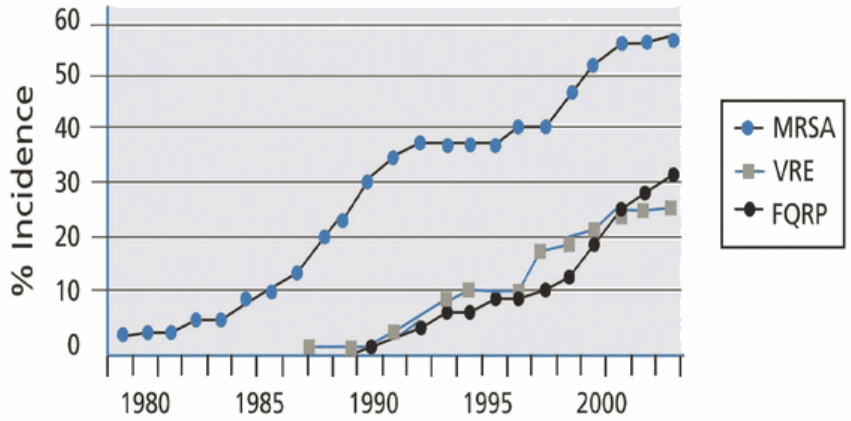
Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs

Magali Jaillard, Maud Tournoud, Leandro Lima, Vincent Lacroix, Jean-Baptiste Veyrieras, Laurent Jacob

LBBE/CNRS, Université de Lyon, bioMérieux

Work performed by Magali Dancette

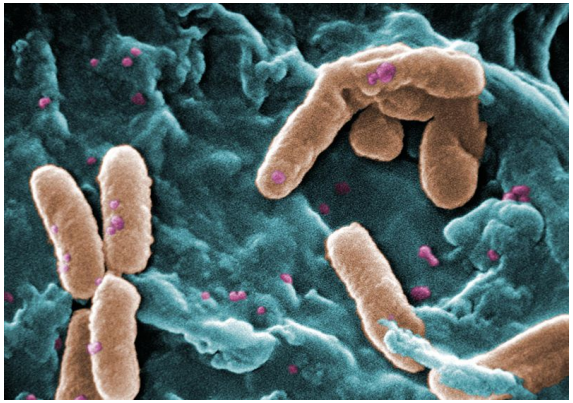
Understanding antibiotic resistance in bacteria



MRSA = methicillin-resistant Staphylococcus aureus
VRE = Vancomycin-resistant enterococci
FQRP = Fluoroquinolone-resistant Pseudomonas aeruginosa

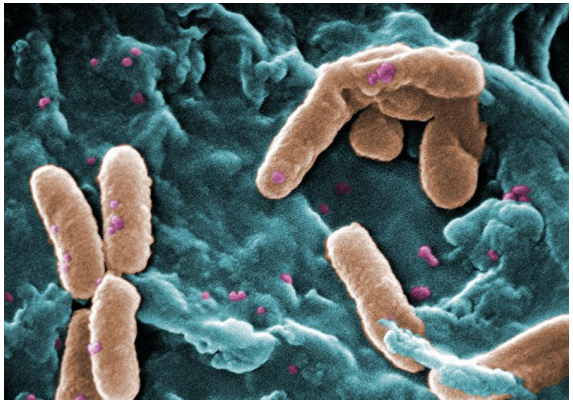
- Antimicrobial resistance has become a major worldwide public health concern.
- Literature on resistance is abundant, however known determinants do not completely explain the phenotype variability.
- Genome-Wide Association Study (GWAS) targeting any region of the genome should help select new candidate markers of resistance.

Pseudomonas aeruginosa



- Ubiquitous bacteria causing a lot of hospital acquired infections.
- Highly variable genome content and size: from 5.5 Mb to 7.5 Mb.
- Long and manifold **accessory genome**, containing about 60% of the known resistance determinants.
- Highest percentage of regulatory genes among Bacteria (>8.5%).

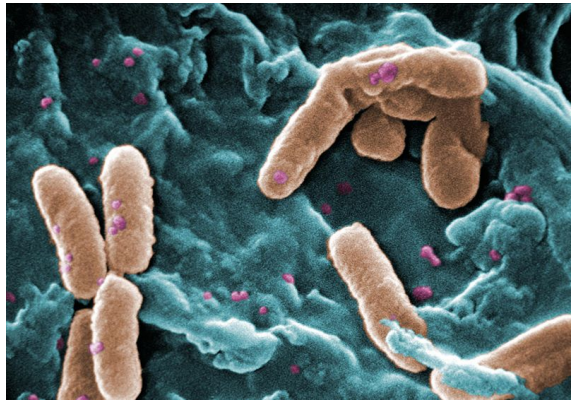
How do we describe such a genome?



Current approaches

- Alignment against a reference genome, SNPs/indels.
- Gene copy number.
- K-mer content (presence/absence or counting). X_{ij} is 1 if the genome of sample i contains the j-th kmer.

How do we describe such a genome?



Current approaches

- Alignment against a reference genome, SNPs/indels.
  Highly variable genome content and size.
- Gene copy number.
  High percentage of regulatory genes: cannot exclude non-coding regions
- K-mer content (presence/absence or counting). X_{ij} is 1 if the genome of sample i contains the j-th kmer.

Fixed-length kmer descriptions are large and redundant

TTCGCTCGTA

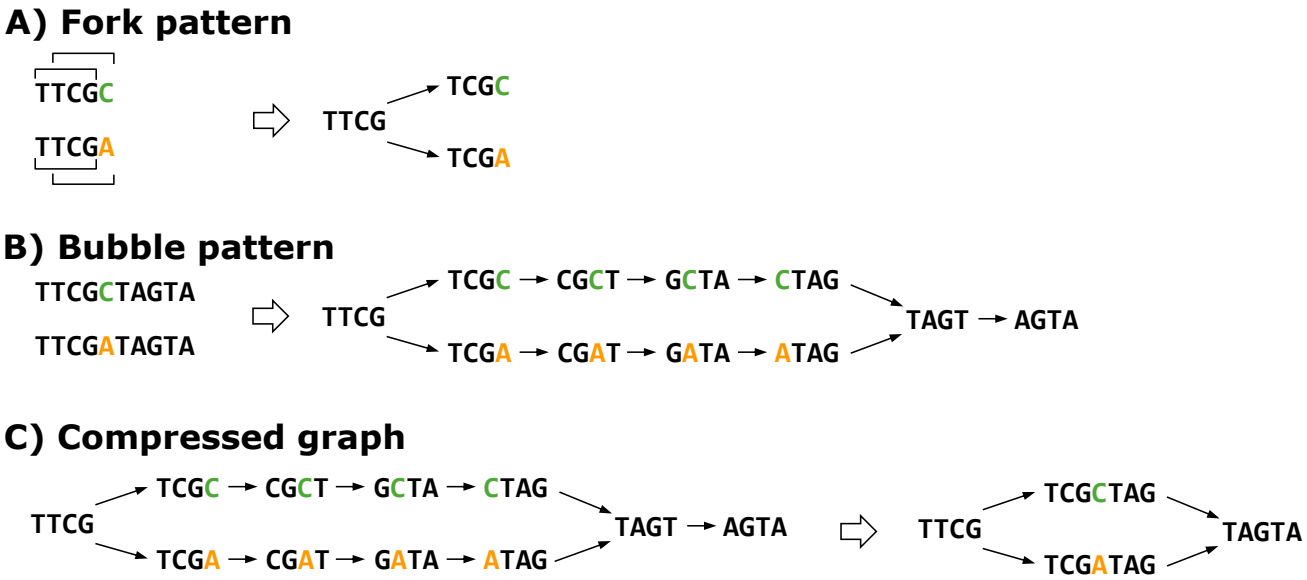
TTCGATCGTAT

Fixed-length kmer descriptions are large and redundant

TTCGCTCGTA
TTCG
TCGC
CGCT
GCTC
CTCG
TCGT
CGTA
GTAT

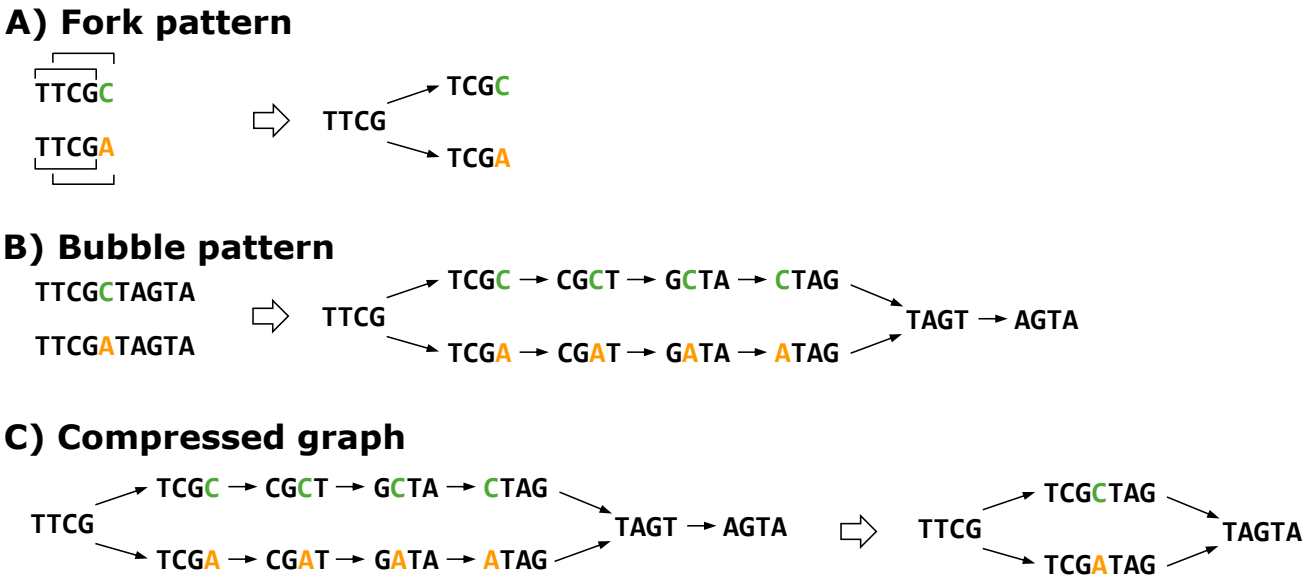
TTCGATCGTAT
TTCG
TCGA
CGAT
GATC
ATCG
TCGT
CGTA
GTAT

De Bruijn Graphs



- Used in most *de novo* assembly methods.
- Compact linear paths.
- Yields lossless, data adaptive, locally optimal resolution.

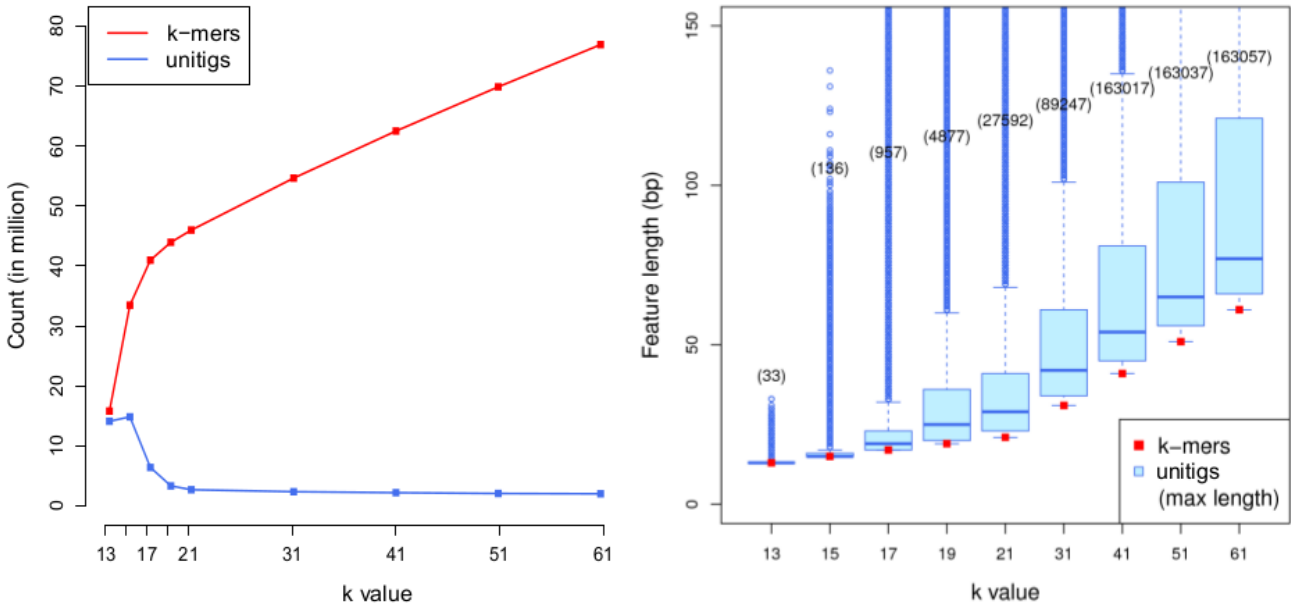
De Bruijn Graphs



- Used in most *de novo* assembly methods.
- Compact linear paths.
- Yields lossless, data adaptive, locally optimal resolution.

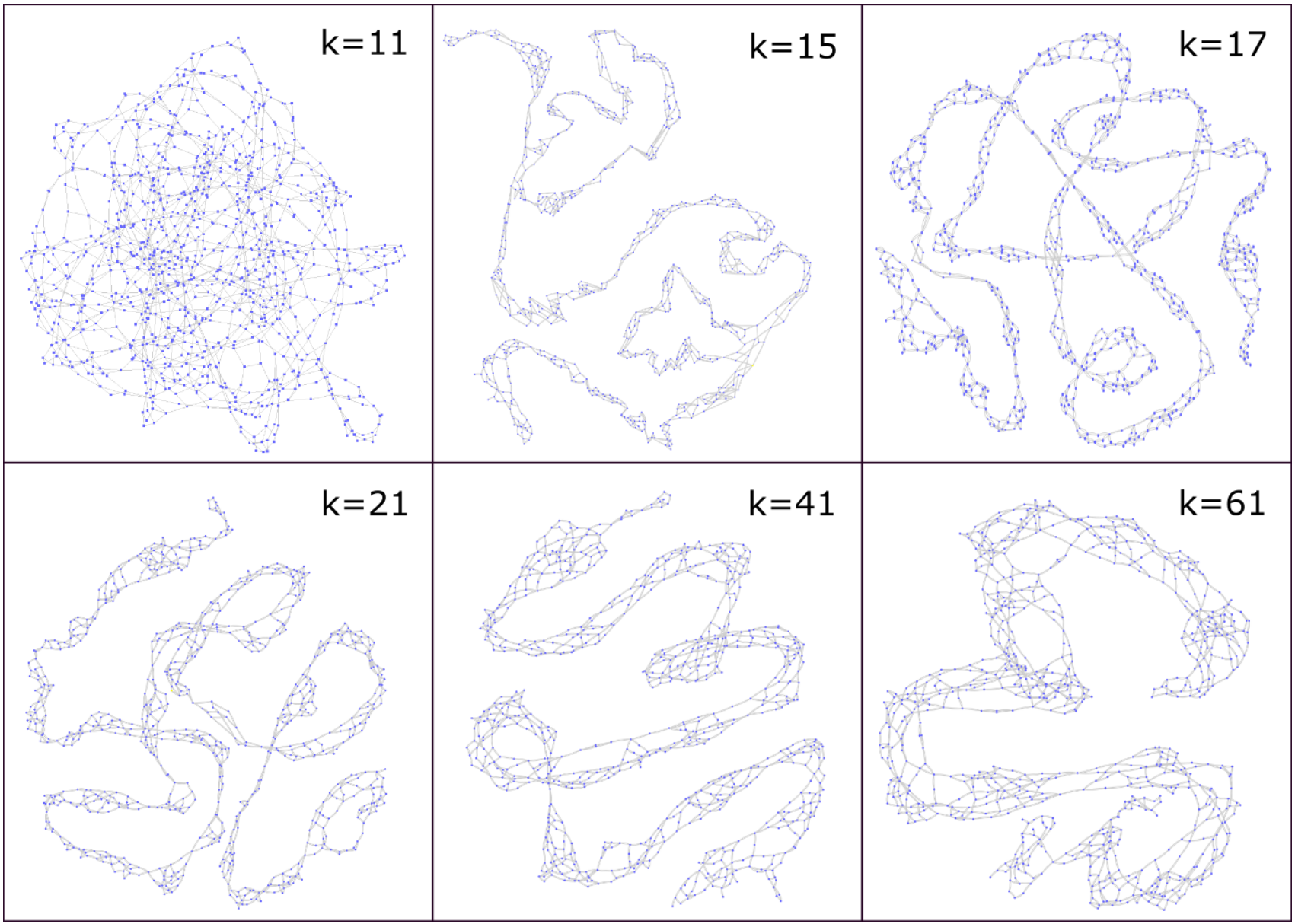
We propose to describe genomes by presence/absence or counting of these variable-length kmers.

Feature count and length

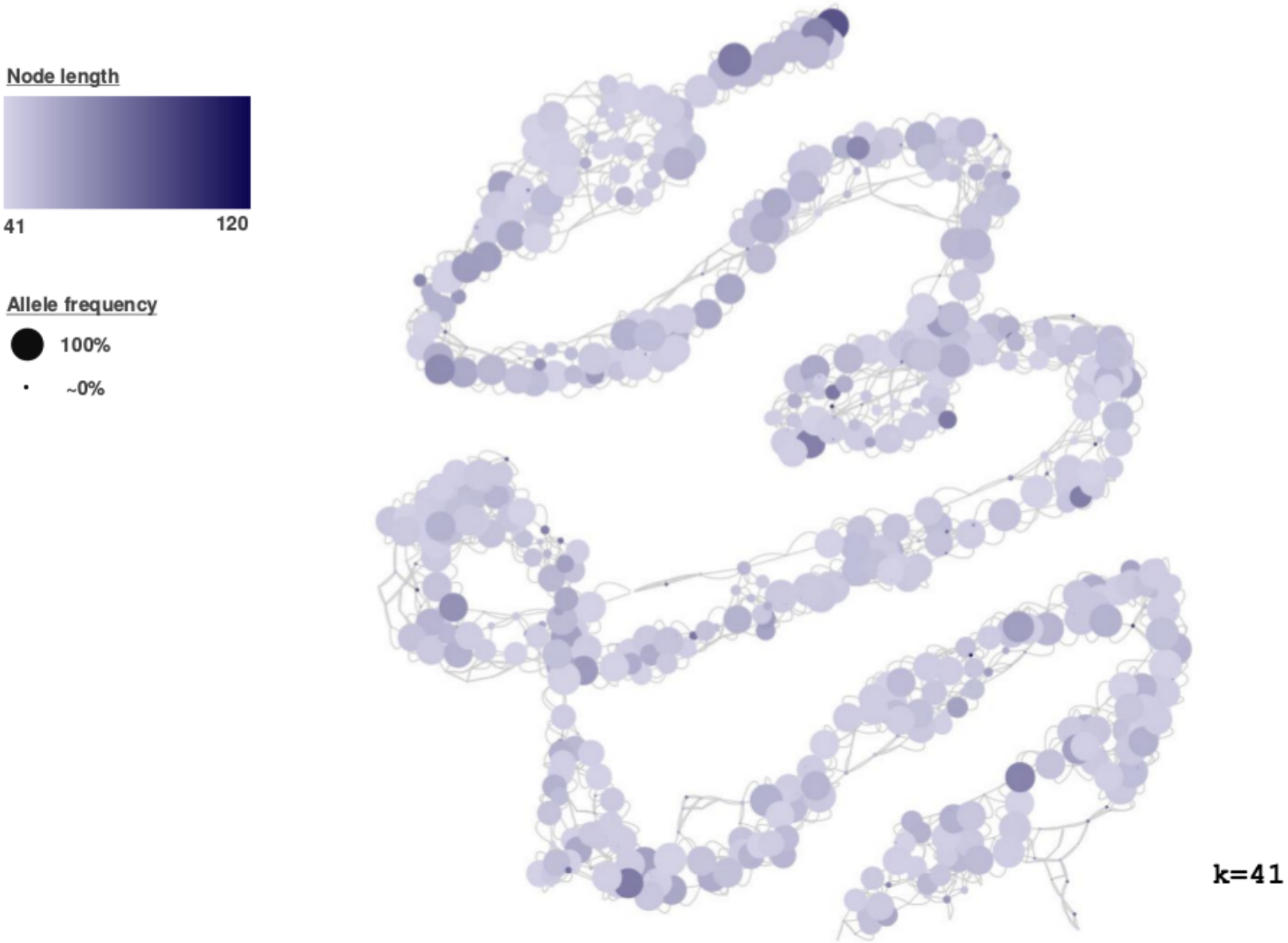


For k=41, median length is 57, max is 163017.

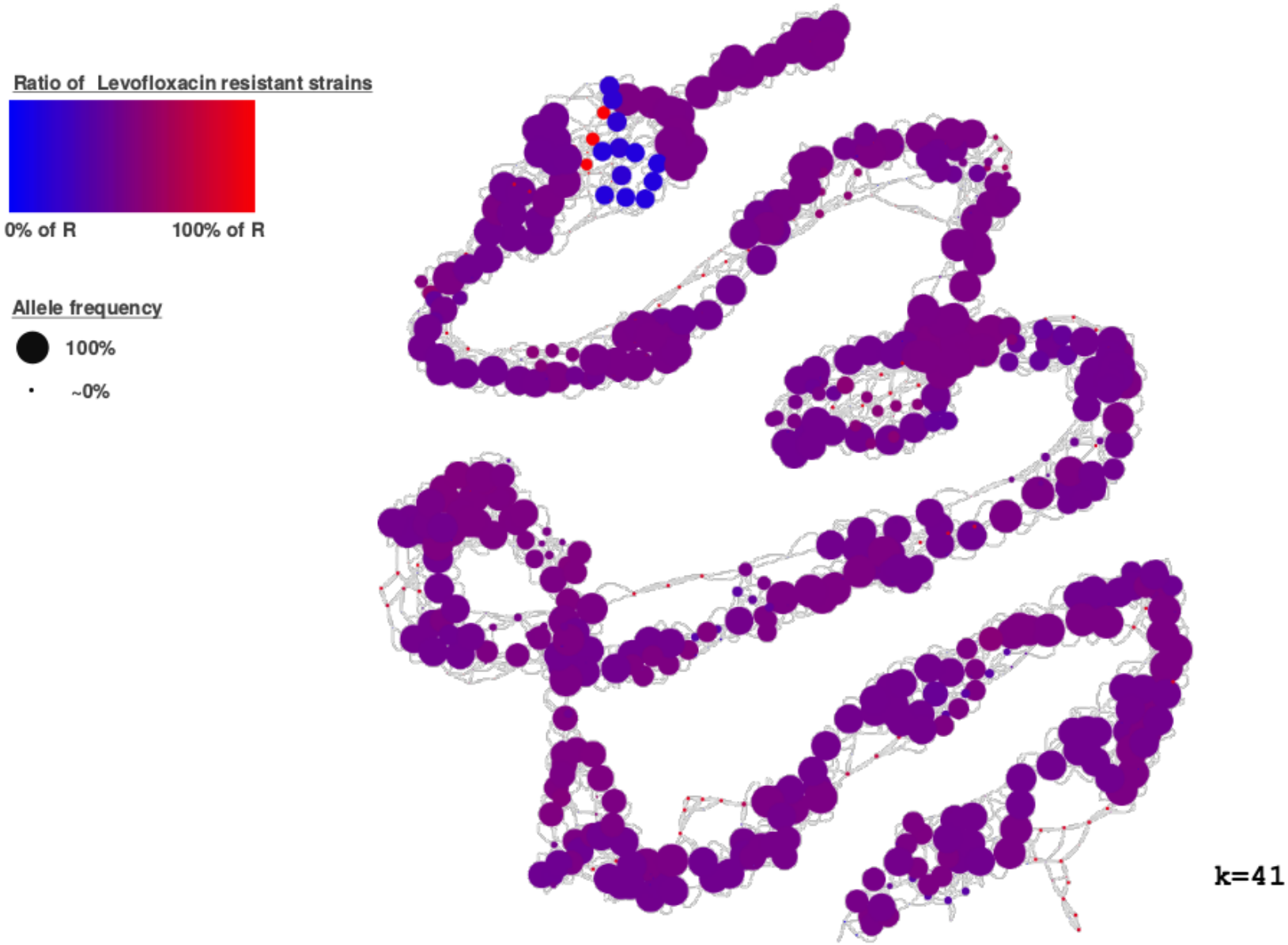
DBG of gyrA gene ($\sim$ 2kb) across 665 *P. aeruginosa* strains



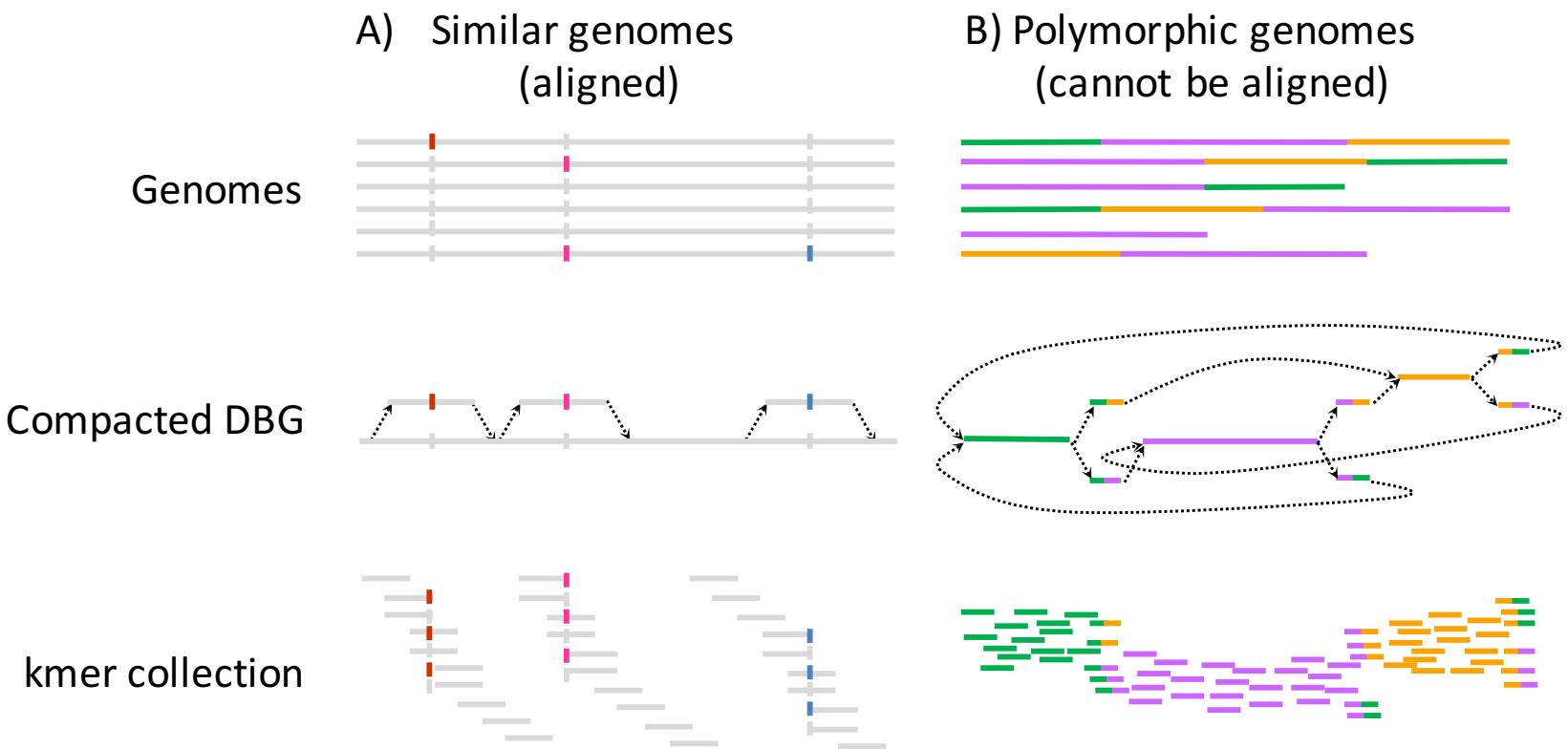
Visualize variable parts



Visualize association with a phenotype



DBG nodes interpolate between SNP and fixed-length kmer representations



We are not introducing new presence/absence patterns

Representation of TTCGCTAGTA with:

- Fixed-length kmer

$$(\text{TTCG } 1, \text{TCGC } 1, \text{CGCT } 1, \text{GCTA } 1, \text{CTAG } 1,$$
$$\text{TCGA } 0, \text{CGAT } 0, \text{GATA } 0, \text{ATAG } 0, \text{TAGT } 1, \text{AGTA } 1)$$

- DBG:

$$(\text{TTCG } 1, \text{TCGCTAG } 1, \text{TCGATAG } 0, \text{TAGTA } 1)$$

We are not introducing new presence/absence patterns

Representation of TTCGCTAGTA with:

- Fixed-length kmer

(TTCG 1, TCGC 1, CGCT 1, GCTA 1, CTAG 1,
TCGA 0, CGAT 0, GATA 0, ATAG 0, TAGT 1, AGTA 1)

- DBG:

(TTCG 1, TCGCTAG 1, TCGATAG 0, TAGTA 1)

We are not introducing new presence/absence patterns

Representation of TTCGCTAGTA with:

- Fixed-length kmer

(TTCG 1, TCGC 1, CGCT 1, GCTA 1, CTAG 1,
TCGA 0, CGAT 0, GATA 0, ATAG 0, TAGT 1, AGTA 1)

- DBG:

(TTCG 1, TCGCTAG 1, TCGATAG 0, TAGTA 1)

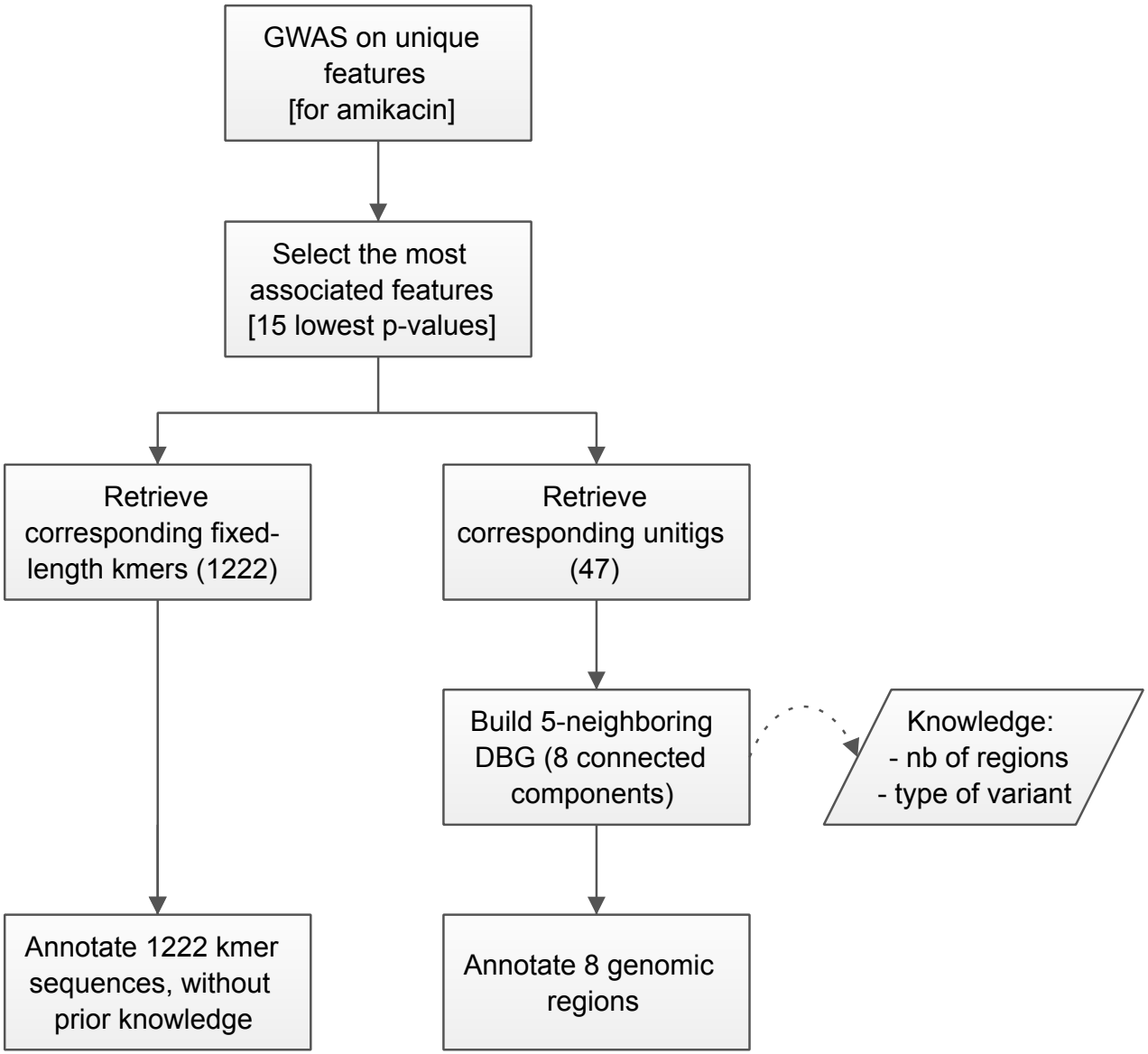
All features of the same color have the same presence/absence pattern.
They will all have the same profile across samples.

We are doing the **same set of tests** for both representations.

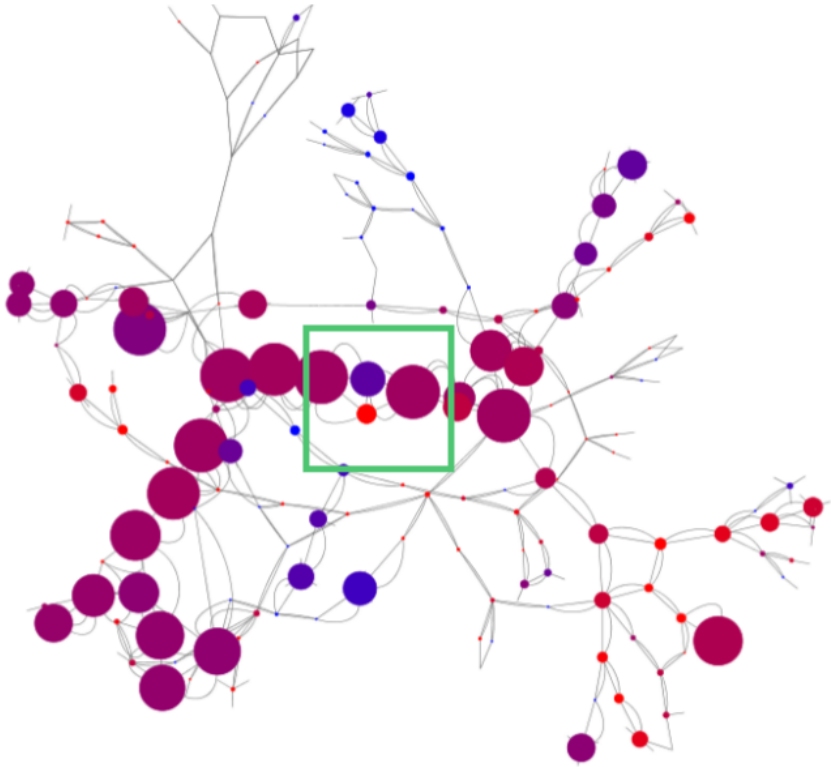
Why use DBG nodes rather than kmers

- kmer redundancy: LD + local redundancy. DBG redundancy: LD only.
- Consequence: fewer sequences to interpret for each feature (*e.g.*, map against all genomes).
- (colored) DBG itself helps us understand the type of genetic feature we selected.
- Could also help estimate population structure.

Postprocessing flowchart (amikacin resistance)

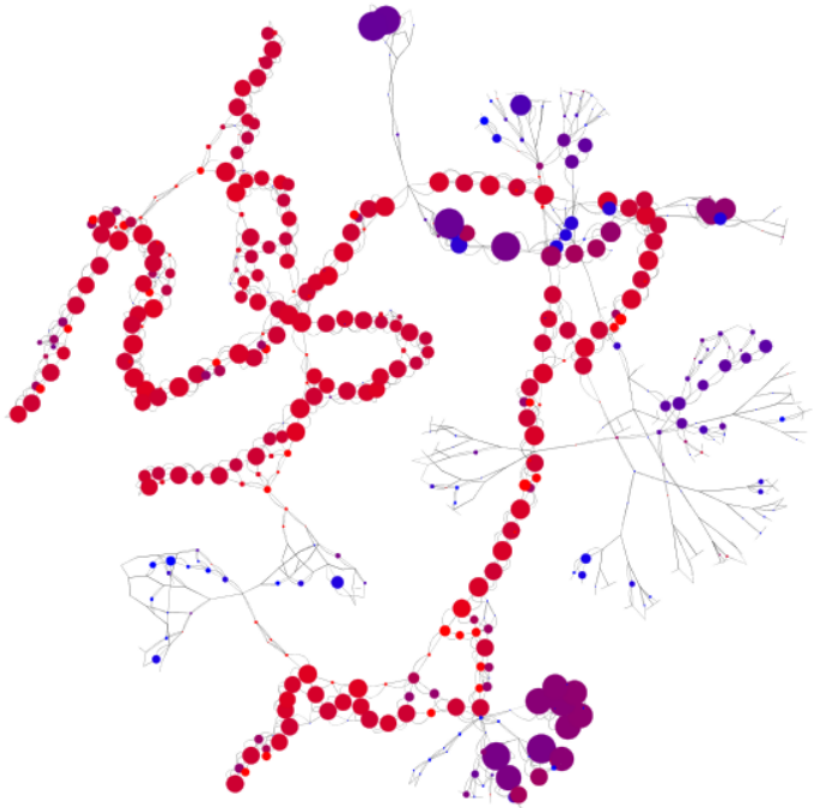


Selected subgraphs: mutation in an accessory gene



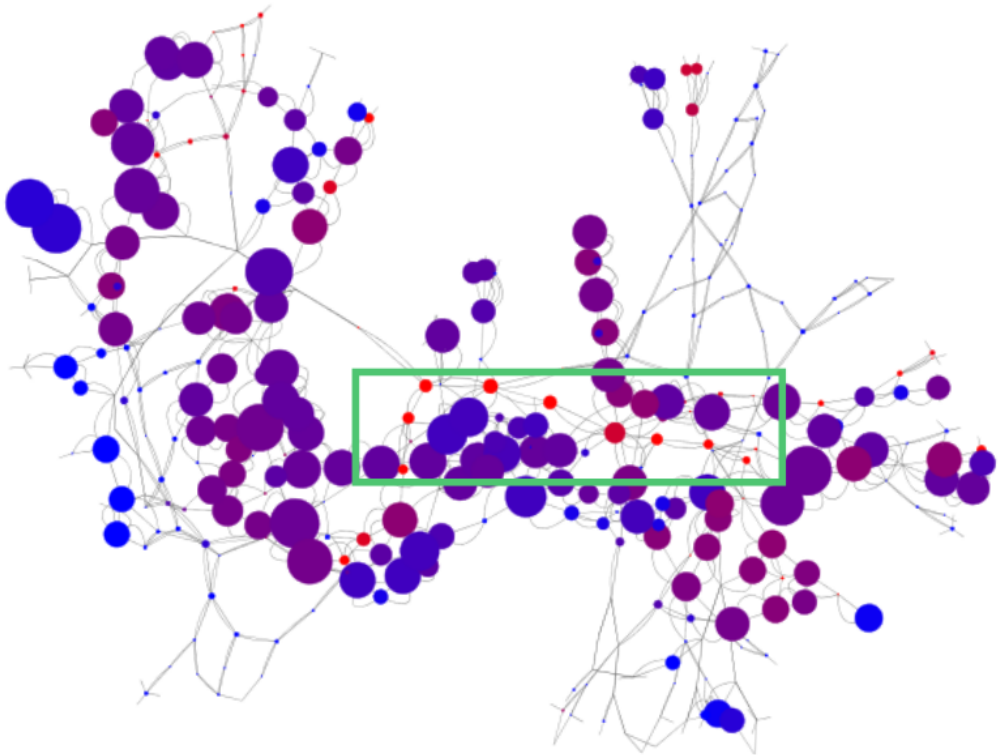
- Mostly linear structure with little difference between resistants and sensitive strains.
- Contains one fork into one blue and one red node, suggesting we found a SNP associated with resistance.
- Mapping to annotation reveals that this structure is the AAC gene.

Selected subgraphs: whole plasmid inclusion



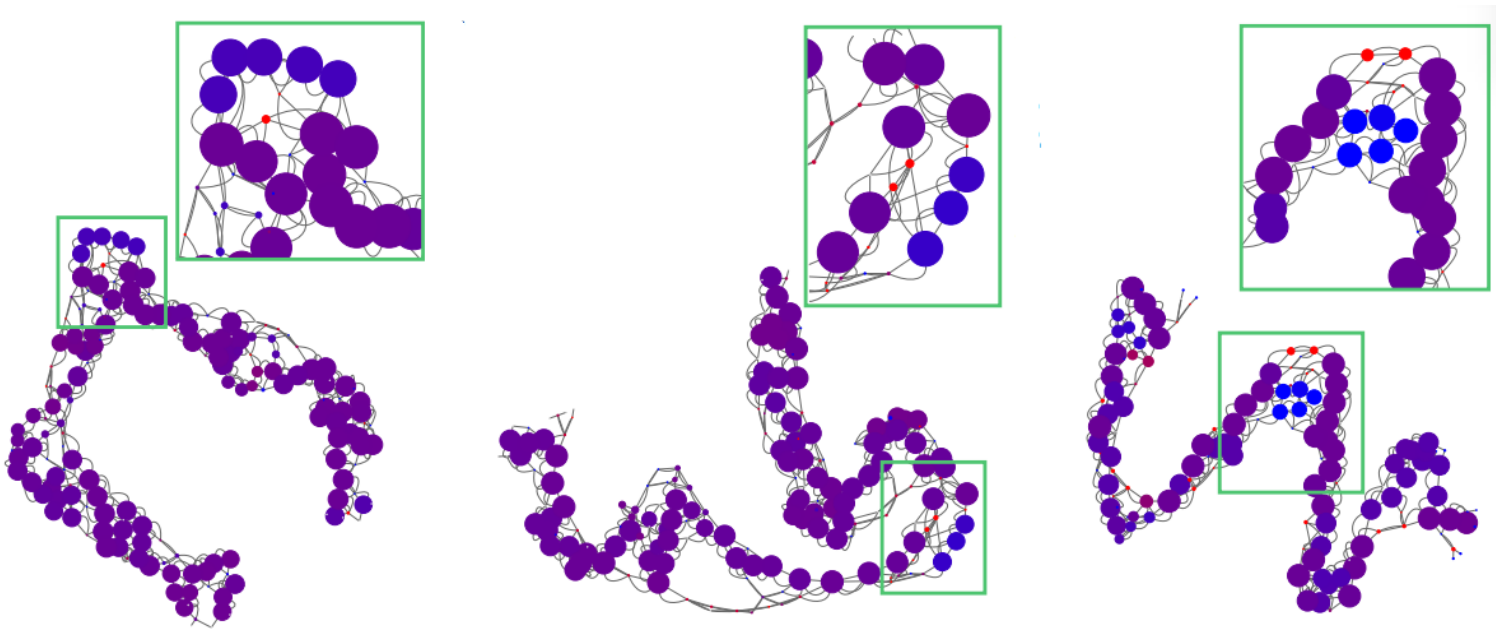
- Linear structure with mostly red nodes: presence of the entire sequence is associated with resistance.
- Maps to pHS87b plasmid recently described as being involved in resistance.

Selected subgraphs: non-coding region



- Connected component mapping to a non-coding region of *P. aeruginosa*.
- Highlights path of red nodes which were not all in the top 15.

Selected subgraphs: SNPs in core genes (levofloxacin)



- Same experiment with levofloxacin: we select components which map to core genes and represent SNPs.
- Two known resistance genes (gyrA, parC). Third one not in our resistance database (could be causal or LD).
- Matches the current knowledge on levofloxacin resistance, mainly based on target alteration (amikacin components all mapped within or near mobile elements).

Perspective: what would alternative approaches do

SNPs in core genome

Would miss all events in accessory genome and non-coding regions (presence/absence, SNPs).

Gene presence/absence

- Would miss all events in non-coding regions.
- Would miss finer events (*e.g.* SNP in AAC gene).

Fixed-length kmers

In the case of pHS87b plasmid, would yield disconnected regions and could miss causal parts.

Discussion

Future work

- Define features based on subgraphs.
- Provide strategies to perform inference on these features.
- Multiple testing correction.

Acknowledgements

Université de Lyon

- Leandro Lima
- Vincent Lacroix
- Vincent Daubin
- Franck Picard

bioMérieux

- Magali Jaillard
- Maud Tournoud
- Jean-Baptiste Veyrieras
- Pierre Mahé
- Stéphane Schicklin