ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data

Achal Dhariwal[1], Roger Junges[1], Tsute Chen[2,3] & Fernanda Cristina Petersen[1*]

[1] Institute of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway

[2] Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA

[3] Department of Oral Medicine, Infection, and Immunology, Harvard School of Dental Medicine, Boston, MA, USA

***** Corresponding author

Abstract

Background

The study of resistomes using whole metagenomic sequencing enables access to the large repertoire of resistance genes usually found in complex microbial communities, such as the human microbiome. Over recent years, sophisticated and diverse pipelines have been established to facilitate raw data processing and annotation. Despite the progress, there are no easy-to-use tools for comprehensive visual, statistical, and functional analysis of resistome data. Thus, exploration of the resulting large complex datasets remains a key bottleneck requiring robust computational resources and technical expertise, which creates a significant hurdle for advancements in the field.

Results

Here, we introduce ResistoXplorer, a user-friendly tool that integrates recent advancements in statistics and visualization, coupled with extensive functional annotations and phenotype collection, to enable high-throughput analysis of common outputs generated from resistome studies. ResistoXplorer contains three modules- the 'Antimicrobial Resistance Gene Table' module offers various options for composition profiling, functional profiling and comparative analysis of resistome data; the 'Integration' module supports integrative exploratory analysis of resistome and microbiome abundance profiles in metagenomic samples; finally, the 'Antimicrobial Resistance Gene List' module enables users to explore antimicrobial resistance genes according to function and potential microbial hosts using visual analytics to gain biological insights. Within these three modules, ResistoXplorer offers comprehensive assistance for ARG functional annotations along with their microbe and phenotype associations based on data collected from >10 reference databases. In addition, it provides support for a variety of methods

for composition profiling, visualization and exploratory data analysis, as well as extensive support for various data normalization methods and machine learning algorithms for identification of resistome signatures. Finally, ResistoXplorer offers also network visualization for intuitive exploration of associations between antimicrobial resistance genes and the microbial hosts, incorporated with functional enrichment analysis support.

Conclusions

ResistoXplorer is a web-based tool with a user-friendly interface that enables comprehensive and real-time downstream analysis of resistome data. As such, it allows for in-depth exploration of metagenomic datasets focusing on the intrinsic networks and correlations of antimicrobial resistance genes and their underlying microbial determinants. ResistoXplorer will assist researchers and clinicians in the field of AMR to facilitate discovery in large-scale and multi-dimensional metagenomic datasets. ResistoXplorer is publicly available at http://www.resistoxplorer.no/ResistoXplorer.

Background

Antimicrobial resistance (AMR) has emerged as a major threat to global public health and the economy [1]. Although antimicrobial resistance is a natural phenomenon, the misuse and overuse of antimicrobial agents in humans, animals and agriculture have accelerated the development and spread of antimicrobial resistance. Consequently, antimicrobials are becoming less effective in the prevention and treatment of infections. In addition, the extensive use of antibiotics has resulted in large amounts being released into the environment. This is a matter of increase concern, not only in relation to antimicrobial resistance, but also for the potential impact on the ecology of human, animal, and environmental ecosystems [2].

Microbial communities have traditionally been studied using culture techniques. Although valuable, culture-based strategies have a low-throughput, and can provide an incomplete microbial profile due to the fact that, in a majority of environments, only a limited proportion of microorganisms are cultivable. Such limitations have been largely circumvented with recent advancements in short-read based high-throughput DNA sequencing (HTS) technologies, using either amplicon targeted or whole metagenomic approaches. Amplicon approaches can be used to map both taxonomic and resistome profiles. However, one of the limitations is that only genes recognized by the specific primers can be identified. Whole metagenomics, on the other hand, enable unbiased taxonomic and antimicrobial resistance information. Such technology has been shown to provide valuable insights into the natural history of antimicrobial resistance genes (ARGs) in humans and nature [3-5], as well as how the resistome develops in early life, and how it can be affected by the use of antibiotics [6, 7]. It is also contributing to the identification of ARGs that may cross environmental and host boundaries [8], and providing unprecedented knowledge into the large reservoir of ARGs in the human, animal and environmental microbial

communities [9-14]. Currently, resistome profiles from complex and diverse microbial metagenomes are primarily generated using whole metagenome shotgun sequencing in which the total DNA extracted from a microbial community is sequenced. The resulting DNA fragments can be analyzed using read or assembly-based approaches to characterize their resistome composition [15]. These derived sequencing datasets are both large (amount) and complex (structure), causing considerable 'big data' challenges in downstream data analysis.

The main computational effort in resistome analysis of metagenomic datasets so far has focused on processing, classification, assembly and annotation of sequenced reads. This has led to the development of a number of excellent bioinformatic pipelines and tools for detecting and quantifying antimicrobial resistance genes in metagenomes [15-17]. However, there is still no clear consensus with regards to standard analysis pipelines and workflows for high-throughput analysis of AMR metagenomic resistome data [17, 18]. Nonetheless, the outputs from most of these pipelines can be summarized as a data table consisting of feature (ARGs) abundance information across samples, *i.e.* resistome profiles, along with their functional annotations and sample metadata. For most researchers, the fundamental challenge in data analysis can often be centered on how to understand and interpret the information in the abundance tables especially within the context of different experimental factors and annotations.

Resistome data analysis can be separated into four main categories: (i) composition profiling- to visualize and characterize the resistome based on approaches developed in community ecology such as alpha diversity, rarefaction curves or ordination analysis; (ii) functional profiling- to assign antimicrobial resistance genes into different functional categories (e.g. Drug class, Mechanism) to gain better insights regarding their collective functional capabilities; (iii) comparative analysis- to identify features having a significant differential abundance between studied conditions and (iv)

integrative analysis- to integrate the resistome and taxonomic data to understand the complex interplay and potential associations between microbial ecology and AMR. The computational methods and approaches to perform such analysis are fairly limited. The first category of analysis can be more straightforward to perform, but the last three are challenging.

First of all, the metagenomic abundance data is often characterized by differences in library sizes, which requires correction. To address this issue, researchers often employ two common normalization approaches prior to analysis: subsampling the reads in each sample to the same number (rarefying) or rescaling the total number of reads in each sample to uniform sum (using proportions). The former may entail the loss of valuable information, while the latter could lead to issues related to data compositionality [19]. In addition to uneven library sizes, metagenomics data are also characterized by sparsity, over-dispersion, zero inflation and skewed distribution [20, 21]. Such unique features make standard parametric tests and most methods established in other omics fields unsuitable for comparative analysis. To address these challenges, non-parametric permutation-based approaches have been adopted to identify significant features in metagenomic abundance data [22, 23]. Even though robust, these approaches are constrained by lack of statistical power, as well as inefficiency to model confounding factors, and inability to accommodate intricate experimental designs.

Overall, development of statistical models that account for features of metagenomic data or use of methods to transform data to have distributions that fit standard test assumptions is recommended. A variety of strategies have propelled empirical development in these directions. For instance, metagenomeSeq algorithm incorporates cumulative sum scaling normalization and a zero-inflated Gaussian (ZIG) mixture model to reduce false positives and improve statistical power for differential abundance analysis [24, 25]. It has also been demonstrated that algorithms developed

for RNA-seq data such as edgeR and DESeq2, along with their respective normalization methods, outperform other approaches used for metagenomic abundance data [25-28]. These standard strategies are widely employed, but have recently been questioned due to the compositional nature of whole metagenomic sequencing data [29, 30]. To address this issue, several Compositional Data Analysis (CoDA) approaches to analyze sequencing datasets have been recently proposed [31, 32]. ALDEx2 and ANCOM are two CoDA methods that integrate log-ratio transformation to explicitly deal with data compositionality while performing differential abundance testing [33, 34]. Naturally, the choice of methodology depends on the research question, and one may be interested in comparing results obtained by different analytical methods. Both standard and CoDA statistical approaches have been implemented as R packages. Although flexible, learning R in order to use these methods can be challenging for most clinicians and researchers.

In addition to the challenges described above, analytical steps have another complexity layer represented by the wide variety of reference databases to identify and characterize the antimicrobial resistance genes from metagenomic data. These differ considerably in the quality of data presented, as well as the scope of resistance mechanisms and the type of information provided [15]. This can influence the accuracy of *in silico* characterization of resistome data, which depends substantially on the comprehensiveness and quality of reference databases [15, 17, 35]. Typically, resistome profiles are analyzed by mapping ARGs either to their respective class of drugs to which they confer resistance (Class-level) or to their underlying molecular mechanism of resistance (Mechanism-level). Analyzing resistomes at such high level categories enable researchers to gain more biological, actionable, and functional insights together with a better understanding of their data. However, these functional levels and categories, along with their classification scheme, vary largely between the databases [18]. Additionally, depending upon the database, users need to

manually collect and curate such information and then generate separate abundance tables for each functional level. Hence, collecting appropriate functional annotation information for hundreds of ARGs in resistomes for functional profiling and further downstream analysis can be confusing, arduous, time-consuming and error-prone.

Finally, some AMR reference databases may also provide information regarding the microbial host that harbor or carry these reference ARGs. Information about such relationship can be complex as one microbe can carry multiple ARGs and single ARGs can in turn be present across multiple microbes. To explore such intricated 'multiple-to-multiple' relations, one option is to use a network-based visualization method. This approach, coupled with suitable functional annotations and enrichment analysis support, would have the potential to provide better interpretation of AMR resistance mechanisms and inform on possible dissemination routes of antimicrobial resistance genes. It is straightforward to identify key players from a network perspective, for instance, by looking for those ARGs that are found in multiple microbes or by identifying those microbes that simultaneously contain multiple ARGs of interest. Currently, there are no web-based tools that provide such functionality.

To address the above challenges as well as to meet recent advances in resistome data analysis, we have developed ResistoXplorer, a user-friendly, web-based, visual analytics tool to assist clinicians, bench researchers, and interdisciplinary groups working in the AMR field to perform exploratory data analysis on abundance profiles and resistome signatures generated from AMR metagenomics studies. The key features of ResistoXplorer include:

- Support of a wide array of common as well as advanced methods for composition profiling, visualization and exploratory data analysis;

- Comprehensive support for various data normalization methods coupled with standard as well as more recent statistical and machine learning algorithms;
- Support of a variety of methods for performing vertical data integrative analysis on paired datasets (*i.e.* taxonomic and resistome abundance profiles)
- Comprehensive support for ARG functional annotations along with their microbe and phenotypes associations based on data collected from >10 reference databases;
- A powerful and fully featured network visualization for intuitive exploration of ARG-microbe associations, including functional annotation enrichment analysis support.

ResistoXplorer also contains manuals to assist users to easily navigate through different analysis tasks. Collectively, these features consist of a comprehensive tool suite for statistical, visual and exploratory analysis of data generated from AMR metagenomics studies. ResistoXplorer is freely available at http://www.resistoxplorer.no/ResistoXplorer.

Implementation

ResistoXplorer is implemented based on Java, R and JavaScript programing languages. The framework is developed based on the Java Server Faces technology using the PrimeFaces (https://www.primefaces.org/) and BootsFaces (https://www.bootsfaces.net) component library. The network visualization uses the sigma.js (http://sigmajs.org/) JavaScript library. Additionally, D3.js (https://d3js.org/) and CanvasXpress (https://canvasxpress.org/) JavaScript libraries are utilized for other interactive visualization. All the R packages for performing back-end analysis and visualization are mentioned in the 'About' section of the tool. At the start of the analysis, a temporary account is created with an associated home folder to store the uploaded data and analysis results. All the analysis results will be returned in real-time. Upon completing their analysis

session, users should download all their results. The system is deployed on a dedicated server with 4 physical CPU cores (Intel Core i5 3.4GHz), 8GB RAM and Ubuntu 18.04 LTS was used as the operation system. ResistoXplorer has been tested with major modern browsers such as Google Chrome, Mozilla Firefox, Safari and Microsoft Internet Explorer.

Results and Discussion

ResistoXplorer consists of three main analysis modules. The first is the **ARG List** module that is designed to explore the functional and microbial host associations for a given list of antimicrobial resistance genes (ARGs) of interest. The second is the **ARG Table** module, which contains functions for analyzing resistome abundance profiles generated from AMR metagenomics studies. Lastly, the **Integration** module enables users to perform integrative analysis on the paired taxonomic and resistome abundance profiles to further explore potential associations coupled with novel biological insights and hypotheses. Figure 1 represents the overall design and workflow of ResistoXplorer. We recommend users to try out our example datasets to get familiar with the basic steps and key features of the tool before proceeding with analysis of their own data. Detailed explanation for all the steps are provided in the corresponding manuals.

Data upload and processing

Overview of data inputs

The three analysis modules (ARG List, ARG Table and Integration) are represented as three interactive buttons in ResistoXplorer. Users must choose an analysis path based on their input. The input can be uploaded in two different ways – by entering a list of antimicrobial resistance genes (ARGs) or by uploading an ARG abundance table along with a sample metadata file containing group information. In the latter case, the files can be uploaded as a tab-delimited text (.txt) or in

comma-separated values (.csv). Further, users must also provide the annotation information of ARGs either by uploading a file (.txt or .csv) or by just selecting the appropriate database used for their annotation. Additionally, the Integration module requires users to also upload a taxa abundance table in the same formats. The taxonomic and functional annotation files are optional in case of integrative analysis. Users can go to the corresponding 'Manuals' and 'Data Format' section, or download the example datasets for more details.

ARG-functional annotations collection

The annotation information and the classification scheme for reference ARGs (or sequences) are collected from 10 well-annotated and widely used AMR databases: ResFinder [36], CARD [37], ARDB [38], ARG-ANNOT [39], MegaRes [18, 40], AMRFinder [41], SARG [42], DeepARG-DB [43], ARGminer [44] and Resfams [45]. Further, this annotation information is organized into tables containing the reference ARGs in rows and their functional annotation levels across columns for each of the databases. The headers (names) for reference ARGs are annotated as in the chosen database. Some of the annotation levels having several multiple functional category assignments for ARGs were removed from the tables to avoid false counts inflation during downstream analysis. Additional manual curation for functional category naming and data redundancies were performed in some of the databases. Further, we also structured functional annotation information from the BacMet [46] database and antimicrobial peptide (AMP) resistance gene dataset [47] to enable users to perform functional profiling and downstream analysis of antibacterial biocides/metals and AMP resistance genes abundance profiles. The tables are stored in RDS file format for less storage space and fast retrieval of data. Users can use such table to annotate their resistome profile directly rather than manually collecting the functional annotation information from their corresponding database. The 'Data Format' and 'Resources' pages provide a detailed

description on the format, structure of annotation table and database statistics, together with the links to allow users to download the example annotation structure available for the individual database.

Data filtration and normalization

Features having very low counts or abundance across all the samples cannot be discriminated from sequencing errors or low-level contamination, and also interfere with some statistical and biological approximations. Hence, such features must be removed from the data before performing downstream analysis. ResistoXplorer automatically removes features corresponding to all the zeros or features only present in one sample. This type of minimally cleaned data is used for alpha diversity analysis. For all other types of downstream analysis, further data filtration is required. By default, features are filtered based on prevalence and abundances levels. Users can also choose to remove low abundant features by setting a minimum count cutoff based on their mean or median value. Conversely, some features remain constant in their abundances throughout the experimental conditions or across all the samples. These features are implausible to be informative in the comparative analysis. Users can exclude such features based on their inter-quantile ranges, standard deviations or coefficient of variations [48]. In case of integrative analysis, users can also choose to apply different data filtration criteria for both taxonomic and resistome abundance data.

After data filtering, users must perform data normalization to remove the systematic variability between samples. Currently, ResistoXplorer offers three categories of data normalization- rarefying, scaling and transformation, based on various widely used methods for metagenomic abundance data. In addition, ResistoXplorer supports other normalization methods like centered log-ratio (clr) and additive log-ratio (alr) transformation to facilitate compositional data analysis.

The normalized data is used for exploratory data analysis including ordination, clustering, comparative and integrative analysis. In the integrative analysis module, taxonomic and resistome datasets are normalized using the same approach. Applying suitable normalization methods is critical for identifying differentially abundant features, and reduce false positive rates [25].

Composition profiling

Visual exploration

ResistoXplorer allows users to visually explore the resistome based on various intuitive visualization approaches used for metagenomic data. For instance, users can visualize resistome abundance data while simultaneously showing the functional hierarchical relationships and connectivity between features using an interactive sankey diagram, zoomable sunburst or treemap graphics (Figure 2B). In case of treemap, users can click a particular rectangular block of interest to further inspect its compositions at a lower functional level. The abundances can be represented as either absolute counts or relative proportions. However, such visualizations are more suitable for resistome profiles having an acyclic and hierarchical functional annotation structure. Abundance profiles can also be viewed at different functional levels using other common visualizations such as stacked area or stacked bar plot (Figure 2A). The plot is organized by experimental factors to help visualize the differences in resistome composition across different conditions.

Diversity profiling

The resistome diversity profiling is implemented mainly based on the R vegan and phyloseq packages [49, 50]. Currently, users can perform alpha diversity (within-sample) analysis using eight common richness and evenness-based (or both) diversity measures. The results are

represented in the form of a dot plot for individual samples and box plots for each sample group (Figure 2C). The corresponding statistical significance is calculated automatically using either parametric or non-parametric tests. The analysis can be performed at different functional levels based on the available annotations. Additionally, the reliability of estimated diversity in samples can be assessed through rarefaction curves in which the number of unique features (ARGs) identified is plotted against the sequence sample size (Fig. 2D).

Ordination analysis

The ordination analysis function allows users to explore and visualize the similarities or dissimilarities between samples based on their composition at different functional levels. The dissimilarity can be calculated using five distance measures. Currently, three widely accepted ordination methods are supported, including principal coordinate analysis (PCoA), non-metric multidimensional scaling (NMDS) and principal component analysis (PCA). In particular, users can follow a CoDA ordination approach by performing PCA on the centered log-ratio transformed data. The corresponding statistical significance is calculated using one of the three multivariate statistical testing methods with Analysis of Group Similarities (ANOSIM) as the default option. The results are represented as both 2D and 3D sample plots. The samples visualized in these ordination plots are colored based on metadata or their alpha diversity measures to help users identify any underlying patterns in the datasets (Figure 2F).

Comparative analysis

Differential abundance testing

This section enables users to perform statistical testing to identify features that are significantly different in abundance across sample groups. ResistoXplorer supports standard tests, such as

edgeR [28], DESeq2 [27], metagenomeSeq [24], LEfSe [23], as well as more recent CoDA univariate analysis approaches such as ALDEx2 and ANCOM [34]. EdgeR and DESeq2, are broadly used and generally considered as robust and powerful parametric statistical approaches. They fit a generalized linear model and assume that read counts follow a negative binomial distribution to account for the features of count data. LEfSe uses the differential abundance between groups. The CoDA methods perform statistical testing on features log-ratio rather than their actual count abundances to deal with the compositional nature of sequencing data. ALDEx2, for instance, performs parametric or nonparametric statistical tests on log-ratio values from a modeled probability distribution of the data and returns the expected values of statistical tests along with effect size estimates. ANCOM tests the log-ratio abundance of all pairs of features for differences in means using nonparametric statistical tests. Both the standard and CoDA methods, except LEfSe, use their own specific normalization procedures while conducting statistical testing. For instance, the relative log expression (RLE) normalization is used for DESeq2, and the centered log-ratio transformation is applied for ALDEx2. The differential analysis can also be performed at different functional levels.

The results from the differential analysis are displayed as a table. By default, the result table will show a maximum of 500 top features ordered according to their P-values. The significant features (if present) are automatically highlighted in orange color. Further, users can also see a boxplot summary for any feature of interest by just clicking the ‘Details’ icon. Since different statistical approaches may generate divergent P-values, it is often recommended to compare and visualize results using more than one statistical approach, as to increase confidence in the interpretation of results.

Machine learning-based classification

Prediction of microbiome signatures using machine learning algorithms has been gaining more recognition and shown to perform well in recent resistome data analyses and classifications [6, 9, 51]. ResistoXplorer provides two such powerful supervised classification methods - Random Forest [52] and Support Vector Machine (SVM). In particular, the Random Forest algorithm uses an ensemble of classification trees (forest) where each tree is grown based on a random subset of features from a bootstrap sample at each branch. The final class prediction is based on the majority vote of the ensemble. The unbiased estimate of classification (OOB) errors is obtained by aggregating cross-validation results using bootstrapped samples while the forest is being constructed. Random forest also measures the importance of each feature based on the increase of the OOB error when it is randomly re-shuffled. Alternatively, the SVM algorithm seeks to find a nonlinear decision function (hyperplane) by projecting data into higher dimensional feature space and separating it by means of a maximum margin hyperplane [53]. The SVM analysis is performed using recursive feature selection and sample classification via linear kernel [54]. Features are selected based on their relative contribution in the classification using cross-validation error rates. The features which hold less variance (least significant) are removed in the successive steps. This procedure generates a series of SVM models (levels). The features used by the best model are considered to be important and ranked based on their frequencies of being selected in the model. Figure 2K shows the classification performance of SVM with regards to decreasing number of variable (levels).

Other features

There are several other valuable functions implemented in ResistoXplorer for exploratory analysis of resistome data. Users can perform core resistome analysis to detect core set of features present in samples or sample groups based on their abundance and prevalence level (|Figure 2E).

ResistoXplorer also supports commonly used correlation analysis and hierarchical clustering. The results from hierarchical clustering can be visualized using heatmaps (Figure 2G) and dendrograms (Figure 2H). For publication purposes, all visualizations can be downloaded as either Scalable Vector Graphics (SVG) or Portable Document Format (PDF) files.

Integrative analysis

The main goal of this module is to explore and unveil potential hidden correlations between the microbiome and resistome using a variety of integrative data analysis approaches. Such analyses have been increasingly used to explore the associations between the bacteria and ARGs in different environments [13, 55, 56]. Currently, ResistoXplorer offers several advanced and commonly used univariate and multivariate statistical methods for data integration and correlation analysis.

Global Similarity analysis

This section allows users to determine the overall similarity between the microbiome and resistome dataset using two multivariate correlation-based approaches - Procrustes analysis (PA) and Coinertia analysis (CIA). The analysis can be performed at various functional and taxonomic levels based on available annotations. These functions currently support five common distance measures. The ordinations from the distance matrices can be calculated using either PCoA, PCA or NMDS. In case of PA, the microbiome ordination is scaled and rotated onto the resistome ordination to minimize the sum of squared differences between the two ordinations. For the CIA, the microbiome and resistome ordination are constrained so that the squared covariances between them are maximized to measures the congruence between datasets. The results are represented using both 2D and 3D ordination plots, where samples are colored and shaped based on the datasets (Figure 3A). Users can also color the samples based on different experimental factors to identify

some patterns or gain biological insights. The corresponding correlation coefficient and P-value from both these analyses are estimated automatically to assess the strength and significance of association between the two datasets. The correlation coefficient ranges from 0 and 1, with 0 suggesting total similarity and 1 total dissimilarity between the two datasets.

Omics data integration approaches

ResistoXplorer offers other multivariate projection-based exploratory approaches such as regularized canonical correlation analysis (rCCA) and sparse partial least square (sPLS) for the integration of microbiome and resistome data. These approaches aim at highlighting the correlations between high dimensional 'omics' datasets. They are implemented primarily based on the R mixOmics package [57]. Users can choose or tune the number of components and regularization parameters for rCCA. In the case of sPLS, all variables are selected in both datasets by default. In addition to sample plots, other variable plots like clustered image maps (Figure 3C) and correlation circle plots (Figure 3D) are displayed to facilitate the interpretation of the complex correlation structure between datasets.

Pairwise microbe-ARG correlation analysis

This section enables users to determine if there are strong relationships (co-occurrence patterns) between individual microbial taxa (microbiome) and ARGs (resistome) using univariate correlation analysis. Users can perform such analysis using four different types of approaches, including Spearman, Pearson, CCLasso and Maximal Information Coefficient. In particular, features (ARGs or taxa) that are that not present in half of the samples are removed from both the datasets to eliminate an artificial association bias before performing Pearson correlation analysis. The analysis can be conducted on taxa at their different taxonomic annotations (i.e., phylum, genus

and species) and on resistome at different functional levels (e.g., class, mechanisms, ARG, etc.) based on available annotations. Users can select strong and significant pairwise correlations using a combination of absolute correlation coefficients and adjusted P-value. The results are represented as a co-occurrence network (Figure 3E) with each node indicating either a microbial taxon or a resistance determinant (ARG). The nodes can be sized based on their network topological measures (degree and betweenness). Users can double click on a node to highlight its corresponding correlated nodes in a network. The width and color of an edge indicate the strength and direction of the correlation between two nodes. The nodes are colored as well as shaped according to the dataset. The underlying correlation matrices and network centrality-based measures are also available to download.

ARGs-microbial host network exploration

ARG-microbial host association data collection

ResistoXplorer currently supports four reference databases (ResFinder [36], CARD [37], ARDB [38] and BacMet [46]) and a recently published AMP dataset [47] for network-based microbial host associations exploration of ARGs. These primary databases contain direct or indirect information regarding the microbial host for the reference ARGs. In the latter case, the information on microbial host associated with each of the reference ARGs has been collected from their corresponding GenBank Accession number using a combination of text mining and manual curation, like in ResFinder and ARDB. Further, this information has been manually annotated to improve name readability and to remove redundancy. Moreover, the available functional annotation information of ARGs were also collected and organized into sets to facilitate enrichment analysis.

ARG–microbial host association table and network creation

The uploaded list of ARGs are searched against the selected target database. This list can comprise significant ARGs detected in differential abundance testing or ARGs identified through high-throughput qPCR. The results will be represented as an association table with each row corresponding to a particular reference ARG (sequence) and its potential microbial host. When available, the table also provides other association information along with hyperlinks to the corresponding GenBank Accession number and PubMed literature. Users can directly remove each row by clicking the delete icon in the last column to keep only high-quality associations supported by literature or experimental evidence. The resulting ARG-host associations are used to build the default networks. Since not all the nodes will be connected, this approach may lead to the generation of multiple networks. The statistics of nodes and edges are provided for users to have an overview of the size and complexity of the generated networks. Further, users can also filter the nodes based on their topological measures (degree and betweenness) in case of large networks for better interpretation.

Network visualization and functional analysis

The resulting networks are visualized using HTML5 canvas and JavaScript-based powerful and fully-featured visualization system. This system is implemented based on a previously published visual analytics tool [58]. It is comprised of three main components - the central network visualization area, the network customization and functional analysis panel on the left, and the right panel containing a node table (Figure 4). Users can intuitively visualize and manipulate the network in the central area using a mouse with a scroll wheel. For example, users can scroll the wheel to zoom in and out the network, hover the mouse over any node to view its name, click a

node to display its details on the bottom-right corner or double click a node to select it. The horizontal toolbar to the top exhibits basic functions to manipulate the network. The first is the color picker to enable users to choose a highlight color for the next selection. Users can also select and drag multiple nodes by using the dashed square icon in the toolbar.

The network customization panel provides various options to configure the general visualization features of the default network or to specify the range of mouse operation. The 'Layout' option enables users to perform automatic network layout using different algorithms; the 'Background' option enables users to select between white and black background. The range of mouse operations during highlighting and dragging-and-dropping can be varied using the 'Scope' option. For instance, in 'single node' mode, only the node which has been clicked or dragged will be highlighted or affected, whereas all the nodes that are being selected by the users will be affected in the 'all highlights' mode. Additionally, the 'Download' option allows users to either save the current network in different formats or to download the network file in GraphML format for visualization in other software. The node table on the right panel displays the ARGs and their microbial hosts along with corresponding network topological measures. The corresponding abundance values, if available, will also be presented in the last column. Users can directly click on any row of interest to select, and the network view will automatically zoom to the related node. The bottom right panel provides more detailed info related to the node(s) being highlighted or currently selected on the network.

To provide some high-level biological insights about data, ResistoXplorer also supports functional enrichment analysis of the ARGs present within the current network using hypergeometric tests. The categories or levels at which enrichment analysis can be performed will be based on the initially selected database. The enrichment analysis results are shown as a table on the left panel.

By clicking on a row of the result table, users can highlight all the nodes related to enriched function within the network. There are also several other options and functionalities present, which allow users to intuitively explore, manipulate and customize the ARG-microbial host association networks.

Use Case

To illustrate the functionality of the tool, we have selected two recent resistome studies with publicly available metagenomic datasets [59, 60]. These datasets have been mounted as an example sets in the 'ARG Table' module of our tool. In Doster et al., 2018 [60], the authors have examined the effects of tulathromycin (antimicrobial drug) on gut microbiome and resistome using commercial feedlot cattle, which we describe here in more detail. Two groups of cattle were used, with one untreated group, and the other treated with metaphylaxis. Fecal samples were collected from 15 cattle within each group at two time points- baseline (Day 1) and after 11 days (Day 11). Shotgun sequencing was performed on the extracted metagenomic DNA, and reads were aligned to MEGARes and custom Kraken database for resistome and microbiome characterization by the authors. The resulting resistome abundance profile is uploaded to the ARG Table module of ResistoXplorer for further downstream analysis and exploration. Since the reads were annotated using the MEGARes database, we have directly selected the precompiled functional annotation information of the corresponding database to annotate and classify ARGs (gene accessions) at higher functional levels. All the ARGs are mapped and classified at three functional hierarchical levels - class, mechanism and group as per the MEGARes classification scheme. We first compared the resistome alpha diversity at the mechanism level; the Shannon diversity indices of the treatment group decrease over time, but the diversity changes are not prominent in the untreated group (Figure 2C). The composition profiling is carried out to explore and represent the gut

resistome of cattle. As shown in Figure 2A, the resistome composition at the class level is dominated mainly by the ARGs that confer resistance to tetracycline and the macrolide-lincosamide-streptogramin (MLS) class of antibiotics in all the samples. The hierarchical composition profiling using the Sankey diagram showed that almost all of the ARGs that belongs to tetracycline-class confer resistance through ribosomal protection proteins. In contrast, most of the ARGs that belong to the MLS-class confer resistance through macrolide efflux pumps. Moreover, the resistome composition at different functional levels was quite similar between treated and untreated cattle (Figure 2B). The ordination analysis at AMR mechanism level using PCA (CoDA-based) and NMDS indicated that the resistome composition of treated and untreated groups were significantly different between time points (ANOSIM: $R = 0.49$, P-value < 0.05; Figure 2F and 2G). We also performed differential abundance testing on fecal resistome profile using metagenomeSeq and ALDEx2 (CoDA-based) at class, mechanism and group level. No significant features were found to be differentially abundant between treatment groups using both these methods at all levels. All these analyses confirm and replicate the previous findings and results of the original publication. We also performed additional analyses of the data to highlight the utility and exploratory capabilities of ResistoXplorer. The rarefaction curve analysis indicated that enough sequencing depth was achieved to describe the true ARG richness in all the metagenomic samples (Figure 2D). The heatmap showed that most of the features (except TETQ, MEFA, TETW and TETO) at group level have a very low abundance and sparse representation across all the 60 samples. The distinct abundance pattern of TET40 group is observed when comparing the Day1 with Day11 for both the treatment groups (Figure 2G). The hierarchical clustering analysis showed that samples belonging to both treated and untreated groups are clustered effectively based on time points (Figure 2H). The ARGs belonging to tetracyclines and

MLS classes comprise the core resistome in cattle, based on their abundance and prevalence level (Figure 2E). Additionally, both Random Forest and SVM algorithms suggested that the treatment groups could not be predicted with high accuracies based on the resistome profiles of fecal samples, confirming the findings that tulathromycin does not seem to influence the gut resistome in cattle (Figure 2I and 2J).

Furthermore, the bacterial and ARG abundance profiles are uploaded to the Integration module of ResistoXplorer to explore the relationship between the fecal microbiome and resistome in cattle. The application of Procrustes analysis suggested that there were no significant overall similarities between the resistome and bacterial abundance profile (M2 = 0.23, P-value > 0.05; Figure 3A). However, the results from Co-inertia analysis indicated that resistome and bacterial composition are moderately correlated with statistical significance (RV coefficient = 0.47, P-value < 0.05; Figure 3B). To deal with the result discrepancies between approaches, we also investigated the pairwise associations between individual taxa (microbiome) and ARGs (resistome) using Spearman and Pearson correlation analysis. The results showed that no significant and strong pairwise correlations (absolute correlation coefficient > 0.7; adjusted P-value > 0.05) were identified between any taxa (phylum level) and ARG (group level) (Figure 3E and 3F), suggesting that the gut resistome was not correlated and structured by bacterial composition.

Comparison with other tools

A variety of tools or pipelines have been developed in recent years to support resistome analysis of metagenomic data [18, 36, 42, 43, 61-63]. Most of these tools have been designed primarily for raw reads processing and annotation, with limited or no support for interactive visual exploration and downstream analysis. ResistoXplorer complements these tools and resources by providing

comprehensive support for statistical, visual and exploratory analysis on the metagenomic resistome data. AMR++ Shiny [18] is one option of web or locally installable application with a graphical user interface. More recently, resistomeAnalysis [64] R package also provides the possibility for simplified visual exploration and downstream analysis of resistome abundance profile. Based on the detailed comparisons among those tools (Table 1), it is clear that ResistoXplorer offers a unique set of features and functions with regards to comprehensive statistical and exploratory data analysis, visualization, integrative data analysis and ARG-microbial network visual analysis.

Limitations and Future directions

The ARG table module can be used for visualization and analysis of resistome profiles characterizing different genetic determinants present within an AMR reference database. However, ResistoXplorer does not allow users to choose multiple precompiled functional annotation databases to analyze the entire ecologies (antimicrobial drugs, biocide, metal and other resistance drivers) of resistance determinants. More importantly, the functional annotations collected in ResistoXplorer mainly depends on the information and classification scheme present in the original databases. Hence, there might still be some acyclic and hierarchical functional annotation structure in databases, which users need to curate for accurate count-based analyses. The biocuration of ARGs and their functional annotation structure for the supported databases is beyond the scope of the study. Also, the ARG-microbial host associations module is intended for visual exploration, and these associations are limited by the type and quality of information available in the databases. Currently, ResistoXplorer offers limited functionalities and features for vertical data integration and pairwise correlation analysis. As most of the advanced approaches for performing this type of analysis on multidimensional datasets are based on computationally

intensive re-sampling (cross-validation) and permutation-based approaches to calculate statistical significance, which adds layers of complexity and computational power demands that often challenging for a real-time interactive web application. In future versions, we plan to continuously update the expand our database and analysis support for exploration of mobilome, virulome and other resistance determinants of relevance in AMR metagenomic studies.

Conclusion

Whole metagenomic sequencing studies are providing unparalleled knowledge on the diversity of resistomes in the environment, animals and humans, and on the impact of interventions, such as antibiotic use [6, 7, 9-14]. The analyses are usually exploratory in nature, and require bioinformatic skills, thus increasing costs, and preventing full data exploration. Therefore, it is critical to assist researchers and clinical scientists in the field to easily explore their own datasets using a variety of approaches, in real-time and through interactive visualization, to facilitate data understanding and hypothesis generation. ResistoXplorer meet these requirements by offering comprehensive support for composition profiling, statistical analysis, integrative analysis and visual exploration. ResistoXplorer will continuously be updated to follow the advancements in approaches for resistome analysis. We believe ResistoXplorer has the potential to find large applicability as a useful resource for researchers in the field of AMR.

Figure Legends

Figure 1: ResistoXplorer flow chart. ResistoXplorer accepts resistance gene list and ARG/taxa abundance tables as input data. Three successive steps are performed - data processing, data analysis, and result exploration. The accompanying web interface proffers a varied suite of options,

and generates several tables and graphics to enable users to intuitively go over the data analysis tasks.

Figure 2: Example outputs from ARG Table module of ResistoXplorer. (A) A stacked bar chart showing class level resistance abundance profiles across samples. (B) A Sankey diagram showing the resistome abundance profile of treated (left) and untreated (right) cattle group at hierarchical functional levels including class, mechanism and group. (C) A box plot summary of the Shannon diversity index at mechanism level in different treatment groups across sampling time points. (D) A rarefaction curve showing the number of unique ARGs identified in each sample as a function of sequence sample size. (E) A heatmap showing the core resistome of cattle analyzed at class level. (F) A 3D PCA plot with sample colors based on time points. (G) A 3D PCoA plot with sample colors with regards to different treatment groups and time points. (H) A clustered heatmap showing the variation of resistome abundance at group level in samples organized based on time point. (I) A dendrogram showing the clustering of samples with colors based on treatment and time point. (J) and (H) A graphical summary of the classification performance on different treatment groups using the Random Forests and SVM algorithm, respectively.

Figure 3: Example outputs from Integration module of ResistoXplorer. (A) A 3D NMDS plot from Procrustes analysis with samples shape and color with regards to datasets. (B) A 3D PCoA plot from Coinertia analysis, with the length of lines connecting two points indicates the similarity of samples between two datasets. (C) A clustered image heatmap showing the correlations between and among taxa (phylum level) and ARGs (group level). (D) A correlation circle plot showing the correlation structure of features (taxa/ARGs) present in two datasets. (E) and (F) A co-occurrence network showing the strong and significant pairwise correlations between taxa (phylum level) and ARGs (class level) identified using Spearman and Pearson correlation analysis, respectively.

Figure 4: A screenshot of ResistoXplorer network visual analytics system, The view is divided into three main compartments with the network visualization (toolbar on top) at the center, the node table on the right and the network customization panel together with functional annotation table on the left. Users can easily highlight and manually organize different groups of nodes based on either their annotations or connectivity patterns. It is straightforward to identify those ARGs that are found in multiple microbes, or those microbes that simultaneously contain multiple ARGs of interest.

Tables

Table 1: Comparison of ResistoXplorer with other web-based tools (except resistomeAnalysis R package) supporting downstream analysis of metagenomic resistome data. Tools exclusively dedicated for sequence annotations are not included. The URL of each tool is provided under the table.

Tools	**ResistoXplorer**	**AMR++ Shiny**	**resistomeAnalysis**	**WHAM!**
Platform	Web-based	Web-based + locally installable	R package	Web-based
Registration	No	No	-	No
Data Processing				
Data Input	Abundance tables	Abundance tables	Abundance tables	Abundance tables (Biobakery + EBI)
Functional annotation	User-defined + collected from >10 AMR databases	MEGARes	CARD	User-defined
Filtering	Abundance, variance	Abundance (Quantile)	-	Variance

Normalization	Scaling, transformation, rarefying	Cumulative Sum Scaling (CSS), rarefying	Total Sum Scaling (TSS), proportion	Proportion, clr
Composition Profiling				
Visual profiling	Stacked bar plot, stacked area, sankey diagram, zoomable sunburst, treemap	Stacked bar plot	Stacked bar plot	Interactive Stacked bar plot
Alpha diversity	Multiple	-	Richness	-
Ordination analysis	PCoA, NMDS & PCA (2D & 3D)	PCA & NMDS (Bray-Curtis) (2D)	PCoA (2D)	-
Comparative analysis				
Differential analysis	DESeq2, metagenomeSeq, EdgeR, LEfSe, ALDEx2, ANCOM	metagenomeSeq	DESeq2	ALDEx2
Classification	Random Forests, SVM	-	-	-
Other functions	Heatmaps, dendrogram, correlation, core resistome, alpha rarefaction curves	Heatmaps, Alpha rarefaction bar plots	Heatmaps, dendrogram, correlation, core resistome	Interactive Heatmaps, correlation
Integrative analysis	Procrustes, Coinertia, rCCA, sPLS, Spearman, Pearson, CCLasso, MIC	-	Spearman	Visual comparisons
Gene (ARG) List exploration	ARG-microbial host network visual analytics & functional analysis	-	-	-

ResistoXplorer: http://www.resistoxplorer.no/ResistoXplorer

AMR ++ Shiny: https://github.com/lakinsm/amrplusplus-shiny

resistomeAnalysis (R package): https://github.com/blue-moon22/resistomeAnalysis

WHAM!: https://ruggleslab.shinyapps.io/wham_v1/

Availability and requirements

Project name: ResistoXplorer

Project home page: http://resistoxplorer.no/ResistoXplorer

Operating system: Platform independent.

Programming Language: Java, R and JavaScript.

Other requirements: None.

License: None.

Any restrictions to use by non-academics: None.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work has been supported and used the computing resources at the University of Oslo.

Funding

This work has been financed by the INDNOR and INTPART programs funded by The Research Council of Norway, grant numbers 273833 and 274867, and the Olav Thon foundation.

Availability of data and materials

The ‘Beef Feedlot cattle’ dataset has been mounted to the ‘ARG Table’ and ‘Integration’ module data upload page as an example dataset and is downloaded from the published article, Doster et al., (2018) Investigating effects of tulathromycin metaphylaxis on the fecal resistome and microbiome of commercial feedlot cattle early in the feeding period. Frontiers in microbiology, 9, 1715. The ‘Pig & Broiler’ example dataset contains a subset of the original data that can be found in the supplementary information of the published article Munk P et al., (2018) Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. Nature Microbiology, 3.8, 898-908. Lastly, the example for ‘ARG List’ module is downloaded from the original article Bengtsson-Palme, J et al., (2014) Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Frontiers in microbiology, 5, 648. All these formatted example datasets can also be downloaded from the ‘Resources’ section of ResistoXplorer.

Authors contributions

AD conceptualized the tool, developed the software and wrote the manuscript. RJ aided in conceptualizing the tool, interface development, and editing the manuscript. TC aided in implementation of the tool and editing the manuscript. FCP supervised the research, and aided in conceptualizing the tool and editing the manuscript. All authors read and approved the final manuscript.

Abbreviations

AMR: Antimicrobial Resistance

ARG: Antimicrobial Resistance Gene

HTS: High-Throughput Sequencing

AMP: Antimicrobial Peptide

CoDA: Compositional Data Analysis

ZIG: Zero Inflated Gaussian

CLR: Centered Log-Ratio

ALR: Additive Log-Ratio

PCA: Principal Component Analysis

PCoA: Principal Coordinate Analysis

NMDS: Non-metric Multidimensional Scaling

RLE: Relative Log Expression

PA: Procrustes Analysis

CIA: Co-inertia Analysis

sPLS: Sparse Partial Least Square

rCCA: Regularized Canonical Correlation Analysis

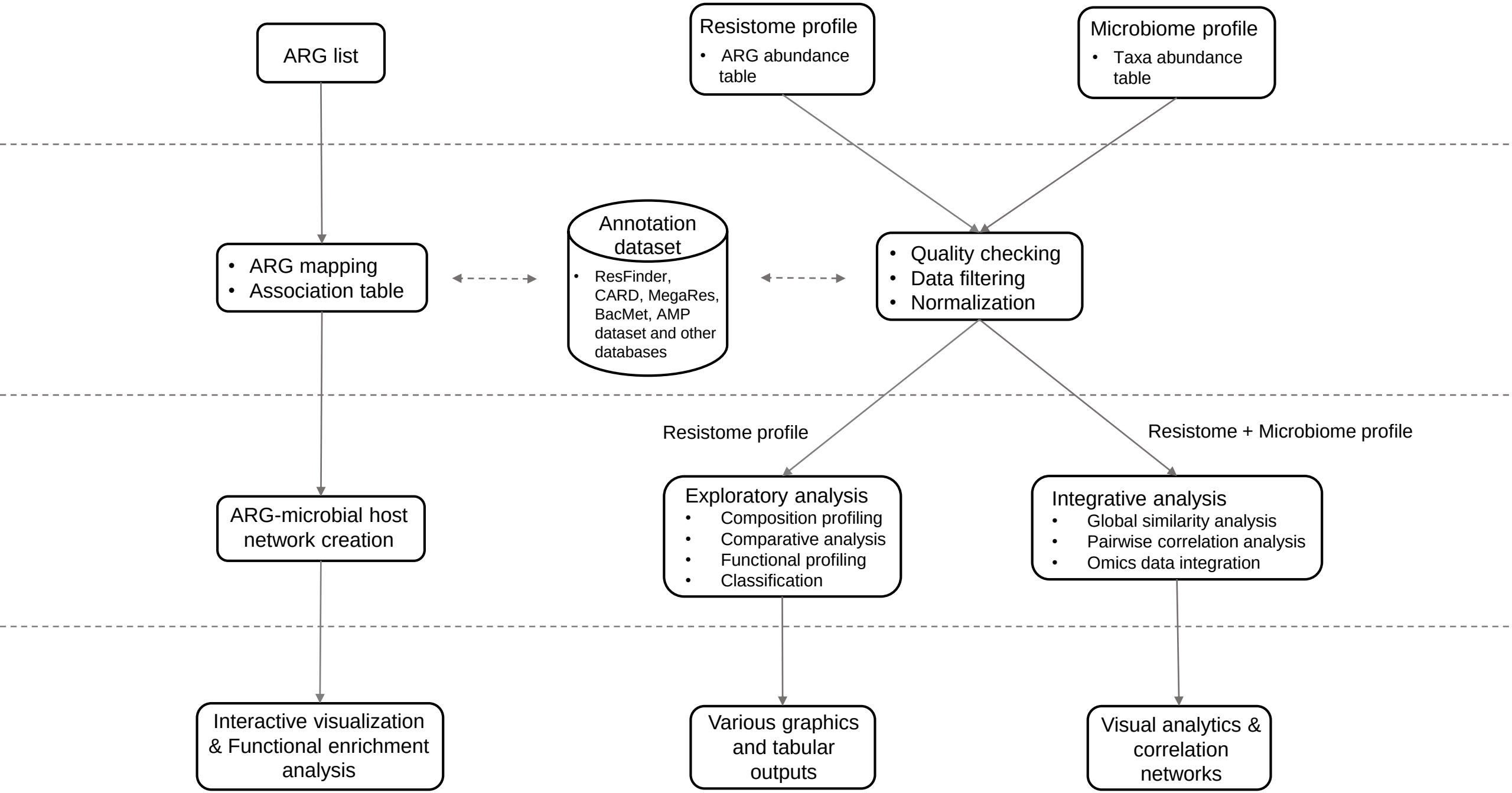
References

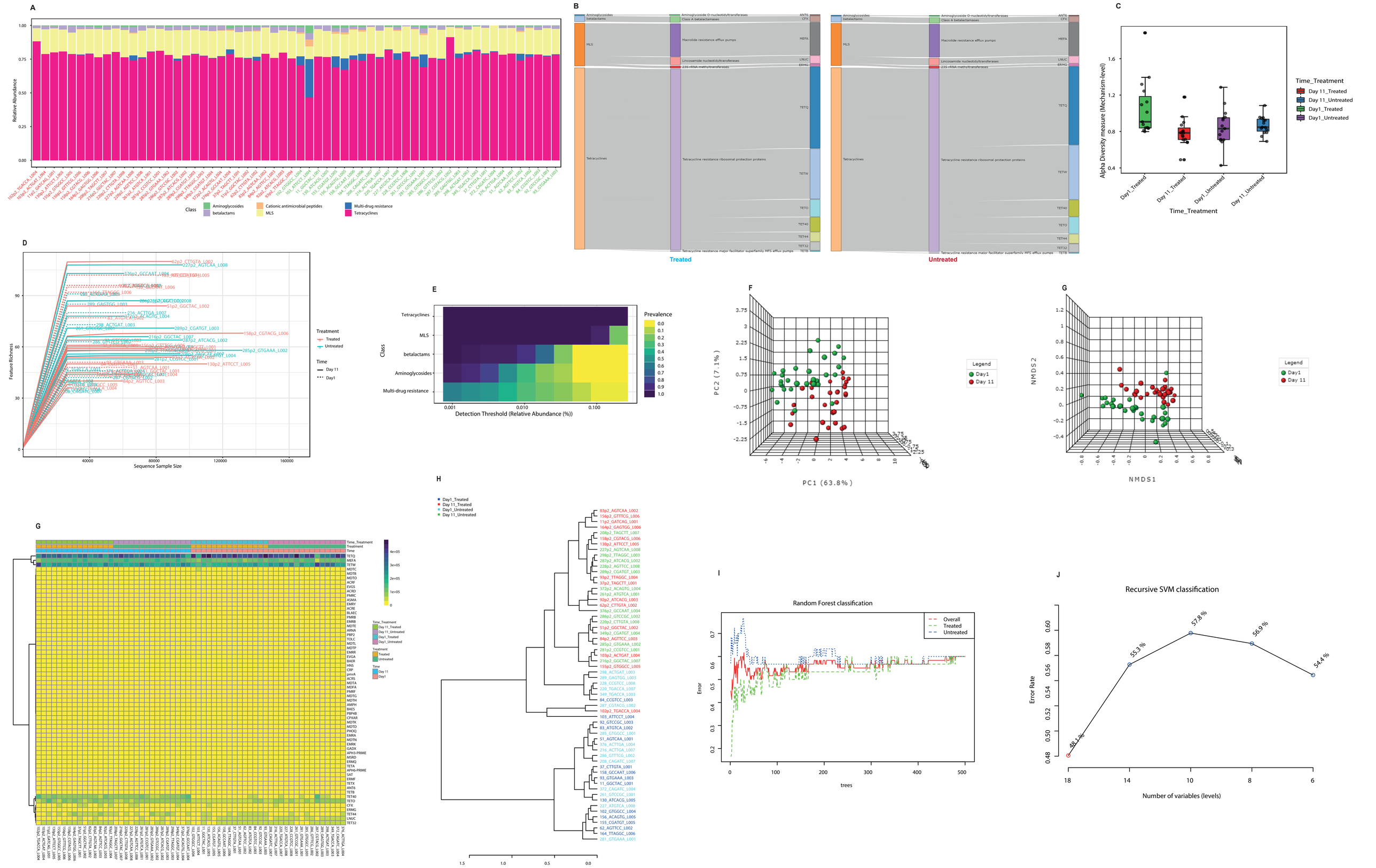
1. United IACG: **No Time to Wait–Securing the Future from Drug-resistant Infections.** *Report to the Secretary General of the Nations* 2019.
2. Kraemer SA, Ramachandran A, Perron GG: **Antibiotic pollution in the environment: from microbial ecology to public policy.** *Microorganisms* 2019, **7:**180.
3. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, Pan Y, Li J, Zhu L, Wang X: **Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota.** *Nature communications* 2013, **4:**2151.
4. D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R: **Antibiotic resistance is ancient.** *Nature* 2011, **477:**457.
5. Nesme J, Cécillon S, Delmont TO, Monier J-M, Vogel TM, Simonet P: **Large-scale metagenomic-based study of antibiotic resistance in the environment.** *Current biology* 2014, **24:**1096-1100.
6. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, Tarr PI, Warner BB, Dantas G: **Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome.** *Nature microbiology* 2019**:**1-13.
7. Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, Dantas G: **Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome.** *Nature microbiology* 2016, **1:**16024.
8. Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, Calderon M, Cabrera L, Hoyos-Arango W, Bertoli MT: **Interconnected microbiomes and resistomes in low-income human habitats.** *Nature* 2016, **533:**212-216.
9. Xia Y, Zhu Y, Li Q, Lu J: **Human gut resistome can be country-specific.** *PeerJ* 2019, **7:**e6389.
10. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, Bork P: **Country-specific antibiotic use practices impact the human gut resistome.** *Genome research* 2013, **23:**1163-1169.
11. Munk P, Andersen VD, de Knegt L, Jensen MS, Knudsen BE, Lukjancenko O, Mordhorst H, Clasen J, Agersø Y, Folkesson A: **A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds.** *Journal of Antimicrobial Chemotherapy* 2016, **72:**385-392.
12. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DJ: **The structure and diversity of human, animal and environmental resistomes.** *Microbiome* 2016, **4:**54.
13. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, Dantas G: **Bacterial phylogeny structures soil resistomes across habitats.** *Nature* 2014, **509:**612-616.
14. Hendriksen RS, Munk P, Njage P, Van Bunnik B, McNally L, Lukjancenko O, Röder T, Nieuwenhuijse D, Pedersen SK, Kjeldgaard J: **Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage.** *Nature communications* 2019, **10:**1124.
15. Boolchandani M, D'Souza AW, Dantas G: **Sequencing-based methods and resources to study antimicrobial resistance.** *Nature Reviews Genetics* 2019**:**1.
16. McArthur AG, Wright GD: **Bioinformatics of antimicrobial resistance in the age of molecular epidemiology.** *Current opinion in microbiology* 2015, **27:**45-50.

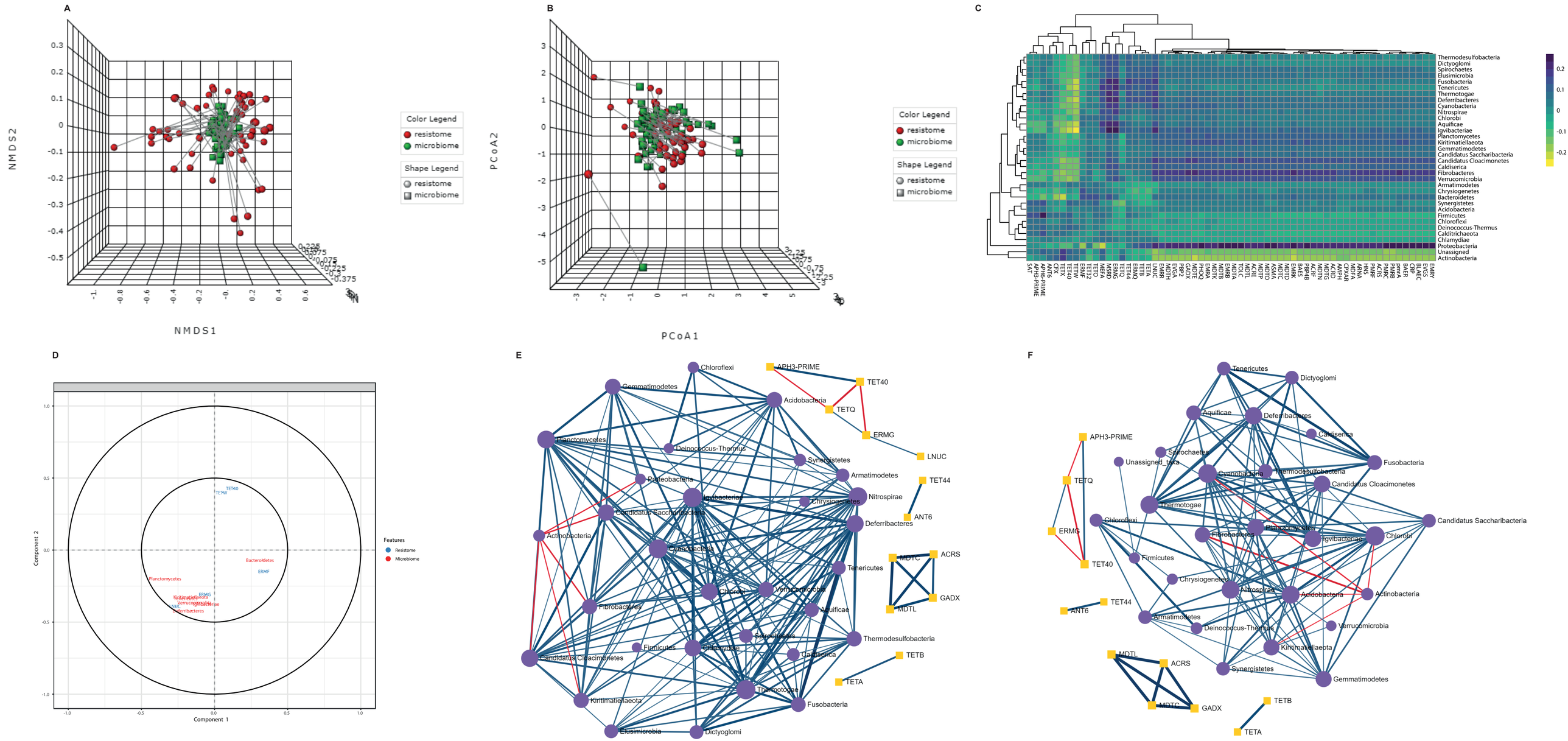
17. Hendriksen RS, Bortolaia V, Tate H, Tyson G, Aarestrup FM, McDermott P: **Using Genomics to Track Global Antimicrobial Resistance.** *Frontiers in public health* 2019, **7:**242.
18. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, Rovira P, Abdo Z, Jones KL, Ruiz J: **MEGARes: an antimicrobial resistance database for high throughput sequencing.** *Nucleic acids research* 2016, **45:**D574-D580.
19. Filzmoser P, Hron K, Reimann C: **Univariate statistical analysis of environmental (compositional) data: problems and possibilities.** *Science of the Total Environment* 2009, **407:**6100-6108.
20. Li H: **Microbiome, metagenomics, and high-dimensional compositional data analysis.** *Annual Review of Statistics Its Application* 2015, **2:**73-94.
21. Calle ML: **Statistical analysis of metagenomics data.** *Genomics informatics* 2019, **17**.
22. Paulson JN, Pop M, Bravo HC: **Metastats: an improved statistical method for analysis of metagenomic data.** *Genome Biology* 2011, **12:**P17.
23. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: **Metagenomic biomarker discovery and explanation.** *Genome biology* 2011, **12:**R60.
24. Paulson JN, Pop M, Bravo HC: **metagenomeSeq: Statistical analysis for sparse high-throughput sequencing.** *Bioconductor package* 2013, **1:**63.
25. Pereira MB, Wallroth M, Jonsson V, Kristiansson E: **Comparison of normalization methods for the analysis of metagenomic gene abundance data.** *BMC genomics* 2018, **19:**274.
26. Jonsson V, Österlund T, Nerman O, Kristiansson E: **Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics.** *BMC genomics* 2016, **17:**78.
27. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**550.
28. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
29. Quinn TP, Erb I, Richardson MF, Crowley TM: **Understanding sequencing data as compositions: an outlook and review.** *Bioinformatics* 2018, **34:**2870-2878.
30. Gloor GB, Reid G: **Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data.** *Canadian journal of microbiology* 2016, **62:**692-703.
31. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ: **Microbiome datasets are compositional: and this is not optional.** *Frontiers in microbiology* 2017, **8:**2224.
32. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM: **A field guide for the compositional analysis of any-omics data.** *GigaScience* 2019, **8:**giz107.
33. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB: **Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis.** *Microbiome* 2014, **2:**15.
34. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD, disease: **Analysis of composition of microbiomes: a novel method for studying microbial composition.** *Microbial ecology in health* 2015, **26:**27663.

35. Bengtsson-Palme J, Larsson DJ, Kristiansson E: **Using metagenomics to investigate human and environmental resistomes.** *Journal of Antimicrobial Chemotherapy* 2017, **72:**2690-2703.
36. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV: **Identification of acquired antimicrobial resistance genes.** *Journal of antimicrobial chemotherapy* 2012, **67:**2640-2644.
37. Alcock BP, Raphenya AR, Lau TT, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S: **CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database.** *Nucleic acids research* 2020.
38. Liu B, Pop M: **ARDB—antibiotic resistance genes database.** *Nucleic acids research* 2008, **37:**D443-D447.
39. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M: **ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.** *Antimicrobial agents chemotherapy* 2014, **58:**212-220.
40. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, Belk KE, Noyes NR, Morley PS: **MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data.** *Nucleic acids research* 2019.
41. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF: **Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates.** *Antimicrobial agents chemotherapy* 2019, **63:**e00483-00419.
42. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, Tiedje JM, Zhang T: **ARGs-OAP v2. 0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes.** *Bioinformatics* 2018, **34:**2263-2270.
43. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L: **DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data.** *Microbiome* 2018, **6:**23.
44. Argoty GAA, Guron GK, Garner E, Riquelme MV, Heath L, Pruden A, Vikesland P, Zhang L: **ARG-miner: A web platform for crowdsourcing-based curation of antibiotic resistance genes.** *bioRxiv* 2019**:**274282.
45. Gibson MK, Forsberg KJ, Dantas G: **Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology.** *The ISME journal* 2015, **9:**207-216.
46. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DJ: **BacMet: antibacterial biocide and metal resistance genes database.** *Nucleic acids research* 2013, **42:**D737-D743.
47. Kintses B, Méhi O, Ari E, Számel M, Györkei Á, Jangir PK, Nagy I, Pál F, Fekete G, Tengölics R: **Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota.** *Nature microbiology* 2019, **4:**447.
48. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J: **MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data.** *Nucleic acids research* 2017, **45:**W180-W188.
49. McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PloS one* 2013, **8**.

50. Dixon P: **VEGAN, a package of R functions for community ecology.** *Journal of Vegetation Science* 2003, **14:**927-930.
51. Rahman SF, Olm MR, Morowitz MJ, Banfield JF: **Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome.** *MSystems* 2018, **3:**e00123-00117.
52. Breiman L: **Random forests.** *Machine learning* 2001, **45:**5-32.
53. Burges CJ: **A tutorial on support vector machines for pattern recognition.** *Data mining knowledge discovery* 1998, **2:**121-167.
54. Zhang X, Lu X, Shi Q, Xu X-q, Hon-chiu EL, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH: **Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.** *BMC bioinformatics* 2006, **7:**197.
55. Ruppé E, Ghozlane A, Tap J, Pons N, Alvarez A-S, Maziers N, Cuesta T, Hernando-Amado S, Clares I, Martínez JL: **Prediction of the intestinal resistome by a three-dimensional structure-based method.** *Nature microbiology* 2019, **4:**112-123.
56. Liu J, Taft DH, Maldonado-Gomez MX, Johnson D, Treiber ML, Lemay DG, DePeters EJ, Mills DA: **The fecal resistome of dairy cattle is associated with diet during nursing.** *Nature communications* 2019, **10:**1-15.
57. Rohart F, Gautier B, Singh A, Lê Cao K-A: **mixOmics: An R package for 'omics feature selection and multiple data integration.** *PLoS computational biology* 2017, **13:**e1005752.
58. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J: **miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis.** *Nucleic acids research* 2016, **44:**W135-W141.
59. Munk P, Knudsen BE, Lukjancenko O, Duarte ASR, Van Gompel L, Luiken RE, Smit LA, Schmitt H, Garcia AD, Hansen RB: **Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries.** *Nature microbiology* 2018, **3:**898-908.
60. Doster E, Rovira P, Noyes NR, Burgess BA, Yang X, Weinroth MD, Lakin SM, Dean CJ, Linke L, Magnuson R: **Investigating effects of tulathromycin metaphylaxis on the fecal resistome and microbiome of commercial feedlot cattle early in the feeding period.** *Frontiers in microbiology* 2018, **9:**1715.
61. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C: **High-specificity targeted functional profiling in microbial communities with ShortBRED.** *PLoS computational biology* 2015, **11:**e1004557.
62. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE: **SRST2: rapid genomic surveillance for public health and hospital microbiology labs.** *Genome medicine* 2014, **6:**90.
63. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN: **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database.** *Nucleic acids research* 2016**:**gkw1004.
64. Carr VR, Witherden EA, Lee S, Shoaie S, Mullany P, Proctor GB, Gomez-Cabrero D, Moyes DL: **Abundance and diversity of resistomes differ between healthy human oral cavities and gut.** *Nature communications* 2020, **11:**1-10.


ARG list
Resistome profile
• ARG abundance table
Microbiome profile
• Taxa abundance table
• ARG mapping
• Association table
Annotation dataset
• ResFinder, CARD, MegaRes, BacMet, AMP dataset and other databases
• Quality checking
• Data filtering
• Normalization
Resistome profile
Resistome + Microbiome profile
ARG-microbial host network creation
Exploratory analysis
• Composition profiling
• Comparative analysis
• Functional profiling
• Classification
Integrative analysis
• Global similarity analysis
• Pairwise correlation analysis
• Omics data integration
Interactive visualization & Functional enrichment analysis
Various graphics and tabular outputs
Visual analytics & correlation networks


A
Relative Abundance
Class
Aminoglycosides
betalactams
Cationic antimicrobial peptides
MLS
Multi-drug resistance
Tetracyclines
B
Aminoglycosides
betalactams
Aminoglycoside O-nucleotidyltransferases
Class A betalactamases
Macrolide resistance efflux pumps
Lincosamide nucleotidyltransferases
23S rRNA methyltransferases
Tetracycline resistance ribosomal protection proteins
Tetracycline resistance major facilitator superfamily MFS efflux pumps
Treated
Untreated
C
Alpha Diversity measure (Mechanism-level)
Time_Treatment
Day 11_Treated
Day 11_Untreated
Day1_Treated
Day1_Untreated
D
Feature Richness
Sequence Sample Size
Treatment
Treated
Untreated
Time
Day 11
Day1
E
Tetracyclines
MLS
betalactams
Aminoglycosides
Multi-drug resistance
Class
Prevalence
Detection Threshold (Relative Abundance (%))
F
PC2 (7.1%)
PC1 (63.8%)
Legend
Day1
Day 11
G
NMDS2
NMDS1
Legend
Day1
Day 11
G
Time_Treatment
Treatment
Time
H
Day1_Treated
Day 11_Treated
Day1_Untreated
Day 11_Untreated
I
Random Forest classification
Overall
Treated
Untreated
Error
trees
J
Recursive SVM classification
Error Rate
Number of variables (levels)
48.1 %
55.3 %
57.8 %
56.9 %
54.4 %


A
NMDS2
NMDS1
Color Legend
resistome
microbiome
Shape Legend
resistome
microbiome
B
PCoA2
PCoA1
Color Legend
resistome
microbiome
Shape Legend
resistome
microbiome
C
Thermodesulfobacteria
Dictyoglomi
Spirochaetes
Elusimicrobia
Fusobacteria
Tenericutes
Thermotogae
Deferribacteres
Cyanobacteria
Nitrospirae
Chlorobi
Aquificae
Igvibacteriae
Planctomycetes
Kiritimatiellaeota
Gemmatimodetes
Candidatus Saccharibacteria
Candidatus Cloacimonetes
Caldiserica
Fibrobacteres
Verrucomicrobia
Armatimodetes
Chrysiogenetes
Bacteroidetes
Synergistetes
Acidobacteria
Firmicutes
Chloroflexi
Deinococcus-Thermus
Calditrichaeota
Chlamydiae
Proteobacteria
Unassigned
Actinobacteria
SAT
APH3-PRIME
APH6-PRIME
ANT6
CAT4
TETX
TETW
TET40
EBRB
TET32
TETO
MSRD
MSRS
MEFA
TETQ
TET44
ERMG
TETB
TETA
LNUC
EMRR
MDTH
KASA
PBP2
GADX
MDTE
FOSA
EMRA
MDTK
MDTB
EMRB
MDTA
TOLC
MDTL
ACRE
MDTP
MDTO
ASMA
MDTC
MDTD
EMRK
BAES
PBP4B
ACRF
MDTN
MDTG
ACRD
AMPH
CAT4
MDFA
HNS
ARNA
MDTF
FMRF
ACRS
PMRC
PMRB
pmrA
BAER
CRP
BLAEC
EVGS
EMRY
D
Component 2
Component 1
Features
Resistome
Microbiome
TET40
TETW
Bacteroidetes
ERMF
Planctomycetes
E
Chloroflexi
APH3-PRIME
TET40
Gemmatimodetes
Acidobacteria
TETQ
ERMG
Planctomycetes
Deinococcus-Thermus
LNUC
Synergistetes
Armatimodetes
Proteobacteria
TET44
Igvibacteriae
Nitrospirae
Chrysiogenetes
Candidatus Saccharibacteria
ANT6
Deferribacteres
Actinobacteria
Cyanobacteria
MDTC
ACRS
Tenericutes
GADX
Chlorobi
Verrucomicrobia
MDTL
Fibrobacteres
Aquificae
Thermodesulfobacteria
Spirochaetes
Firmicutes
Dictyoglomi
Caldiserica
Candidatus Cloacimonetes
TETB
TETA
Thermotogae
Kiritimatiellaeota
Fusobacteria
Elusimicrobia
Dictyoglomi
F
Tenericutes
Dictyoglomi
Aquificae
Deferribacteres
Caldiserica
APH3-PRIME
Spirochaetes
Unassigned_taxa
Fusobacteria
TETQ
Cyanobacteria
Thermodesulfobacteria
Candidatus Cloacimonetes
Thermotogae
Chloroflexi
Candidatus Saccharibacteria
ERMG
Fibrobacteres
Planctomycetes
Igvibacteriae
Chlorobi
Firmicutes
TET40
Chrysiogenetes
Nitrospirae
Acidobacteria
Actinobacteria
TET44
ANT6
Armatimodetes
Deinococcus-Thermus
Verrucomicrobia
Kiritimatiellaeota
MDTL
ACRS
Synergistetes
Gemmatimodetes
MDTC
GADX
TETB
TETA

Network Customization

Network: module1
Background: Black
View: Topology
Layout: Default Layout
Scope: Single node
Download: -- Specify -- View Options

Network customization

Function Explorer

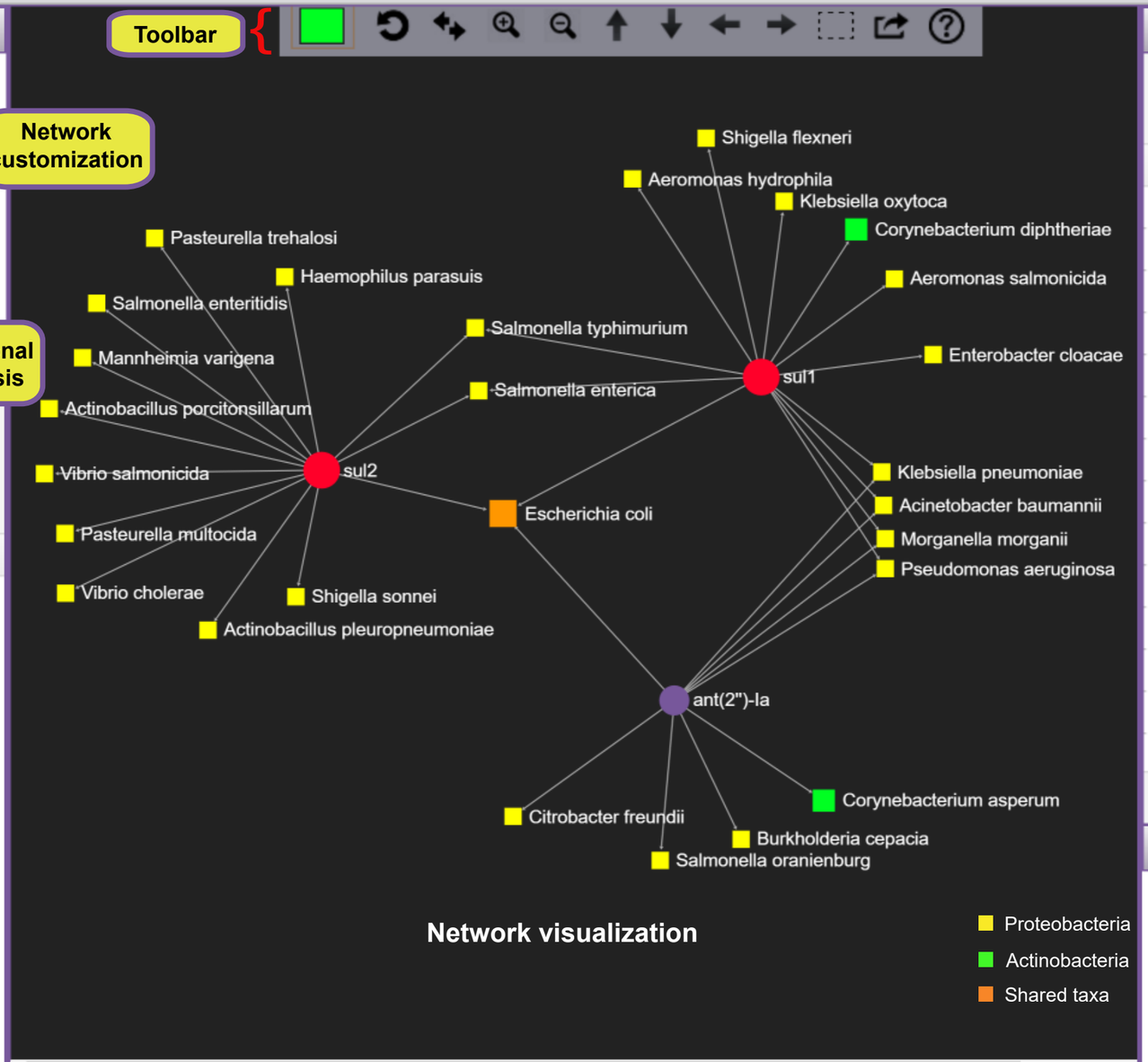
Functional analysis

Query: All genes
Algorithm: Hypergeometric test
Functional category: Class Submit

Name	Hits	Pval	Color
Folate pathway antagonist	3	0.038	
Aminoglycoside	3	0.51	
Macrolide		0.51	
Quinolone		0.764	
Beta-lactam	2	1	

Enriched Functions

Toolbar



Network visualization

Node Explorer

ID: Search Delete
Highlight: All interactions Submit

Node table

ID	Degree	Betweenness	Value
sul1	13	365.2799	713
sul2	13	391.3584	37332
dfrA1		250.1944	36
ant(2")-Ia	9	207.7581	27
Escherichia coli	8	390.4271	0
ere(A)	5	110.3828	45
aph(6)-Id	4	30.64881	3221
Klebsiella pneumoniae	4	80.18447	0
Pseudomonas aeruginosa	4	30.23452	0
Salmonella enterica	4	88.85584	0
aph(3")-Ib	3	22.3776	2595
Acinetobacter baumannii	3	54.73225	0
Shigella flexneri	3	48.23333	0
Salmonella typhimurium	3	74.51525	0

Page 1 of 2

Current Selections

- **Node**: Corynebacterium diphtheriae
- **Link**: NCBI
- **Link**: NCBI Pathogens

Node info