

CHALMERS



UNIVERSITY OF GOTHENBURG



Exploratory Metagenomic Analysis of Antibiotic Resistance Genes in Bacterial Communities

A statistical approach for classification of the bacterial resitome

Paula Andrea Martinez Villegas

MSc. Bioinformatics and Systems Biology
Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden, 2012
Report No. xxxx

Exploratory Metagenomic Analysis of Antibiotic Resistance Genes in Bacterial
Communities

A statistical approach for classification of the bacterial resitome

© PAULA A. MARTINEZ, 2012.

Technical report no xxxx.xx
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Exploratory Metagenomics Analysis of Antibiotic Resistance Genes in Bacterial Communities

A Statistical Approach for Classification of the Bacterial Resistome

PAULA A. MARTINEZ

Department of Mathematical Sciences

Chalmers University of Technology

Abstract

The increasing prevalence of antibiotic-resistant bacteria has become a notorious threat to human health. Bacteria become resistant through resistance genes that can move between cells using horizontal gene transfer. Antibiotics are naturally produced by microorganisms in the environment and therefore bacterial communities maintain a large collection of resistance genes (the resistome). The diversity and mobility of the environmental resistome is however not well studied and further research into these issues is warranted.

The aim of this project is to explore the environmental resistome and to characterize the abundance of known resistance genes in the environment, using metagenomic data. We collected 98 gigabytes of publicly available data from “The Community Database for Metagenomic Data” CAMERA, including more than 650 study sites around the world. Based on this data, we identified several common antibiotic resistance genes families spread in different environments, where the beta-lactamase TEM was the most abundant (having 41.7 % occurrence between 347 sites). We also compared the sites with clustering, and found that the resistome is highly variable. However, similarities were found also in geographically close sites, and between sites from similar environments. For instance, environments contaminated with antibiotics showed similarities in their resistome abundance. Additionally, we also clustered the resistome, observing groups of antibiotic resistance genes with similar abundance patterns between the sites. Several of these groups could be associated with genetically linked co-resistance through known horizontally transferred elements.

We conclude that metagenomics is a powerful tool for identifying antibiotic resistance genes in uncultured bacteria.

Keywords: metagenomics, environmental bacterial communities, antibiotic resistance, resistome, next generation sequencing NGS.

The report is written in English.

Acknowledgments

May 2012

I do not want the reader to bypass this page without knowing my sincere recognitions to all the people who had helped me fulfilling this stage.

Thank you Erik, (Erik Kristiansson, supervisor) for that spark during all this time in the guidance of my project, your endless ideas and a great quality of teaching.

Thank you Fredrik (Boulund) and Viktor (Jonsson) for being actively part of all the development of this project. Your input and teamwork in this thesis project made me feel part of the group and I enjoyed it.

I would also like to thank the great teachers and classmates, I met from Bioinformatics, Systems Biology, Molecular Biology programs (from Chalmers and GU), and other colleagues at the Department of Mathematical Statistics at Chalmers and the Department of Biological and Environmental Sciences at GU, by giving special thanks for sharing incomparable experiences, that help me through all this time.

My sincere gratitude to all of you who read my report, had such interesting conversations about the topic with me, and gave me so valuable comments to improve it.

Last but not least, to the members of my close family here in gbg, those beloved persons who help me every single day, by giving a reason to smile and continue, from the bottom of my heart thanks to Pablo and dear friends. And to my family that proudly awaits of my return, at home.

Personally, I can say that this project took more than just “a little bit of extra work” to be accomplished, but I am completely happy about all the process and the outcomes. It has been absolutely worthwhile and rewarding.

Thank you all!

Paula

Contents

Chapter 1 - Introduction.....	1
Why this study?	1
Aim and Objectives.....	1
How to do it?	1
Document Overview	2
Chapter 2 - Background	3
About the antibiotic resistance problem	3
The resistome	4
Metagenomics	4
Chapter 3 - Data Collection.....	6
Metagenomic data.....	6
AR Profiles.....	8
Chapter 4 - Methods and Analysis	9
SS: Software to Succeed.....	9
Data Abstraction	9
Bioinformatics Pipeline with Python	9
Final reduction	11
Statistical Analysis.....	13
Relative abundance	16
Chapter 5 - Results.....	17
1) How common are antibiotic resistant genes in the environment?	17
2) What differences can we find between sites?	19
3) Which AR genes have co-resistance?	22
Chapter 6 - Discussion and Conclusions.....	26
Discussion	26
Conclusions	28
References	30
Appendices.....	1
Interesting Facts	1
Additional graphics.....	3
Data reduction	7

Chapter 1 - Introduction

Why this study?

With the increasing use of antibiotics, bacteria show a nonstop prevalence of antibiotic-resistance mechanisms to survive. This remarkable and constantly evolving problem has negative implications to human health, ecology and economy. Briefly, bacteria become resistant through interchange of resistance genes by Horizontal Gene Transfer, HGT. Since antibiotics are naturally produced by microorganisms in the environment, bacterial communities maintain a large collection of resistance genes, called the resistome. The diversity and mobility of resistance genes in the environment is however not well studied and further research into these issues is then warranted. From this study is expected to determine the feasibility to use metagenomics in the identification of Antibiotic Resistance Genes, AR genes, in uncultured bacteria.

Aim and Objectives

This research project aims to explore the environmental resistome and to characterize it. A comprehensive effort to collect, filter and gather the data for statistical analysis will be described in this report. This is of significant importance for primary research in metagenomic analysis, for new hypothesis formulation and for future studies. Overall, the main objectives are set to answer three biological questions:

- (1) First, to quantify the abundance of AR genes relative to the number of DNA fragments studied per site. In order to appreciate how common are AR genes in different environments.
- (2) Then, to make a comparative analysis of the distribution of genes between those environments studied, including clean and polluted sites.
- (3) Finally, to explore correlations (co-resistance) between the known AR genes to speculate about how these genes may move together.

How to do it?

The analysis uses bioinformatics know-how as instrument to discern about the biological questions involved. The main focus resides on offering new methodologies to analyze complex metagenomics data. To begin with, we established a methodology for identification of known resistance genes in existing metagenomics data, using i.e. HMMER. After careful revision of the results of the homology comparison a series of reducing steps were performed with the goal of making data abstraction. Then, we provided with novel statistical methods for exploring and quantifying this resulting data, for further biological interpretation. At last, we expect that by using powerful bioinformatics tools, it is possible to use metagenomics data for identifying antibiotic resistance genes in uncultured bacteria.

Document Overview

This document is organized into six chapters, as follows:

First, Chapter 1 - Introduction, gives the motivation and clear aims of the project. Followed by, Chapter 2 - Background which contains essential biological concepts for this analysis. In Chapter 3 - Data Collection, a detailed description of the data used and its characteristics is given. In Chapter 4 - Methods and Analysis, a description of the preprocessing steps, with the purpose of abstraction are explained, and afterwards the statistical methodology used for obtaining the last results is described. In Chapter 5 - Results, a selected set of illustrations are presented to answer to the three main objectives. Finally, in Chapter 6 - Discussion and Conclusions, several points focusing on what has been part of the challenge are debated; and then thoughts after this research are presented.

In Appendices, some of the extended results are shown.

Chapter 2 - Background

This project studies biological information from a bioinformatics, statistical and computational biology approach. Some concepts and background information are required to create a good connection between the different aspects of the study. With this purpose a short review of the background of antibiotic resistance genes and metagenomics will be given in this chapter.

About the antibiotic resistance problem

Initially antibiotics were designed, developed and produced to be prescribed against bacterial infections. When used in other circumstances or inappropriate doses, antibiotics can be very harmful. Nowadays antibiotics misuse is a global problem, for example, the common misprescription of antibiotics for viral infections, is wide extended throughout the world, also the excessive use of antibiotics for animal husbandry and agriculture. Whereas, the “inconvenient truth” is that after the production of antibiotics their final destination is to be released into the environment. Here antibiotics are widely disseminated, contributing to the maintenance and new appearance of *antibiotic resistance genes*, *AR genes* (Davies & Davies, 2010).

Bacterial antibiotic resistance is the capability of bacteria to defend themselves or withstand the harmful effects of antibiotics. This capacity is transmitted through AR genes which act like packages of information that bacteria can transfer to one another. The acquisition and spread of antibiotic resistance in bacteria can occur through different mechanisms, and those are continuously and quickly evolving. Bacteria may (1) inherit natural resistance, by chromosomal mutations and then inherent those through Vertical Gene Transfer (VGT). Today is recognized as most common mechanism of transfer (2) the acquisition of resistance. This consists of genome interchange from new genetic material coming from another source, mechanism so called Lateral or Horizontal Gene Transfer (HGT) (Martinez *et. al.* 2007) (Kenneth, 2008-2012).

In turn, this second mechanism, HGT, can take place through at least three different processes to exchange genetic material in bacteria. These are transduction, transformation and conjugation. The main mechanism of HGT is thought to be *conjugation*, which occurs when two closely related bacteria have direct cell-to-cell contact, and they transfer small pieces of DNA called plasmids. *Transformation* is a process where parts of DNA, released from a dead microorganism, are taken up by the bacteria from the external environment. *Transduction* occurs when viruses living in bacteria (bacteriophages) transfer DNA between two closely related bacteria (Wright, 2007), (Park Talaro & Chess, 2011), (Kenneth, 2008-2012). A particular focus of this research project is to study the processes involving mobile elements in HGT.

HGT mechanisms of resistance can be spread through mobile elements which can also be divided in four categories (Schimieder & Edwards, 2012): (1) *enzymatic antibiotic inactivation*, is the most common category, for example beta-lactamases AR genes that inhibit beta-lactams (Penicillin). (2) Various kinds of *efflux pumps* which allow bacteria to expel different kinds of antibiotics and thus become multi-resistant. (3) Alteration of target site,

for example Penicillin Binding Proteins, PBP's which also inhibit beta-lactams, and (4) Modification of metabolic pathways, i.e. sulfonamide family of AR genes. These mechanisms will be studied along with other families of AR genes.

The resistome

Most antibiotics have been naturally produced by microorganisms from the environment. Thus one hypothesis is that resistance is a natural inherited trait which is gained from the environment and spread between bacterial communities through HGT. Although, the origin of antibiotic resistance is not deeply studied yet, some recent researches recognized that bacterial communities maintain a large collection of resistance genes, *the resistome* (D'Costa, *et al.* 2006), (Wright, 2007), (Martínez, 2008), (Davies & Davies, 2010). One can understand the resistome as all genetic elements that can potentially confer resistance (Martinez *et al.* 2007). Not surprisingly, multiple mechanisms of resistance have been observed in different natural environments, (Alonso *et al.* 2001), (D'Costa *et al.* 2006), (D'Costa *et al.* 2007), (Martinez *et al.* 2007), (Allen *et al.* 2010). These research studies show that the resistome is remarkably large in environmental bacterial communities, but still the exact size and diversity have not yet been systematically studied (Hugenholtz & Tyson, 2008).

This remains a gap of information demanding further studies to understand where resistance genes come from, how many there are, and how those evolve and spread. Some of these questions will be addressed in the results of this research. Nevertheless, for deeper understanding about the spread, acquisition and effects of antibiotic resistance genes from the environment, I recommend to use the reference list for more insights especially the journal paper "The origins and Evolution of Antibiotic Resistance" by Davies & Davies 2010 and the newly edited book "Antimicrobial resistance in the environment" by Keen & Montforts 2012. These are both excellent research studies which assess the human and ecological impacts and risks associated with the presence of AR genes.

Metagenomics

With traditional molecular biology one of the impediments to study microorganisms is that only 1% of them could be cultured by standard techniques (Riesenfeld *et al.* 2004). Metagenomics is the culture independent method, that serves to explore the vast amount of microbial information not yet studied. With metagenomics, the samples are collected directly from the environment and then sequenced to form a collection of metagenomes, assorted genomic data generated from environmental (i.e. uncultured) samples (Hugenholtz & Tyson, 2008). With metagenomics, microorganisms can be studied by observing random fragments of their genomes with a gene-centric approach (Hugenholtz & Tyson, 2008), (Kristiansson *et al.* 2009) (CAMERA, 2011). Without cultivation of individual species, metagenomics can be used to examine the interaction of numerous microbial species and communities present in different ecosystems.

This project was conducted by analyzing metagenomes, which is suitable to identify which genes are involved in cooperation and competition (Handelsman, 2004). Additional, importance of metagenomics is highlighted since it offers remarkable potential to understand better the ecology and evolution of microbial ecosystems, (Hugenholtz & Tyson, 2008), and the different roles of resistance genes in the environment (Allen *et al.*

2010). Since there is so vast information in the micro-ecosystems, metagenomics is considered especially dependent on massive amounts of sequenced data. Thus, the opportunities go hand in hand with the improvement of next generation sequencing (NGS) techniques (Desai, 2012). Metagenome-based techniques are now widely available and developing thanks to NGS techniques becoming more accessible. This arise a future for deeper understanding of the biosphere, from sole organisms to ecosystems, and ecological interactions and evolution (Quing Yun & YuHe, 2011). In this project we deal with such large complex metagenomic data, and we face the computational and statistical challenges associated with this kind of data analysis.

Chapter 3 – Data Collection

Large amount of data, retrieved from various sources, was used for this analysis. The approximate size of all the initial data constitutes ~99 GB, divided in two parts. First, metagenomic data ~98 GB and second the AR genes profiles ~61MB. A short description of the data and how it was characterized will be given in this chapter with the corresponding acknowledgments to scientist and their institutions involved.

Metagenomic data

CAMERA

About ninety-eight percent of the metagenomic data for this analysis has been retrieved from The Community Cyber infrastructure for Advanced Microbial Ecology Research and Analysis, *CAMERA* (CAMERA, 2011), (Seshadri, *et al.* 2007). This portal is a reference source of datasets, which possesses a rich and distinctive data repository of metagenomic data frequently updated. Nowadays, CAMERA is enabling scientists to consider each gene in the context of its ecology, and that is why this is the most suitable source of data for this study.

Publicly available data from CAMERA have been assessed. The size of this dataset sums up a total of 98 GB, including 49 different projects which contain more than 650 study sites around the world. For each project (collection of sites) CAMERA includes metadata associated with: habitat, sample type, location, date, and time of collection; along with other more specific details according to the project, like sequencing techniques, number of DNA fragments, type of DNA etc. Consequently, we face with important differences in the experimental design of the projects. Most important and for future discussion: time of collection, sequencing techniques, number of samples, amount of genetic information available, and others.

Some details about the metagenomic data collected from CAMERA are described in three parts, as follows.

- 1) Type of data and sequencing technology: All raw data files are nucleotide sequences in FASTA format. Each project retrieved from CAMERA contain from one to many study sites. Each of these files itself contains information about the DNA fragments from bacterial communities collected from all sites in a project with their corresponding metadata. These research projects used shotgun sequencing methods to determine the DNA from thousands of microbes simultaneously. And the two sequencing techniques used were traditional Sanger sequencing and also NGS 454 pyrosequencing.

Table 0-1 Comparison between Sanger and 454 Sequencing: a few differences among the two sequencing techniques.

Traditional Sanger	NGS 454 pyrosequencing
~700 read length	~300 read length
Less data	More data
More expensive	Less expensive
< 1 % error rate	> 1 % error rate
Process in serie	Process in parallel

- 2) Locations: Project locations are far and wide, involving the five continents: The Americas, Antarctica, Asia, Australia with Oceania, and Europe. Projects can be situated geographically close, such as in the same country, or in the same region, or in the same park, or lake, etc. For example the project Antarctica Aquatic has the largest number of sites ~160 sites, all of these had been collected through the same region. However, if we consider the large geographic extent of this region we might find vast differences. As well, we might find some slight differences in a project, for example Project HOT which has collected samples from different depths and different temperatures in the same aquatic location at Pacific Subtropical Gyre. Additionally, we have samples of bacterial communities living in host-organisms (humans and animals), which can be from the same place (country). Furthermore, some samples study similar environments in different geographically distant locations, for example project EBPRSludge studies waste water treatment plants, with one site located in US and the other in Australia.
- 3) Environmental classification: there are many possible ways of classifying all the data collected, which can be further discussed. For this analysis bacterial environmental communities of non-clinical environments were studied. The preliminary classification of bacteria includes: water, soil and air environments. Then these bacterial ecosystems can be divided by habitat in three environmental categories (National Academy of Sciences (US), 2009).

Table 0-2 Classification of environments: A short description of the metagenomic data collection in three different environments.

Natural environments	Host-associated environments	Managed environments
Aquatic environments (Marine microbial communities, open ocean, lakes, coastal, estuary, saline water, (hydro)thermal vents, hot springs, coral, mangrove, fresh water, Antarctic, bays) Soil (Sediments)	Human microbiota (gut and gastrointestinal tract) Animal microbiota (Dog, chicken, worm, whale, biofilms, and bioplankton, fossil)	Waste water treatment plants (WWTP) (among common urban waste water, and antibiotic-contaminated discharge) Untreated sewage, harbor, sludge, acid mine drainage Animal – agriculture related Air of densely populated urban buildings.

+ One

One extra project from Kristiansson, *et. al.*, 2011 was included for the analysis. This study is part of the extreme environments studied, (waste water from the production of antibiotics). Kristiansson, *et. al.* 2011 reported that high levels of antibiotic pollution increases risks for mobilization of resistance genes from the environment. This is the only study in our dataset in which antibiotic concentrations confer a potential selection pressure for antibiotic resistance; moreover this dataset can contribute to the comparison between environments. The data is available at the NCBI Sequence Read Archive under accession number SRP002078 (Kristiansson, *et. al.*, 2011).

As short description, river sediments samples were collected up and downstream from an Indian waste water treatment plant (WWTP). The location is Patancheru, Hyderabad, India, discharge zone which process releases from drug manufacturers. Also, as a control, additional samples were collected up and downstream from a Swedish sewage effluent treatment plant (located in Skövde), which is not connected to any production of pharmaceuticals, (Kristiansson, *et. al.* 2011). In total this project has 9 sites, where 7 are from India and 2 reference sites from Sweden. All DNA samples were analyzed with (parallel sequencing) 454 pyrosequencing techniques.

AR Profiles

In order to identify AR genes between the different sites a reference of AR genes is needed. For this purpose antibiotic resistance HMM profiles were used. These are a repository describing 404 multiple alignments of known resistance genes families, wrapping ~61 MB. This is an updated version of the profiles previously used in (Kristiansson, *et. al.* 2011). The HMM profiles provide information of AR genes that can be associated with resistance to several classes of antibiotics. Six major classes of antibiotics which can be inhibited by AR genes (Schimieder & Edwards, 2012), are contained within these profiles, such as beta-lactams, aminoglycosides, tetracyclines, sulfonamides, amphenicols, and quinolones. Additionally, some mobile elements for Horizontal Gene Transfer genes, such as introns, and transposons are also included.

The release 0.1-2 or AR profiles used for this project was updated on February, 2012 by Erik Kristiansson. To shortly describe how this release was generated a short explanation of the input data will be given here. The resistome profiles were created from (1) the Antibiotic Resistance Data Base (ARDB) (Liu & Pop, 2009), and (2) the nucleotide sequences, available in the database INTEGRALL (Moura *et al.* 2009). These sources are automatically generated based on sequence similarity using BLAST against NCBI RefSeq database and GenBank nucleotide database, thus the accuracy of the initial data can be further discussed. To expand this database, additional information from literature review has also been included to the database. Currently, this version contains approximately 120 manually curated profiles for resistance genes, which is a valuable step towards the cleaning of the previously mentioned databases.

Chapter 4 – Methods and Analysis

Schimieder & Edwards, 2012 unmistakably stated that “the data analysis requires both expertise in bioinformatics, and advanced computational infrastructure” for this analysis this statement was put into practice, successfully overcoming the difficulties of the data analysis. In this section all the steps performed will be shortly described, starting with the software used, the data abstraction concept, respective data reduction paces with the final gathering of the data and to end with the statistical analysis and the description of a new similarity measure.

SS: Software to Succeed

The new methodology proposed used freely-available software, such as:

- ✓ HMMER 3, High quality bioinformatics software, with significant accuracy and improved speed to do sequence search and alignment, using Hidden Markov Models (Eddy, 2011).
- ✓ Python, useful scripting programming language, having advantages such as readability, quick response and scalability (Python Software Foundation, 1990-2012).
- ✓ R, Programming language and software environment for statistical computing and graphics (R Development Core Team, 2012).
- ✓ SQL powerful statements embedded for convenience into R (sqldf package) creating automatically database layouts/schema. Remarkably fast and flexible (Grothendieck, 2011).

To create and apply an effective bioinformatics pipeline for data analysis these tools are effective in achieving the formulated goal.

Data Abstraction

Abstraction is a particular concept used to refine data and reduce it. For the purpose of this analysis, the concept is defined from the computer science perspective as follows: If an object is defined as anything that possesses certain attributes or characteristics, then abstraction can be conceptualized as the recognition of important characteristics of the object filtering out the un-wanted characteristics. In order to make this concept useful, systematization is needed with the purpose of reducing the complexity of the data. Notice that there are numerous possible abstractions depending on the data structure and the ultimate goal. For the purpose of this characterization, the three objectives of this research become the pre-defined criteria for the abstraction.

Bioinformatics Pipeline with Python

The bioinformatics pipeline was prepared using GNU/Linux environment version 2.6.32. Distributor ID: Red Hat Enterprise Server, release 6.1 (Santiago). And all tasks were accomplished using scripting capabilities of Python version 2.6.6 (sept 12, 2011).

Data preparation

To be able to use HMMER 3 potential as a sequence similarity comparison tool; the first step was to translate all DNA fragments collected from CAMERA and Kristiansson *et al.* 2011, into proteins. Since the correct reading frame for this translation is unknown, the translation was made for all 6 reading frames. In this step, for each project a new FASTA file with protein information was generated, making the original data size approximately three and a half times larger (from ~98 GB to ~348GB).

Sequence Similarity Comparisons

HMMER (*hmmsearch*) was the tool used to identify the resistome of the metagenomes studied, through an intensive search to find signatures of known AR genes. HMMER software implements probabilistic models, called profile hidden Markov models to perform sequence homology comparisons with high accuracy (Eddy, 2011). HMMER detected all sequences from the metagenomic data collection (translated into proteins) that had similarities with the AR profiles sorted according to the E-value (expectation values) (Eddy, 2011). The results of these findings are also managed with scores which represent the probability of finding certain amino acid in a given position of the protein sequence.

Systematic calls to generate HMMER outputs were built up with python scripts. To suppress the default voluminous alignment output, *--tblout* option was used in the *hmmsearch* function. Using this formatting, the results are retrieved with space-delimited tabular form, summarizing in each line the output of the detected sequence (Eddy, 2011). This organizing not only reduces the default size of the output but also facilitates the upcoming parsing. Each sequence was assed according to its E-value which is the number of hits one can expected to score by chance in a sequence database; the use of thresholds will be explained in the following.

In addition, the sequence comparison was the most time consuming task, due to the massive amount of data to handle. As input HMMER received ~348GB of protein information to be compared to ~61MB of AR profiles. After the comparison the results made up ~404 GB. The resulting data was organized in a single folder for each project, containing 404 files, one for each resistance gene profile. These were then saved as starting point of further analysis, in order to parse and filter later as convenient. These processes were run using multiple shell sessions, with *screen* command, as a first attempt to improve the speed of the comparison with a parallel call. More discussion about the time and difficulties of this task will be given in chapter 6.

Thresholds

The initial controlling reporting threshold used for the *hmmsearch* was set to 10^{-50} . This measures the statistical significance of the output, meaning that on average less than one false positive sequence will be reported per query; restating the thresholds are used to filter all those sequences that can be found just by chance. In this case the *hmmsearch* option *-E* was used to retrieving all the sequences with E-value $\leq 10^{-50}$. After this initial threshold one can decide if more filtering is needed (Eddy, 2011).

Deciding which threshold to use to test significance is always a difficult task. For this analysis three thresholds (10^{-50} , 10^{-100} , 10^{-150}) were carefully considered. This selection was also implemented with python scripts with the results of the *hmmsearch -tblout* and this was the first stage towards a substantial reduction of the data size. All three threshold (cut offs) results were saved and used for the upcoming analysis. However, in Chapter 5 Results only the results obtained after the analysis using the data reduced with threshold 10^{-50} will be shown. This threshold was decided upon sensitivity (to find all true positive resistance genes) over sensibility (with less concern over the false positives). Given the amount of data, we considered that to keep at least half of the studied projects will give better results. With this threshold 30 projects were kept for further assessment, with the next threshold 10^{-100} only 15 projects were kept and then with the stricter threshold 10^{-150} only 8 projects resulted to have AR genes. Thus, we considered that 10^{-50} threshold was a good overview of all the output files, considering that the maximum E-value was $\sim 10^{-250}$ and anything below 10^{-50} was considered irrelevant.

Gathering

Having the results from the cut off thresholds, the data was extremely reduced. The next step was to gather all the files that correspond to a project folder. For the gathering an individual file per project containing all the resistance genes profiles with matching similar sequences was created. Each file has a header with the most representative information, such as E-values, their corresponding sequence id, the site from where those sequences were collected and the resistance gene profile. Using the results of the dataset selected with threshold 10^{-50} , the optimal data reduction generated $\sim 8.75\text{MB}$. This size is $\sim 0.001\%$ of the data size used for the sequence similarity comparison, and $\sim 0.008\%$ of the initial data.

Final reduction

The final part of the data reduction, and all statistical calculations were completed using R-project version 2.14.1, released on Feb 29th, 2012, for windows environment. (R Development Core Team, 2012).

Filtering

Once having the gathered files as a summary of all results from each project, two more filtering steps were required. First, in the case that one sequence matched more than one resistance gene profile, only the sequence with the best E-value was kept for each site (in order to account all the sequences per profile only once). The second filtering was to remove all null rows and columns. This means that we additionally filtered all rows (AR genes) which did not have any presence in the studied sites, and all the sites which did not have any presence of AR genes.

The best way to accomplish these tasks was to use SQL commands. For convenience and improved speed the sqldf R package was used. Sqldf package uses SQLite embedded by default, making its use very suitable. Sqldf compiles SQL statements in R, using data frames instead of tables. The database layouts/schema is automatically created behind the

scenes, and this process is transparent and fast (Grothendieck, 2011). At this point the data size was only 3.84MB. See in the appendices the results of the filtering steps: counts per project, for all three reporting thresholds after filtering.

Quantification

At this point, having accumulated all unique sequences, the quantification started. Each project has an individual file containing information of all its sites which have matching AR genes profiles with the database. The headers of these files contain three columns, site id, profile id and the number of sequences matching the specific profile. This collection of files was named “counts per project” and in total has a size of 156 KB. This task was also completed using the Sqldf package. For example:

Table 0-1 Example of counts per project

Site id	Profile	Nr of Sequences
Site 1	AR gene A	1
Site 2	AR gene A	2
Site 4	AR gene A	10
Site 5	AR gene A	1
Site 4	AR gene B	1
Site 2	AR gene C	3
Site 3	AR gene C	1
...

Reshape

To organize the data after the counts, a very useful function from R, called reshape was used. This function reshapes a data frame transposing the rows into a total and organizing it by columns. In this case all profiles had been gathered in the first column and then for each site of a project separated columns showed the number of matching sequences. As it is understandable not all the profiles have matching sequences in all the sites and vice versa. In total this collection of files has a size of 212 KB. This task was also completed using the Sqldf package. For example using Table 3 as starting point the results of the reshape will be:

Table 0-2 Example of reshaped counts

	Site 1	Site 2	Site 3	Site 4	Site 5
AR gene A	1	2	0	10	1
AR gene B	0	0	0	1	0
AR gene C	0	3	1	0	0
...

Joined Table

The final step of data gathering was to collect all reshaped files and merge them together, in order to make a new table. This table consists of all sites containing AR genes, grouped by project in column-wise arrangement, and all AR genes as row-wise arrangement. This new table is a sum up of all the previous steps completed with the cut off of 10^{-50} . The dimensions of this so called table50 were 404 AR genes profiles (rows), and 347 sites from 30 projects out of 50. After removing those (rows) AR genes which had no matching sequences in any site, the dimensions were: 229 x 347. This table was saved in tab separated text file format, with size 171Kb, and 12 kb as RData format. (The heat map of this table can be seen in Figure 3 , results section).

This task was also divided in shorter steps. Initially to make the frame of the table all the sites ids were collected and grouped by project from the original files. This was completed using dictionaries in python. Due to differences in the naming of all sites, some manual adjustments had to be done, and it was another time consuming task. At the same time, these python dictionaries were also used to count the total number of fragments per site. This last information was necessary for the normalization of relative abundance.

Statistical Analysis

When all the processing steps had been performed, a single table with all the abstracted data was ready for the statistical analysis. The statistical methods had been implemented in R-project.

Normalization

For the normalization of number of AR genes found in each site (counts), two approaches were first suggested, but then only the first was implemented with the new similarity measure. First, having matrix A , with dimensions $m \times n$, denoted by $\mathbf{A} = [a_{i,j}]$ where $i = 1, \dots, n; j = 1, \dots, m$ (See Appendices for examples of these two normalizations)

- 1) Normalization using the ratio of DNA fragments per site.
Each count was divided by the total number of DNA fragments per site. Here a new matrix B with dimensions $1 \times m$ was used; where m is the total number of sites (a_j) and vector $\mathbf{B} = [b_{1,j}]$ where $j = 1, \dots, m$ and $b_{1,j}$ refers to the total number of DNA fragments in site j

$$\text{Normalized abundance per total nr. DNA fragments} = \frac{A}{B}$$

- 2) Normalization using the total abundance of AR genes per site. (This is included in the similarity measure, see n_j and n_j , in Equation 6). We normalized the counts in each column by the column sum. With this normalization is possible to compare the sites and vice versa also to compare the relative abundance within AR genes (when using the transposed matrix).

$$\text{Normalized abundance per colSum} = \frac{A}{\text{colSum}(A)}$$

Similarity measure

Since the proposed methodology includes the use of clustering techniques, a similarity measure that accounts for fundamental features of the dataset, must be chosen carefully. Instead of relying on well-known distance measures, a new measurement was motivated by the type of data to be analyzed. First of all, this data does not have continuous features, and also possess high amount of null counts, which will drive the mean to a lower outcome. For these known distance measures will not lead to the right conclusions, mainly due to the effect of the mean parameter. The aim of the new measurement is then to analyze and understand better this very sparse kind of data.

We a clear aim we offer a new similarity measure to analyze the metagenomic data under the previously mention conditions. This similarity measure can be seen as a variation of the standard correlation with significant transformations (Equation 1). First of all, the similarity measure uses the abundance of AR genes for each site (called counts) as input. Additionally to this abundance counts a following normalization is included in the final equation (Equation 7). To start with, the matrix to be studied should be of discrete size, having dimensions $n \times m$, where n and m do not need to be of the same size. Notice that larger dimensions will entail more computational effort.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

A benefit of this measure is an effort to assess how well two variables show a comparable relationship, even when having different sequencing depths. For example, a 0 count in a site where more ~1000 DNA fragments had been checked does not reflect the same quantity as a 0 count in another site where ~100000 fragments had been checked. This new similarity measure is a column wise combination of all the variables in the matrix, and the relationship is measured emphasizing all non-zero counts. To do this a conditional statement, called non-zero pairs (Equation 2), has been used to denote the influence of having present values in final equation. This aims to add more sensitivity to present counts in the data matrix. In other traditional measures, the zeros strongly affect the mean of the data, which is essentially avoided here.

Non-zero pairs (Equation 2): is a sum of the results of a pair column wise comparison. This identity function compares every row of two columns (j, j') giving a value of one when both compared counts are greater than zero, and a value of zero when one or both counts are zero. The result of this condition is a scalar for each two positions compared. Next, all of these scalars will be summed up in the column, resulting in the total number of non-zero pairs when comparing two columns in the matrix.

$$k(j, j') = \sum_i I(x_{ij} > 0, x_{ij'} > 0) \quad (2)$$

Additionally, we used a **flag** condition that checks the pairs that are zero, called **Flag Zero-pairs** (Equation 3). The result is a one-column-matrix with the same number of rows as the original matrix. This function compares two counts from the same row, giving a value of one when both compared counts are zero, and giving a value of zero when the counts are unlike.

$$F(j, j') = I(x_{ij} = 0, x_{ij'} = 0) \quad (3)$$

Stressing the importance of non-zero pairs along with the flag, we used a scaling factor, called \tilde{n} (Equation 4). This results in a matrix of n rows and one column, which represents the sum over all the present counts in a row of (all those that have a count greater than zero, get a value of one). The judgment for this factor is as follows, once comparing two counts it is possible to have them both with zero presence; but we need to evaluate these zeros in relation to all other counts in the same row (the relative abundance of a specific AR gene among all sites). That pair of zeros will not mean that this AR gene counts are more or less similar before comparing the counts spread in other environments. Only, if that pair of zeros lack a gene that is present in many other sites, then those counts will be set apart from others and those two should be considered more similar. As a final step for scaling, we used the term \tilde{n} , which is equal to the amount of non-zero values for each row, divided by term M , equal to the total number of columns to be compared (Equation 5).

$$\tilde{n} = \sum_j I(x_{ij} > 0) \quad (4)$$

At last the **flag** could be summarized into vector v_i (Equation 5)

$$v_i = \left\{ \begin{array}{l} \frac{\tilde{n}}{M} \text{ if } I(x_{ij} = 0, x_{ij'} = 0) \\ \text{or else } 0 \end{array} \right\} \quad (5)$$

Coming back to the original standard correlation coefficient (Equation 1), the normalization factor was omitted, removing the entire denominator squared function. Thus, this measure can have a resulting matrix which contains values in the range of all positive real numbers. In addition, the mean value part (\bar{x}, \bar{y}) was also removed, as stated before, in order to take away the influence of non-zero counts to the mean value. In this step, considerable more significance has been given to present counts in relation to non-present counts. All of this resulted in Equation 6, the squared root of a ratio between the normalized counts (normalizing by the column sum). Here each row count is multiplied by its row neighbor in the compared column, then divided by the column sum of each compared column. (Note that if one of the counts is zero then all the result will be zero, which is why we need the outcome of the flag.)

$$\sqrt{\frac{x_{ij} x_{ij'}}{n_j n_{j'}}} \quad (6)$$

The resulting Equation for the similarity measure is expressed in Equation 7

$$Similarity(j, j') = \sum_i \left[k \sqrt{\frac{x_{ij} x_{ij'}}{n_j n_{j'}}} + v_i \right] \quad (7)$$

The measure is such that the scheme is applicable to a wide variety of matrix sizes and that could be of aid to solve practical problems involving complex dimensional data.

Clustering

Clustering is a common statistical technique for exploratory data analysis, used in Bioinformatics. After measuring the similarities with the new Equation 7, we applied agglomerative hierarchical clustering analysis. To perform this step, standard functions from the stats package in R-project were used. First, *as.dist* function was used to return the distance matrix computed between the rows distances of the data matrix. Here the negative similarity matrix was the input to produce the dissimilarity distance, needed for the clustering. Then, function *hclust* performed a hierarchical cluster analysis using the dissimilarity matrix.

We used the complete linkage method to find similar objects that cluster together. With complete linkage we merge in each consecutive stage the two clusters with the maximum object-to-object distance, where the objects belong to two different clusters. In agglomerative clustering, we start with all objects as individual clusters, and then the individuals will sequentially group with the neighbors according to the linkage criterion, forming in every step a new larger cluster. At the end the distance between clusters is determined by the two most distant points in different clusters. In addition, complete linkage criterion is favorable because is less susceptible to noise.

Relative abundance

AR gene abundance in relation to the total number of fragments

The abundance in relation to the number of fragments is the first result that helps to understand how common genes are in the environment over all the metagenomic information. To obtain this, we summed up all the counts for each gene (total sum of the rows) and then divide over the total sum of the number of fragments for all sites, the formula is using matrix A and B previously specified. For each gene we will obtain the relative abundance and afterwards, the percentage of highest abundances will be shown with a bar plot (See results).

$$\text{Relative Abundance of AR genes} = \frac{\text{sumRows}(A)}{\text{Total num of fragments}}$$

Sites with the highest AR genes relative abundance

Similarly the abundance of AR genes per site can be calculated to understand the differences in the resistome of the study sites. For this analysis the total amount of AR genes found in each site is summed up (sum of the columns) and then divided by the number of fragments in that specific site. In this way is possible to retrieve the percentage of AR genes normalized by the amount of information per site (See results).

$$\text{Distribution of AR genes in different sites} = \frac{\text{sumColumns}(A)}{\text{Num of fragments per site}}$$

Chapter 5 - Results

Based on this metagenomic data we obtained many results which can lead to further biological interpretation. For the purpose of this report we selected three main results along with the aim and objectives of this project.

1) How common are antibiotic resistant genes in the environment?

We studied the abundance of AR genes in relation to the total number of fragments over all 347 sites. This analysis helps to identify how common are AR genes in the environments studied. We identified several common resistance genes families spread in different environments, where the beta-lactamase TEM class A, was the most abundant resistant variant (having 41.72% presence over 347 sites). As shown in Figure 1 the top ten most abundant genes in the environment differ considerably, being TEM overall most abundant. Additionally two PBP (Penicillin binding proteins) and SHV were found between the top most abundant, these AR genes provide mechanisms to inhibit Beta-lactams. Also we notice that AR genes sul 2 which inhibits sulfonamides is also very abundant and that is caused by the impact of the WWTP in India. More details are described in Table 5 to complement to Figure 1 and also in the next results.

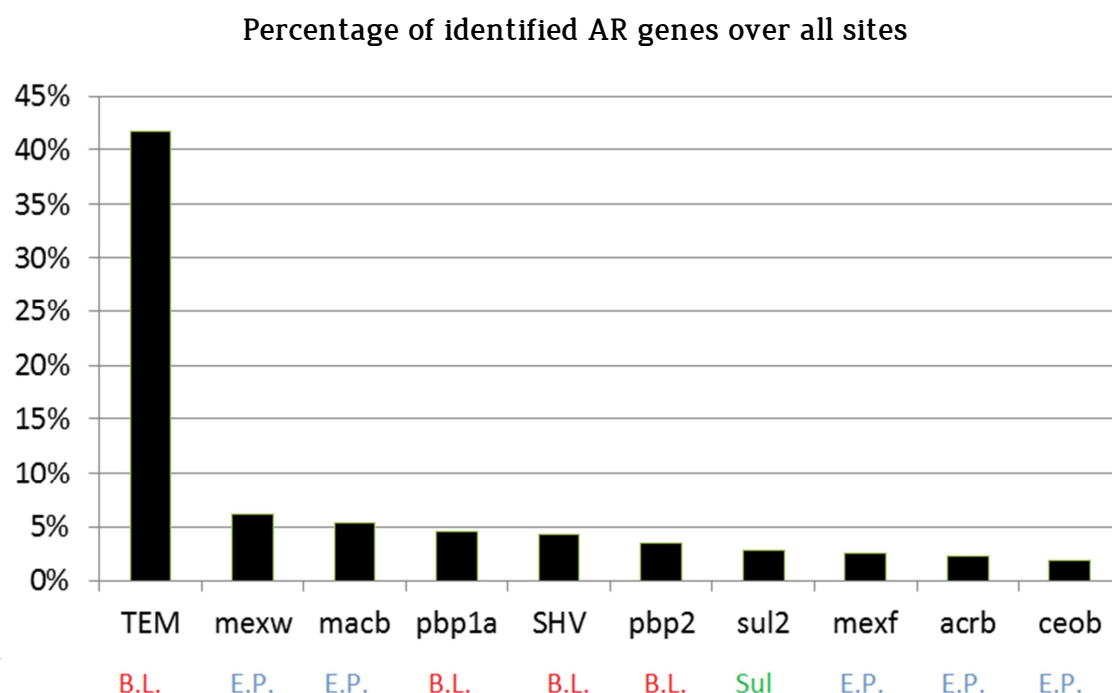


Figure 1 Top 10 relative abundance of Antibiotic Resistance Genes. The percentage in the y-axis is the ratio between the total abundance of the AR gene and the total number of fragments studied. Below the name of the AR gene the abbreviation of the family of antibiotics that are inhibited by these.

Table 0-1 Top ten most abundant resistance genes

AR gene	Description	Counts (presence over all sites)	Relative Abundance (Counts in ratio with total num of fragments)	Percentage of relative abundance
TEM	Extended Spectrum Beta lactamase, TEM, Ambler class: A, Functional class: 2b. AR gene prevalent for several years	14715	0,000121867	41,72%
mexw	Multidrug resistance efflux pump. Resistance-nodulation-cell division transporter system	2167	1,79467E-05	6,14%
macb	Multidrug resistance efflux pump. Resistance-nodulation-cell division transporter system. Macrolide-specific efflux system	1914	1,58514E-05	5,43%
pbp1a	Penicillin-binding proteins (PBPs). Resistant to β -lactams	1617	1,33917E-05	4,58%
SHV	Beta lactamase, SHV, Ambler class: A, Functional class: 2b. The predominant ESBL type in Europe and the United States, also found worldwide	1536	1,27209E-05	4,36%
pbp2	Penicillin-binding proteins (PBPs). Resistant to β -lactams	1259	1,04268E-05	3,57%
sul2	Sulfonamide-resistant dihydropteroate synthase, which cannot be inhibited by sulfonamide (sul)	1014	8,39778E-06	2,88%
mexf	Multidrug resistance efflux pump. Resistance-nodulation-cell division transporter system. Resistance to chloramphenicol, fluoroquinolone	904	7,48678E-06	2,56%
acrb	Multidrug resistance efflux pump. Resistance-nodulation-cell division transporter system. Resistance to acriflavin, aminoglycoside, beta lactam, glycylicycline and macrolide	813	6,73313E-06	2,31%
ceob	Multidrug resistance efflux pump. Resistance-nodulation-cell division transporter system. Resistance to chloramphenicol	691	5,72275E-06	1,96%

Given the details in table 5 the relative abundance of AR genes is presented as a percentage in the last column and also the y-axis of the bar plot, shown in Figure 1. In the analysis is worth to mention that many efflux pumps mechanisms are very commonly spread in the environment and that is why we see five of these AR genes among the top ten. These results are also very significant since the efflux pumps can inhibit different families of antibiotics

(which is why are called multidrug resistance efflux pumps). The environments which had contributed more to these AR gene abundances will be shown in the next Figure 2.

2) What differences can we find between sites?

We studied differences between sites corresponding to the abundance of AR genes in relation to the number of fragments per site. We identified the top ten sites having the major abundance of AR genes, encountering these as very extreme environments (See Figure 2 and details in Table 6). Figure 2 shows top ten sites according to the percentage of AR genes in a site in ratio to its total number of DNA fragments. Subsequently, in table 6 a short description of the sites and the projects they belong to is also given. This shows the high variation of AR genes between different environments and the diversity of the resistome.

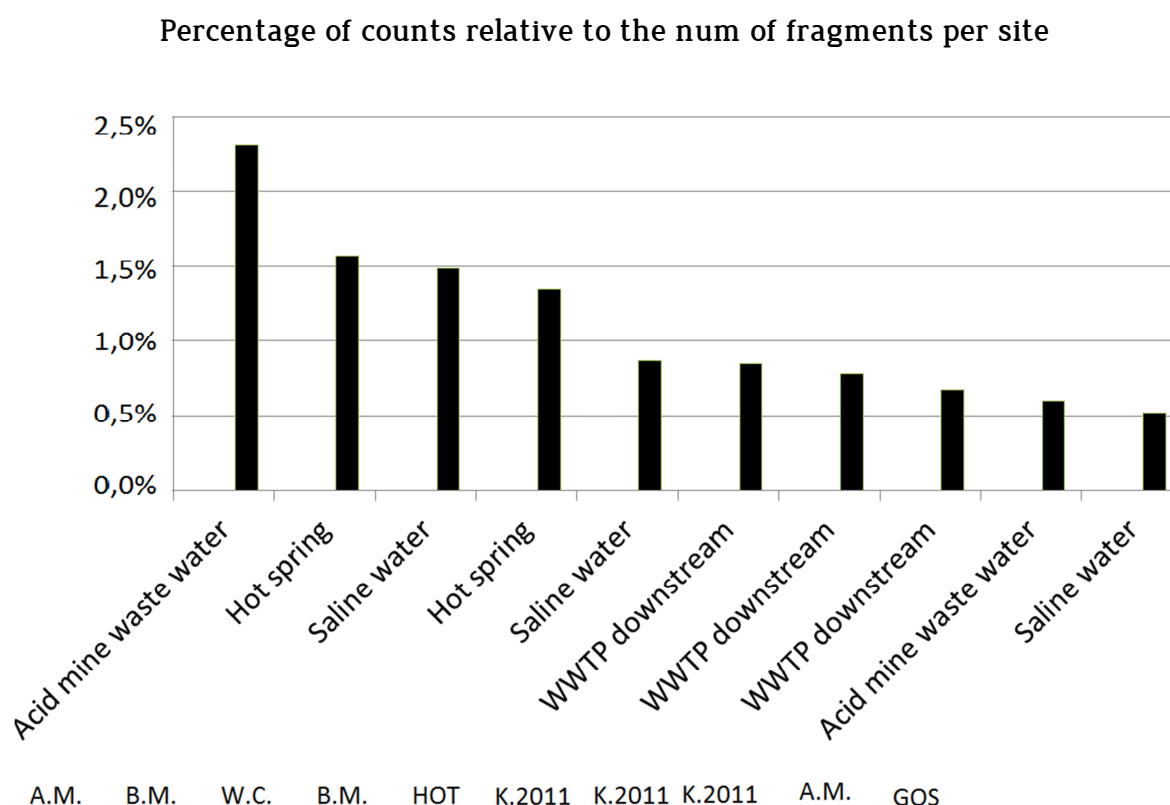


Figure 2 Top 10 sites according to the relative abundance of AR Genes in that site. The percentage shown is the amount of AR genes found in the total number of fragments studied for this specific site. Clearly Acid Mine has the highest abundance of AR genes (position 1 and 9). In the horizontal axis the descriptive name of the site and below the abbreviation of project this site belongs to, More details of the sites and project are given the next table 6.

Table 0-2 Top ten sites according to the most abundant AR genes

Project	Site	Type of Environment	Country	Sequencing method
Acid Mine Drainage Metagenome	Acid mine 02	Acid Mine Waste Water	United States	Sanger
Metagenome from Yellowstone Bison Hot Spring	Pool 3	Hot spring	United States	Sanger
Western English Channel	152.Jan Night Prefilter	Saline water	United Kingdom	454
Metagenome from Yellowstone Bison Hot Spring	Pool 4	Hot spring	United States	Sanger
HOT aloha	216 20M	Sargasso Station	Bermuda	Sanger
Kristiansson et al. 2011	KN.1766_R2	WWTP Downstream	India	454
Kristiansson et al. 2011	KN.1766_R1	WWTP Downstream	India	454
Kristiansson et al. 2011	KN.1766_R3	WWTP Downstream	India	454
Acid Mine Drainage Metagenome	Acid mine 05	Acid Mine Waste Water	United States	Sanger
Global Ocean Sampling	GS000a	Sargasso Station	Bermuda	Sanger

Figure 2 points up the top ten environments with highest abundance of AR genes. For example the Acid Mine Drainage Metagenome (Richmond Mine), located at Iron Mountain, CA in United States, has the most abundant counts of AR genes, especially TEM contributing to at least 1% of the AR genes encountered in relation to the number of fragments studied. Also there are three samples from downstream of the discharge site of antibiotics polluted environment in India, which have very high abundance of AR genes. In the next result, three highly correlated elements will be described, due to high appearance in these Indian sites. Although the number of fragments studied in this project were scarce in comparison to other major projects; due to the polluted condition of these sites it was clearly identifiable the high presence of AR genes.

In addition to corroborate both latest results a heat map of the table will show the patterns of AR genes distributed in different environments. Figure 3 shows the table 50, with 229 AR genes in the y-axis order by type of antibiotic that can be affected by the AR genes, and in the x-axis all 347 sites grouped by project.

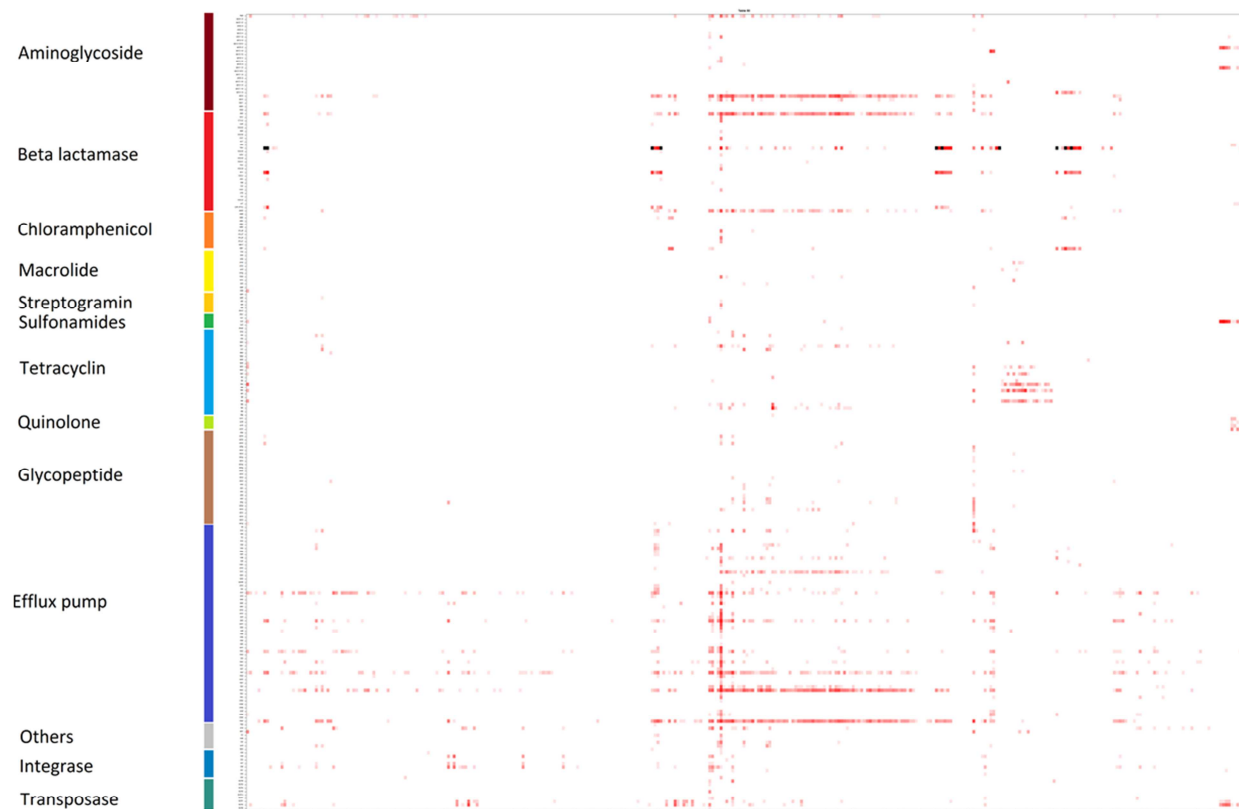


Figure 3 Heat map of Table 50 showing the distribution of AR genes in the environment. The input data for the statistical analysis is this matrix, obtained after the bioinformatics pipeline. The dimension of this matrix is 229 genes in the y-axis with genes ordered by type of antibiotic that they can inhibit. In the x-axis 347 sites are shown ordered by project name. This table shows very sparse data, all counts that are zero are shown with white and as the amount of counts increases the color also increases in the intensity, having darker red those that are more abundant. Additionally in color black the most abundant gene TEM (Beta lactamase) is highlighted as the most abundant AR gene in different environments.

From this table we can also observe, in the group of Beta lactamase, four darker lines which correspond to TEM, pbp1a, pbp2 and SHV, found in the top ten AR genes Figure 1. Additionally if we look at the columns, we can particularly observe in the beginning of the table (left side) the two acid mine sites having the highest presence of TEM. We can also note a broad range of white in the left side of the table; this is due to the presence of Antarctica sites in this region of the table, which have significantly low amount of AR genes. Other example is the clear middle red column which belongs to the western channel sites from UK. Finally, the further right sites are the India samples which also possess high abundance of AR genes. Many more details can be described from the table but we will now focus on the results of the clustering for similarities between environments.

Moreover, we also compared the similarities between sites by clustering, using the new similarity measure proposed. To show that even though the resistome has a high diversity we can find interesting similarities between the studied sites. Similarities were found in geographically close sites and sites with similar environments even if they are located far

away from each other. These sites will then cluster together when having similarities in their abundance of AR genes. Five examples are described in the following Figure 4 showing particular similarities in clusters between sites.

First, see in purple, two sites from project EBPR sludge (enhanced biological phosphorus removal), This is an example of managed environment, WWTP, urban polluted environment of two geographically apart sites clustering together, one site located in US and the other in Australia. The project had used traditional Sanger sequencing, and the data is from 2004. Then in yellow, eighteen sites from the project Twin Study are clustered together, all those samples were sequenced using 454 pyrosequencing. This is an example of a host-associated environment, from human microbiota of monozygotic twins and their mothers, which characteristics in health condition were to be lean or obese. See in orange, all sites from India discharge site are very well clustered together, even differentiating upstream and downstream sites. Example of managed environment, WWTP, antibiotic polluted environment, all sequenced with 454 NGS. Additionally, we found similarities between other sites from similar environments (even though they are physically apart). As a last example, see in light green, many saline water environments had been clustered together from the project Global Ocean Sampling Expedition GOS, the most right cluster has two sites, from two different environments, one in Panama City and the other between Madagascar and South Africa. (Example of semi natural environments) sequenced samples with traditional Sanger sequencing.

3) Which AR genes have co-resistance?

Also with clustering, we observed groups of antibiotic resistance genes with similar abundance patterns between sites, meaning these AR genes co-occur in many sites (This co-occurrence could be in the same bacteria or in different species of bacteria which cohabit in the same community). Several of these groups could be linked to known mobile elements and genetically linked co-resistance. To characterize this co-resistance we used the transposed matrix of counts (Figure 3, transposed) to first calculate the similarity between the AR genes and then with clustering significant groups were found for several AR genes. These genes are associated with resistance to several classes of antibiotics, including beta lactamases, sulfonamides, aminoglycosides, between others. Additionally, with some elements such as transposases and integrases which facilitate the mobility of these AR genes.

In Figure 4 two groups are highlighted to show examples of co-resistance. First in blue a well know group of AR genes sitting in the same plasmid. The *sul2* gene which encodes sulphonamide resistance, commonly found in *Escherichia coli* from different hosts (Trobos *et al.* 2009), is found in the same group cluster with APH(3")-Ib and APH(6)-Id, two resistance genes encoding for the aminoglycoside phosphotransferases which inactivates streptomycin by phosphorylation (Shakil *et al.* 2008). These two AR genes are also known as *strA* and *strB*. These three elements are found in the RSF1010 plasmid (Carattoli, 2009), (Yau *et al.* 2010), (Kristiansson *et al.* 2011). Furthermore, but not shown very close to this example, it is know that this group is particularly mobilized by the transposase ISCR2. From figure 4, ISCR1 and ISCR2 are both hanging from the cluster that mobilizes a broad range of other genes (to the left of this example).

The second highlight in Figure 4 is shown in green. A similar group, *sulI* also grouped with two other three genes providing resistance to aminoglycosides, and some mobile elements. There are many examples of *sulI* encountered in close relation with ISCR3 and ISCR5. As before this also a known group, but characterized more recently by Coyne *et al.* 2010 indicating that *sulI* is found in the same plasmid with ISCR3 suggesting horizontal acquisition. Also from GenBank: AB207867.1 where ISCR3 is called putative transposase in the *E. coli* transposon Tn2610, found together with *sulI*. For ISCR5 the GeneBank: DQ517526.1 plasmid from *E. coli* APEC O1 named pAPEC-O1-R shows sulfonamide resistance *SulI*, together with ISCR5. Moreover *aac(6')-Ib* (aminoglycoside) is known to be in the group of multiple AR genes together with *sulI* (Dubois *et al.* 2002) and others to provide resistance to quinolones (Robicsek *et al.* 2006). For further inference of the clusters in Figure 4 one should carefully have some experimental verification with other collaborations with additional extensive search in the literature.

Locations-Table50fFT

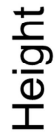


Figure 4 Cluster of sites with 4 highlights

Genes-TranspTable50fFT

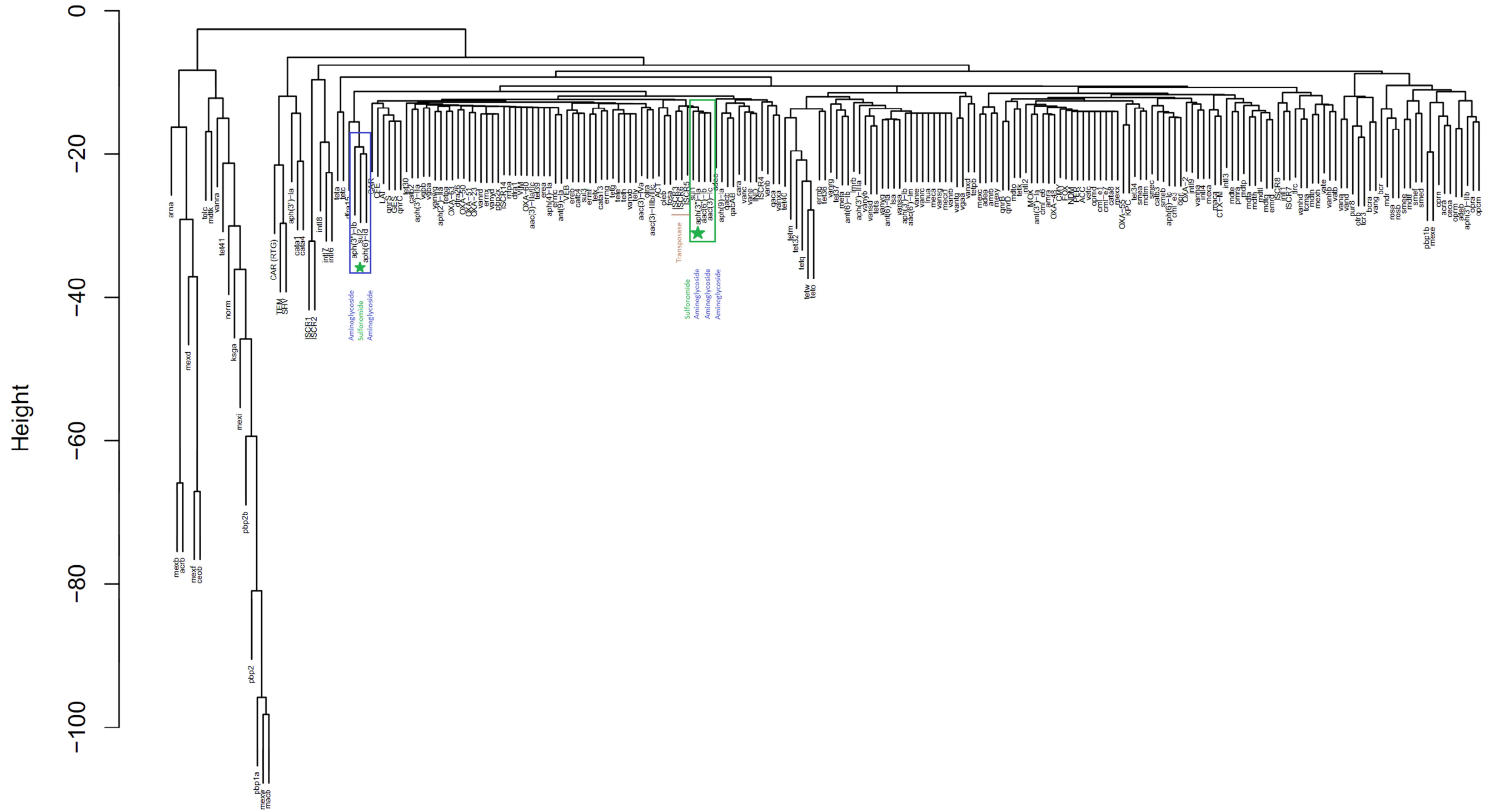


Figure 5 Cluster of Antibiotic resistance genes

Chapter 6 - Discussion and Conclusions

Discussion

Studying complex dimensional data has inherent challenges to overcome (Ruder, 2004). Having this extensive collection of data our study had multiple levels of complexity. However, as an exploratory approach to use metagenomics in search for a better understanding of the resistome, our aim was successfully fulfilled. The project had defined clear questions to be answered, driven by the initial hypothesis of having a large and diverse resistome in environmental micro communities. In order to answer these questions a computational biology scheme was put into practice first to reduce the dimensionality of the data and then to focus on the specific biological questions from a statistics perspective. We had been able to assess the data with the formulation of a new similarity measure and find out interesting results about the resistome in these environments.

At present, metagenomic analyses are in the spotlight for studying microbial communities with promising results, especially along with the improvements in NGS techniques. Essentially in data mining, of genomes of microorganisms collected from an entire microbial community, possess some obvious challenges. Also stated by Pop, 2009 metagenomics relies on large-scale sequencing efforts of entire microbial communities which undoubtedly leads to new computational challenges. Interesting, is that instead of looking to isolated genomes, metagenomics focus on the functions of the genes collected from communities. At this point the differences in the sequencing read length from the initial data will be discussed.

As starting point, approximately half of the projects used traditional Sanger sequencing and the other half used 454 pyro-sequencing. Some important differences that affect directly the analysis are, Sanger sequencing gives a read length, between 500-1000 base pairs long. With a very low error rate (which is lower than 1%). Sanger sequencing is still considered more accurate sequencing technique for analyzing complete genomes. However, 454 sequencing offers less read length, up to 300-500 base pair, it sequences in parallel and thus provides considerable larger quantities of sequences with a relative low price in comparison to Sanger sequencing. In addition, this technique does not require DNA cloning before the sequencing, removing one of the main biases in metagenomics. Read length is unquestionably crucial when it comes to assembling accurate sequences. On the other hand, the gene-centric analysis of metagenomics flairs to obtain sufficient information from a microbial community that ensures the specificity needed to compare sequenced reads against DNA or protein databases.

Also in relation to the amount of data studied the collection differs in data size. The major division is the collection of metagenomic data from 50 different projects. Then, some of projects had only one site, while others can have more than 100. When having less information (number of sites) thus little information could be retrieved from those (also depending on the sequencing technology and the amount of sequences and the read length). After the preprocessing of data, many results where encountered between larger projects or those that were more polluted (as expected). The results showed that the most abundant AR genes are very spread, but located in extreme environments (Figure 2).

Additionally to the amount of data one can discuss how the abstracted data was assessed. This was another challenging task, due to the un-limited possibilities to define the criteria of generalization. In the simplest aspect we defined the amount of sequences found for each AR gene per site. With a more complex view we could have divided this sequences first by the type of sequence methodology or even regroup the projects by similar locations, etc. Add information from metadata, which in this case is important to notice that the quality of this was not the best and it added some difficulties in the systematization.

Another point that deserves to be justified is that while undertaking the analysis, one of the projects "Gutless worm" (host-associated environment) with only one site, had to be removed from the data. Regarding the literature review (Woyke, 2006) it was found that the plasmids were amplified with AR genes encoding chloramphenicol resistance. During the cloning step it can be inferred that some error had made these AR genes to be amplified and probably account for higher counts than normal. To prevent deviations in our results, it was consider best to exclude this project (and its unique site) from the analysis.

Significantly, these sorts of metagenomic analyses are totally dependent on the quality and amount of AR genes to be used as a reference. In this case from our 404 AR-profiles (version 0.1-2 of the AR database) the quality is being improved with some literature review and the amount is also in process to add some other data from collaborators and new findings of the group. One can notice that with this approach no new AR genes will be found; and only those that are part of the data base will be found and characterized. Also to clearly identify the suggested co-resistance one need to perform the experimental validation which is why in bioinformatics the collaboration with experimentalist is so important.

Also as relevant the different thresholds are determinants when analyzing the results from the homology comparison. For this preliminary analysis three different thresholds were selected in a pure exploratory approach, selected in judgment with the global output. These thresholds can be considered ad hoc for the purpose of this analysis, in which the data itself was the main driver. After the careful judgment to select the thresholds we analyzed the data with the clustering. From these three, we decided that threshold E-50 was sensible enough to encounter all possible true positives AR genes, even though we could lose some significance while adding some false positives to the end results.

The last point of the discussion corresponds to the statistical analysis. The need of a new measure was emphasized by the sparse matrix. The main goal of this measure was to affect the statistical power of detecting less abundant resistance genes as well as assessing other slightly less counts of resistance. One of the discussions was derived from the presentation in which it was observed the importance of null-pairs to add the scaling factor, but perhaps another beneficial improvement could be to add the importance to those pairs that possessed one null and a low value. Nevertheless the significance of this approach was proven effective when finding the expected outputs about those projects from which we had more information. The similarity measure has a pragmatic formulation and it is mathematically tractable. The main drawback of this method, as other similar implementations, is that the data size has a direct exponential relationship with the computational time required for the analysis.

Conclusions

Our effort to analyze the resistome by using metagenomics data was proven very powerful. A new workflow to investigate this large metagenomic data was developed and it helped to reveal such interesting results and also validate our findings with previous research. We used HMMER as software for sequence comparison which uses statistical models to find the close related alignments giving to the outcome high statistical accuracy. Also we refined the results by filtering the output with python scripts. This, to ensure the quality of the first step, identification of AR genes with the reference data base of AR profiles.

After making a notorious reduction of the data we developed a novel statistical measure aimed to fit the sparse matrix in the analysis considering the large variety of data involved. This step was verified to be useful when validating the results with the clustering analysis and finding expected patterns. Undoubtedly, significant differences in the resistome of the investigated sites were shown. Yet we conclude that the resistome is extensive, adaptable and it can be characterized. We also conclude that is feasible to use metagenomics to identify antibiotic resistance genes in uncultured bacteria, in various environments.

Future work

During the forthcoming years it is predicted that next generation sequencing and metagenomics are going to generate massive amounts of data. Our effort to be pioneers in developing new tools and methods to study this data put us in a standing position for others to benefit and improve from our research. Additionally, significant value is given to exploratory analyses for the future generation of new hypotheses. There are many depth levels to gain with the prospectus of metagenomics.

In a turned direction, a common belief about antibiotic resistance is simply associated to the defense response to the presence of a given antibiotic. However, recent work indicates that the situation is more complex. Other alternative functional roles for resistance elements are now being proposed, and need further studies. For example signal molecules to be aware of antibiotic molecules that could shape the structure of microbial communities. This important behavior could be used in bioremediation and similar applications. Also, other research could focus on the adaptation of microbial communities for nutrient poor environments and the consequences of these changes.

Thoughts

Over the years antibiotic resistance mechanisms have been increasing worldwide, and our generation faces this problem with potential damage to human health, and to the environment. Remarkably, this study revealed that multidrug resistant genes are among the most common AR genes found in the environment. This could give us insights about the (miss) use of antibiotics and the consequences of the increment of antibiotics and antibacterial in the last decades. Significantly, we showed that multiple classes of resistance genes are promoted by the selective pressure of highly contaminated sites.

Finally, the study of the resistome requires not only gene-sequence analyses but also rigorous follow-up experimental validation. The practice obtained in this project has fundamental importance due to the future applications of the methodology applied, which can be used for metagenomic analyses with other sources of information and with other hypotheses.

References

- 1) Allen, H., Donato, J., Wang, H., Cloud-Hansen, K., Davies, J., & Handelsman, J. (2010). Call of the wild: antibiotic resistance genes in natural environments. *Microbiology* , 251-259.
- 2) Alonso, A., Sanchez, P., & Martinez, J. (2001). Environmental selection of antibiotic resistance genes. *Environ Microbiology* , 3, 1-9.
- 3) CAMERA. (2011). *CAMERA Education - The Community Cyber infrastructure for Advanced Microbial Ecology Research and Analysis*. Hämtat från CAMERA stands for Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: <http://camera.calit2.net/about/index.shtm> den 17 January 2012
- 4) Carattoli, A. (June 2009). Resistance Plasmid Families in Enterobacteriaceae. *Antimicrobial Agents Chemotherapy* , 53 (6), ss. 2227-2238.
- 5) Coyne, S., Courvalin, P., & Galimand, M. (May 2010). Acquisition of multidrug resistance transposon Tn6061 and IS 6100-mediated large chromosomal inversions in *Pseudomonas aeruginosa* clinical isolates. *Microbiology* , 156 (5), ss. 1448-1458.
- 6) D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W., Schwarz, C., o.a. (2011). Antibiotic resistance is ancient. *Nature* (477), 457-461.
- 7) Davies, J., & Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. (A. S. Microbiology, Red.) *Microbiology and Molecular Biology Reviews* , 74 (3), 417-430.
- 8) D'Costa, V. M., McGrann, K. M., Hughes, D. W., & Wright, G. D. (2006). Sampling the antibiotic resistome. *Science* , 374-377.
- 9) D'Costa, V., Griffiths, E., & Wright, G. (2007). Expanding the soil antibiotic resistome: exploring. *Current Opinion in Microbiology* , 481-489.
- 10) Desai, N. A. (2012). From genomics to metagenomics. *Current opinion in biotechnology* , XXIII (1), 72-76.
- 11) Dubois, V., Poiriel, L., Marie, C., Arpin, C., Nordmann, P., & Quentin, C. (March 2002). Molecular Characterization of a Novel Class 1 Integron Containing bla_{GES-1} and a Fused Product of aac(3)-Ib/aac(6")-Ib" Gene Cassettes in *Pseudomonas aeruginosa*. *Antimicrobial Agents Chemotherapy* , 46 (3), ss. 638-645.
- 12) Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology* , VII (10), 1-16.
- 13) European Commission Directorate General for Research. (1995-2012). *European Commission > Research > European Research in Action > Antibiotics resistance*. Hämtat från European Commission > Research: http://ec.europa.eu/research/leaflets/antibiotics/page_32_en.html den 25 April 2012
- 14) GeneBank. (den 20 December 2010). *Escherichia coli transposon Tn2610, complete sequence*. Hämtat från GeneBank Nucleotide: <http://www.ncbi.nlm.nih.gov/nucleotide/AB207867>
- 15) GeneBank. (den 2 July 2007). *Escherichia coli APEC O1 plasmid pAPEC-O1-R, complete sequence*. Hämtat från GeneBank Nucleotide: <http://www.ncbi.nlm.nih.gov/nucleotide/DQ517526>
- 16) Grothendieck, G. (2011). *sqldf: Perform SQL Selects on R Data Frames*. Hämtat från <http://CRAN.R-project.org/package=sqldf>
- 17) Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol* , 68:669-685.
- 18) Hugenholtz, P., & Tyson, G. W. (2008). Metagenomics. *Microbiology. Nature* , 455 (25), 481-483.
- 19) J. Craig Venter Institute. (2012). *Research*. Hämtat från J. Craig Venter Institute: <http://www.jcvi.org/cms/research> den 2 April 2012
- 20) Keen, P. L., & Montforts, M. H. (2012). *Antimicrobial Resistance in the environment*. Hoboken, New Jersey, US simultaneously in Canada: Jhon Wiley & Sons, Inc.
- 21) Kenneth, T. (2008-2012). *Bacterial Resistance to Antibiotics*. Hämtat från Todar's Online Textbook of Bacteriology: http://textbookofbacteriology.net/resantimicrobial_3.html den 25 January 2012

- 22) Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutegersson, C., Weijdegård, B., o.a. (2011). Pyrosequencing of Antibiotic-contaminated River Sediments Reveals High Levels of Resistance and Gene Transfer Elements. (F. Rodriguez-Valera, Red.) *PLoS ONE* , 6 (2), 1-7.
- 23) Kristiansson, E., Hugenholtz, P., & Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* , 2737-23738.
- 24) Liu, B., & Pop, M. (2009). ARDB–Antibiotic Resistance Genes Database. *Nucleic Acids Research* , 443–447.
- 25) Martínez, J. L. (2008). Antibiotics and Antibiotic Resistance Genes in Natural Environments. *Science* , 365-367.
- 26) Martinez, J., Baquero, F., & Andersson, D. (December 2007). Predicting antibiotic resistance. *Nature Reviews Microbiology* .
- 27) Martinez, J., Fajardo, A., Garmendia, L., Hernandez, A., Linares, J., Martínez-Solano, L., o.a. (2009). A global view of antibiotic resistance. *FEMS Microbiology Reviews* , XXXIII (1), 44-65.
- 28) Moura, A., Soares, M., Pereira, C., Leitão, N., & Henriques, I. a. (2009). INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* , XXV (8), 1096-1098.
- 29) National Academy of Sciences (US). (2009). *Metagenomics*. Hämtat från Metagenomics and out Microbial Planet: <http://dels-old.nas.edu/metagenomics/> den 17 January 2012
- 30) National Research Council based on the Metagenomic Committee's report. (2007). *The New Science of Metagenomics. Report in Brief*. Washington, D.C., United States: Oresident's Circle Communications Initiative of the National Academies.
- 31) Park Talaro, K., & Chess, B. (2011). *Foundations in Microbiology* (8th uppl.). McGraw-Hill.
- 32) Philippon, A., Labia, R., & Jacoby, G. (1989). Extended-spectrum beta-lactamases. *Antimicrob Agents and Chemotherapy* , XXXIII (8), 1131–1136.
- 33) Piddock, L. (August 2006). Multidrug-resistance efflux pumps - not just for resistance. *Nat Rev Microbiol.* , 629-636.
- 34) Pop, M. (den 29 may 2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics* , 10 (4), ss. 354-366.
- 35) Python Software Foundation. (1990-2012). *About Python*. Hämtat från Python Programming Language – Official Website: <http://www.python.org>
- 36) Quing Yun, Y., & YuHe, Y. (2011). Metagenome-based analysis: A promising direction for plankton ecological studies. *Sci China life Sci* , 54 (1), 75-81.
- 37) R Development Core Team. (2012). *R Project, Language and Software for Statistical Computing*. Hämtat från <http://www.r-project.org>
- 38) Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. *Annual Review of Genetics* , 325-352.
- 39) Robicsek, A., Strahilevitz, J., Sahm, D. F., Jacoby, G. A., & Hooper, D. C. (August 2006). qnr Prevalence in Ceftazidime-Resistant Enterobacteriaceae Isolates from the United States. *Antimicrobial Agents Chemotherapy* , 50 (8), ss. 2872-2874.
- 40) Ruder, K. (den 4 March 2004). *Exploring the Sargasso Sea*. Hämtat från Genome News Network: <http://www.genomenewsnetwork.org/articles/2004/03/04/sargasso.php> den 15 March 2012
- 41) Schimieder, R., & Edwards, E. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiology* , VII (1), 73-89.
- 42) Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., & Frazier, M. (2007). CAMERA: A Community Resource for Metagenomics. *PLoS Biology* , V (3), 0394-0397.
- 43) Shakil, S., Khan, R., Zarrilli, R., & Khan, A. U. (2008). Aminoglycosides versus bacteria – a description of the action, resistance mechanism, and nosocomial battleground. *Journal of Biomedical Science* , 15 (1), ss. 5-14.
- 44) Toleman, M. A., Bennett, P. M., & Walsh, T. R. (2006). ISCR elements: novel gene-capturing systems of the 21st century? *Microbiol. Mol. Biol. Rev.* , 70, 296–316.
- 45) Trobos, M., Lester, C. H., Olsen, J. E., Frimodt-Møller, N., & Hammerum, A. M. (2009). Natural transfer of sulphonamide and ampicillin resistance between *Escherichia coli* residing in the human intestine. 63 (1), ss. 80-86.

- 46) WHO . (2001). *GLOBAL STRATEGY FOR CONTAINMENT OF ANTIMICROBIAL RESISTANCE*. Geneva, Switzerland: World Health Organization WHO/CDS/CSR/DRS/.
- 47) Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., o.a. (den 26 October 2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* , 443, ss. 950-955.
- 48) Wright, G. D. (2007). The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology* , 5, 175-186.
- 49) Yau, S., Liu, X., Djordjevic, S. P., & Hall, R. M. (December 2010). RSF1010-Like Plasmids in Australian *Salmonella enterica* Serovar Typhimurium and Origin of Their *sul2-strA-strB* Antibiotic Resistance Gene Cluster. *Microbial Drug Resistance* , 16 (4), ss. 249-252.

Appendices

Interesting Facts

"Therapeutic use of antibiotics in humans, accounts for less than half of all applications of antibiotic produced commercially" (Davies & Davies, 2010).

"Only 40% of Europeans know that antibiotics are ineffective against viruses". (Eurobarometer 2001), (European Commission Directorate General for Research, 1995-2012)

"Lack of knowledge is a major factor responsible for inappropriate antimicrobial use globally." (WHO , 2001)

TEM was named extended spectrum β -lactamases (ESBLs) by (Philippon, Labia, & Jacoby, 1989)

"Multidrug-resistance efflux pumps can confer resistance to natural substances produced by the host. In addition, some efflux pumps have been shown to have a role in the colonization and the persistence of pathogenic bacteria in the host." (Piddock, 2006)

"Genome sequencing endeavors to reveal a plethora of resistance genes that are present in all bacteria." (Wright, 2007)

One example of antibiotic resistance determinants with a functional role other than affording resistance is provided by multidrug resistance (MDR) efflux pumps, other functions are to contribute to virulence, to maintaining homeostasis and to the detoxification of intracellular metabolites, among other functions (Martinez *et al.* 2009).

J. Craig Venter, President of (J. Craig Venter Institute, 2012), has invaluable contributions in genomic research. For this project some of these research data had been used, such as:

- 1) Global Ocean Sampling Expedition, circumnavigating the globe and collecting metagenomic samples throughout.
- 2) The pilot project, conducted in the Sargasso Sea.
- 3) The Human Microbiome Project (HMP) initiated by National Institutes of Health, NIH US now continued with the help of JCVI.

In the AR data base, we had some transposons, integrons and plasmids which are mobile elements (Keen & Montforts, 2012). These elements help to spread resistance, transposition includes the well-known transposons flanked by inverted repeat insertion elements and other elements such as ISCR elements that only require one insertion element for gene mobilization (Toleman, Bennett, & Walsh, 2006).

The remarkable ability of bacteria to mobilize genes and the selective pressure provided by antibiotics, conspire to facilitate the distribution of antibiotic resistance genes throughout microbial populations. As a result, the resistome expands even in the absence of continuous selection. (Wright, 2007)

The integrons (Stokes & Hall, 1989) are modular structures with a major role in the development of antibiotic resistance. These gene-recruitment determinants have a high impact in clustering antibiotic resistance genes, which may later be transferred to pathogens. (Martinez *et al.* 2009) Type 1 integrons contain a conserved gene coding for resistance to sulphonamides (Stokes & Hall, 1989).

It is important to note that sulphonamide resistance is still widespread in the United Kingdom despite national prescribing restrictions (Enne *et al.*, 2001), probably because of the success of these integrons in recruiting other antibiotic resistance genes.

Additional graphics



Figure 6 some of the sites around the world

In this plot, I joined together more than 150 sites out of the 650 study sites. Just with the sites which had clear named locations in the metadata. Many of the points that are not shown in this picture have either coordinates (which need another type of plotting, and thus cannot be combined in this plot) or have no location at all. Additionally, some of the points have very few information of the sites, for instance the project Antarctica has 160 sites where most of those just have a location "Antarctica" this continent has a size of 14,000,000 km² which of course makes the plotting more difficult, because of lack of specific metadata. Very interesting, is to note all the sites that do not have any collection of metagenomes, and this could be an starting point to plan new projects.

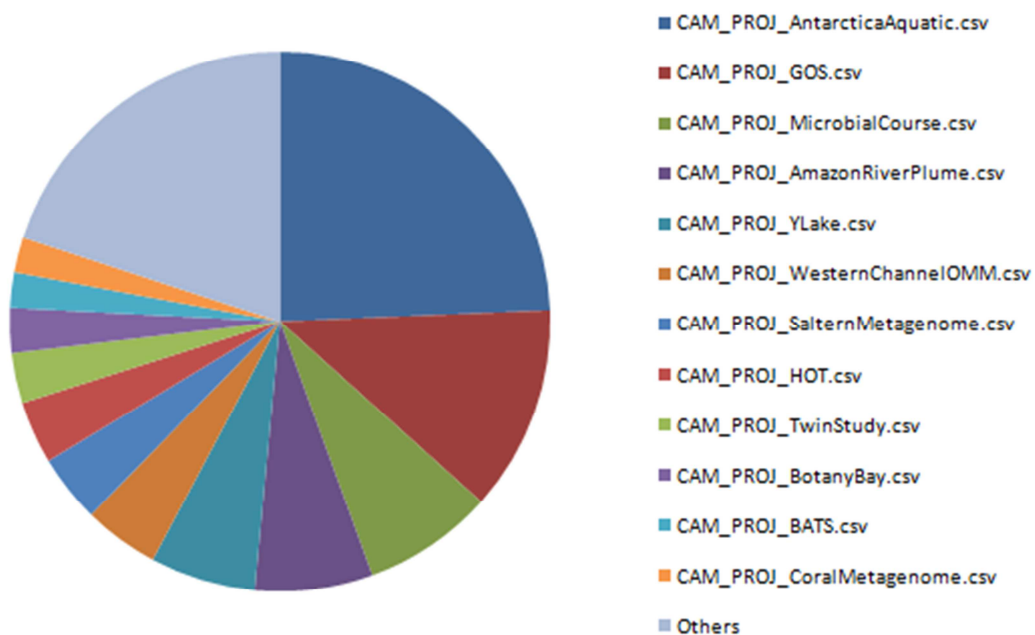


Figure 7 Sites by number of projects

As mentioned in the previous figure, Antarctica accounts for almost 25% of all the data collection, being used as a reference project. We believed that Antarctica is the cleanest environment that we had studied, considering this continent is permanently covered with a thick layer of ice and no significant (anthropogenic) pollution has affected these lands. The second largest project is the Global Ocean Sampling Expedition, voyage of the Sorcerer II visiting 23 different countries and island groups on four continents. Funding for this came from the J. Craig Venter Science Foundation (now JCVI). As can be seen 10 of the 50 projects provide almost 45% of all the metagenomic information studied.

Similarly the next bar plot (Figure 9) shows the projects which provide more information to the study, this time by summing up all the sequence information. Again Project Antarctica has the highest amount of sequences (which we always keep in mind is totally dependent on the sequence technology). Figure 9 goes together with Figure 10 which reflects the amount of AR genes found in the project and one can see that even when having such a large sample like Antarctica not much AR genes are found, or on the other hand a small sample like India WWTP, being a polluted site has found a vast amount of AR genes.

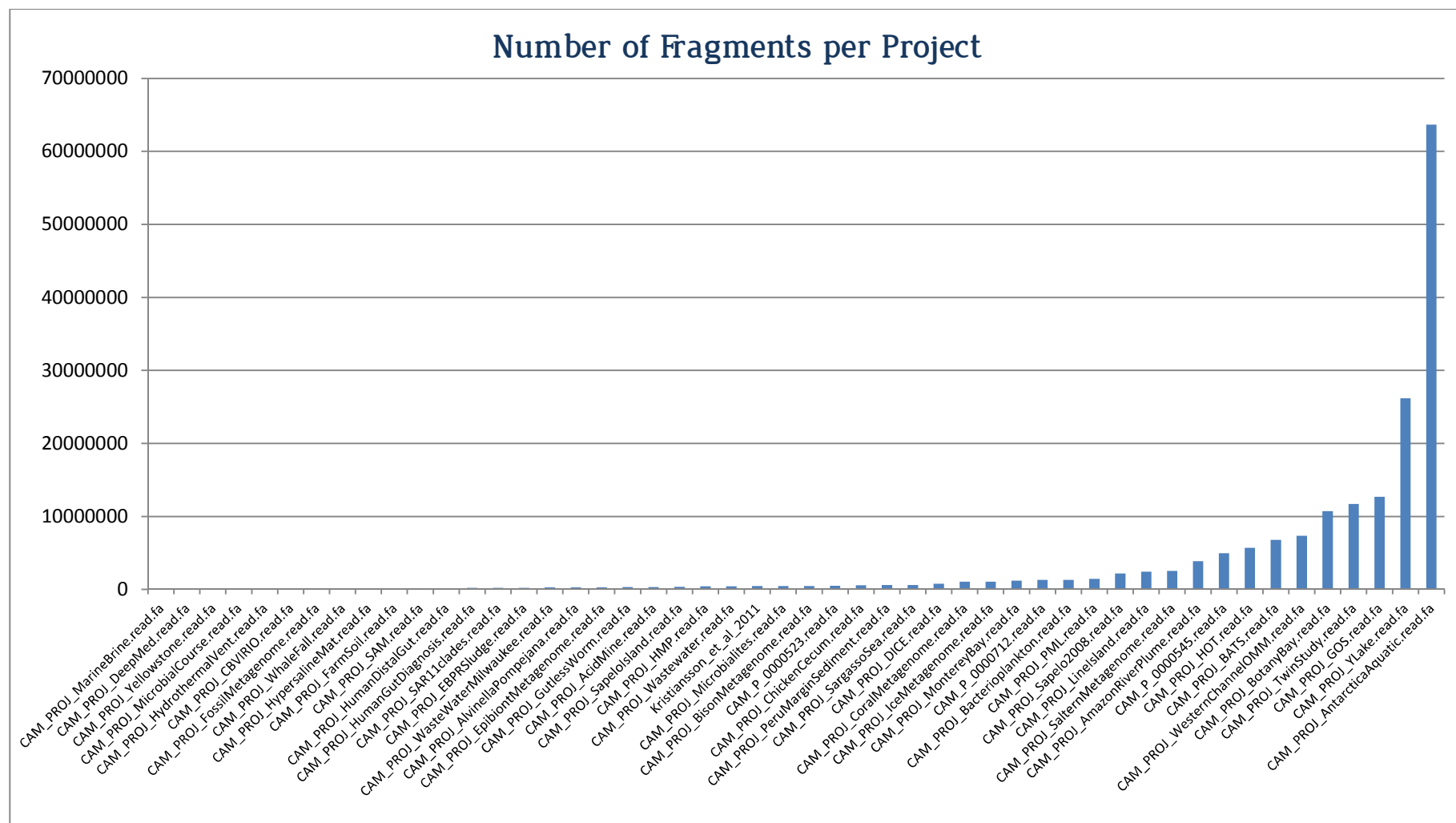


Figure 8 Amount of data by Nr of Fragments per project

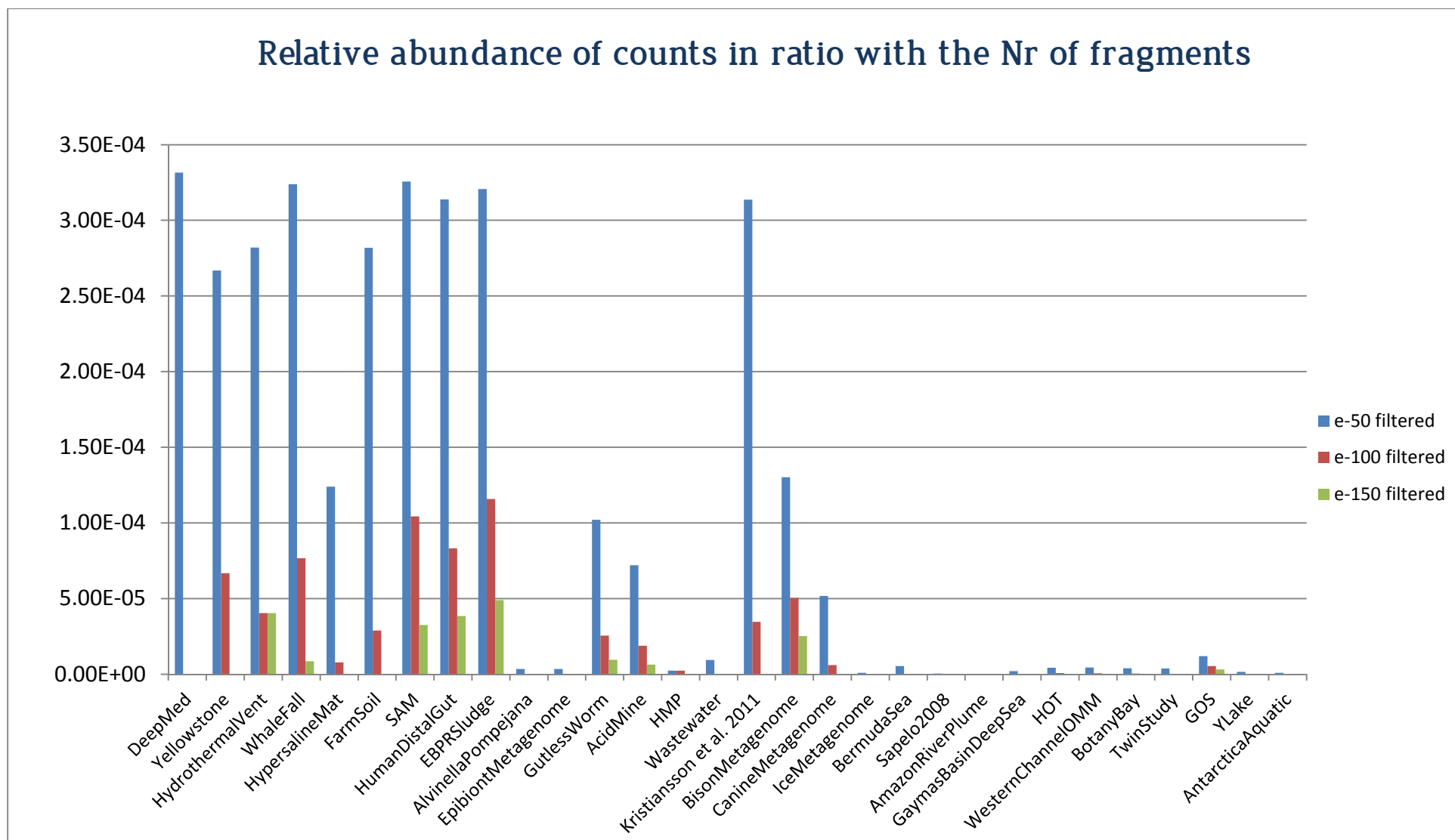


Figure 9 Thresholds in 30 projects filtering out those that do not show any AR genes in higher threshold that E-50

Data reduction

Initial data 98 GB after data preparation, 355,840 MB

Just as example, the normalization methodologies used in the data abstraction. As stated in the methods sections, these normalization methods were the first attempt, but later only normalization by abundance per site was used, included in the similarity measure.

Table 0-1 Example of an abstracted table

Profile	Site 1	Site 2	Site 3
ARG1	1	2	4
ARG2	0	2	3
ARG3	25	0	0
ARG4	2	11	2

Table 0-2 Example of table with total number of fragments per site

Total number of fragments per site	Site 1 1500	Site 2 3000	Site 3 800
	1+0+25+2=28	2+2+0+11=15	4+3+0+2=9

Table 0-3 Example normalized by the number of fragments per site

Profile	Site 1	Site 2	Site 3
ARG1	1/1500	2/3000	4/800
ARG2	0/1500	2/3000	3/800
ARG3	25/1500	0/3000	0/800
ARG4	2/1500	11/3000	2/800

Table 0-4 Example Normalized by abundance per site

Profile	Site 1	Site 2	Site 3
ARG1	1/28	2/15	4/9
ARG2	0/28	2/15	3/9
ARG3	25/28	0/15	0/9
ARG4	2/28	11/15	2/9

Table 0-5 Results of normalization per total num. of fragments

Profile	Site 1	Site 2	Site 3
ARG1	6,666e-4	6,666e-4	0,005
ARG2	0	6,666e-4	0,00375
ARG3	0,01666	0	0
ARG4	0,00133	0,00366	0,0025

Table 0-6 Results of normalization per abundance per site

Profile	Site 1	Site 2	Site 3
ARG1	0,0357	0,1333	0,4444
ARG2	0	0,1333	0,3333
ARG3	0,8928	0	0
ARG4	0,0714	0,7333	0,2222