 11,691 consensus sequences, we saw that the majority of the sequences contained 75%–90% unknown bases (see Supplementary Figure S2). We could have included more sequences, but if the cutoff had been at 50%, we could have had templates where no phylogenetic distance could be calculated, because none of the positions were covered in each template, by only having two “bad sequences”.

We chose to use Neighbor-Joining (NJ) to infer the phylogeny, as we have shown in previous studies, that when using a distance-based approach for phylogeny, the NJ algorithm performs as well as Maximum Likelihood-based approaches^{13,22,24–26}. Furthermore, the speed gained by Neighbor-Joining is significant.

Phylogenetic trees representing unique bacterial templates in the samples were obtained from 79 sewage samples. For each tree, the tendency to regional clustering was found by calculating the statistical significance of the average genetic distance between samples from the same region versus samples from different regions. The environmental samples clustered significantly more than the pathogenic bacteria using the EID2p classification of the bacteria (commensal, environmental, pathogen), when using the WB-R or the WHO-R classification of the countries of the world. The 5CC classification expanded the bacterial classification with the following categories: opportunistic pathogens and colonizing opportunistic pathogens (COP). Using these classifications, the differences between environmental and pathogenic bacteria were still significant. Significant differences could be observed between both commensal and environmental bacteria as compared with pathogens as well as opportunistic pathogens for either the WB and WHO regions. Furthermore, it could be observed that COPs were less regionally clustered than the environmental, however, this difference was only statistically significant for the WB regions and for WHO health impact. In general, the results were comparable using the two different definitions of regions, at tendency if not significance level.

Commensal and environmental bacteria displayed the highest level of regional clustering. Commensal bacteria were in general more significantly different than environmental bacteria from both pathogens and opportunistic pathogens. COPs showed a tendency to be less regionally clustered than commensals and environmental bacteria, but more than the pathogens and opportunistic pathogens for the two regional groupings (the only statistically significant difference was to the environmental using the WB regions). We only found significant differences in clustering between two groups (COP and Pathogen) of the microorganism classifications and country groupings according to income (WB-IL). This could indicate that bacteria are spread, or are selected globally, dependent of geographic distances but independent of income status.

The bacteria were fairly equally divided in the different categories. When dividing the species by the 5CC classification, the average number of samples per distance matrix are the following: Commensal 18.4, COP 12.3, Opportunistic pathogen 10.9, Pathogen 17.5, and Environmental 13.1.

Some microorganisms in each classification have a ratio that is noticeably different from the median. Three *Klebsiella* genomes (two classified as COPs, one as commensal) and one *Acinetobacter* genome (classified as an environmental strain) have high ratios, indicating low regional clustering. These four genomes behaved more like pathogens. Two *Areomonas* genomes and one *Enterobacter* genome classified as opportunistic pathogens have a low degree of regional clustering with very high ratios (>8 for the two *Areomonas* and 2.26 for the *Enterobacter*),

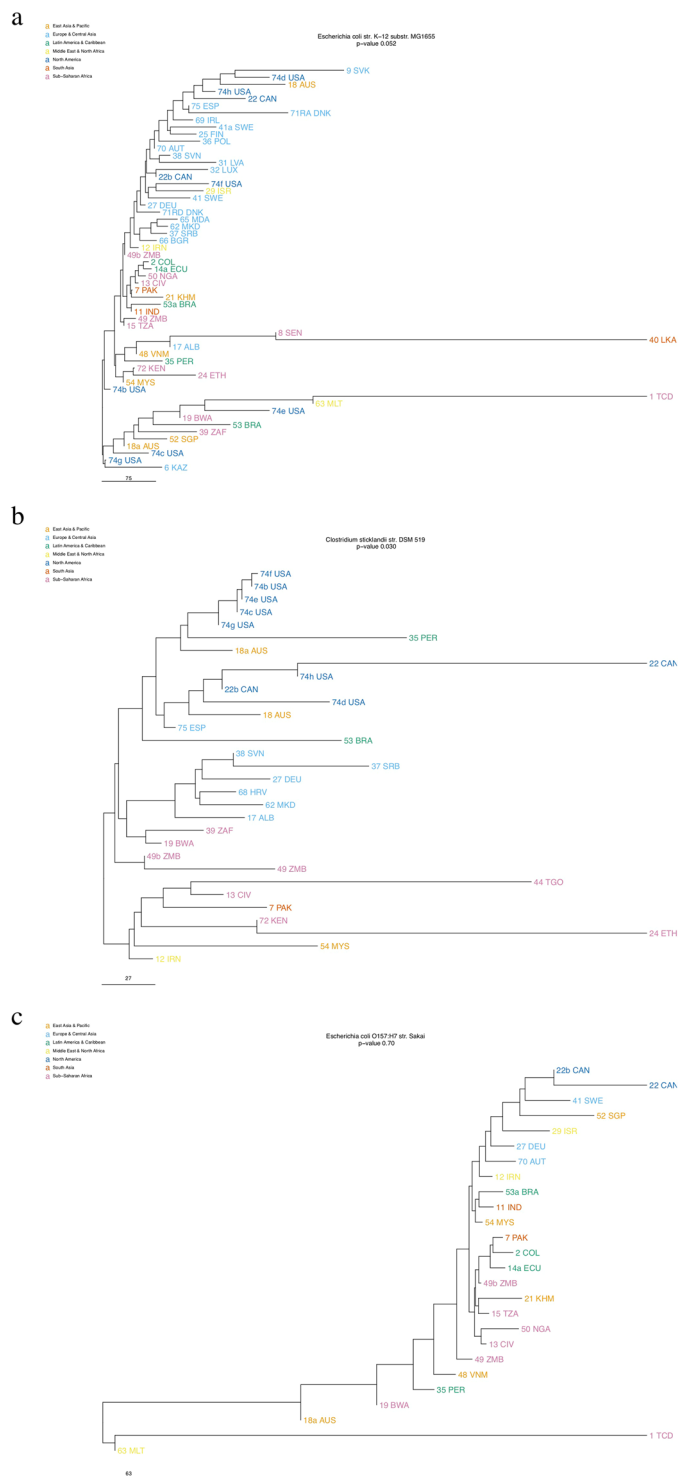


Figure 5. Phylogenetic tree for (a) the commensal *Escherichia coli* K-12 strain (b) the environmental *Clostridium sticklandii* (c) the pathogen *Escherichia coli* O157:H7 str. Sakai. The samples are coloured by World Bank regions.

even for this group, where the median is 0.37. The annotation by the 5CC is ambiguous for these three strains. The first is an industrial strain as well as a zoonotic pathogen, the second is a fish pathogen and possibly also commensal bacterium, and the third could also be classified as environmental bacterium. Three *E. coli* classified as pathogens also stand out as they all have a high level of regional clustering. The first is an AIEC (Adherent-invasive *Escherichia coli*), this *E. coli* pathotype is related to Crohn’s Disease where it colonizes the ileum of the patient^{27,28}, and although it is a pathogen, the colonization may still add to the regional clustering of this particular type, as the bacteria will develop together with the infected host. The second is a uropathogenic ExPEC (Extra-intestinal

pathogenic *Escherichia coli*)²⁹, which may be classified as a COP rather than a pathogen. The last one is an *E. coli* with the O157:H7 serotype, which most likely is a pathogen.

Although we see a pattern of clustering, not all species follow this pattern. Even some strains of the same species, with the same serotype, such as the two *E. coli* O157:H7 strains, do not have the same pattern, one has a ratio of 1.33 (WB-R) or 1.10 (WHO-R) (not regionally clustered), the other 0.48 (WB-R) or 0.38 (WHO-R) (regionally clustered). However, these two particular strains support that the mapping method we use is specific, since we get different clustering for two closely related strains.

Most global phylogenetic or phylogenomic studies have been conducted on pathogenic bacterial species or clonal sub-groups, and it has been shown for several species that such pathogenic clones can spread worldwide. This includes: *Shigella*³⁰, *Staphylococcus aureus*, *Klebsiella*³¹, *Streptococcus agalactiae*³², *E. coli*^{33,34} and *Acinetobacter baumannii*³⁵. Such studies have very rarely been conducted on colonizing or environmental bacterial species. However, studies have suggested that the fruit pathogen *Pseudomonas syringae* pv. *actinidiae*^{36,37} and the environmental *Streptococcus thermophilus*³⁸ have regional phylogenies. These results are in line with the findings of this study.

Conclusion

In general, we find that for environmental and commensal bacteria in particular, and to a lesser extent for COPs, there is a selection or barriers to spread based on geographical regions. For pathogens and opportunistic pathogens less regional clustering is observed. Income level and health impact were less correlated with the spread of the bacteria than geography-based clustering.

Methods

Pipeline. The workflow of the pipeline is depicted in Fig. 6 and consists of a template database creation step and six data analysis steps (A-F), including four major steps and two minor pre-processing steps.

Reference template database. The pipeline requires a database of unique bacterial reference template genomes. This database is both used for identification of templates in the metagenome samples and as template sequences to map the reads against in step C. This database was created by downloading all complete bacterial genomes from the NCBI RefSeq database²³. A Hobohm 1¹⁵ homology reduction with a similarity threshold of 98% was performed on the database. This was done by *kma_index*³⁹, which, in addition to homology reduction, also created a database of unique 20-mers with the prefix “ATGAC” for each genome⁴⁰. The genomes in the homology-reduced database are representative strain templates and here referred to as unique bacterial templates.

- (A) **Template identification:** The first analytic step in the pipeline is to identify all bacterial templates present in each of the metagenome samples with *KMA*³⁹. *KMA* was used with its sparse mapping option and the “winner takes all” scoring method.
- (B) **Mapping pre-processing:** After the bacterial template identification, the results from *KMA* were sorted on the total mapping depth, i.e. how many times the bacterial template had been covered by the raw reads. The cut-off was set to 1.
- (C) **Mapping:** After discarding the bacterial templates with low depth, the raw reads for each sample were mapped to each of the remaining bacterial templates identified in that sample. This was done using the *Assimiler* mapping tool, which is also employed in *NDtree* (<https://cge.cbs.dtu.dk/services/NDtree/>), with default settings. *NDtree* has been validated in a number of studies^{13,22,24–26}. This step yields consensus sequences for each bacterial template.
- (D) **Distance calculation pre-processing:** This pre-processing step checks the consensus sequences for their content of unknown bases. If a consensus sequence contains less than 40% unknown bases, they will be used for distance matrix calculation, otherwise they are discarded.
- (E) **Distance calculation:** For each unique bacterial template, the genetic distance is calculated between all the consensus sequences for the bacterial templates that passed the pre-processing step. The distance calculation is done by the same distance algorithm, which is employed by *NDtree*, with the use of the “all called” option that only uses the positions that are known in all genomes for the distance calculation. The output from this step is a distance matrix for each of the remaining unique bacterial templates.
- (F) **Phylogenetic inference:** The phylogeny was inferred by the Neighbor-Joining algorithm⁴¹ on each of the distance matrices from the previous step, if there were more than three samples participating in the matrix. The phylogenetic inference was done by the program *Neighbor* from the *Phylib* package⁴².

Statistical analysis. For the statistical analysis a Distance-based multivariate Welch *t*-test based on Alekseyenko’s multivariate Welch *t*-test on distances⁴³, modified to test the difference between distances within groups (intra-regional) compared with distances between groups (inter-regional) was developed (see `bitbucket` repository `distance_matrix_tests.R` for R script). The modified Welch *t*-test is used to calculate a value for the intra- vs inter-regional clustering. Each distance matrix was shuffled 999 times, and a *p*-value was calculated to assess whether the original clustering was significant compared with the randomly occurring clustering. Furthermore, the ratio of intra- vs inter-regional distances was calculated by dividing the sum of squared distances for the intra-regional cluster by the sum of squared distances for the inter-regional clusters. These ratios were used to test if there were any significant differences in regional clustering of trees with different types of organism classifications. This is performed using a Wilcoxon rank sum test¹⁶.

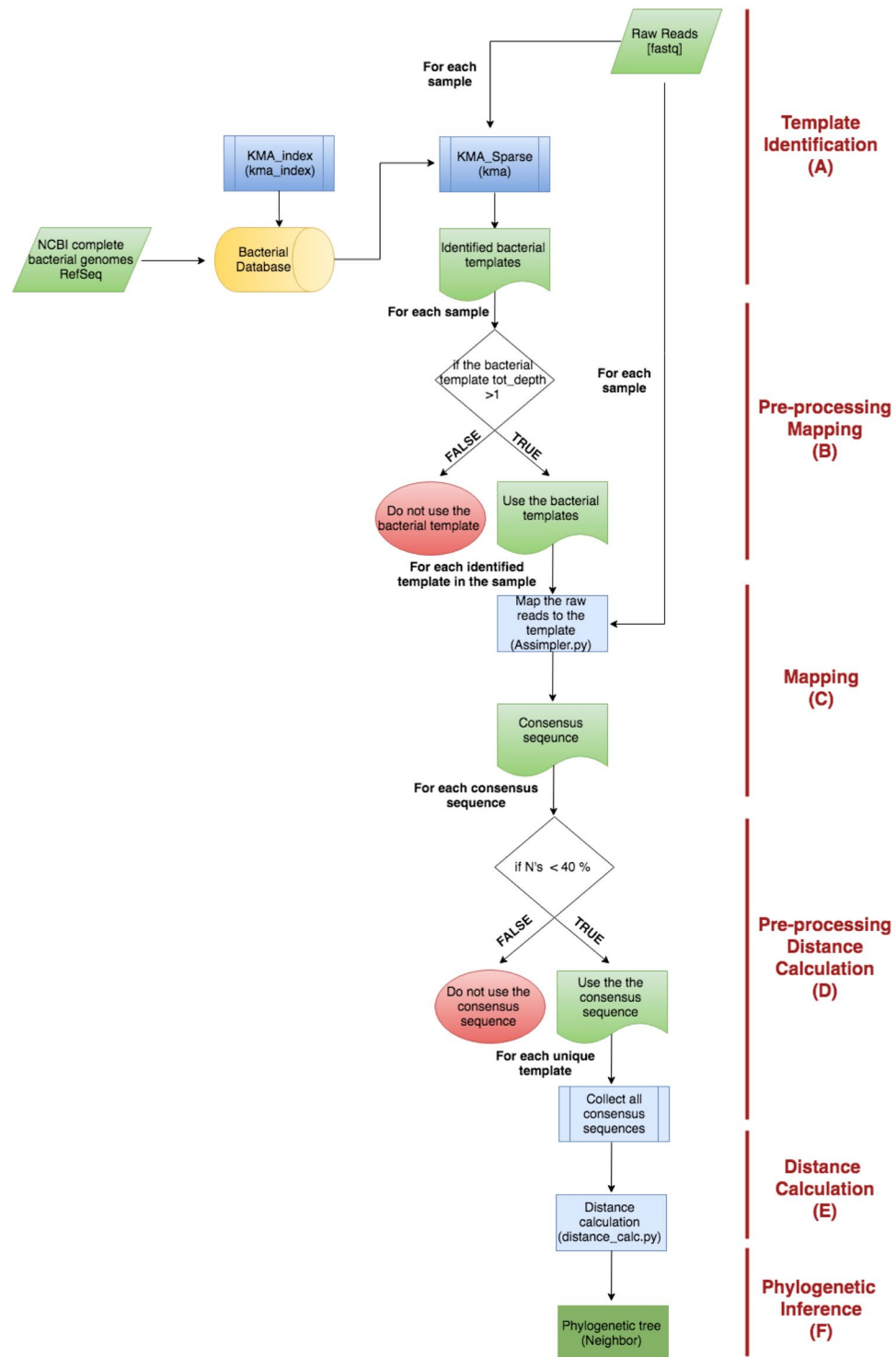


Figure 6. The workflow of the MetaPhygeny pipeline for bacterial phylogeny. Blue boxes indicate algorithms and scripts used in the pipeline. Green boxes are input and output files. The red spheres illustrate discarded templates. The yellow cylinder is the database.

The scripts for the statistical analysis, the distance matrices from the pipeline, together with the tree files and the metadata for the statistical analysis can be found on Bitbucket. (https://bitbucket.org/genomicpidemiology/metaphygeny_paper)

Visualization. The visualizations of the statistical analysis and of the trees were done in R (version 3.6.0)⁴⁴ with the packages ggplot2 (3.2.1)⁴⁵, ggtree (1.16.6) and treeio (1.8.2)⁴⁶. Other used packages include cowplot (1.0.0)⁴⁷, dplyr⁴⁸, ggbeeswarm (0.6.0)⁴⁹ and tidy (0.8.3)⁵⁰.

Materials

Regional classification was obtained from the World Bank⁵¹ regarding region and income. Regional classification was obtained from WHO⁵² regarding region and health impact. All regional data can be found in the supplementary information (Supplementary Table S1 and Supplementary Dataset 2).

The annotation of bacterial templates was done according to two classification schemes: EID2 plus (EID2p) and 5CC. The EID2p was built from the EID2 database⁵³ and Taylor *et al.*⁵⁴. The 5CC is based on EID2p classification, which further classified pathogens into COP (colonizing opportunistic pathogen) and OP (opportunistic pathogen) according to Price *et al.*⁵⁵. The full list can be seen in Supplementary Dataset 3 and the workflow of the annotation is further described in the Supplementary Methods (Classifications of bacterial templates).

Data

The data was obtained from the Global Sewage project for the first 79 samples collected¹⁴.

Data availability

The scripts for the statistical analysis, the distance matrices from the pipeline, together with the tree files and the metadata for the statistical analysis can be found on Bitbucket. (https://bitbucket.org/genomicpidemiology/metaphylogeny_paper)

Received: 27 June 2019; Accepted: 22 January 2020;

Published online: 20 February 2020

References

- O'Malley, M. A. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Stud. Hist. Philos. Biol. Biomed. Sci.* **39**, 314–25 (2008).
- de Wit, R. & Bouvier, T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ. Microbiol.* **8**, 755–8 (2006).
- Hendriksen, R. S. *et al.* Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* **2**, e00157–11 (2011).
- Harris, S. R. *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* (80-). **327**, 469–474 (2010).
- Nicolas-Chanoine, M.-H. *et al.* Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J. Antimicrob. Chemother.* **61**, 273–81 (2008).
- He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109–113 (2013).
- Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–4 (2016).
- Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol. Int.* **66**, 515–522 (2017).
- Lin, A. *et al.* Distinct Distal Gut Microbiome Diversity and Composition in Healthy Children from Bangladesh and the United States. *PLoS One* **8**, e53838 (2013).
- Mueller, S. *et al.* Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: a Cross-Sectional Study. *Appl. Environ. Microbiol.* **72**, 1027–1033 (2006).
- Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
- Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457 (2017).
- Joensen, K. G. *et al.* Evaluating next-generation sequencing for direct clinical diagnostics in diarrhoeal disease. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 1325–1338 (2017).
- Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).
- Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein Sci.* **1**, 409–417 (1992).
- Bauer, D. F. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690 (1972).
- Turco, R. F., Carrero-Colón, M. & Wickham, G. S. In *The Fecal Bacteria* 23–38 (American Society of Microbiology). <https://doi.org/10.1128/9781555816865.ch2> (2011).
- Tannock, G. W. *et al.* Analysis of the fecal microflora of human subjects consuming a probiotic product containing *Lactobacillus rhamnosus* DR20. *Appl. Environ. Microbiol.* **66**, 2578–88 (2000).
- Repizo, G. D. *et al.* The Environmental *Acinetobacter baumannii* Isolate DSM30011 Reveals Clues into the Preantibiotic Era Genome Diversity, Virulence Potential, and Niche Range of a Predominant Nosocomial Pathogen. *Genome Biol. Evol.* **9**, 2292–2307 (2017).
- Dogan, B. & Boor, K. J. Genetic Diversity and Spoilage Potentials among *Pseudomonas* spp. Isolated from Fluid Milk Products and Dairy Processing Plants. *Appl. Environ. Microbiol.* **69**, 130–138 (2003).
- Byrne-Bailey, K. G. *et al.* Completed genome sequence of the anaerobic iron-oxidizing bacterium *Acidovorax ebreus* strain TSPY. *J. Bacteriol.* **192**, 1475–6 (2010).
- Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O. & Aarestrup, F. M. Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. *PLoS One* **9**, e87991 (2014).
- O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
- Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. *PLoS One* **9**, e104984 (2014).
- Joensen, K. G. *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **52**, 1501–10 (2014).
- Ahrenfeldt, J. *et al.* Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. *BMC Genomics* **18**, 19 (2017).
- Martinez-Medina, M. & Garcia-Gil, L. J. *Escherichia coli* in chronic inflammatory bowel diseases: An update on adherent invasive *Escherichia coli* pathogenicity. *World J. Gastrointest. Pathophysiol.* **5**, 213–27 (2014).
- Palmela, C. *et al.* Adherent-invasive *Escherichia coli* in inflammatory bowel disease. *Gut* **67**, 574–587 (2018).
- Touchon, M. *et al.* Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genet.* **5**, e1000344 (2009).
- Holt, K. E. *et al.* *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44**, 1056–1059 (2012).
- Bowers, J. R. *et al.* Genomic Analysis of the Emergence and Rapid Global Dissemination of the Clonal Group 258 *Klebsiella pneumoniae* Pandemic. *PLoS One* **10**, e0133727 (2015).

32. Skov Sørensen, U. B., Poulsen, K., Ghezzi, C., Margarit, I. & Kilian, M. Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. *MBio* **1**, 1–9 (2010).
33. Petty, N. K. *et al.* Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc. Natl. Acad. Sci.* **111**, 5694–5699 (2014).
34. von Mentzer, A. *et al.* Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat. Genet.* **46**, 1321–1326 (2014).
35. Sahl, J. W. *et al.* Phylogenetic and genomic diversity in isolates from the globally distributed *Acinetobacter baumannii* ST25 lineage. *Sci. Rep.* **5**, 15188 (2015).
36. McCann, H. C. *et al.* Genomic Analysis of the Kiwifruit Pathogen *Pseudomonas syringae* pv. *actinidiae* Provides Insight into the Origins of an Emergent Plant Disease. *PLoS Pathog.* **9**, e1003503 (2013).
37. Darcy, J. L., Lynch, R. C., King, A. J., Robeson, M. S. & Schmidt, S. K. Global distribution of *Polaromonas* phylotypes - evidence for a highly successful dispersal capacity. *PLoS One* **6**, (2011).
38. Delorme, C. *et al.* Study of *Streptococcus thermophilus* population on a world-wide and historical collection by a new MLST scheme. *Int. J. Food Microbiol.* **242**, 70–81 (2017).
39. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
40. Larsen, M. V. *et al.* Benchmarking of Methods for Genomic Taxonomy. *J. Clin. Microbiol.* **52**, 1529–1539 (2014).
41. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25 (1987).
42. Felsenstein, J. PHYLIP (phylogeny inference package) version 3.69. (2013).
43. Alekseyenko, A. V. Multivariate Welch *t*-test on distances. *Bioinformatics* **32**, btw524 (2016).
44. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
45. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).
46. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
47. Wilke, C. O. cowplot: Streamlines Plot Theme and Plot Annotations for 'ggplot2'. (2017).
48. Wickham, H., Francois, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2017).
49. Clarke, E. & Sherrill-Mix, S. ggbeeswarm: Categorical Scatter (Violin Point) Plots. (2017).
50. Wickham, H. & Henry, L. tidy: Easily Tidy Data with 'spread()' and 'gather()' Functions. (2018).
51. World Bank Country and Lending Groups – World Bank Data Help Desk. at <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>.
52. WHO | Country groupings. WHO (2014).
53. Wardeh, M., Risley, C., McIntyre, M. K., Setzkorn, C. & Baylis, M. Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data* **2**, 150049 (2015).
54. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356**, 983–9 (2001).
55. Price, L. B., Hungate, B. A., Koch, B. J., Davis, G. S. & Liu, C. M. Colonizing opportunistic pathogens (COPs): The beasts in all of us. *PLoS Pathog.* **13**, e1006369 (2017).

Acknowledgements

This work was supported by the European Union Horizon 2020 research and innovation program under grant agreement 643476 to the COMPARE project (<http://www.compare-europe.eu>), and The Novo Nordisk Foundation (NNF16OC0021856) - Global Surveillance of Antimicrobial Resistance.

Author contributions

J.A., O.L. and F.M.A. developed the idea and framework of the project. M.W. made the phylogenetic pipeline with help from J.A., J.S., R.A. and P.T.L.C.C. and carried out the data analysis with help from J.A. I.C.L. did the statistical analysis. R.S.H. provided the data and guidance in the interpretation of the results. J.A., I.C.L., M.W. an O.L. drafted the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-59292-w>.

Correspondence and requests for materials should be addressed to O.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020