Supplementary materials to: Boyd A, El Dani M, Roula Ajrouche R, et al. Gut microbiome diversity and composition in individuals with and without Extended-Spectrum β-Lactamase-Producing Enterobacterales carriage: a matched case-control study

Table of contents:

SUPPLEMENTARY METHODS

Procedure of coarsened exact matching

The coarsened exact matching approach has been detailed elsewhere [1,2]. In brief, this method attempts to remove any distributional differences (i.e., multivariate nonlinearities, interactions, moments, etc.) between individuals with and without ESBL-E carriage by using a method free from model dependence or model misspecification (as found in propensity score matching, e.g.).

First, each variable is recoded (i.e., coarsened), leading to a dataset from which indistinguishable values can be grouped in strata and assigned the same value. In this study, we included all matching variables in the coarsened dataset: sexual group (5 categories), ESBL-E prevalence of countries traveled in the previous 12 months (3 categories), number of sexual partners in the previous 6 months [log(n+1)-transformed], geographic origin (4 categories), and any antibiotic use in the previous 6 months (dichotomous). We coarsened the only continuous variable by intervals of one log(n+1)-transformed number of partners. All categorical variables remained the same, as their levels could yield more than 150 matched pairs.

Second, a matching algorithm is applied whereby one individual with ESBL-E carriage is matched to an individual without carriage on an exact stratum of coarsened values in the coarsened dataset (i.e., exact matching). The coarsened data are then removed, leaving only the matched pairs and the original, uncoarsened data.

Description of the variables used for matching

We asked questions on whether participants had sex with a steady or casual partner and if so, the gender(s) of their partner(s). HIV-status was self-reported and in the event of unknown or negative HIV status, confirmed with the HIV Ag/Ab Combo assay (Alinity, Abbott). Current use of pre-exposure prophylaxis (PrEP) for HIV prevention (daily or intermittent) was also self-reported. From these questions, we were able to characterize 5 sexual exposure groups: HIV-negative men who have sex with men (MSM) on PrEP, HIV-negative MSM not on PrEP, HIV-positive MSM, HIV-negative other men, and HIV-negative woman. The limited numbers of transgender persons and those with other genders precluded a separate group of these individuals.

We asked questions on any travel in the previous 12-months and if so, the countries visited and the dates of entry and exit of each country. We then used estimates from a previous study on the prevalence of ESBL-E carriage among travelers [3] to assign each country a level of prevalence: low (<10%), moderate (10-25%), and high (>25%). Participants were categorized in the following groups: no travel, low prevalence, moderated prevalence and high prevalence. Those traveling to countries with multiple prevalence levels were categorized according to the highest prevalence category.

In participants who reported any sex with a steady or casual partner, we asked the number of partners in the previous 6 months for each type of partner. We added the number of steady and casual partners to obtain the total number of partners. We then added 1 and log-transformed values to ensure normal distributions.

We asked the country of birth for all participants. We defined geographic origin based on the country of birth and categorized them on World Health Organization regions. Given the distribution of variables, we regrouped regions as follows: European, Eastern Mediterranean/African, the Americas/Western Pacific Region [which includes *département outre mer* (overseas department) and *territoire outre mer* (overseas territory)], and South-East Asian.

We asked questions on any antibiotic use in the past 12 months and if so, the names of each antibiotic along with their start and stop dates. From this information, we could establish whether someone used any antibiotics in the past 6 months.

Measuring imbalance between cases and controls

Given that matches of cases and controls are made with strata of like covariate patterns, a multivariate imbalance measure can be made by summing the absolute differences in frequencies of each covariate between cases and controls in the multivariate space. This is called an L_1 distance metric and its calculation and statistical properties have been detailed elsewhere [4]. In brief, L_1 ranges between 0 (i.e., exact match) and 1 (i.e., exact mismatch) and is a relative measure representing 1 – the proportion that the multivariate histograms between cases and controls overlap. For this study, no value was used to conclude that the matching procedure was successful. Nevertheless, the L_1 distance was 0.154 in the 181 initially matched cases and control (Figure 1), suggesting very close overlap between the two groups.

Bioinformatic analysis

Briefly, bacterial 16S paired reads were first merged to acquire the full V3-V4 amplicon region. Low quality reads were removed based on their Q scores (if the expected number of errors in the read was > 1). Reads were dereplicated to keep only unique sequences. The cluster size of each representative sequence was stored for later diversity analysis. Unique sequences were clustered into operational taxonomic units (OTUs) using an identity threshold of 97%. For each sample, an OTU table was generated after reads mapping on OTUs. In order to compare samples to each other, they were normalized to the same number of reads using random subsampling. Normalization was fixed to 5000 reads. Alpha diversity metrics were calculated from the normalized OTU table. Taxonomic analysis was performed using the SINTAX algorithm14 with OTUs alignments to the RDP (Ribosomal Database Projet) databases, designed for 16S bacteria (v16). The taxonomy was predicted using a bootstrap confidence value of 0.8. OTUs and diversity metrics generated were then used for statistical analysis.

Targeted maximum likelihood estimation

Briefly, this method makes an initial estimate of the conditional mean of the outcome given a specific determinant and other covariates (outcome regression). The mean within level of the determinant is then modelled as a function of the other covariates (propensity score regression). The estimate of the outcome regression is updated using information from the propensity score regression in an iterative manner until convergence is reached. The target parameter is then estimated from outputs of the outcome and propensity score regression and variance estimators obtained based on an efficient influence curve. To avoid further model misspecification, both the outcome and propensity score regression are optimised using an ensemble of machine learning techniques, referred to as a 'super learner' [5]. The weighted combination of predictions by which the cross-validated mean square error is minimised is selected through the super learner.

For this study, a target parameter was calculated for ESBL-E status, while accounting for all covariates used in the matching procedure.

Estimates were constructed using the 'tmle' and 'SuperLearner' packages in R. The ensembles included the following: generalised linear models (with and without interactions), generalised additive models, regression trees, random forests (minimum node sizes of 50, 100, 150 and 200 individuals), extreme gradient boosting (with the same node specifications as in the random forests with combinations of shrinkage parameters at 0.001, 0.01 and 0.1), and elastic net regression (alpha at 0, 0.2, 0.4, 0.6, 0.8 and 1). Standard errors were corrected for strata of matched participants. Observations with missing values were excluded from the analyses.

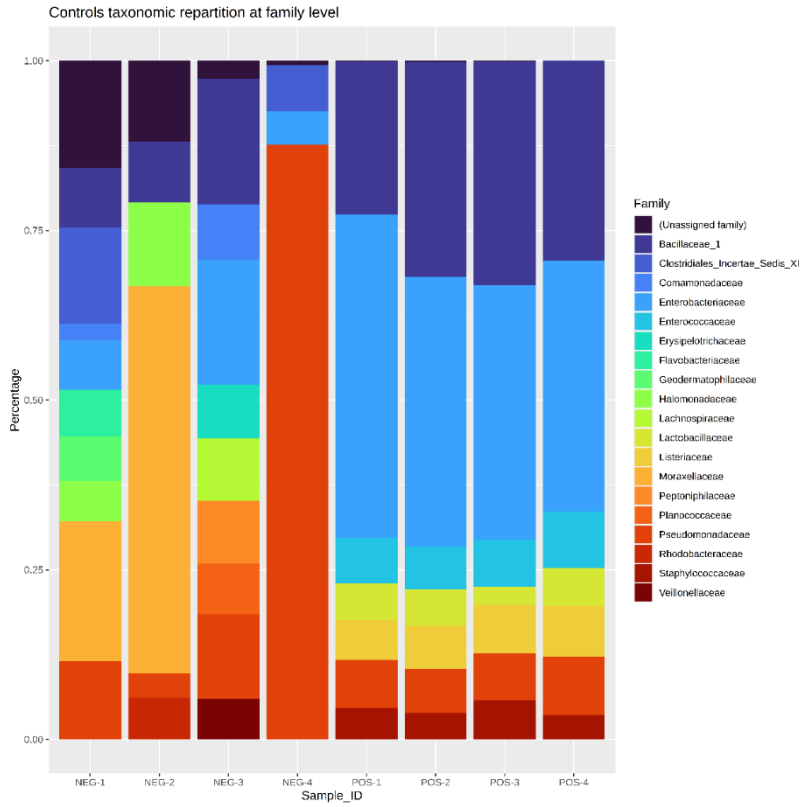
References:

1. Blackwell M, Iacus S, King G, Porro G. Cem: Coarsened Exact Matching in Stata. The Stata Journal. 2009 Dec;9(4):524-46
2. Iacus SM, King G, Porro G. Causal Inference without Balance Checking: Coarsened Exact Matching. Political Analysis. 2012;20(1):1-24.
3. Arcilla MS, van Hattem JM, Haverkate MR, et al. Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. Lancet Infect Dis 2017; 17:78-85.
4. Iacus SM, King G, Porro G. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. J Am Stat Assoc 2011; 106:345-361.
5. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J of Epidemiol 2017 185, 65-73

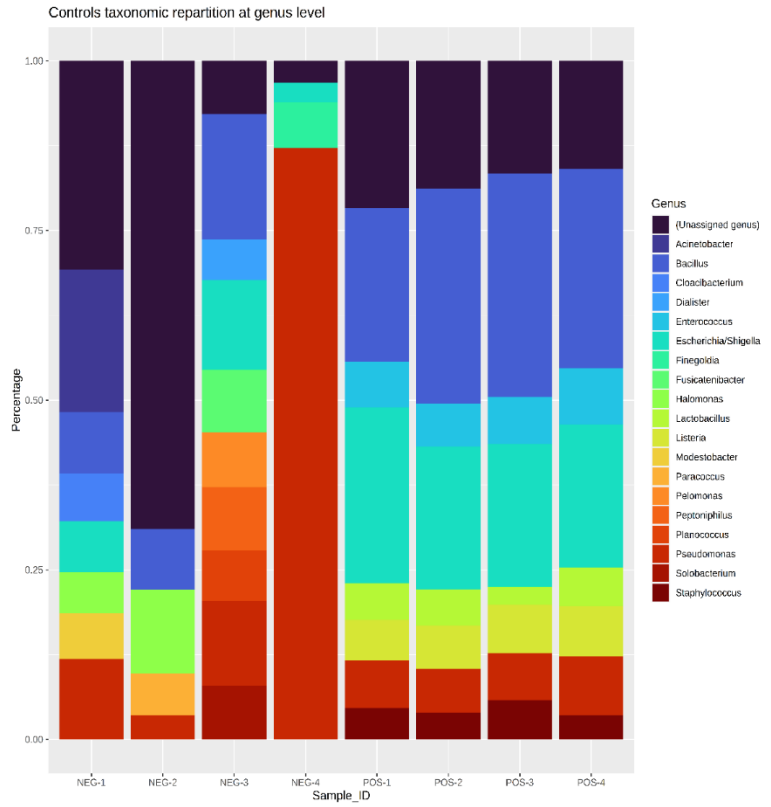
SUPPLEMENTARY FIGURES

Supplementary Figure 1. Results of the negative (NEG-1 to NEG-4) and positive (POS-1 to POS-4) controls used in the 16S sequencing method

A



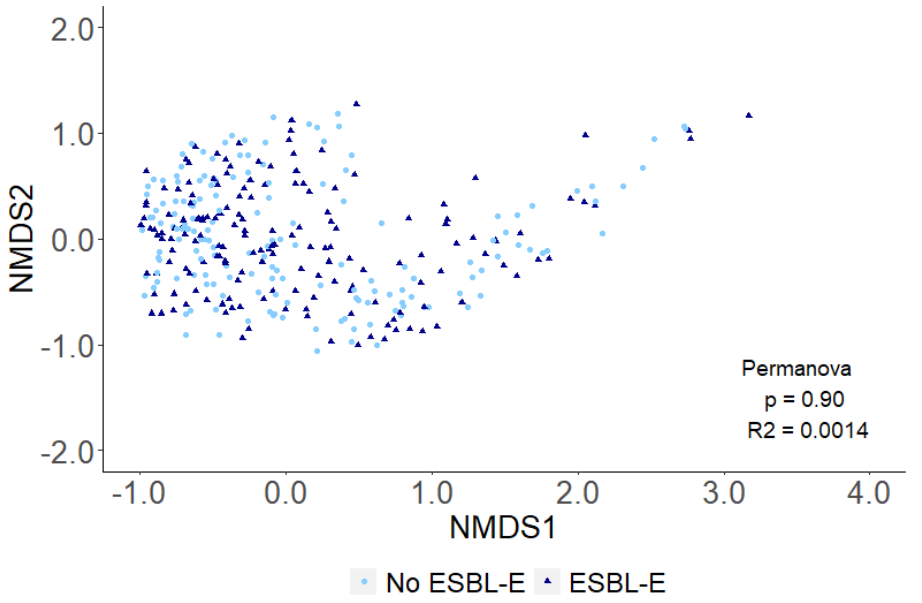
B



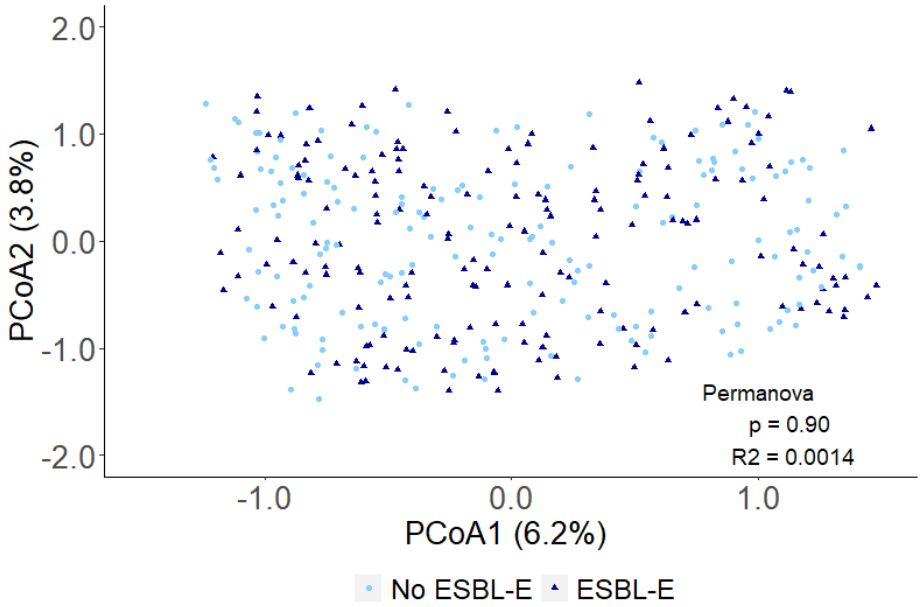
The positive control was represented by a microbial community standard (Zymobiomics®, Zymo Research, Irvine, CA, USA). Results are displayed at the family (**A**) and the genus level (**B**).

Supplementary Figure 2. Analysis of β-diversity between those with and without ESBL-E carriage (Canberra method)

A

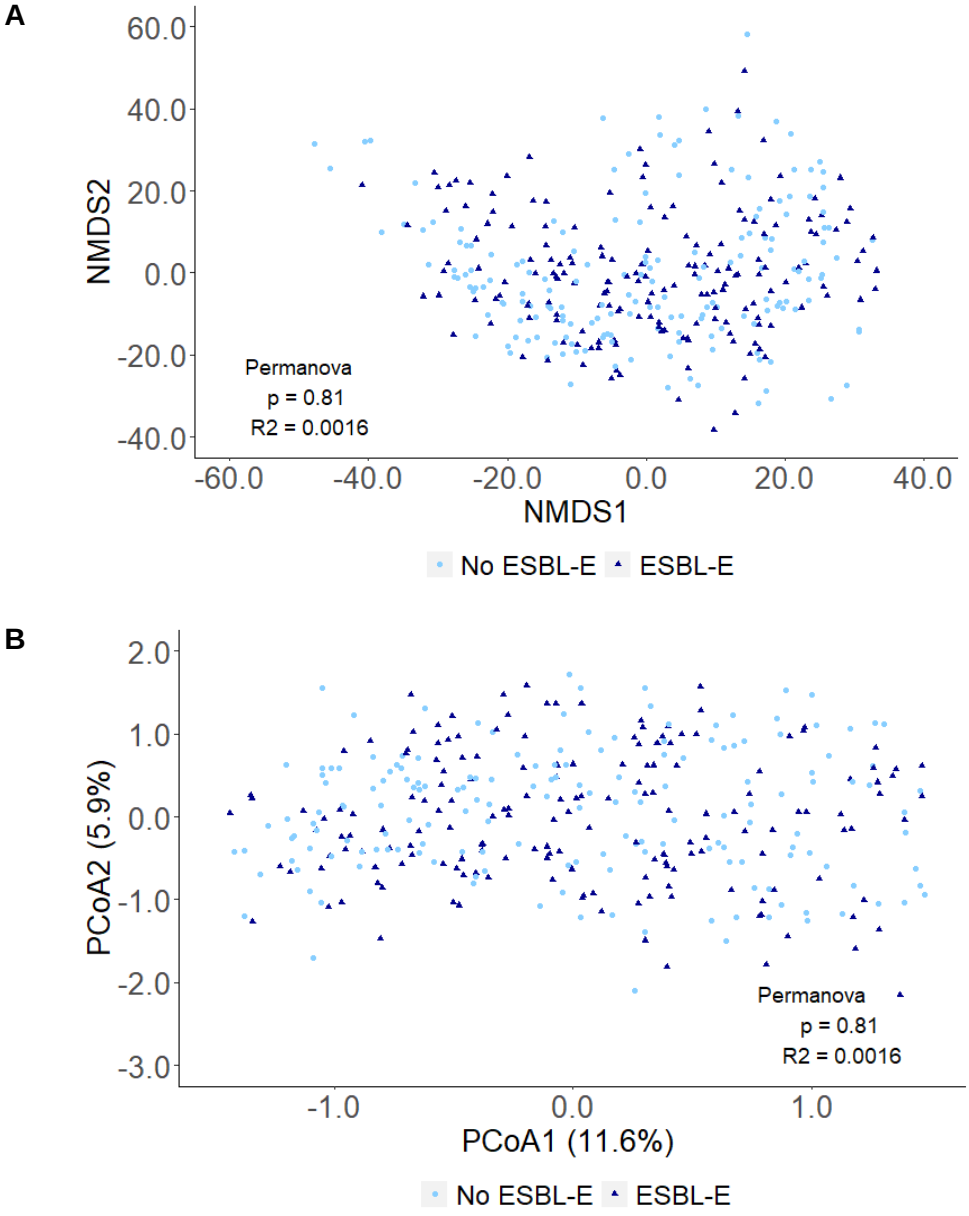


B



Non-metric multidimensional scaling ordination (**A**) and Principal Coordinates Analysis (**B**) plot of dissimilarities for microbial communities between individuals with and without ESBL-E carriage. Dissimilarities were calculated using the Canberra method. Analyses were based on relative abundances.

Supplementary Figure 3. Analysis of β-diversity between those with and without ESBL-E carriage (robust Aitchison method)



Non-metric multidimensional scaling ordination (**A**) and Principal Coordinates Analysis (**B**) plot of dissimilarities for microbial communities between individuals with and without ESBL-E carriage. Dissimilarities were calculated using the robust Aitchison method. Analyses were based on relative abundances.