Juan José Rubio Guillamón - s162166
Jakub Czerny - s161200

Antimicrobial resistance, what do we know?

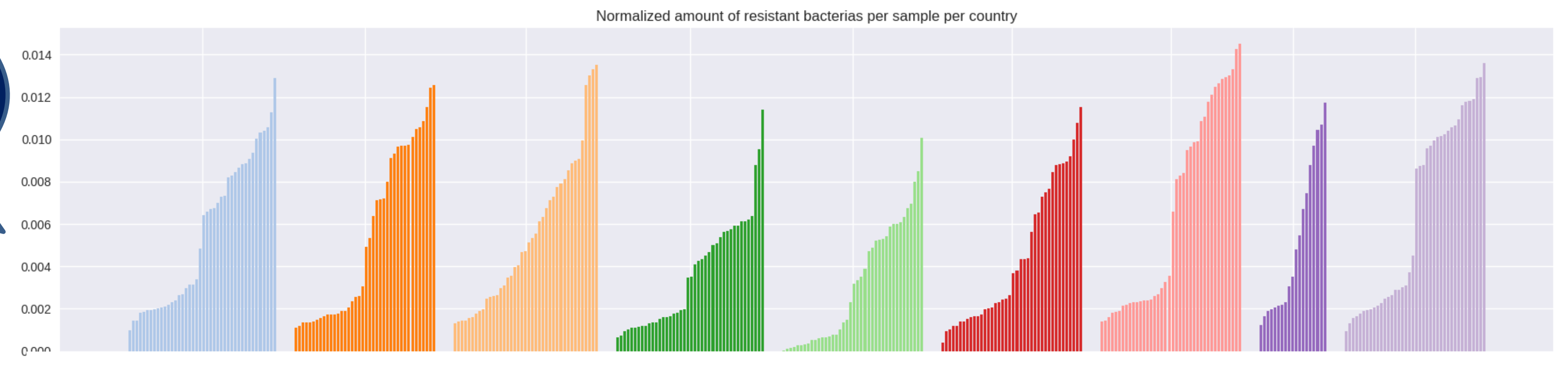
Data exploration

Even after performing the normalization, the samples within the countries are not very uniform. All the plots however, have similar shapes, and are composed of 2 different looking parts. This was a trigger to analyze samples for both, swine and poultry separately as they two segments of the plots may correspond to different animals.

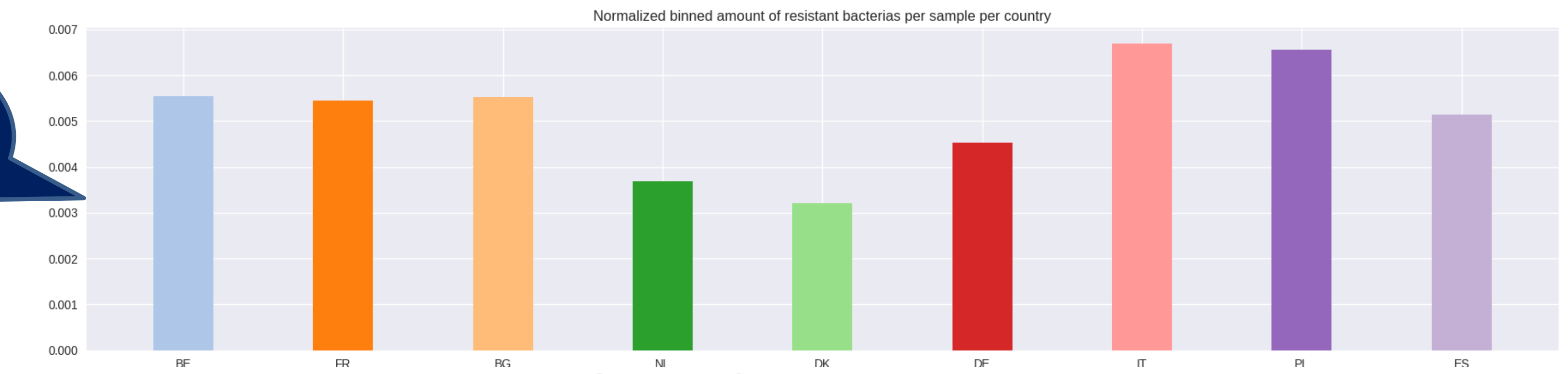
To get a better overview on the shape of the animals in the countries, we also looked at the average amount of bacteria over all the samples within each country. Clearly, Denmark and the Netherlands have animals with the least number of AMRs whereas Italy and Poland perform the worst.

The second plot may be considered relevant because we only have average amount of administered drugs per country with no distinction on farms of even animals. Interestingly, Denmark and the Netherlands not only have the lowest rate of AMRs but also usage of the drugs. On the other end of the scale are Spain and Italy.

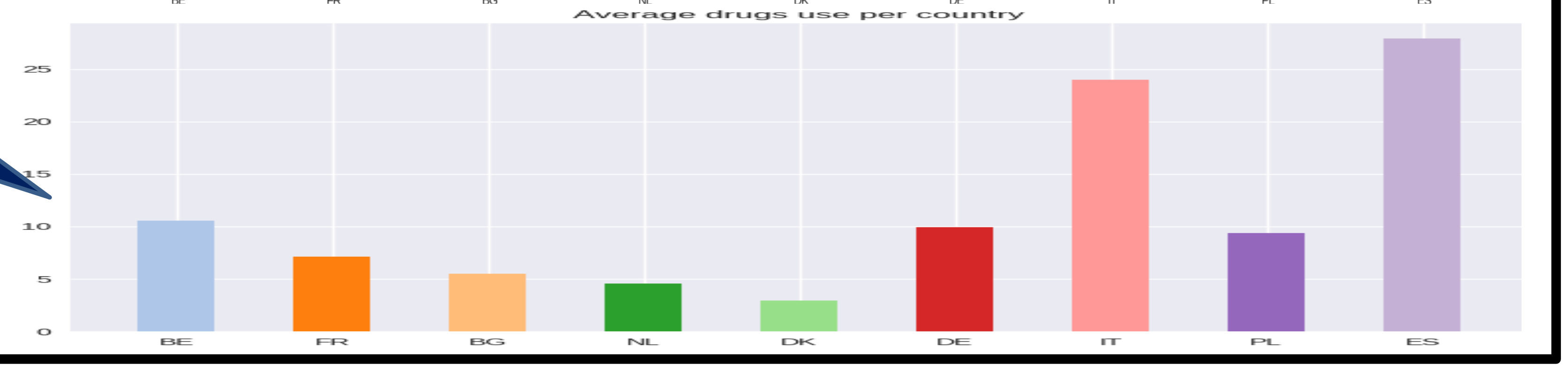
How different are the samples between and within the countries



Well, that's a lot of data, what's average value for each country?



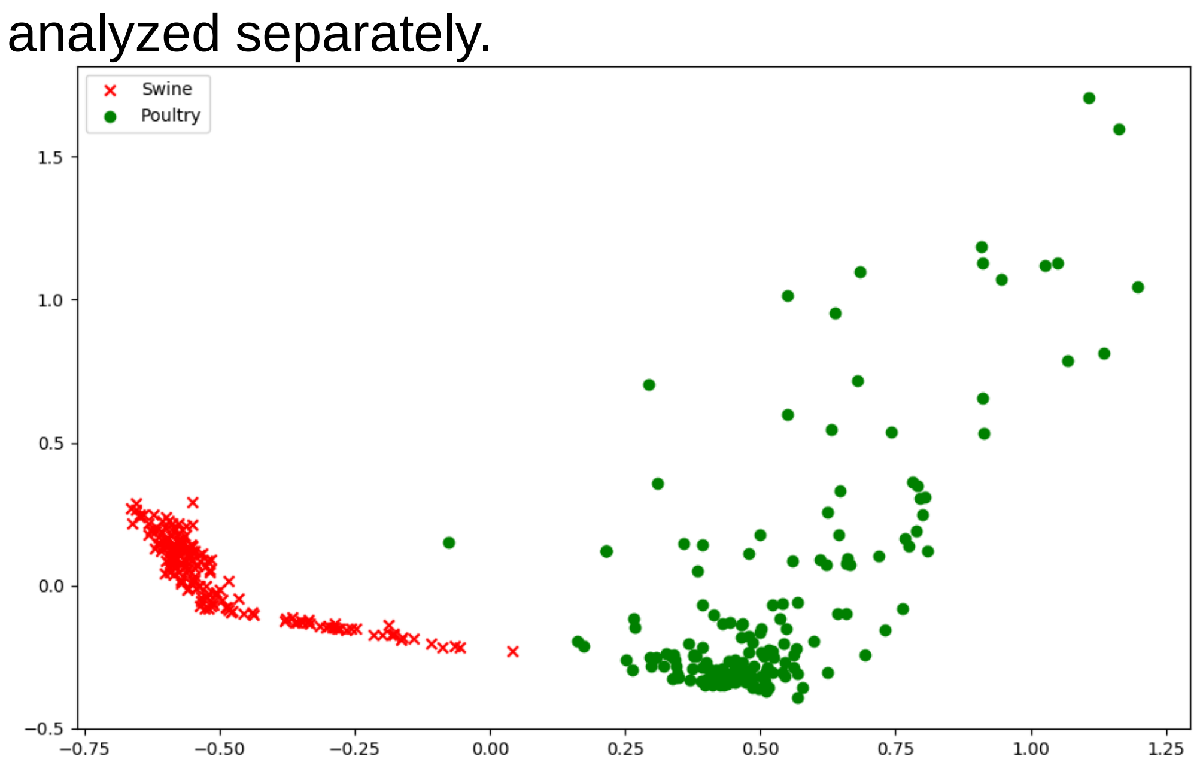
Drugs, drugs, drugs..



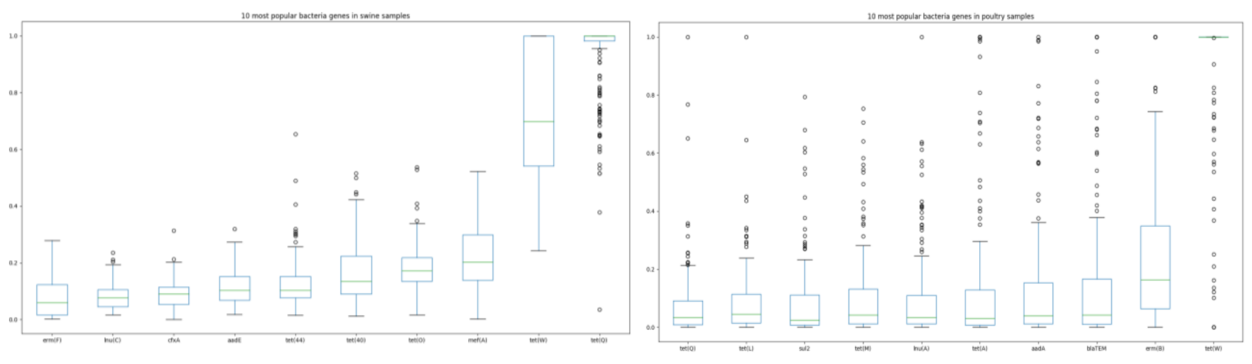
Analysis

PCA

Performing PCA and extracting only 2 components showed that the swine and poultry samples are significantly different, thus should be analyzed separately.



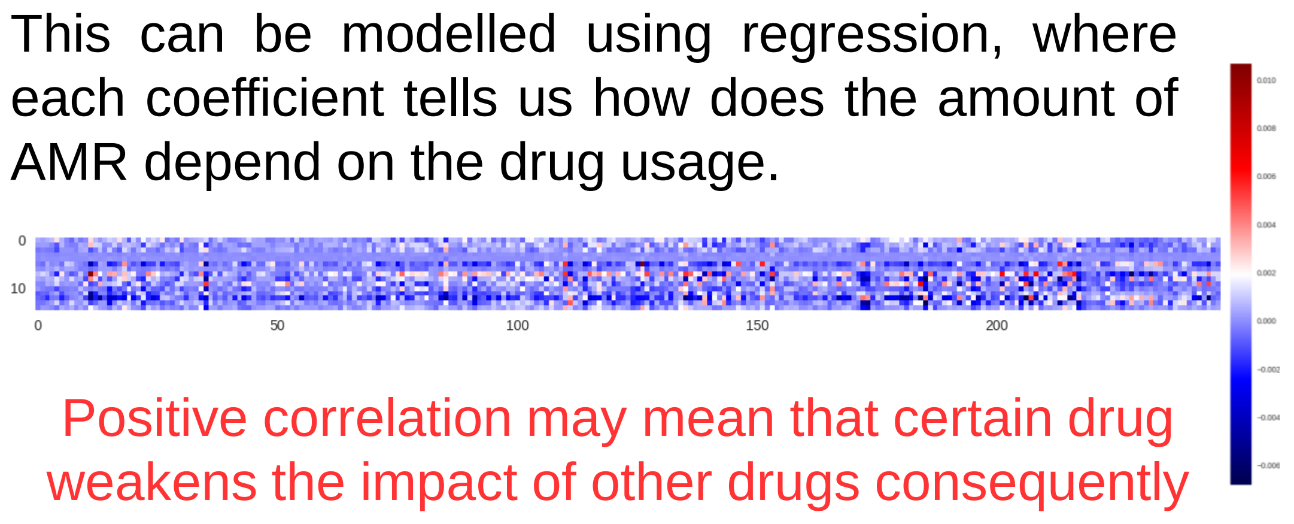
10 most common AMRs



Interestingly, tet(W) is the most common AMR for swine and second most popular for poultry. Moreover, tet(Q) also appears among the most common AMRs for both animals.

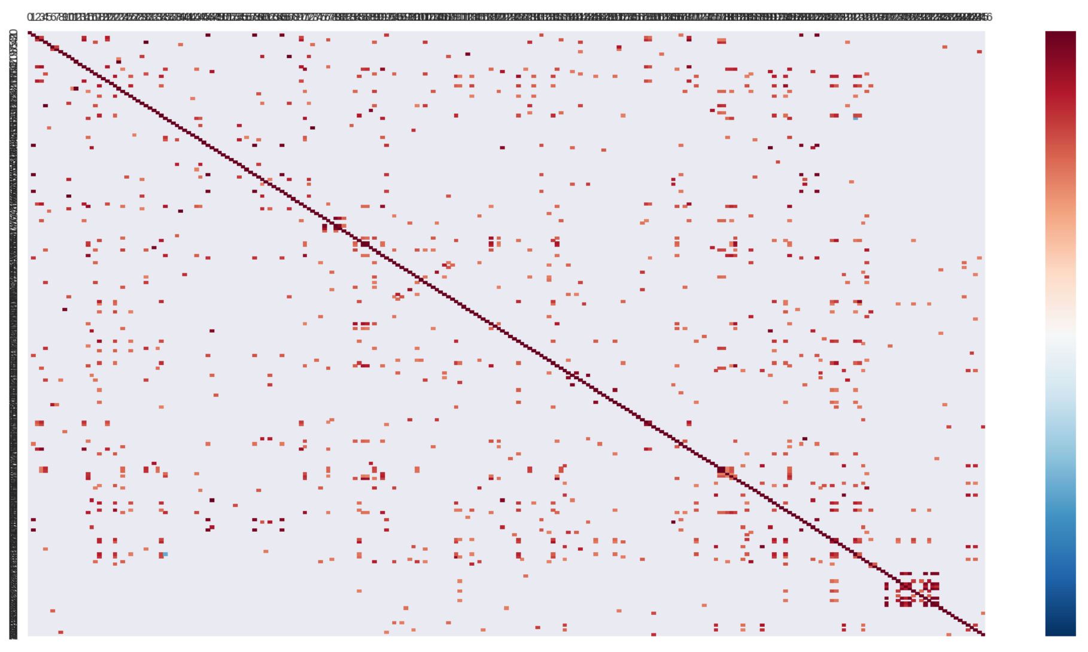
Regression - Lasso

So what's the relationship between amount of each drug and presence of the AMRs? This can be modelled using regression, where each coefficient tells us how does the amount of AMR depend on the drug usage.



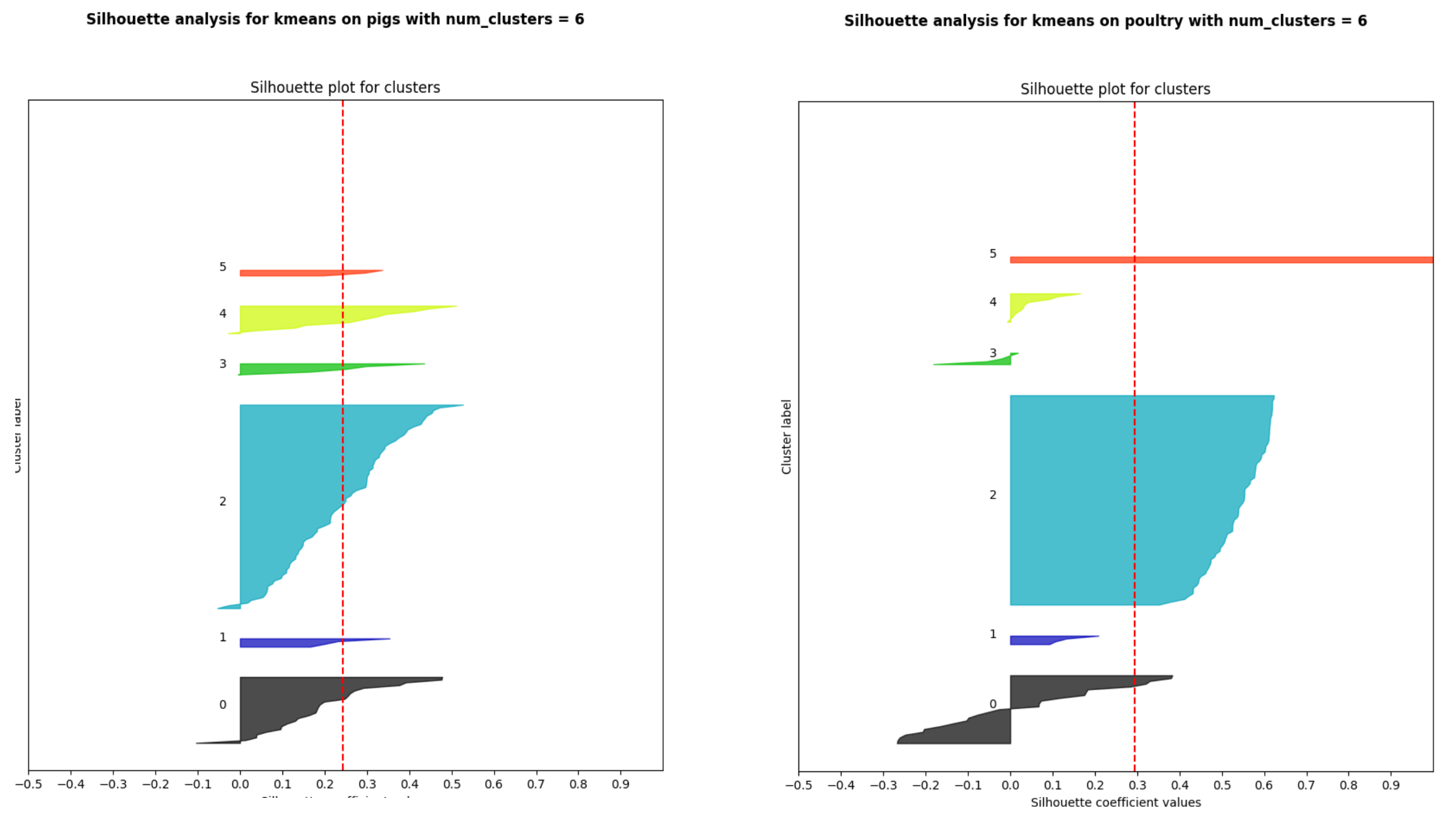
Positive correlation may mean that certain drug weakens the impact of other drugs consequently leading to higher number of bacteria.

To render this analysis more relevant, it should be combined together with an analysis of multiple correlation between the AMRs which is presented by the following figure. Here, the red colour denotes that two genes are coexisting symbiotically, whereas blue (less present) the opposite.



Clustering – k means / GMM

Since even the samples for one animal from the same country varied a lot between each other we decided to perform grouping using clustering. Afterwards we inspected the clusters to find in what sense the samples were similar. As an evaluation criterion for number of clusters we used Silhouette coefficients that measures the intra-cluster distance.



We tested different number of clusters ranging from 3 to 12 so that the number 9 which corresponds to the number of countries is enclosed within the range. Eventually, we picked 6 cluster for both animals as they yielded optimal average Silhouette values for all the samples.

Conclusion & results

Due to the drug usage data limitation, our study has focused mainly on data exploration based on unsupervised learning considering the difficulty and unreliability of fitting prediction models for this case. This data exploration derived some conclusions regarding the AMR distribution in each subset of individuals:

- Difference in AMR gene pool based on animal species. tet(X) (W,Q,0,40,44) are predominant homogeneously in pigs being top 10 AMRs in no less than 95% of pigs for all the clusters. In poultry, tet(W) is top 1 for all the individuals, while the remaining tet(X) are in the top 10 for the average 50%.
- Higher variability in terms of AMR presence in poultry compared to swine. Verified both in statistical and cluster analysis.
- Relation between country and AMR genes. Top 2 countries defining a particular country represent on average at least 50% of the individuals.

cluster	0		1		2		3		4		5	
# samples	30		37		46		33		23		4	
10 majority AMRs (%)	tet(Q)	100.0	tet(Q)	100.0	mef(A)	100.0	tet(Q)	100.0	mef(A)	100.0	tet(Q)	100.0
	tet(W)	100.0	tet(W)	100.0	tet(Q)	100.0	tet(W)	100.0	tet(Q)	100.0	aadE	100.0
	tet(O)	100.0	tet(40)	100.0	tet(W)	100.0	tet(O)	100.0	aadE	100.0	tet(44)	100.0
	aadE	100.0	tet(O)	100.0	tet(O)	100.0	tet(40)	100.0	tet(44)	100.0	tet(W)	100.0
	mef(A)	96.7	mef(A)	97.3	aadE	100.0	mef(A)	97.0	tet(W)	100.0	ant(6)-I	100.0
	tet(44)	96.7	aadE	97.3	tet(44)	97.8	aadE	97.0	tet(O)	100.0	tet(40)	100.0
	tet(40)	93.3	erm(F)	94.6	tet(40)	95.7	tet(44)	97.0	tet(40)	95.7	tet(O)	100.0
	lnu(C)	76.7	tet(44)	91.9	cfxA	89.1	lnu(C)	78.8	cfxA	87.0	mef(A)	75.0
	erm(F)	70.0	lnu(C)	86.5	lnu(C)	87.0	cfxA	66.7	lnu(C)	60.9	lnu(C)	50.0
	cfxA	56.7	cfxA	70.3	erm(F)	32.6	erm(F)	63.6	erm(F)	47.8	tet(32)	50.0

3 majority countries (%)	DE	26.7	IT	35.1	NL	34.8	ES	27.3	BG	30.4	DE	100.0
	ES	26.7	FR	29.7	DK	23.9	DE	21.2	PL	26.1		
	FR	16.7	BE	16.2	BG	15.2	BE	12.1	DK	17.4		

cluster	0		1		2		3		4		5	
# samples	31		95		12		22		23		3	
10 majority AMRs (%)	tet(W)	100.0	tet(W)	98.9	tet(W)	100.0	tet(W)	100.0	tet(W)	100.0	tet(W)	100.0
	erm(B)	93.5	erm(B)	80.0	erm(B)	83.3	erm(B)	72.7	erm(B)	100.0	strB	100.0
	tet(L)	58.1	blaTEM	64.2	tet(A)	66.7	lnu(C)	59.1	tet(Z)	66.7	aadA	66.7
	tet(Q)	54.8	aadA	54.7	lnu(C)	58.3	blaTEM	59.1	strB	66.7	tet(M)	66.7
	tet(M)	54.8	tet(Q)	50.5	tet(Q)	50.0	aadA	59.1	aac(3)-IV	66.7	sul2	66.7
	tet(O)	51.6	tet(M)	49.5	tet(M)	50.0	lnu(A)	54.5	cmx	66.7	tet(A)	66.7
	lnu(A)	51.6	tet(A)	47.4	lnu(A)	50.0	sul2	50.0	aph(4)-I	66.7	strA	66.7
	blaTEM	51.6	tet(L)	45.3	blaTEM	50.0	tet(M)	45.5	lnu(A)	33.3	erm(B)	66.7
	lnu(C)	48.4	lnu(A)	43.2	aadA	50.0	aadE	40.9	tet(Q)	33.3	tet(Q)	33.3
	aadE	45.2	lnu(C)	43.2	aadE	50.0	tet(A)	40.9	aadE	33.3	lnu(C)	33.3

3 majority countries (%)	BE	32.3	FR	16.8	BG	25.0	BG	27.3	ES	66.7	BG	33.3
	IT	19.4	DK	14.7	DE	25.0	DK	18.2	FR	33.3	NL	33.3
	NL	12.9	NL	12.6	PL	16.7	DE	18.2			ES	33.3