*PREDI*dicting bacterial *PATH*ogenicity on plant: *PREDIPATH*

Felipe LIRA[1], Gilles HUNAULT[2], Martial BRIAND[1], Perrine PORTIER[1], Claudine LANDES[1], and Marion FISCHER-LE SAUX[1]

[1]IRHS, INRA, Université d'Angers, Agrocampus-Ouest, SFR 4207 QuaSaV, 49071, Beaucouzé, France.

[2]Hémodynamique, Interaction Fibrose et Invasivité Tumorales Hépatiques Laboratory, Unité Propre de Recherche de l'Enseignement Supérieur 3859, Structure Fédérative de Recherche 4208, Bretagne Loire University, Angers, France.

Corresponding Author: felipelira3@gmail.com

Backgrounds

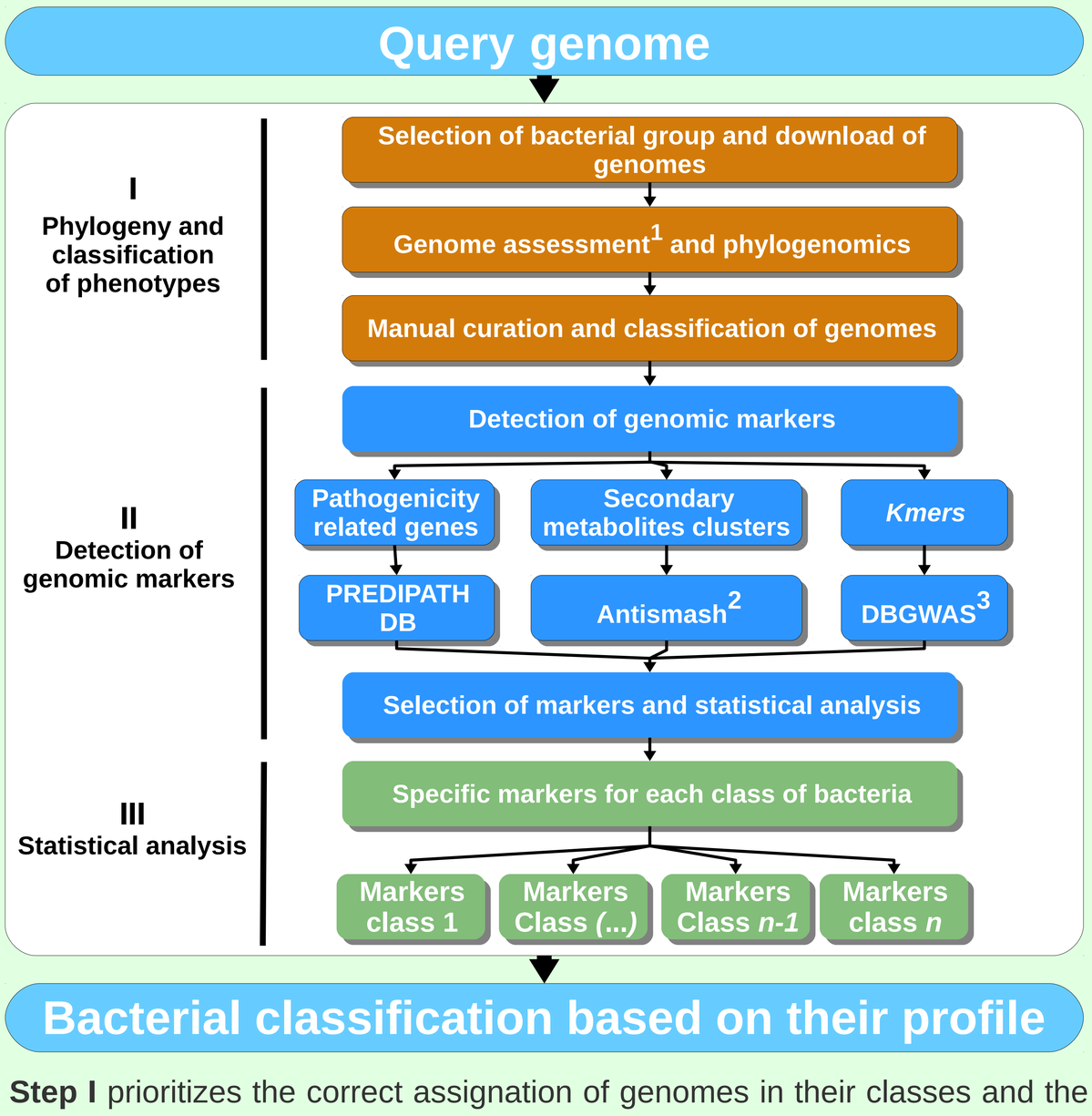
Prediction of bacterial pathogenicity commonly relies on microbiological methods. **Comparative genomics** emerges as a efficient method for distinction and detection of genomics elements able to distinguish two or more classes of organisms (**pathogenic vs. non-pathogenic; commensal vs. free-living organisms**). Genes, genes clusters, and operons, are closely associated with the bacterial survival and spread. Most of them are exclusive and determinants to characterize bacterial groups. In order to facilitate the prediction of potential bacterial plant-pathogenicity of plant-associated bacteria, we propose the **PREDIPATH workflow**.

Objective

Creation of specific datasets of genes, clusters and sequence markers (*kmers*) to discriminate bacterial species based on their genomic sequences.

Development

The PREDIPATH methodology relies on the detection of genome-based markers and creation of specific datasets of markers enabling to identify potential pathogenic organisms based on their genomes. PREDIPATH pipeline was developed using Python programming language and external bioinformatics tools in the process. Our methodology for detection of markers is summarized in three major **Steps**:



Step I prioritizes the correct assignation of genomes in their classes and the correction of their nomenclature when needed. **Step II** consisted in to create a customized database to detect potential genes to be used as markers - **PREDIPATH Database** (*a priori* approach); the detection of differential secondary metabolites clusters, and small DNA fragments, such as *Kmers* exclusive for each class of organisms described in **Step I. Step III** gave support for the results obtained in **Step II**.

Benchmarking

PREDIPATH Database was compiled clustering the data from public repositories[4,5,6] comprising a non-redundant dataset of sequences close-related with bacterial virulence and antimicrobial resistance.

PREDIPATH Workflow was tested using genomes from genus *Erwinia.*

68 Genomes downloaded

Completeness of genomes[1] >=85% of genes for Enterobacteriaceae

Elimination of duplicated genomes

Manual annotation of phenotypes

Phylogeny using bacterial core genes[7]

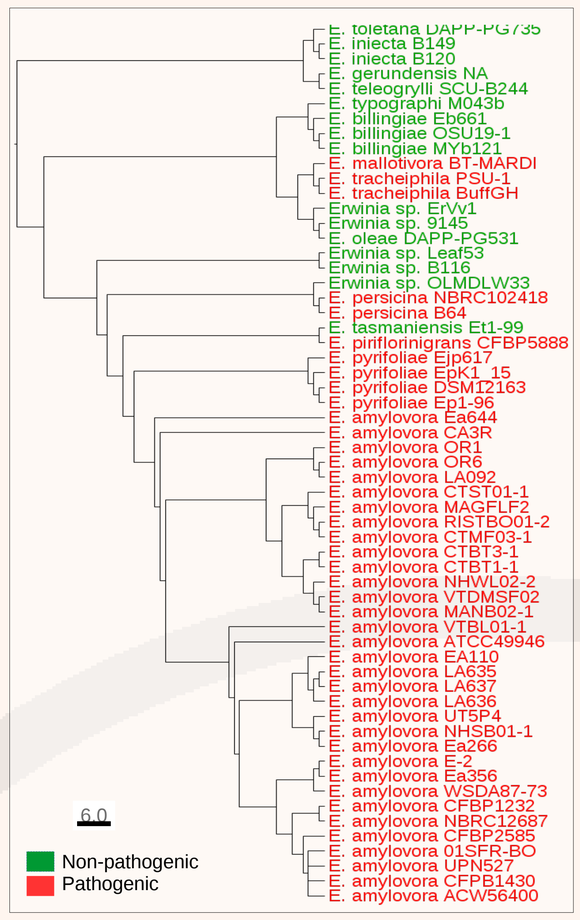
59 Genomes

Pathogenic 43 genomes

Species	*N. Genome*
E. billingiae	(3)
E. gerundensis	(1)
E. iniecta	(2)
E. oleae	(1)
E. tasmaniensis	(1)
E. teleogrylli	(1)
E. toletana	(1)
E. typographi	(1)
Erwinia sp.	(5)

Non-pathogenic 16 genomes

Species	*N. Genome*
E. amylovora	(33)
E. mallotivora	(1)
E. persicina	(2)
E. piriflorinigrans	(1)
E. pyrifoliae	(4)
E. tracheiphila	(2)



After processing and detection of genes, secondary metabolites clusters, and k*mers*, simple and multiple binary logistic regressions were applied to identify specific markers.

	PREDIPATH DB	*Sec.Metabolites Clusters*	*Kmers*	
Total	14.248	24	1.143.473	Genomic elements
Detected	213	9	512	Total of markers detected
Non-pathogenic	-	-	51	Predictors* of classes
Pathogenic	5	9	12	Predictors* of classes

*statistically significants

Simple binary logistic regression with PREDIPATH DB results were able to define a profile to predict the potential pathogenicity of plant-associated species:

- ***fur*** transcriptional repressor of iron-responsive genes
- ***hrp*T** type III secretion lipoprotein
- ***hrp*F** type III secretion protein
- ***hrp*J** Hypersensitivity response secretion protein
- ***par*E** fluoroquinolones resistance gene

- A complete multiple binary logistic regression was able to predict the class using 9 variables only: thiopeptide, HSER, HSER arylpolyene, nrps, siderophore, terpene, butyrolactone, arylpolyene t1pks.
- *Kmers* exclusive to NP class were present from 19 to 100% of genome; exclusive kmers in class P were distributed between 7 to 53% of the genomes.

Conclusions

- Phylogenetic distribution was not able to distinguish between pathogenic and non-pathogenic organisms in genus *Erwinia*.
- Our approach enable the compilation of a complete genomic datasets, composed by **genes**, **clusters** and ***kmers***.
- Detection of exclusive markers by comparative genomics using the **PREDIPATH** workflow allowed the creation of exclusive datasets of predictors to diagnostic potential pathogenicity of plant-associated bacteria.

1.RM Waterhouse, M Seppey, FA Simão, M Manni, P Ioannidis, G Klioutchnikov, E V Kriventseva, EM Zdobnov, BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics, Molecular Biology and Evolution, Volume 35, Issue 3, March 2018, Pages 543–548

2.Blin K, Wolf T, Chevrette MG, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*. 2017;45(W1):W36-W41.

3.Jaillard M, Lima L, Tournoud M, Mahé P, *et al*. (2018) A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between kmers and genetic events. *PLoS Genetics* 14(11).

4.Chen LH, Yang J, Yu J, Yao ZJ, Sun LL, Shen Y and Jin Q, 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 36 (Database issue):D539-D542.

5.Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2016;45(D1):D566-D573.

6.Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., Larsson, DGJ. (2014) BacMet: antibacterial biocide and metal resistance genes database, *Nucleic Acids Res.*, 42, D737-D743