Extracting Genetic Determinants from De Bruijn Graphs in Bacterial GWAS

2015 November 19-20

STATISTICAL ANALYSIS OF MASSIVE GENOMIC DATA

Magali Jaillard[1,2*], Leandro Ishi[2], Maud Tournoud[1], Vincent Lacroix[2], Jean-Baptiste Veyrieras[1] and Laurent Jacob[2]

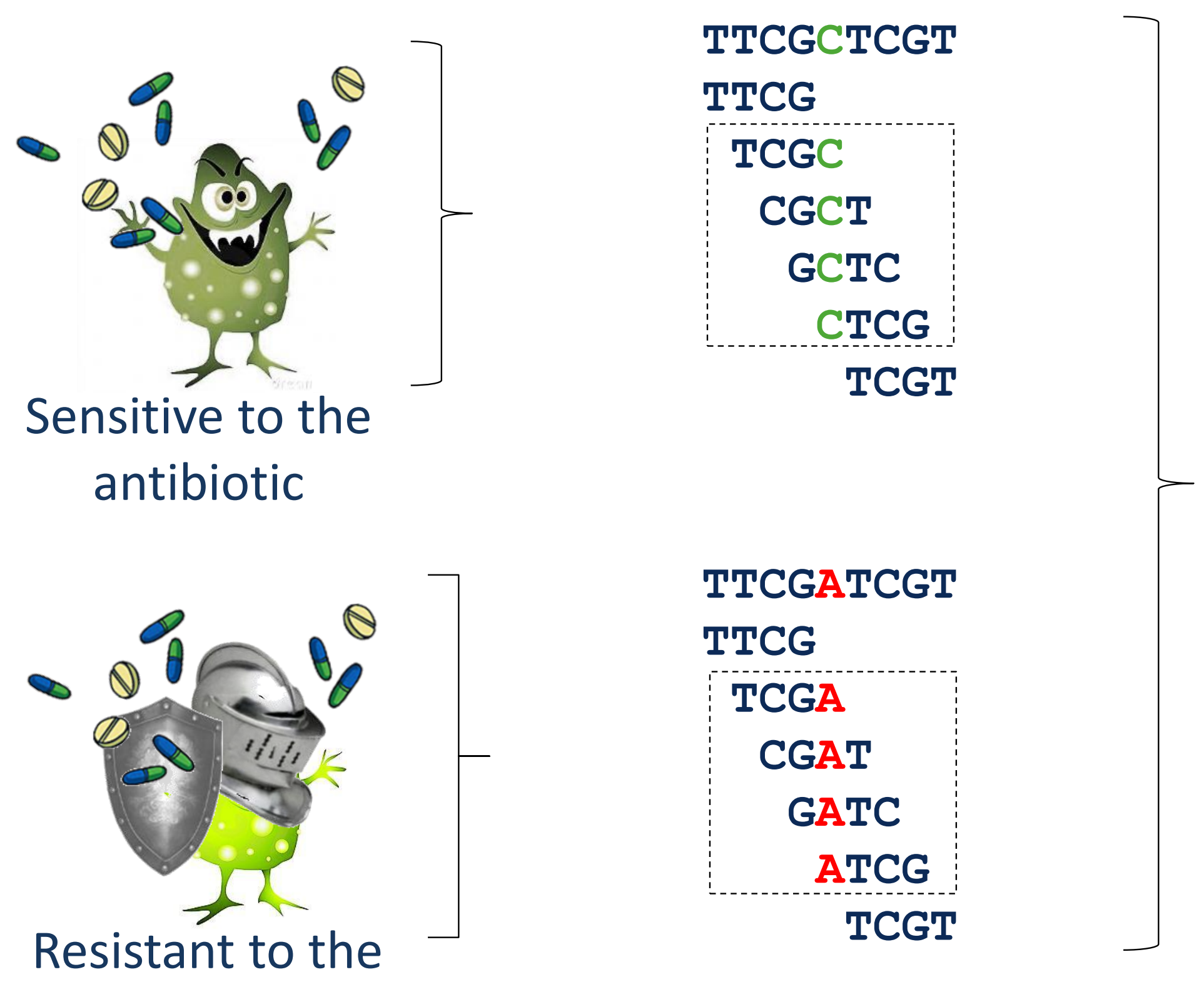
[1] Bioinformatics Research Department, bioMerieux, 69280 Marcy L'Etoile, France
[2]LBBE, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne, France
*Corresponding author : magali.dancette@biomerieux.com

Introduction

Antimicrobial resistance has become a major public health concern, calling for a better definition of existing and novel resistance mechanisms and the **discovery of novel resistance markers**. Most existing GWAS approaches for bacterial genomes either look at SNPs obtained by sequence alignment or consider sets of k-mers, whose presence in the genome is associated with the phenotype of interest. We present an **alignment-free GWAS method**, targeting any region of the genome and selecting haplotypes of variable length associated to the resistance phenotype. The exploitation of De Bruijn graph structure, implicitly containing all genomes k-mers of all sizes, results in a **drastic reduction of the number of explored features** without loss of information, thus increasing the statistical power of the tests.

Methods

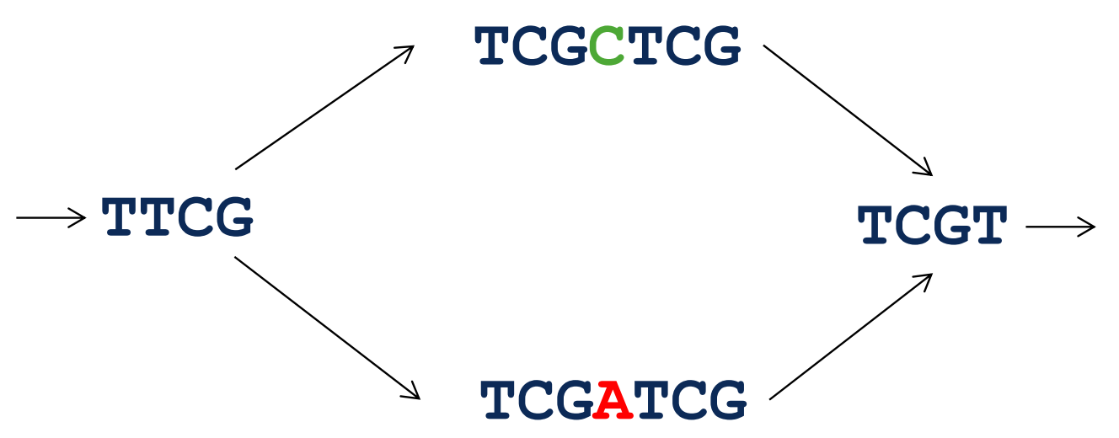
Genomes are split into k-mers



*Features = **k-mers***
***2xk** features describing the mutation*

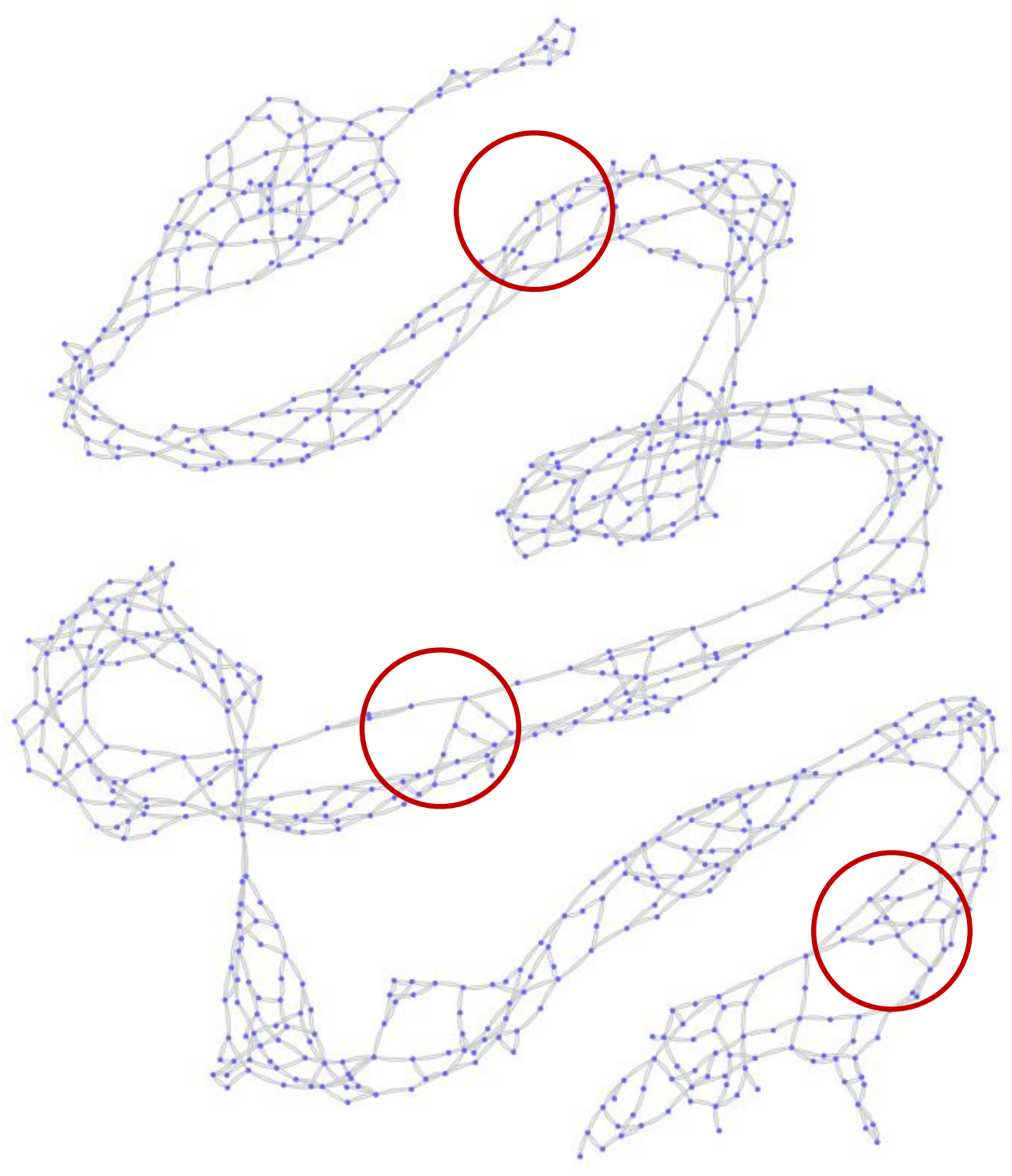
k-mers are connected in a De Bruijn Graph

De Bruijn graphs are widely used for *de novo* assembly. They represent overlaps between sequences



Reduction of the number of features by compression of linear paths

*Features = graph **nodes***
***2** features describing the mutation*



Features are selected by GWAS

$$Y = X\beta + W\alpha + \epsilon$$

Where :

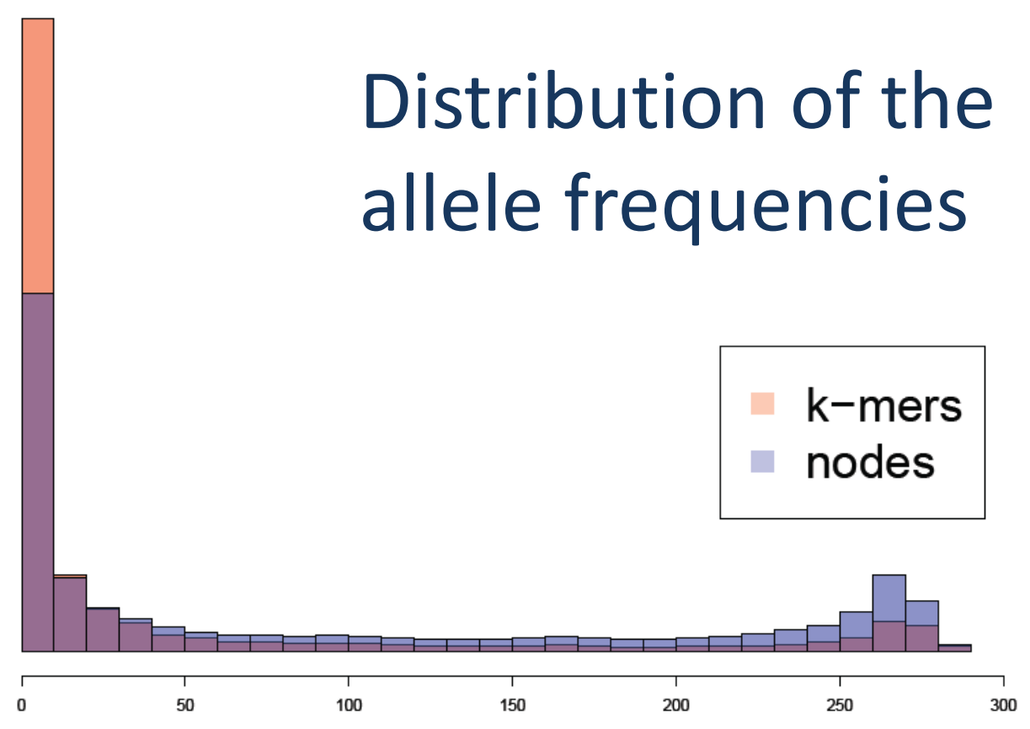
- Y = phenotype vector: status of each strain for the antibiotic
- X = genotype matrix : presence or absence of each feature for each strain
- W = population structure: matrix representing between-strains correlation

Results

Data: 280 *Pseudomonas aeruginosa* strains from all phylogroups

- Genotype: computed from genome assemblies
- Phenotype: MIC (Minimum Inhibitory Concentration) for Amikacin

Evaluation: a list of markers described in the literature for the Amikacin is used to test for low p-value enrichment.



Distribution of the allele frequencies

Nodes describe all haplotypes (of different lengths) found in the strain panel.

Min (=k)	Median	Max
41 bp	57 bp	104,553 bp

Modeling choice

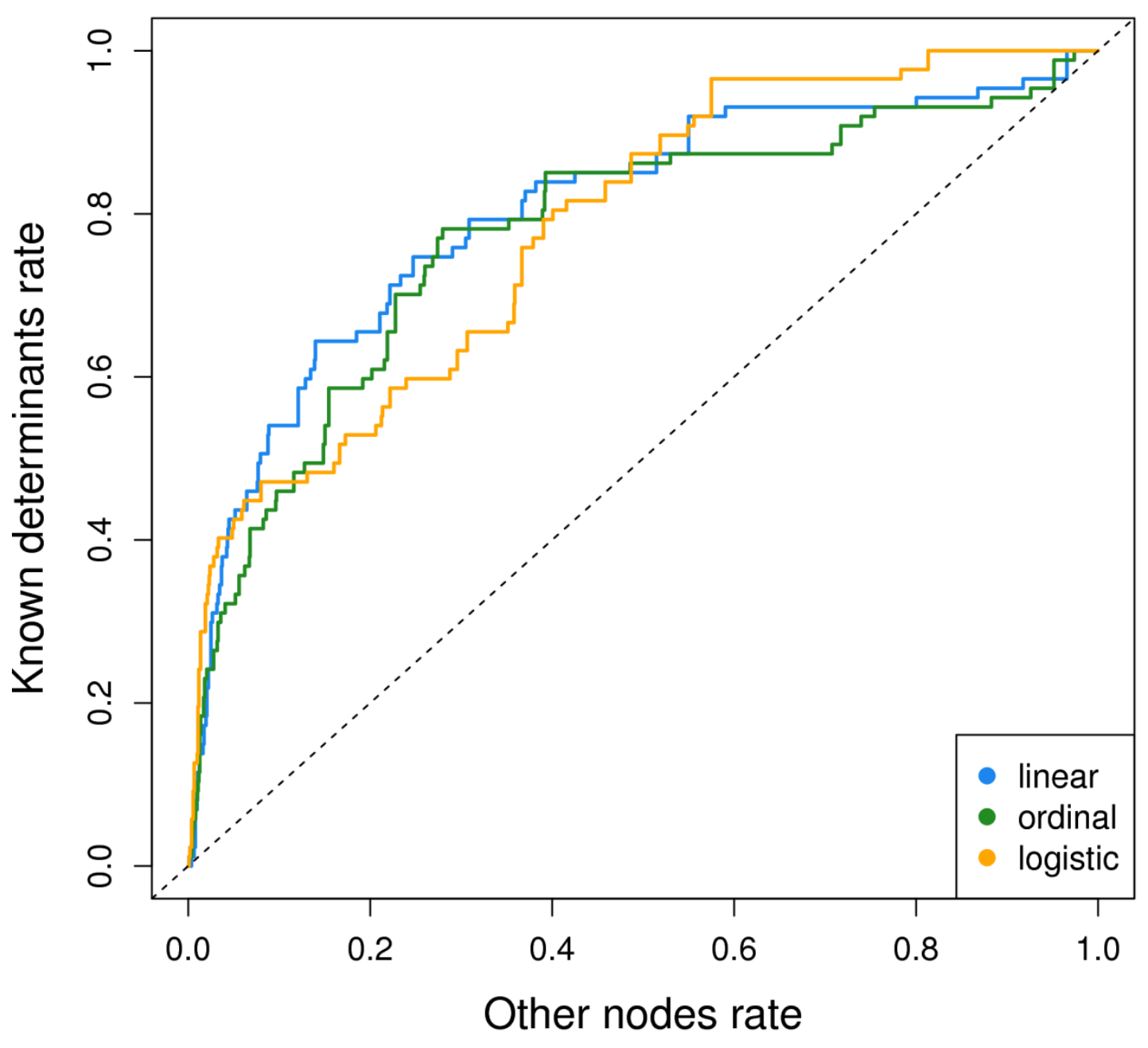
Three models of phenotype were tested with the graph-nodes as genotype :

Logistic → Binary phenotypes obtained using CLSI thresholds on MIC data

Linear → A linear model is applied to the logarithm of the MIC values

Ordinal → MIC values are encoded as ordered categories

The linear model is retained



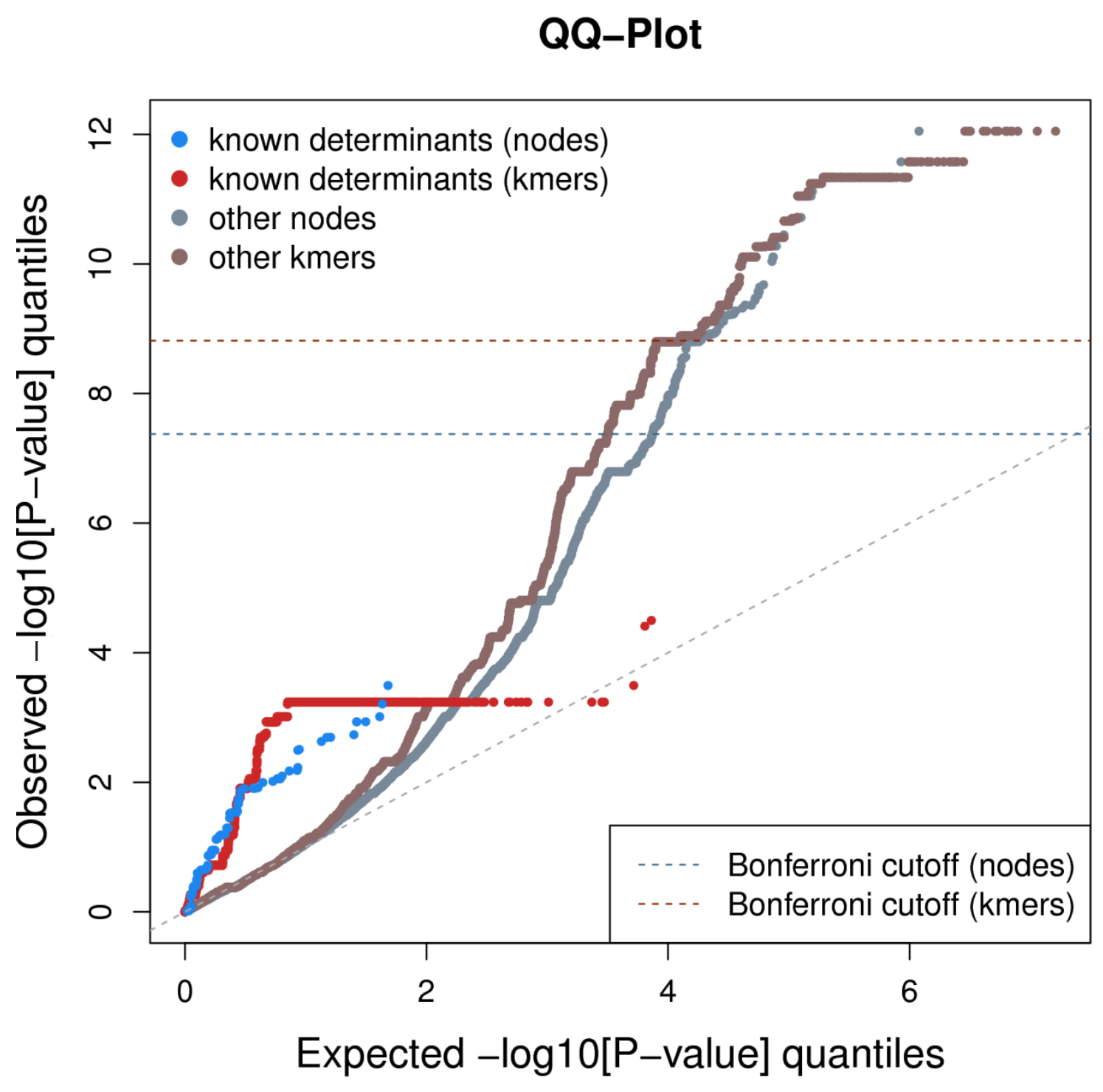
Nodes versus k-mers

Using words of length k=41, we obtain for the 280 genomes:

k-mers	nodes
41,187,547	1,353,852

We thus reduce **30 times** the parameter space by using De Bruijn Graphs structure.

The QQ-plot shows an enrichment of low p-values for known determinants.



Conclusion

These encouraging results suggest **De Bruijn Graphs nodes** are well suited to describing genetic determinants of bacterial resistance and can be used for GWAS on bacterial **species with high plasticity**. Extracting significant subgraphs composed of several nodes is a natural next step. We also plan to adapt the resolution of our determinants to take linkage disequilibrium into account.

Earle et al. arXiv:1510.06863v2 [q-bio.GN]

Sacomoto et al. BMC Bioinformatics (2012) 13(Suppl 6):S5

Dehman et al. BMC Bioinformatics (2015) 16:148

We warmly thank Vincent Daubin and Franck Picard for the fruitful discussions and advices on this work.