Whole genome typing methods for differentiating between epidemiologically related and unrelated *E. coli* ST131 isolates

Freek de Kreek[1,2]

J. J. M. Stohr[2], J.J. Verweij[2], S. Pas[2], E. Schrauwen[1]

[1]Academy of Life Sciences and Technology, Avans University of Applied Sciences, Breda, Noord-Brabant, Netherlands

[2]Microvida, Tilburg, Noord-Brabant, Netherlands

Objectives

- Determine the discriminatory power of whole genome typing methods: SeqSphere, chewBBACA, PopPUNK and Snippy for distinguishing between strictly defined epidemiologically related and unrelated *E. coli* ST131 genomes
- Define 3 cutoffs values for each whole genome typing method:
    1. Cutoff at the lowest threshold where all epidemiologically related cases are genetically related
    2. Cutoff at the highest threshold where all epidemiologically unrelated cases are genetically unrelated
    3. Cutoff at the maximum accuracy
- Determine the discriminatory power of whole genome typing methods at 3 different cutoffs for distinguishing between *E. coli* ST131 genomes

Materials & methods

Data acquisition

The dataset consists of *E. coli* ST131 isolates with strictly related (same patient within 3 months) and unrelated (different hospital or nursery home). Draft *E. coli* ST131 genomes were obtained by sequencing on the Illumina Miseq platform using Nextera XT and assembled *de novo* in CLC genomics workbench.

Whole genome typing comparison

Whole genome typing was performed using SeqSphere, chewBBACA, PopPUNK, and Snippy for related and unrelated *E. coli* ST131 isolates. The medians of genetic distances of related and unrelated were compared using the Mann-Whitney U test and the monotonic relationship between the typing methods was determined with the Spearman Rank correlation. The percentage misclassifications (un)related and threshold for each typing method was determined for the 3 previously mentioned cutoffs.

Data availability

Data analysis of the genetic distances was performed with Python and available on GitHub:

The FASTA's of the draft genomes, and exact parameters for running the whole genome typing tools can be found on GitHub as well.

GitHub

Results

Distribution of genetic distances

All whole genome typing tools showed a significant genetic difference between epidemiologically related and unrelated based on the Mann-Whitney U test.

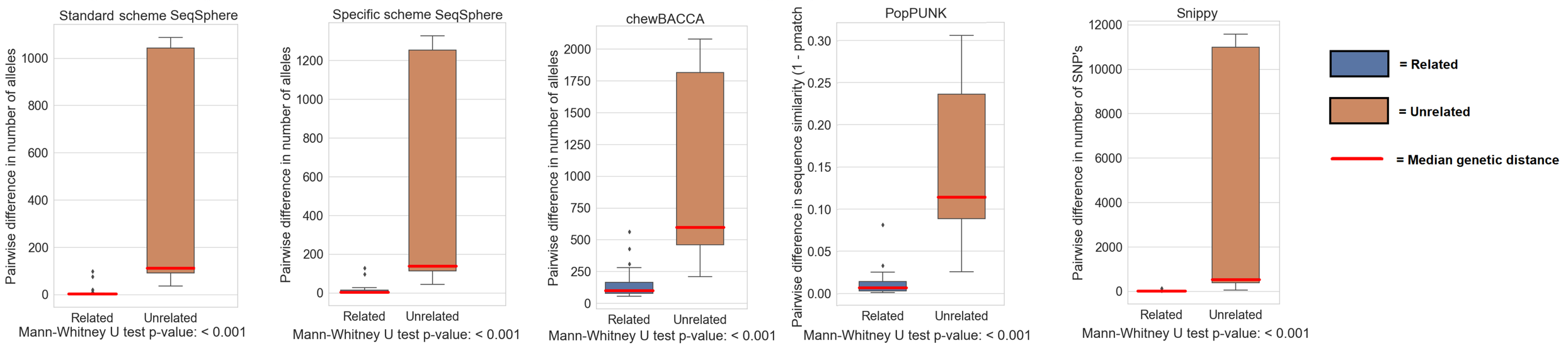

Figure 1, each boxplot displays the distribution of the pairwise genetic differences in either alleles, Single Nucleotide Polymorphisms (Snippy) or the probability of a k-mer will match between a pair of sequences (PopPUNK) for samples labeled as related and unrelated shown in a boxplot. The median is indicated in red and the probability is significant (the p-value is smaller than 0.05).

Spearman Rank Correlation

The high Spearman rank correlation coefficients indicate a high correlation between the ranking of genetic distances between the *E. coli* ST131 isolates across the different whole genome typing methods.

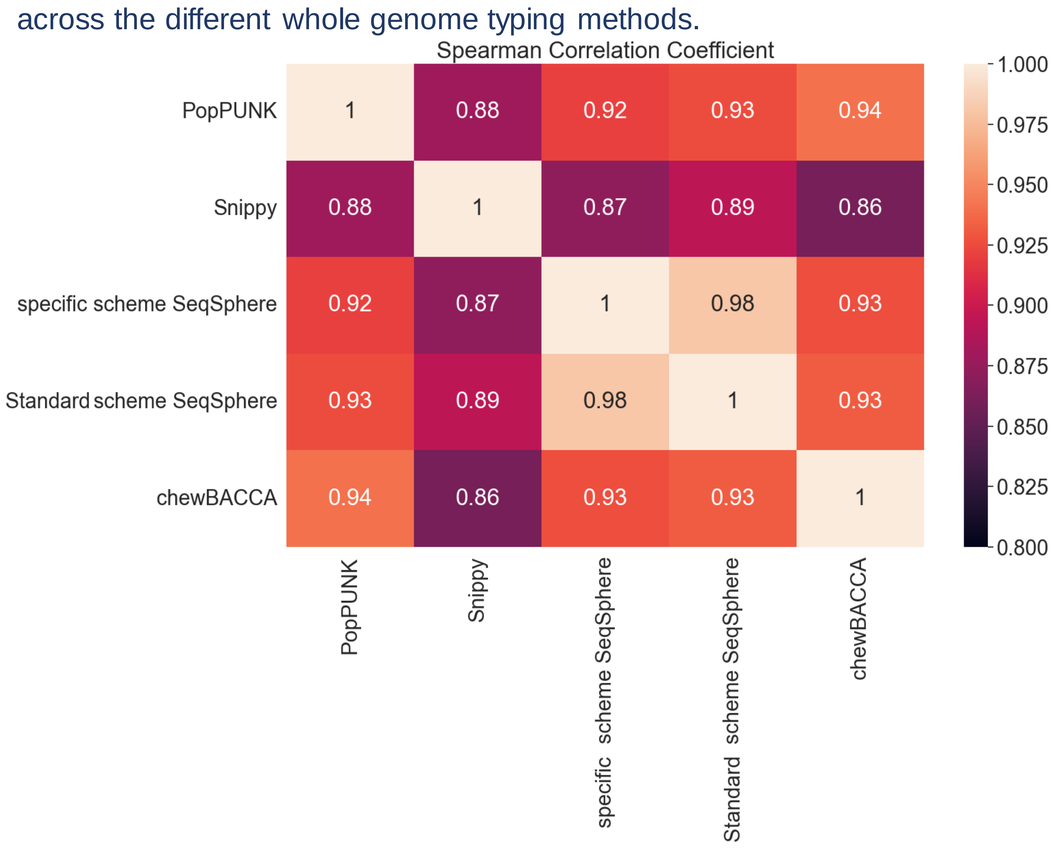

Figure 2, a heatmap of the Spearman rank correlation coefficient between the whole genome typing methods.

Determination of cutoff values

The percentage of misclassifications as related at the cutoff where all related typed as genetically related (cutoff 1) was between 35.9% - 47.1% for the wgMLST schemes, 18.3% for PopPUNK and 3.3% for Snippy. When outliers (two related cases) were excluded, the misclassification was less than 1%, but chewBBACA remained unchanged.

The percentage of misclassifications as related was equal for cutoffs 2 & 3 was 8.0% for all typing tools except chewBBACA. The percentage of misclassifications as related was approximately double for both cutoffs.

Typing method	Cutoff 1	Misclassified as related	Cutoff 2	Misclassified as unrelated	Cutoff 3	Misclassified as unrelated	Misclassified as related
Stable scheme Seqsphere	97	35.9%	34	8.0%	34	8.0%	0.0%
Specific scheme Seqsphere	128	43.1%	42	8.0%	42	8.0%	0.0%
chewBACCA	561	47.1%	208	20.0%	237	16.0%	0.7%
PopPUNK	0.0810	18.3%	0.0256	8.0%	0.0339	8.0%	0.0%
Snippy	120	3.3%	44	8.0%	44	8.0%	0.0%

Table 1, the percentage of misclassifications as either related or unrelated for each cutoff per typing method.

Discussion & conclusion

The whole genome typing methods have enough discriminatory power to distinguish strictly related from strictly unrelated *E. coli* ST131 isolates. The Mann-Whitney U test shows that there is a significant difference between the genetic distances between related and unrelated *E. coli* ST131 isolates. The ranking of the pairwise distances is highly similar as indicated by the Spearman Rank correlation coefficients being close to 1. Cutoff 1, at which epidemiologically related are genetically related, was influenced considerably by two related cases.

For future study, more insight might be gained by evaluating clustering for each typing method at the three determined cutoffs on a more 'mixed' dataset (less distinction between related and unrelated *E. coli* ST131 isolates.

Contact:
f.dekreek@student.avans.nl