Identification of the most versatile mediators of the horizontal gene transfer by assessing the genomic signature distances between plasmids and bacteria

Vicente Arnau[1,2,3] , Wladimiro Diaz-Villanueva[1,2,3], Mária Džunková[1], Andrés Moya[1,2,3]

1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
3. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain

INTRODUCTION

Plasmid are among the principal mediators of the horizontal gene transfer (HGT) in microbial communities. In order to assess their impact on evolution, including the HGT across distant bacterial taxa, we need to link them with their hosts.

Linking plasmids with their host is straightforward in the case of genomes of isolated strains, however, it is extremely complicated in the case of mixed bacterial communities present in natural environments due to the difficulties to link the plasmids with their producers and recipients

METHODS

In this study, we analyzed the possibility to link the plasmids with their host strains by genomic signature distances.

A) The Euclidean distance between the relative frequencies of short length k-mers (in our case k = 4) and for a given k-mer w, its occurrence in a contig X is defined as X_w and the relative frequency of this k-mer is defined as:

$$f_w^X = \frac{X_w}{\sum_w X_w}$$

Following the guidelines of Vinga & Almeida (2003), we calculated the Euclidean distance between the pairs of genomes:

$$Eu(X,Y) = \sqrt{\sum_{w \in S^k} |f_w^X - f_w^Y|^2}$$

B) Alignment-free comparison based on exact k=14 oligonucleotide matches. We proposed a new distance of similarity for high values of k (k > 14) [3], where the value of 4k is two orders of magnitude larger than the size of the largest genome.

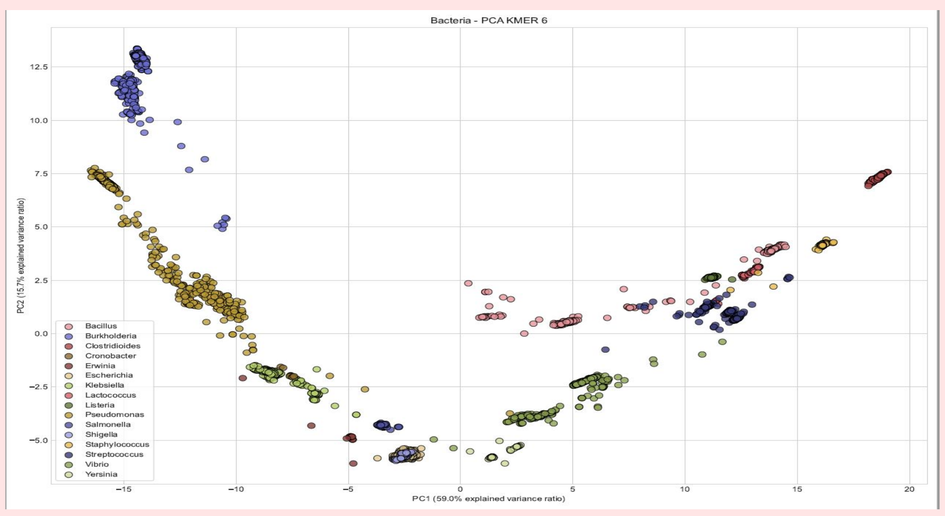
$$SX = \sum_w X_w \qquad SY = \sum_w Y_w$$

If SX < SY, we define the similarity function SIM between two sequences as:

$$SIM(X,Y) = \frac{\sum_w X_w}{SX} \qquad \forall (Y_w > 0)$$

Finally, we define the distance measure *DSW* as the inverse of the Similarity function, as follows:

$$DSW(X,Y) = (1 - SIM(X,Y))$$

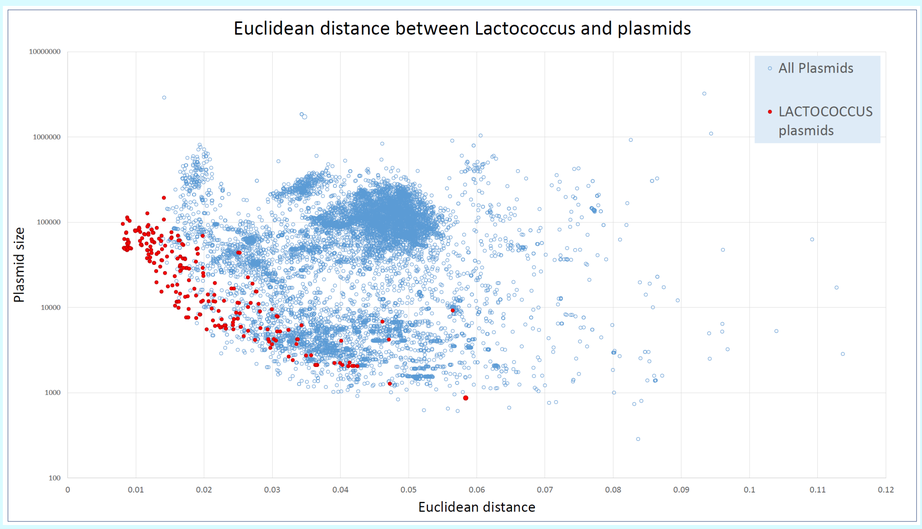


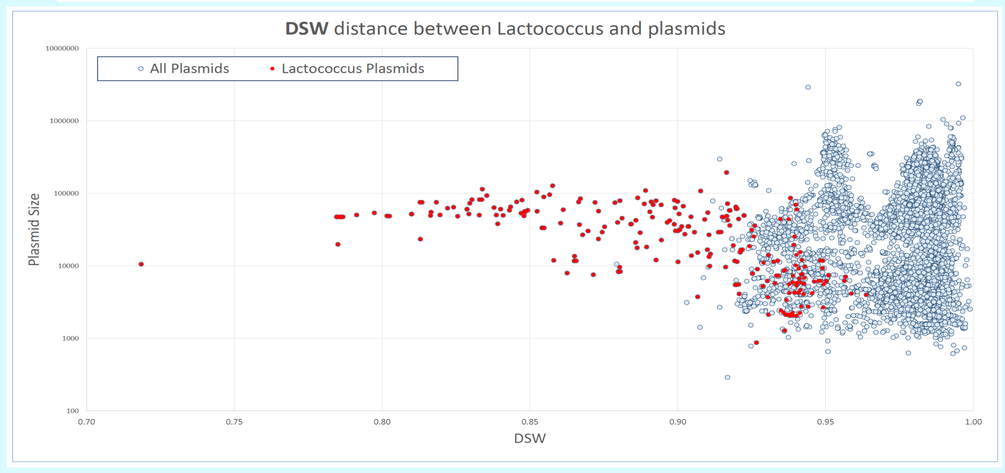
PCA of the 16 bacteria with clinical interest

RESULTS

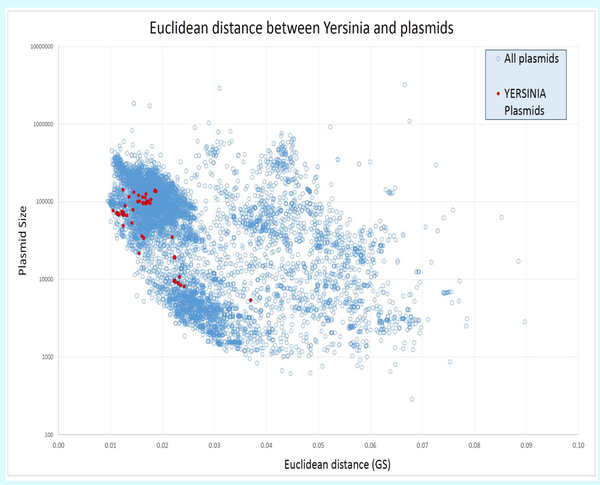
We analyzed 12,061 plasmids from 8,715 strains from 16 bacterial genera from the RefSeq NCBI database, which come from genomes of bacterial isolates with assigned plasmids. The results showed that the plasmids assigned to their bacterial host have significantly shorter genomic distance to their assigned host that to the non-related bacteria. Nevertheless, for each bacterial species we identified a small set of promiscuous plasmids that had shorter genomic distance to non-related bacterial genomes belonging to different genera.

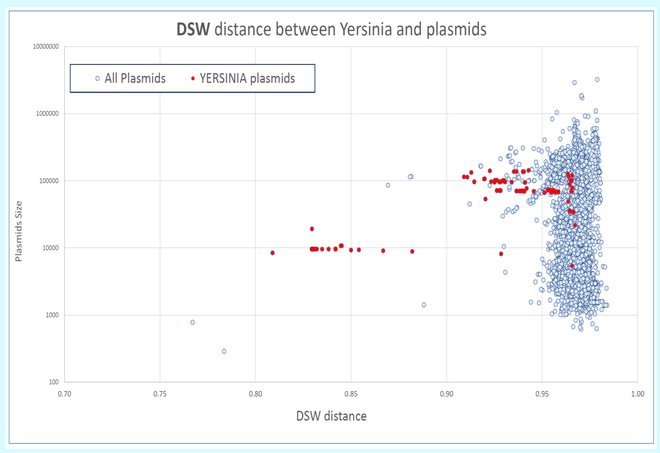
The results suggest that such plasmids are versatile drivers of the HGT across genera.





We compare the two distances defined between the Lactoccocus plasmids and the rest of the plasmids to the 74 strains of the Lactococcus bacteria





We compare the two distances defined between the Yersinia plasmids and the rest of the plasmids of the 88 strains of the Yersinia bacteria.

CONCLUSIONS & NEAR FUTURE

The herein developed methods can be used for identification of promiscuous plasmids in metagenomic assemblies, which can shed light on evolution of complex microbial communities.

CONTACT

Vicente.Arnau@uv.es or Maria.Dzunkova@uv.es

@Vicent_Arnau @MDzunkova

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation [PID2019-105969GB-I00], the Generalitat Valenciana ([CIPROM/2021/042] and [CDEIGENT/2021/008]) and the Biomedical Research Networking Centre for Epidemiology and Public Health (CIBEResp).

The computations were performed on the HPC cluster *Garnatxa* at the Institute for Integrative Systems Biology (I2SysBio). I2SysBio is a mixed research center formed by the University of Valencia (UV) and Spanish National Research Council (CSIC).

REFERENCES

1. Pride DT et al. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8.
2. Deschavanne P et al. (2010) The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virol J 7, 163.
3. Swain, M.T. et al. Interpreting alignment-free sequence comparison: What makes a score a good score? NAR Genom. Bioinform. 2022, 4, lqac062.
4. Vinga, S.; Almeida, J. Alignment-free sequence comparison—A review. Bioinformatics 2003, 19, 513–523.