Overview of code and data for “Sequence type diversity amongst antibiotic-resistant bacterial strains is lower than amongst antibiotic-susceptible strains”

Data

To begin with we created sequence type level data for three of the data sets (Yamaji1999, Yamaji 2016 and Adams-Sapper datasets) directly from data in the journal articles (Yamaji et al. 2018; Adams-Sapper et al. 2013).

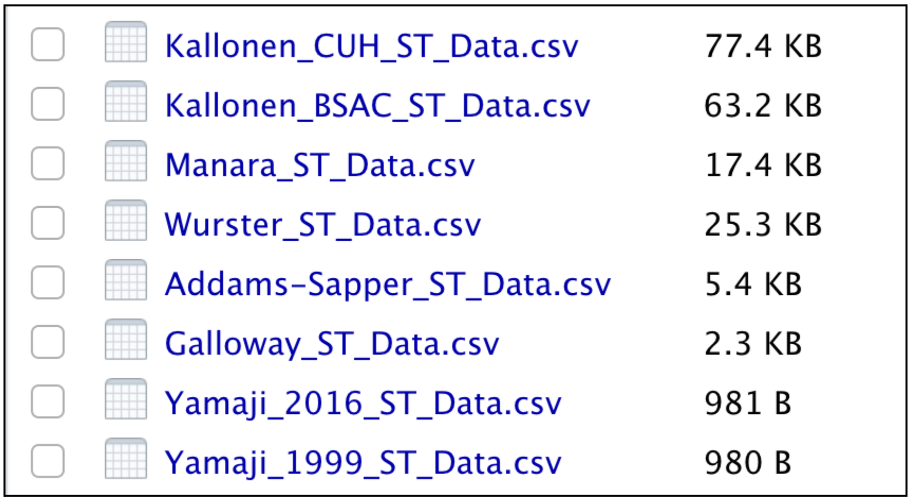
For five of the data sets (Kallonen_CUH, Kallonen, Kallonen_BSAC, Galloway, Manara, and Wurster datasets) the original papers provided data at the level of patients (Galloway Peña et al. 2009; Kallonen et al. 2017; Manara et al. 2018; Wurster et al. 2018).

We saved these datsets as patient level data (“Kallonen_E_coli_Patient_Data.csv”, “Manara_S_aureus_Patient_Data.csv”, “Wurster_S_aureus_Patient_Data_Revised”, and “Galloway_Patient_Data.csv”). Next, we used custom R scripts to convert those patient level data into sequenc type level data. The ST level data and the respective Rscripts are named as following;

	Name of the datasets	**Name of the RScripts**
	Datasets directly from paper (ST level csv files)	
1.	Yamaji_1999_ST_Data.csv	N/A
2.	Yamaji_2016_ST_Data.csv	N/A
3.	Addams-Sapper_ST_Data.csv	N/A
	Dataset we converted to ST level (ST level csv files)	
4.	Kallonen_E_coli_Patient_Data.csv → Kallonen_BSAC_ST_Data.csv	PatientLevelToSeqTypeLevelData_Kallonen_2.R
5.	Kallonen_E_coli_Patient_Data.csv → Kallonen_CUH_ST_Data.csv	PatientLevelToSeqTypeLevelData_Kallonen_2.R
6.	Wurster_S_aureus_Patient_Data_Revised → Wurster_ST_Data.csv	PatientLevelToSeqTypeLevelData_Wurster.R
7.	Manara_S_aureus_Patient_Data.csv →	PatientLevelToSeqTypeLevelDat

	Manara_ST_Data.csv	a_Manara.R
8.	Galloway_Patient_Data.csv → Galloway_ST_Data.csv	PatientLevelToSeqTypeLeveData_Galloway.R

Screenshot of the resulting ST data files



☐	Kallonen_CUH_ST_Data.csv	77.4 KB
☐	Kallonen_BSAC_ST_Data.csv	63.2 KB
☐	Manara_ST_Data.csv	17.4 KB
☐	Wurster_ST_Data.csv	25.3 KB
☐	Addams-Sapper_ST_Data.csv	5.4 KB
☐	Galloway_ST_Data.csv	2.3 KB
☐	Yamaji_2016_ST_Data.csv	981 B
☐	Yamaji_1999_ST_Data.csv	980 B

Analysis

- After we created ST level csv files for each of the data sets, we then proceeded with our analysis;

Pie charts

- "Rscript_MakesPieCharts.R" takes ST level csv files, and creates pie charts for resistant and susceptible based on Sequence Types, there are 60 pie charts stored in "Output/PieCharts" as png files.

Diversity indices and permutation tests

- Diversity_Indices_WithFunctions.R takes all ST level csv files, calculates the Gini-Simpson Index, the Inverse Simpson Index and the Shannon Index and writes to "DivIndices.csv".
- Diversity_Indices_WithFunctions.R unifies some of the drug names between the datasets and removes "minor", "ND" and "-".

- Diversity_Indices_WithFunctions.R also carries out the permutation tests and stores the p values in "DivIndices.csv".
- Diversity_Indices_WithFunctions.R also visualizes simulated histograms for all three diversity indices and stores them in "DivIndices_Histograms" folder.

Table 1

- We manually created the "Antibiotic_Classification.csv" file which contains all the antibiotics based on all the ST level csv files, and we classified each of the antibiotics.
- "OverviewDatasets.R" takes two files "Output/DivIndices.csv" and "Data/Antibiotic_Classification.csv" files and creates a combined csv files "Output/OverviewDatasets.csv"
- Table 1 is made by hand, based on these csv files.

Figure 1

- Rscript_Figure1_Bar_Pie.R creates Fig_1_Bar_PieCharts.png

Figure 2

- "Rscript_Figure2_GSI_Datasets.R" takes "OverviewDatasets.csv", and the R script and makes two barplots "Figure2A_GiniSimpson.png" and "Figure2B_FracRes.png" which are based on Drug Vs Simpson's index and fraction ratio of resistant count.

Figure 3

- GSI_normalized_linearmodel.R.R should uses DivIndices.csv and creates following files: Figure3_GSINormalized_values_Res.png text file with model test: "Output/Figure3_lm_output_GSINormalizedR.txt".
- Writes data to "Output/DataNormelizedGSI.csv"