Title: Correlation among antibiotic resistance and virulence traits in human gut microbiomes

Pedro Escudeiro¹, Joël Pothier², Francisco Dionisio¹,³ and Teresa Nogueira¹

¹cE3c – Centro de Ecologia, Evolução e Alterações Ambientais, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

²Atelier de Bioinformatique, ISYEB, UMR 7205 CNRS MNHN UPMC EPHE, Muséum National d’Histoire Naturelle, CP 50, 45 rue Buffon F-75005 Paris, France

³Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal.

Corresponding author:

Teresa Nogueira
cE3c – Centro de Ecologia, Evolução e Alterações Ambientais, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
Phone. +351 217500577
Fax. +351 21750002
teresainogueira@gmail.com

Keywords: antibiotic resistance – virulence – microbiome - metagenomics – human gut – co-occurrence
Abstract:

Human beings have used large amounts of antibiotic, not only in medical contexts but also, for example, as growth factors in agriculture and livestock, resulting in the contamination of the environment. Even when pathogenic bacteria are the targets of antibiotics, hundreds of non-pathogenic bacterial species are affected as well. Therefore, both pathogenic and non-pathogenic bacteria have gradually become resistant to antibiotics. We tested whether there is still co-occurrence of resistance and virulence determinants. We performed a comparative study of environmental and human gut metagenomes issuing from different individuals from different human populations across the world. We found a great diversity of antibiotic resistance determinants (ARd) and virulence factors (VFd) in metagenomes. Importantly, even after correcting for metagenome sizes, there is a correlation between ARd and VFd. In the human gut there are less ARd and VFd than in more diversified environments, and yet the ARd/VFd correlations are stronger. They can vary from very high in Malawi, where antibiotic consumption is unattended, to non-existent in the uncontacted Amerindians. We conclude that there is co-occurrence of resistance and virulence determinants, suggesting a possible co-selective phenomenon. Therefore, by selecting for resistant bacteria, we may be selecting for more virulent strains, as a side effect of antimicrobial therapy.
Introduction:

Antibiotics are present in microbial communities, not only as a result of the natural lifestyle of microorganisms but also due to the usage of these drugs in agriculture, food industry, livestock, or in human health (Castanon 2007). Therefore, antibiotics can affect bacterial communities as a whole, comprising both pathogenic and non-pathogenic bacteria. Take, for example, the human microbiome, defined as the set of microorganisms that colonize humans. This microbiome is composed of about 3.8e+13 bacterial cells (Sender et al 2016) spanning thousands of taxa and colonizing our body’s surfaces and bio fluids, including tissues such as skin, mucosa and most importantly, the gastrointestinal tract. Thus, even when virulent bacteria are the targets of antibiotics, the administration of these drugs may also affect many non-pathogenic mutualistic or commensal bacterial species present in individuals undergoing treatment (Francino 2015).

The resistome (the collection of all antibiotic resistance (AR) genes, which exist in both pathogenic and non-pathogenic bacteria (Wright 2007)) is frequently encoded on mobile genetic elements. Similarly, the virulome, the set of genes encoding virulence can also be encoded on the mobilome (Svara and Rankin 2011; Nogueira et al 2012; Nogueira et al 2009; Smith 2001). Therefore, many bacterial virulence factors (VF) are easily spread in bacterial populations by horizontal gene transfer, converting mutualistic or commensal bacteria into potential opportunistic pathogens. Several examples of highly virulent and multi-resistant disseminated clones have already been reported throughout the literature (for a review see (Beceiro et al 2013)).

A given bacterial population may constitute a life insurance to other bacterial populations present in the same microbiome through at least two different mechanisms. First, if antibiotic resistance genes are coded by plasmids or other mobile genetic elements, they may transfer to pathogenic cells and save them from the negative effects of antibiotics. Indeed, these genetic elements can spread into the bacterial community by horizontal gene transfer, even crossing species (Sommer et al 2010; Sommer et al 2009; Andersson and Hughes 2014; Vega and Gore
2014). Curiously, some bacterial strains from different species have been shown to be extremely good donors of certain plasmids. As such, they are able to amplify the number of plasmids in a bacterial community and spread those plasmids to other bacterial cells (Dionisio et al. 2002), a phenomenon probably explained by interactions between different plasmids (Gama et al. 2017a, 2017b).

Secondly, even bacterial cells not coding for antibiotic resistance determinants nor receiving mobile elements may be protected by cells coding for certain drug resistance genes. Indeed, gene products that inactivate antibiotics by degrading or modifying antibiotic molecules are also decreasing their concentration in the local environment (for a review about mechanisms of possible indirect resistance see (Wright 2005). This mechanism of indirect resistance has been shown to occur in different systems and to be pervasive (Nicoloff and Andersson 2016; Sorg et al. 2016; Domingues et al. 2017).

In this paper we ask whether there is co-occurrence of antibiotic and virulence among bacterial populations. For this, we have chosen to study metagenomes. There are three main reasons to use metagenomes to address this question. First, it is known that, for millions of years, bacteria had to cope with the presence of many other species, mostly competing with them (Foster and Bell 2012), but also cooperating, both phenomena relying on the full set of genes of the metagenome (Ponomarova and Patil 2015; Rosenthal 2011; Foster and Bell 2012). Secondly, horizontal gene transfer promotes the genetic relationship between species thus enforcing cooperation (Nogueira et al. 2009) avoiding the emergence of cheaters within the microbiome. Third, the study of metagenomes gives us access to the repertoire of genes involved in adaptation to the environment, given that many of these traits are often encoded in the mobilome and thus can be shared by different, eventually unrelated bacteria. In this context, mining for genes coding for antibiotic resistance and virulence traits in metagenomes is a reliable way to access the selective pressures the population is subject to, as well as the co-occurrence of genetic traits of the whole microbiome.
The present study aims at understanding the relationship between antibiotic resistance genes and those coding for virulence. Here we show that there is indeed a linkage between the dissemination of virulence factors and genes coding for antibiotic resistance, within natural microbiomes, and that this relationship is influenced by the behaviours of human populations, spanning from very different geographical locations across the world.
Materials and methods:

Metagenomic datasets

Our human gut query cohort included 110 previously studied, and publicly available metagenomes pertaining to individuals from different regions of Venezuela, Malawi, and the USA, as well as a broad age span (0.05 to 53 years) (Yatsunenko et al. 2012). All the metagenomes files were generated by the same team and project, and by using the same bioinformatics pipeline. Our environmental cohort comprised 64 previously selected, and publicly available environmental metagenomes, belonging to 12 different biomes (Delmont et al. 2011). Although Delmont’s team report using a dataset comprised of 77 metagenomes, there are only 70 MG-RAST accession numbers present in the article’s appendix, of which only 64 are publicly available. Both project’s metagenomes were downloaded from MG-RAST (Meyer et al. 2008) under FASTA format, making use of successive calls to its Application Programming Interface (API) (Wilke et al. 2015) using the respective MG-RAST’s accession numbers available in the aforementioned bibliography. Each FASTA file comprised clustered protein-coding sequences, retrieved from MG-RAST’s file-formating pipeline (550.cluster.aa90.faa files). The protein sequences enclosed in these files are clustered at 90% identity, containing non-redundant translated sequences (Yatsunenko et al. 2012). These protein-coding formatted FASTA files contain the translations of one representative from each cluster. Thus, the numbers of protein sequences for a given metagenome used here represent its richness.

BLAST, VFDB and Resfams

For every metagenome present in our query, a BLASTP (Altschul et al. 1997) search was performed against the VFDB database for bacterial virulence factors families (Chen et al. 2012) and the Resfams AR Proteins database of bacterial antibiotic resistance protein families (Gibson et al. 2014). An alternative approach made use of the Antibiotic Resistance Genes Database (ARDB) (Liu and Pop 2009), but several hindrances concerning its sub-classification by functional antibiotic resistance protein families made us discard the possibility of using such database to address of AR and VF protein family co-occurrence in metagenomes.
The BLAST+ executables package was downloaded on the 17th of November 2015 (ncbi-blast-2.2.31+ version). The VFDB was downloaded on the 11th of November 2013 (31 classified FASTA files of bacterial virulence factor sub-families), and the Resfams database was downloaded on the 29th of January 2016, (123 classified FASTA files of bacterial antibiotic resistance protein subfamilies. Every protein enclosed in each of the addressed metagenomes, was used as a query in order to search for similarities to either AR or VF protein-coding traits. Hence, we aimed at retrieving the best hit (best scored alignment) that enabled us to assign an AR or VF function to each of the aforementioned metagenomic proteins. Every BLASTP search was performed with a very stringent e-value cut-off of 1e-15 (10 orders of magnitude lower than the conservative commonly used E-value of 1e-5). Next, we filtered the resulting output files, as to only retrieve the alignment hits with >60% coverage of the query size, and whose query and subject length ratio is between 0,75 and 1,5. When using a combination of this very stringent e-value of 1e-15 together with coverage, and length ratio filters, we expect to only retrieve true homologues, avoiding false positives. Furthermore, from all the generated alignments between our metagenomic query cohort and the preceding databases, we computed the representative counts for the different gene families that coded for either AR or VF traits. Thus, the amounts of different classes (gene families) that are present in a given metagenome represent its diversity in terms of AR or VF traits.

We have also removed hits for proteins that aligned with both antibiotic resistance and virulence factor proteins (25.9% of hits against the VFDB, and 29,9% of the hits against the Resfams database). The abovementioned filters and algorithms were implemented making use of Unix scripting languages (GNU Awk version 4.0.1, and Z-Shell version 5.0.2), under a Linux environment.
Statistical Analysis

To test for relationships in metagenomes between the presence of both antibiotic resistance and virulence factors traits we proceeded as follows. Our expectation is that the diversity number for homologues of antibiotic resistance protein families (henceforth denoted as ARd) and the diversity number for homologues of virulence factors protein families (denoted as VFd) in each metagenome (clustered at 90% identity) increases with the protein family richness of the latter (being that the protein family richness here means the total number of cluster representative proteins. Given the potential diversity of these protein families, one does not expect ARd and VFd to level off with a metagenome’s protein family richness. Therefore, we assume a linear relationship between ARd and metagenome’s protein family richness, with a fixed 0 intercept. The same for VFd. Our results will show that these assumptions are reasonable. Thus, to avoid spurious correlation, we corrected the diversity of ARd and VFd in a metagenome by taking into account its protein family richness: we have standardized the different representative counts that issued from these samples according to the protein family richness of their respective metagenomes, therefore escaping “statistical inevitability”.

Hence we define α as the slope of the regression line of ARd on the protein family richness (henceforth denoted as Size solely in the equations) of the metagenomes (with a fixed 0 intercept) and β as the slope of the regression line of VFd on the protein family richness of the metagenomes (also with a fixed 0 intercept):

\[ ARd = \alpha \cdot Size \]  \hspace{1cm} \text{Equation 1A}

\[ VFd = \beta \cdot Size \]  \hspace{1cm} \text{Equation 1B}

Thus a metagenome \( i \) of a given protein family richness (Size(\( i \))) is expected to have \( ARd(i) = \alpha \cdot Size(i) \) antibiotic resistance protein families’ homologues and \( VFd(i) = \beta \cdot Size(i) \) virulence factors protein families’ homologues. Or:

\[ \frac{ARd(i)}{\alpha \cdot Size(i)} = 1 \]  \hspace{1cm} \text{Equation 2A}
\[
\frac{VFd(i)}{\beta \cdot Size(i)} = 1
\]  
Equation 2B

Naturally, it would be the case if ARd and VFd were only correlated with the protein family richness of the metagenomes. This is our null hypothesis, H0: data points are evenly present in the four quadrants around point \((\frac{ARd(i)}{\alpha \cdot Size(i)}, \frac{VFd(i)}{\beta \cdot Size(i)})=(1,1)\). However, some of the metagenomes do not match these predictions. Therefore, one may ask whether a given metagenome with more ARd than expected for its protein family richness has less or more VFd than expected for its protein family richness. There are two alternative hypotheses: 1) data points distribute over upper right and lower left quadrants around \((\frac{ARd(i)}{\alpha \cdot Size(i)}, \frac{VFd(i)}{\beta \cdot Size(i)})=(1,1)\) (i.e. the more diverse are AR genes, the more diverse are VF genes); or 2) they distribute over upper left and lower right quadrants around \((\frac{ARd(i)}{\alpha \cdot Size(i)}, \frac{VFd(i)}{\beta \cdot Size(i)})=(1,1)\) (i.e. the more diverse are AR genes, the less diverse are VF genes, and vice versa) This is the alternative hypothesis H1: data points distribute over two quadrants around.

The Spearman rank correlation coefficient \((r_s)\) was used to test the association between VFd and ARd. All possible associations between VFd and ARd hits for sub-families were also generated \((n = 123 \times 31 = 3813)\). As to control the expected proportion of rejected null hypotheses, the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was applied to the Spearman’s \(r_s\) p-values generated this way. The correlation coefficient \((r)\) and the slope of the linear fits were also calculated as to access the strength and direction of the relationship between the standardized ARd and VFd on the generated scatterplots, as well as to address the ratio between these variables. Associations with \(r_s > 0.5\), that also had \(r > 0.5\) and a \(r_s\) P-value < 0.001 were considered valid. All statistical analysis was conducted with the R programming language (version 3.2.2). Scatterplots and marginal boxplots were generated using R’s ggplot2, grid, gtable and scales packages.
Results:

Antibiotic resistance (AR) gene families in the metagenomes

We started by evaluating the quantity of different AR genes present in the metagenomes under study. We wanted to know how does the diversity of antibiotic resistance gene families’ homologues (that is, ARd, the number of different families) vary with the metagenome protein family richness (equation 1A). To answer this question, we used a data set of environmental metagenomes issuing from diverse ecosystems and biomes, such as oceans, coral atolls, deep oceans, Antarctic aquatic environments and snow, soils, hyper saline sediments, sludge’s, microbial fuel cell biofilms, and animal microbial populations (Delmont et al 2011). In this dataset, there is a broad variation in AR gene diversity (ARd) (Fig. 1.A), even for a given fixed metagenome protein family richness. For example, when the number of sequences of a given metagenome is about 1.75e+5, the number of ARd can range from almost zero (chicken cecum or cow rumen) to about 3000 (acid mine biofilms). However, there is a close to linear relationship between ARd and the protein family richness of the metagenomes. Indeed, we draw a regression line, with \( r = 0.7543, r_s = 0.6426, P = 1.037e^{-08} \) (\( r_s \) is the Spearman rank correlation coefficient, slope = \( \alpha \approx 0.0048 \)).

We also draw a fit line for the subset of metagenomes belonging to the human faeces biome. Another dataset composed by 110 human gut metagenomes, sampled from healthy individuals with ages ranging from 0.05 to 53 years of age, spanning different regions of the world such as: USA, Malawi, and the Venezuelan Amazon was also studied (Yatsunenko et al 2012). In this dataset (see Supplemental Fig. A) the frequency of antibiotic resistance protein families is strongly dependent on the protein family richness of the metagenome \( (r = 0.9142, r_s = 0.9424, P = 2.2e-16, \text{slope} = \alpha_g \approx 0.0052) \).
Virulence factor (VF) gene families in the metagenomes

Bacterial virulence factors are proteins that enable pathogenic bacteria to parasitize a host, including gene products involved in adhesion and invasion, secretion, toxins, and iron acquisition systems (Chen et al 2012).

The environmental microbiomes of our dataset reveal a great diversity of virulence factors protein families' (VFd) densities (Fig. 1.B). For example, when the number of sequences of a metagenome is about 3.7e+5, VFd varies from about 1000 (for human faeces) to more than 7000 (ocean). However, there is a close to linear relationship between VFd and the protein family richness of the metagenomes. Indeed, we draw a regression line, with \( r = 0.6694, r_s = 0.5219, P = 9.745 \times 10^{-06} \), where \( r_s \) is the Spearman rank correlation coefficient, slope = \( \beta_e \approx 0.0083 \)). Since virulence may also be associated with the colonization of different types of biomes (some virulent traits are involved in adaptation to new environments, by allowing bacteria to adhere and colonize substrates, to access to resources such as iron, among other functions) besides the context of infection, one can expect different types of these genes in environmental microbiomes. Moreover, in human gut microbiomes (see Supplemental Fig. B), we observed a good correlation between VFd and the protein family richness of the metagenomes \( (r = 0.6331, r_s = 0.7626, P = 2.2 \times 10^{-16}) \), where \( r_s \) is the Spearman rank correlation coefficient, slope = \( \beta_e \approx 0.0058 \)).

The AR / VF correlations

The main purpose of the present work is to evaluate the relationship, if any, between ARd and VFd in microbiomes. Therefore, we excluded from our analysis all the gene products that are both homologues to AR and VF determinants. Thus we avoid the introduction of a bias in the correlation analysis. We computed \( \frac{ARd(i)}{a.Size(i)} \) and \( \frac{VFd(i)}{\beta.Size(i)} \) for all metagenomes. In Fig. 2 one can see that the two variables are positively correlated: a metagenome with more (less) than expected VFd also has more (less) than expected ARd. Although there is a large variation of ARd (Fig. 1.A), and VFd (Fig. 1.B), in environmental metagenomes, there is a strong correlation...
between ARd and VFd \( (r = 0.7547, r_s = 0.8256, P = 2.2e^{-16}, \text{Spearman rank correlation coefficient}) \) (Fig. 2). When we draw the regression line for the human gut microbiomes from this dataset (human gut), we can see that the ratio ARd/VFd is even higher than for the case of environmental metagenomes. Despite the fact that the values for ARd and VFd are lower in gut metagenomes (see Fig. 1), one can witness a higher ratio ARd/VFd (larger slope) of metagenomes pertaining to the human faeces biome relatively to the slope calculated for all environmental metagenomes (Fig. 2). This suggests that there is a greater accumulation of ARd over VFd for this particular biome than in environmental ones. In Fig. 2, one can see that just a few points fall in the second and fourth quadrants. Points that fall into the second quadrant correspond to metagenomes for which there is more ARd than expected, given the VFd. The points that fall into the fourth quadrant pertain to metagenomes for which there is less ARd than expected, given the VFd.

Fig. 3 shows AR diversity versus VF diversity in the human gut metagenomes for each population (country) and for the three populations together. All the statistics are summarized in Table I and Supplemental Table (for the study using the ARDB database (Liu and Pop 2009)). The correlation between ARd and VFd is lower in the human gut metagenomes dataset \( (r = 0.5572, r_s = 0.6289, P = 2.2e^{-16}, \text{Spearman rank correlation coefficient}) \) (Fig. 3.A) than in the environmental samples \( (r = 0.7547, r_s = 0.8256, P = 2.2e^{-16}) \) (Fig. 2). This correlation is not dependent on the protein family richness of the metagenomes (Fig. 3.A). In order to better understand the results, we can divide the graph into quadrants, using the axes set by coordinates (1,1)(See Equations 2.A and 2.B). We can see that there are many points (metagenomes) located on the first and third quadrant following the regression line. These points represent the metagenomes for which there is a correlation between ARd and VFd; that is, ARd and VFd are both either in excess or in deficit (compared to their expected values given the protein family richness of the metagenome). Nevertheless, some points fall in the second quadrant, (corresponding to metagenomes for which there is more ARd than expected given VFd), and some other points fall into the fourth quadrant (corresponding to metagenomes for which there is less ARd than expected given VFd.
However, we can distinguish different trends upon geographical localization of the human populations under study. The largest contribution to this graph comes from the North American samples, which account for 66/110 (60%) of the individuals (Fig. 3.B).

We have compared these results with those of the gut metagenomes issuing from non-western human populations, having very contrasting lifestyles and exposure to antibiotics: the uncontacted Amerindians, harbouring a resistome similar to that of pre-antibiotic era; and those of the people from Malawi with a high prevalence of infectious diseases and widespread use of unprescribed antibiotics (WHO s.d.; UNICEF s.d.; Forslund et al 2013).

While studying metagenomes collected from Amerindians (Venezuela) (21/110, or 19% of the metagenomes of the dataset) and Malawians (23/110, or 21%), we can see two completely different scenarios (Fig. 3.C and 3.D). In the Amerindian gut metagenomes there is no statistically significant correlation between ARd and VFd ($r = 0.326$, $r_s = 0.4519$, $P = 0.04111$, Spearman rank correlation coefficient). This result is relevant, when we compare it to the Malawian gut metagenomes where there is a very strong correlation between ARd and VFd ($r = 0.9088$, $r_s = 0.8706$, $P = 2.533e^{-06}$). Interestingly, the slope of the regression line portrayed by the association between ARd and VFd belonging to Malawian human gut metagenomes (Fig. 3.D) is twice as large than those issuing from the Amerindian (Fig. 3.C) and USA populations (Fig. 3.B). That is, the Malawian metagenomes accumulate two times more ARd per VFd than its Amerindian and North-American counterparts.

**The co-occurrence of AR and VF belonging to the cell envelope**

Our results suggest that co-occurrence of AR and VF might be taking place amidst bacterial communities. We wondered, however, which were the genetic traits that were more prone to this effect. All the AR and VF protein families associations were plotted and analysed. Thus, after all possible associations between subfamilies of AR and VF pairs had been generated (123 * 31 = 3813) a Spearman’s rho ($r_s$) cut-off of 0.5 was applied as to filter the best correlations. The vast majority of associations falls into the functional category of multi-drug efflux pumps (AR’s) associated with either secretion systems, as well with iron uptake and adhesion...
mechanisms (VF’s), respectively (Table II). Amongst the most representative associations between AR and VF traits are those belonging to the cell envelope and the general secretion mechanisms. On the other hand, the AR protein families for beta-lactamases, and general beta-lactam resistance mechanisms, bore no statistically significant correlations whatsoever, attaining very weak Pearson and Spearman correlation coefficients as well.
Discussion:

Antibiotic resistance among environmental and human gut metagenomes

As expected, we found different homologues for antibiotic resistance (AR) gene products belonging to different protein families (AR diversity or ARd) among environmental metagenomes (Fig. 1.A). We have shown that, in the environmental samples, ARd varies a lot from metagenome to metagenome. This may result from the differential microbial community composition of the metagenomes, whose genetic diversity can be grouped according to the adaptation to the environment in question (Delmont et al. 2011); but also from the fact that the selective pressure for the maintenance of antibiotic resistance genes in environmental microbiome varies widely from environment to environment (Martinez 2008; Martinez and Baquero 2002; Fitzpatrick and Walsh 2016). Human gut metagenomes, on the contrary, have a less diversified repertoire of AR determinants (lower ARd) than that of their environmental counterparts.

We found a very strong correlation between the diversity of AR determinants and the metagenome’s protein family richness (Supplemental Fig. A), independently of the geographic origin of the human populations. These similar densities of ARd can indicate that, in human gut microbiomes, the number of different AR genes is not influenced by the human lifestyle, such as diet, medical care, access to antibiotics or other cultural habits (Ghosh et al. 2013), and/or that the adaptation to the intestinal tract shapes microbial AR diversity as well.

Antibiotic use can vary from country to country (CDDEP s.d.), and, as consequence, individuals from different human populations are under different antibiotic exposition. Forslund and co-workers have demonstrated that there are robust differences in the antibiotic resistance arsenal between countries and that these differences follow the veterinary and human antibiotic use (Forslund et al. 2013).
Virulence among environmental and human gut metagenomes

In what concerns virulence, we can assert that there is a wide diversity and density differences of VFs in environmental metagenomes, which poses as evidence of the plasticity portrayed by environmental bacteria in order to adapt to different hosts and niches (Fig. 1.B). On the other hand, human gut microbiomes harbour a less diverse VF repertoire, especially in the USA samples (Supplemental Fig. B), which seem to indicate an evolution towards adaptation to the human gut, or lower contact with pathogens, eventually due to vaccinations and sanitation.

Association of AR / VF

According to Yatsunenko and colleagues, once the human gut metagenome is established at the age of 3 years old, it does not diverge much from individual to individual, in what concerns both phylogenetic diversity and functional richness (Yatsunenko et al 2012).

The most relevant results shown here are that, when corrected for metagenome’s protein family richness, ARd and VFd are strongly correlated both in the environmental samples (Fig. 2 and 3) and in the human gut samples (Fig. 3), with special emphasis on the human gut.

The North American intestinal samples (Fig. 3.B) show a wide variety of associations between ARd and VFd, always presenting a statistically significant correlation between these genetic traits. This result in itself reinforces our hypothesis that antibiotic resistance and virulence are in fact co-associated. In particular there are 5/66 (7.5%) points in the second quadrant of Fig. 3.B representing metagenomes where there is an accumulation of ARd by VFd, thus suggesting that the mobilization of mobile genetic elements as gene cassettes, driving to multirresistance, might be taking place. The USA population, like other industrialized countries, is culturally exposed to antibiotics from health care facilities such as hospitals, antibiotic therapy, and the use of antibiotics in agriculture and livestock.
Amerindians from Venezuela and Malawians share, phylogenetically, more similar human gut microbiomes than the North Americans (Yatsunenko et al 2012), which belong mainly to a different enterotype (Lozupone et al 2012). Yet, whereas there is no ARd/VFd correlation among the Venezuelan individuals, there is a very strong correlation among the Malawian ones. This result highlights the potential effect of the exposure to antibacterial drugs on promoting the dissemination of antibiotic resistance by horizontal gene transfer, and the potential co-selection of virulence traits within bacterial communities.

The uncontacted Amerindian microbiome represents a “frozen” relic of a pre-antibiotic era of the human resistome, while the Malawian gut microbiome is much more exposed, both to antibiotics, and to colonization by pathogens (WHO s.d.; UNICEF s.d.; Forslund et al 2013).

Amerindians have no known access to pharmaceutical drugs, as they usually make use of the traditional indigenous medicine. Malawi is one of the poorest countries in Africa, where most people live with less than one dollar a day, many people cope with AIDS and bacterial infections, and where many children suffer from severe malnutrition (WHO s.d.). Nutrition has been reported to have a big impact on both the human gut microbiome composition and resistome (Ghosh et al 2014; 2013; Pehrsson et al 2016). In Malawi, UNICEF has been implementing a program of Ready-to-Use-Therapeutic-Food (RUTF) in order to reduce mortality among children. RUTF often contains antibiotics such as co-trimoxazole. It has been questioned, however, whether the success of this therapy is due to re-nutrition or to the combination with antibiotics (Alcoba et al 2013; Makoka et al 2012) that could have an effect on the microbial gut composition. It has been reported that there is widespread resistance to almost all of antibiotics that are empirically used in Malawi due to the lack of routine microbiological culture and sensitivity testing (Makoka et al 2012), but also due to self-medication. Human gut samples from the Malawi people of the dataset used in the present study are described as having a high overall resistance potential with an over representation of cephalosporin and tetracycline resistance genes which may suggest extensive use of old, broad-spectrum antibiotics, a known problem in many developing countries (Forslund et al 2013).
The co-representation of AR and VF targeting the cell envelope

Between all the possible statistical AR and VF protein families’ correlations, the strongest ones are among proteins belonging to the bacterial cell envelope (Table II). This result is not surprising, as many proteins that belong to the secretory system or to the secretome itself, and thus target the bacterial envelope, are frequently encoded on mobile genetic elements (Nogueira et al 2012; Nogueira et al 2009). As such, this co-occurrence may also be regarded as a direct consequence of the dynamics of these genetic elements coding for both types of determinants.

In human gut the strongest correlations involve efflux pumps that can extrude antibiotics non-specifically, and adhesion and iron scavenging mechanisms of pathogenicity. One possible explanation for the fact that the most frequent AR and VF associations in human gut involve efflux pumps could be that they allow a fast and efficient response to new man-made antibiotic molecules, while in environmental genomes, specific resistance mechanisms targeted to particular antibiotic molecules may have had more time to evolve to specific strategies.

Another possible explanation is that extrusion can also be implicated in pathogenic mechanisms of human associated pathogens as well as bacteria-host interactions. For example, biofilm formation in a Staphylococcus aureus methicillin resistant (MRSA) strain has been known to be essentially reliant on the activity of fibronectin binding proteins (McCourt et al 2014), and multi-drug efflux pumps have direct implications with the formation and maintenance of such biofilms (Soto, 2013). Furthermore quinolone resistant S. aureus strains up regulate the production of fibronectin binding proteins when subjected to sub-lethal dosages of ciprofloxacin (Bisognano et al 2004). It has also been acknowledged that physiological levels of some cations present within the host promote the up regulation of genes encoding putative efflux transporters, oxidoreductases, and mechanisms of iron uptake either in Acinetobacter baumannii (Hood et al 2010) as in Burkholderia cenocepacia (Drevinek et al 2008), which could explain the co-association of iron acquisition systems with those of multi-drug efflux pumps.
We show here that there is a correlation between the diversity of antibiotic resistance and virulence traits in human gut microbiomes and in environmental microbiomes. Interestingly, this correlation varies from a human population to another. Human antibiotic exposure due, either to therapy or to the environment and food, can have an effect on selecting for potentially pathogenic bacteria in the human gut microbiomes. It can also drive and shape changes on the gene pool of microbiomes. This means that, by selecting for resistant bacteria, we might be also selecting for more virulent strains, as a side effect of antimicrobial therapy.
Funding

T.N. was a recipient of a grant from Fundação para a Ciência e a Tecnologia (SFRH/BPD/41863/ 2007).

Acknowledgements:

We thank Eduardo P. C. Rocha for critical reading of the manuscript and constant interest and suggestions since the beginning of this work, Guillaume Sapriel for suggestions to improve the methods, Pedro David, Eric Duchaud, Guillaume Achaz, and all the members of our teams for the fruitful discussions, Octávio Paulo, Francisco Pina-Martins and Sophie Brouillet for the technical help, and granting access to the servers.

Conflict of Interest:

The authors declare no conflict of interest.

References


Delmont TO, Malandain C, Prestat E *et al.* Metagenomic mining for microbiologists. *ISME J* 2011, **5**: 1837-1843.


Drevinek P, Holden MT, Ge Z et al. Gene expression changes linked to antimicrobial resistance, oxidative stress, iron depletion and retained motility are observed when Burkholderia cenoccepacia grows in cystic fibrosis sputum. *BMC Infect Dis* 2008, 8: 121.


Gama JA, Zilhão R; Dionisio F. Conjugation efficiency depends on intra and intercellular interactions between distinct plasmids: Plasmids promote the immigration of other plasmids but repress co-colonizing plasmids. *Plasmid* 2017a 93: 6-16.

Gama JA, Zilhão R; Dionisio F. Multiple plasmid interference - Pledging allegiance to my enemy’s enemy. *Plasmid* 2017b 93: 17-23.


Nicoloff H, Andersson DI. Indirect resistance to several classes of antibiotics in cocultures with resistant bacteria expressing antibiotic-modifying or -degrading enzymes. *Journal Antimicrob Chemother*, 2016 **71**(1):100-110


Rosenthal AZ. RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* 2011, **5**: 1133–1142.


WHO (n.d.) *Global Health Observatory (GHO) data*. Retrieved 7 20, 2016 from

http://www.who.int/gho/countries/mwi/en/


Yatsunenko T, Rey FE, Manary MJ. Human gut microbiome viewed across age and geography.

Legends:

Figure 1. Distribution of the diversity number of antibiotic resistance and virulence factor protein families by environmental metagenomes’ protein family richness

The vertical axes represent the total diversity count of AR protein families’ homologues (ARd) (1.A) and VF protein families’ homologues (VFd) (1.B), present in metagenomes. The horizontal axes represent the protein family richness of the metagenome, that is, the number of cluster representative sequences - see Methods. In 1.A each dot represents one of the 64 metagenomes. The black line represents the simple linear regression of best fit for all of the metagenomes, where the grey shading is the 95% confidence interval. The points are scattered showing that the diversity of AR gene families can vary greatly from metagenome to metagenome. The points’ dispersal illustrates different VFd frequencies when taking distinctive environmental metagenomes into comparison. The light line represents the linear regression of best fit for the human faeces metagenomes subset. The Correlation coefficient (r), Spearman’s rank correlation coefficient (rs), P-value obtained from the latter, Coefficient of determination (r^2) and respective P-value are shown on Table I.

Figure 2. Distribution of AR by VF diversity in environmental metagenomes.

Scatter plots with marginal boxplots of $\frac{ARd(i)}{\alpha_e \cdot \text{Size}(i)}$ versus $\frac{VFd(i)}{\beta_e \cdot \text{Size}(i)}$ of each metagenome where $\alpha_e$ and $\beta_e$ are the slopes for the environmental metagenomes, calculated in Fig. 1.A and 2.A, respectively. The black line represents the simple linear regression line of best fit, where the grey shading is the 95% confidence interval. The black line in the box plots represents the median, and the black diamonds represent the mean. Each dot represents a metagenome, and the grey lines parallel to the axes represent the (1,1) reference coordinates. The correlations seem to depend on the biome. The light line represents the linear regression of the human faeces metagenomes subset. There is a significant correlation involving the 12 different environmental biomes ($n = 64$, $r = 0.7547$, $r_s = 0.8256$, $P < 0.001$, BH post hoc procedure, lm slope = 0.6283), and that the slope is greater in metagenomes issuing from the human faeces.
subset, than in the environmental ones. These statistics along with the Coefficient of
determination \((r^2)\) and respective P-value are shown on Table I.

**Figure 3. Distribution of AR by VF diversity in Human gut metagenomes.**

Scatter plots with marginal boxplots of \(\frac{ARd(i)}{\alpha_g Size(i)} \) versus \(\frac{VFd(i)}{\beta_g Size(i)}\) of each metagenome where \(\alpha_g\) and \(\beta_g\) are the slopes calculated in Supplemental Fig. A and B, respectively. A simple linear regression line of best fit is represented in each graph, where the grey shading is the 95% confidence interval. The black line in the box plots represents the median and the black diamonds represent the mean. Each dot represents a metagenome, and the grey lines parallel to the axes represent the (1,1) reference coordinates. A) Significant correlation in metagenomes of all human populations – Malawi, USA, Venezuela, respectively \((n = 110, r = 0.5572, r_s = 0.6289, P < 0.001, BH post hoc procedure, lm slope = 0.2075)\), coloured according to the metagenome protein family richness. B) Significant correlation in gut microbiomes of USA individuals \((n = 66, r = 0.5304, r_s = 0.6367, P < 0.001, BH post hoc procedure, lm slope = 0.1897)\). C) Non-significant correlation in gut microbiomes of Venezuelan individuals \((n = 21, r = 0.326, r_s = 0.4519, P > 0.001, BH post hoc procedure, lm slope = 0.1371)\). D) Strong significant correlation in gut metagenomes of Malawian individuals \((n = 23, r = 0.9088, r_s = 0.8706, P < 0.001, BH post hoc procedure, lm slope = 0.3725)\). These statistics along with the Coefficient of determination \((r^2)\) and respective P-value are shown on Table I.

**Supplemental Figure. Distribution of the diversity number of antibiotic resistance and virulence factor protein families by human gut metagenomes’ protein family richness**

The vertical axes represent the total diversity count of AR protein families’ homologues (ARd) (A) and VF protein families’ homologues (VFd) (B), present in metagenomes. The horizontal axes represent the protein family richness of the metagenome, that is, the number of cluster representative sequences - see Methods. Each dot represents one of the 110 human gut metagenomes. The lines depict the linear regressions of the metagenomes from Malawi, USA and Venezuela, and the black line depicts the regression line for all metagenomes, where the grey shading is the 95% confidence interval. The Correlation coefficient \((r)\), Spearman’s rank
correlation coefficient ($r$), and P-value obtained from the latter, Coefficient of determination ($r^2$) and respective P-value are shown on Table I.
1. A

Antibiotic Resistance diversity vs. Size of the metagenome

Human gut

1. B

Virulence diversity vs. Size of the metagenome

Human gut

biome
- acid mine biofilms
- antarctic aqua. env.
- chicken cecum
- cow rumen
- mouse cecum
- P rem. sludges
- sediments
- soils

Human gut

0e+00  2e+05  4e+05  6e+05
0  2000  4000  6000  8000
0e+00  2e+05  4e+05  6e+05
2.

Global slope $\approx 0.6283$

Human gut slope $= 0.9156$

Virulence diversity

Antibiotic Resistance diversity

biome

- acid mine biofilms
- coral attols
- human feces
- P rem. sludges
- antarctic aqua. env.
- cow rumen
- mouse cecum
- sediments
- chicken cecum
- deep oceans
- oceans
- soils
3. A

3. B

3. C

3. D

The copyright holder for this preprint (which was this version posted April 17, 2018): https://doi.org/10.1101/298190

do: bioRxiv preprint
## Resfams vs VFDB

### Human

<table>
<thead>
<tr>
<th>Statistic</th>
<th>All Countries</th>
<th>USA</th>
<th>Malawi</th>
<th>Venezuela</th>
<th>All Countries</th>
<th>USA</th>
<th>Malawi</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>0.9142</td>
<td>0.9034</td>
<td>0.9413</td>
<td>0.9689</td>
<td>0.9331</td>
<td>0.9768</td>
<td>0.8251</td>
<td>0.6288</td>
</tr>
<tr>
<td>$r_s$</td>
<td>0.9424</td>
<td>0.9394</td>
<td>0.9202</td>
<td>0.9468</td>
<td>0.7626</td>
<td>0.7211</td>
<td>0.8429</td>
<td>0.8987</td>
</tr>
<tr>
<td>$r_s$ p-value</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>5.19E-010</td>
<td>5.16E-006</td>
<td>2.20E-016</td>
<td>8.61E-012</td>
<td>1.73E-006</td>
<td>2.98E-006</td>
</tr>
<tr>
<td>slope</td>
<td>0.055192</td>
<td>0.005039</td>
<td>0.005494</td>
<td>0.005587</td>
<td>0.005825</td>
<td>0.005259</td>
<td>0.006937</td>
<td>0.007301</td>
</tr>
<tr>
<td>$r$-squared</td>
<td>0.9658</td>
<td>0.9443</td>
<td>0.9816</td>
<td>0.9958</td>
<td>0.8301</td>
<td>0.7859</td>
<td>0.9482</td>
<td>0.9247</td>
</tr>
<tr>
<td>F-statistic p-value</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
<td>2.20E-016</td>
</tr>
</tbody>
</table>

### Environmental

<table>
<thead>
<tr>
<th>Statistic</th>
<th>All Biomes</th>
<th>ARd(i) ~ Size(i)</th>
<th>VFd(i) ~ Size(i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>0.7543</td>
<td>0.6694</td>
<td>0.7547</td>
</tr>
<tr>
<td>$r_s$</td>
<td>0.6426</td>
<td>0.5219</td>
<td>0.8256</td>
</tr>
<tr>
<td>$r_s$ p-value</td>
<td>1.04E-006</td>
<td>9.75E-006</td>
<td>2.20E-016</td>
</tr>
<tr>
<td>slope</td>
<td>0.004762</td>
<td>0.008337</td>
<td>0.6283</td>
</tr>
<tr>
<td>$r$-squared</td>
<td>0.72131</td>
<td>0.6241</td>
<td>0.9156</td>
</tr>
<tr>
<td>F-statistic p-value</td>
<td>2.20E-016</td>
<td>5.19E-015</td>
<td>5.05E-013</td>
</tr>
<tr>
<td>AR mechanism</td>
<td>VF mechanism</td>
<td>Correlation coefficient</td>
<td>Spearman’s r_s</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>-------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td><strong>Environmental:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC (ATP-binding cassette superfamily) Antibiotic Efflux Pump</td>
<td>Heme-mediated Iron Uptake</td>
<td>0.8785</td>
<td>0.8598</td>
</tr>
<tr>
<td><em>msbA</em> (Lipid A export ATP-binding/permease protein MsbA)</td>
<td>Heme-mediated Iron Uptake</td>
<td>0.8602</td>
<td>0.8738</td>
</tr>
<tr>
<td><em>macB</em> (Macrolide export ATP-binding/permease protein MacB)</td>
<td>Heme-mediated Iron Uptake</td>
<td>0.8415</td>
<td>0.8889</td>
</tr>
<tr>
<td><em>macB</em> (Macrolide export ATP-binding/permease protein MacB)</td>
<td>Type VI secretion system &amp; effectors</td>
<td>0.8017</td>
<td>0.8367</td>
</tr>
<tr>
<td><em>macB</em> (Macrolide export ATP-binding/permease protein MacB)</td>
<td>Fibronectin-binding proteins</td>
<td>0.7463</td>
<td>0.7549</td>
</tr>
<tr>
<td>TetM TetW TetO TetS (Tetracycline resistance proteins)</td>
<td>Fibronectin-binding proteins</td>
<td>0.7375</td>
<td>0.7620</td>
</tr>
<tr>
<td><em>macB</em> (Macrolide export ATP-binding/permease protein MacB)</td>
<td>Type III secretion system &amp; effectors</td>
<td>0.7307</td>
<td>0.8205</td>
</tr>
<tr>
<td><em>msbA</em> (Lipid A export ATP-binding/permease protein MsbA)</td>
<td>Fibronectin-binding proteins</td>
<td>0.7304</td>
<td>0.7545</td>
</tr>
<tr>
<td><em>msbA</em> (Lipid A export ATP-binding/permease protein MsbA)</td>
<td>Others (Adhesion &amp; Invasion proteins)</td>
<td>0.7005</td>
<td>0.7216</td>
</tr>
<tr>
<td>TetM TetW TetO TetS (Tetracycline resistance proteins)</td>
<td>Others (Adhesion &amp; Invasion proteins)</td>
<td>0.6594</td>
<td>0.7308</td>
</tr>
<tr>
<td><strong>Human Gut:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFS (major facilitator superfamily) Antibiotic Efflux Pump</td>
<td>Type VI secretion system &amp; effectors</td>
<td>0.8101</td>
<td>0.6806</td>
</tr>
<tr>
<td>MFS (major facilitator superfamily) Antibiotic Efflux Pump</td>
<td>Siderophore-mediated Iron Uptake</td>
<td>0.7489</td>
<td>0.7353</td>
</tr>
<tr>
<td>ABC (ATP-binding cassette superfamily) Antibiotic Efflux Pump</td>
<td>Heme-mediated Iron Uptake</td>
<td>0.6477</td>
<td>0.6307</td>
</tr>
<tr>
<td>ABC (ATP-binding cassette superfamily) Antibiotic Efflux Pump</td>
<td>Fibronectin-binding proteins</td>
<td>0.6390</td>
<td>0.6603</td>
</tr>
<tr>
<td><em>mexW mexI</em> (multidrug resistance efflux pump)</td>
<td>Type IV Pili</td>
<td>0.6189</td>
<td>0.5899</td>
</tr>
<tr>
<td>ABC (ATP-binding cassette superfamily) Antibiotic Efflux Pump</td>
<td>Others (Adhesion &amp; Invasion proteins)</td>
<td>0.6012</td>
<td>0.5639</td>
</tr>
<tr>
<td><em>msbA</em> (Lipid A export ATP-binding/permease protein MsbA)</td>
<td>Fibronectin-binding proteins</td>
<td>0.5538</td>
<td>0.5958</td>
</tr>
</tbody>
</table>