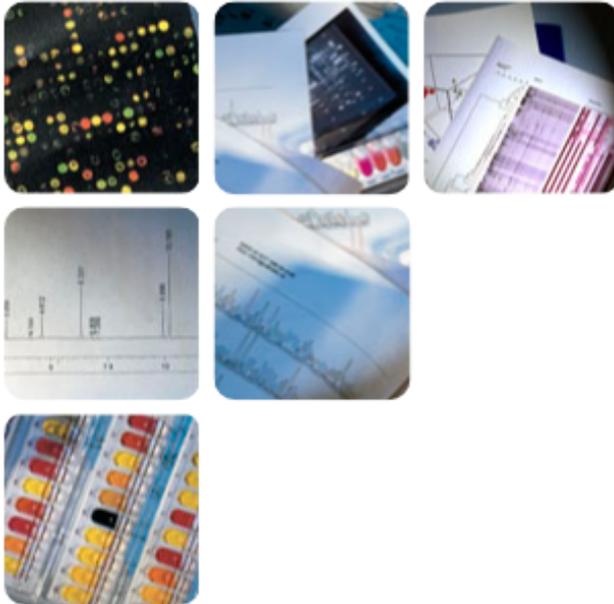


Whole genome sequence analysis using **Bio**Numerics

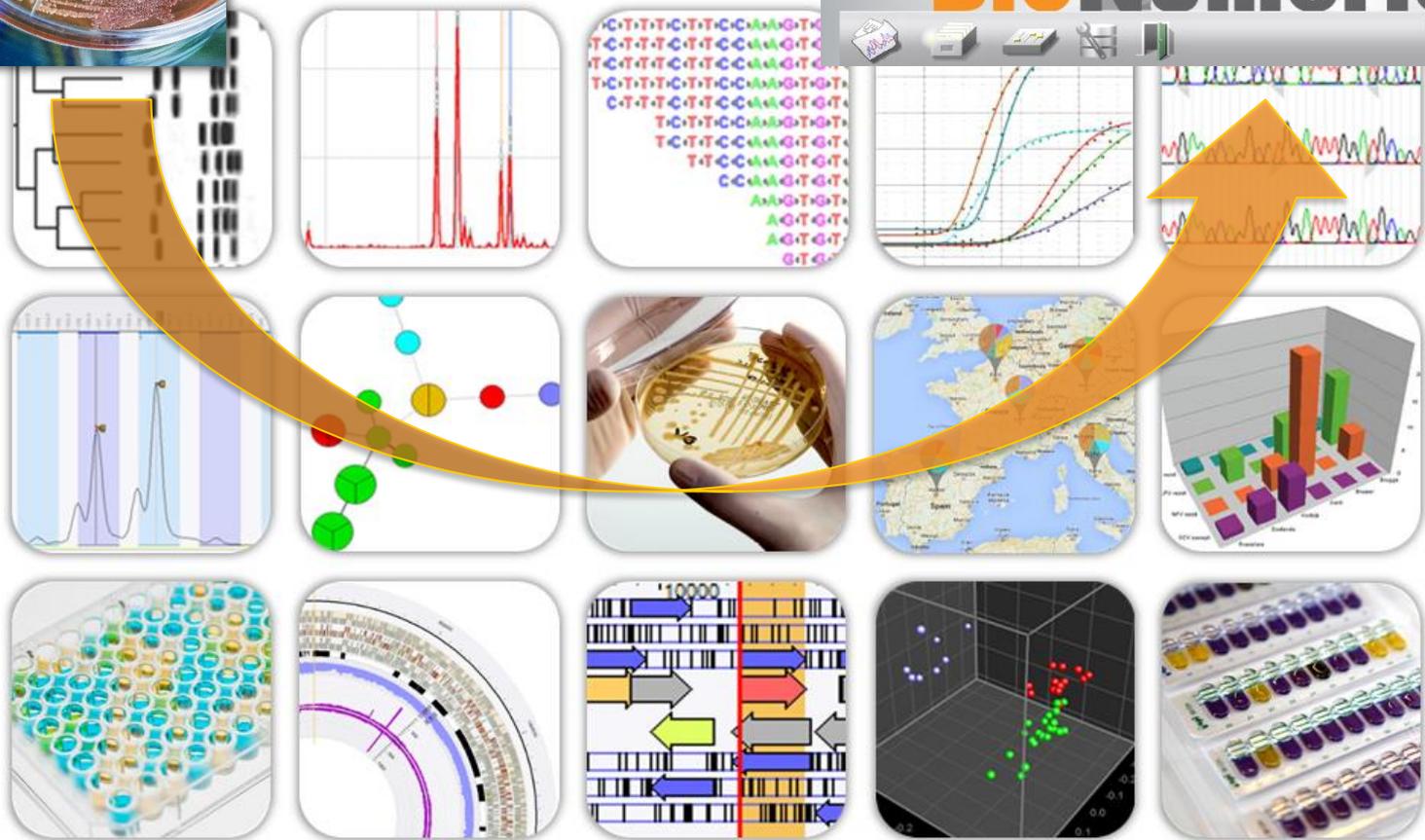


Katrien De Bruyne



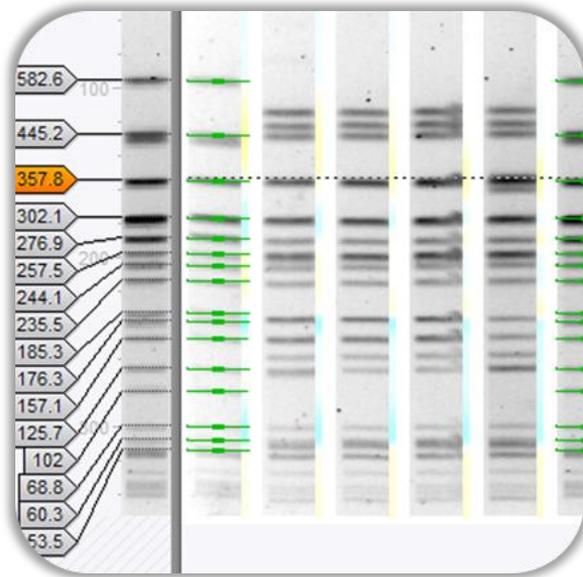
A BIOMÉRIEUX COMPANY

All biological data and metadata integrated in one software platform,

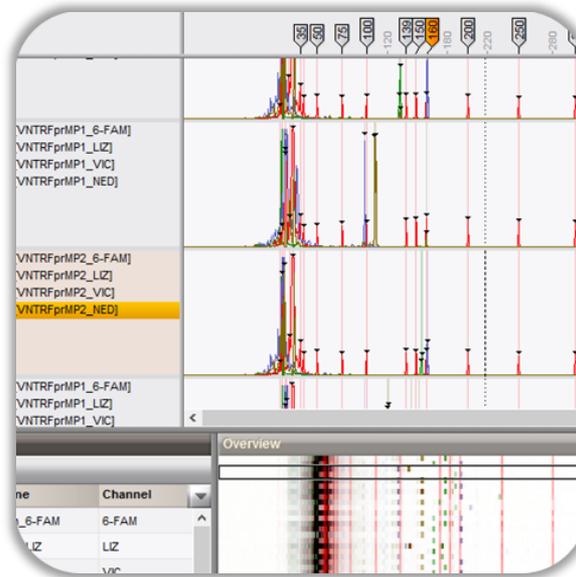


both classical techniques...

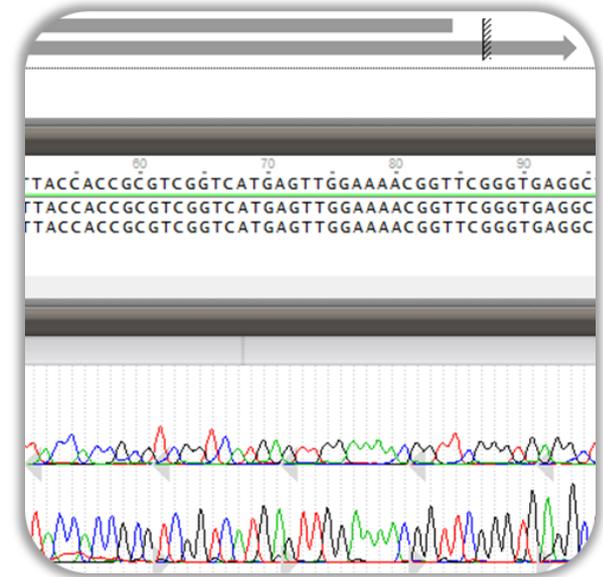
PFGE



MLVA

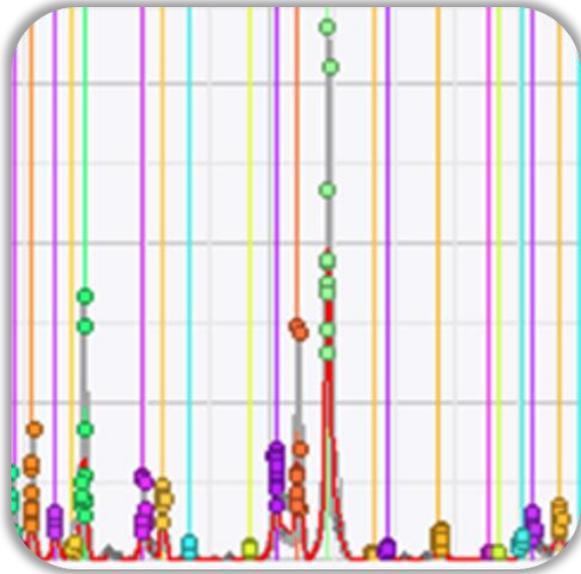


MLST



... and new technologies !

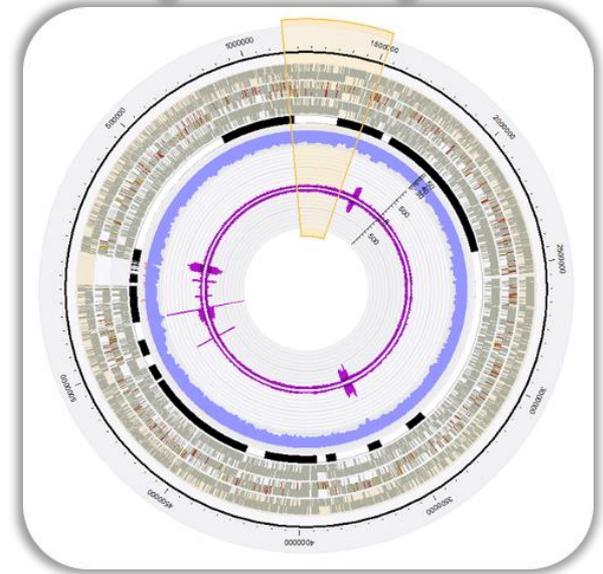
MALDI-TOF MS

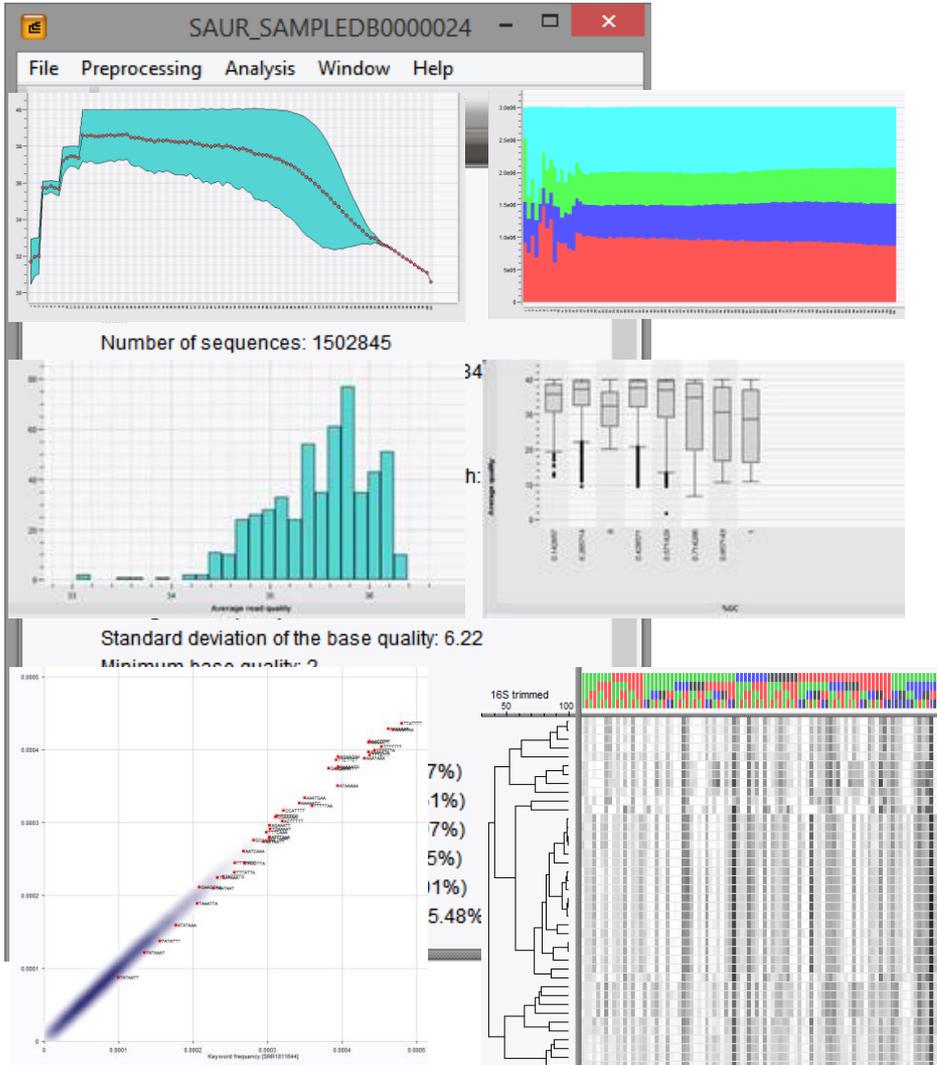


NGS



wgMLST / wgSNP



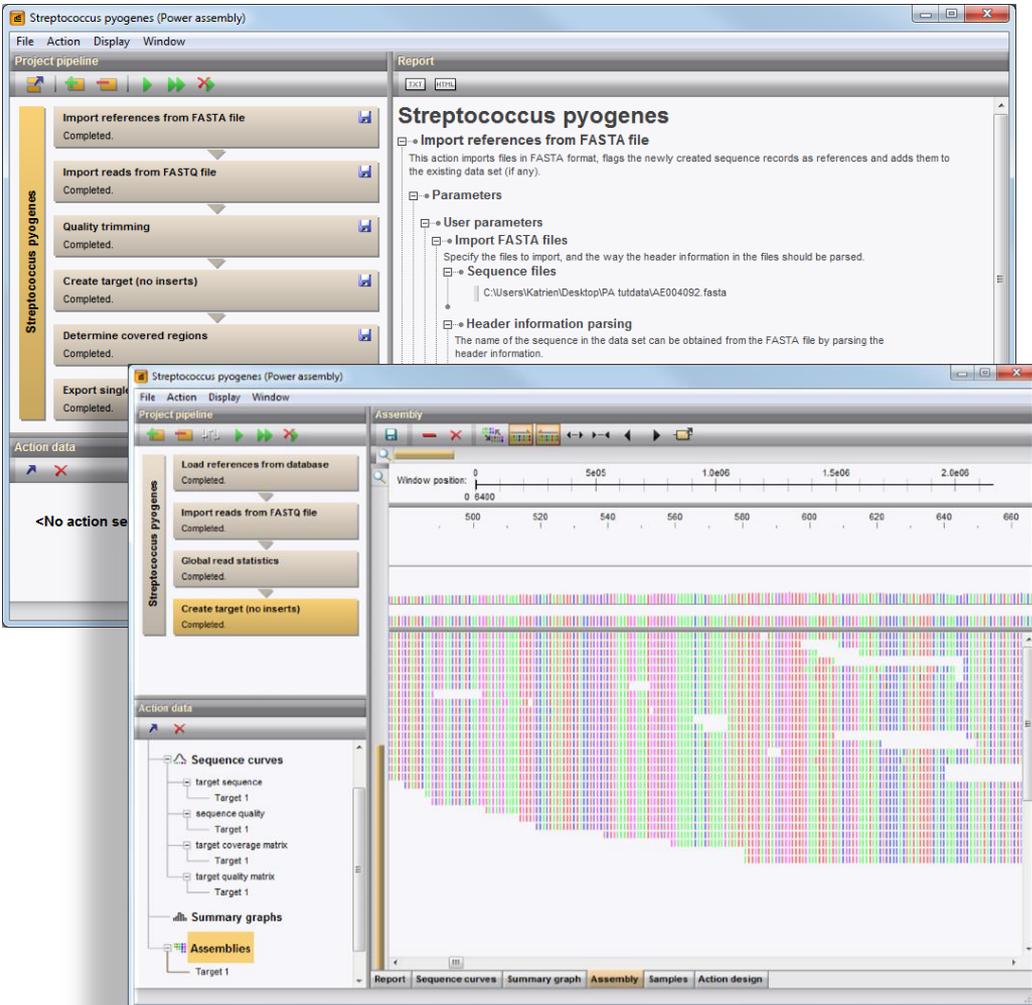


- Raw data statistics
- Per-position statistics, per-sequence statistics, per-oligo statistics for data exploration & quality assessment
- Pairwise & multi-sample comparison of keyword profiles:
 - Reproducibility
 - Identify contamination
 - Find most suitable reference

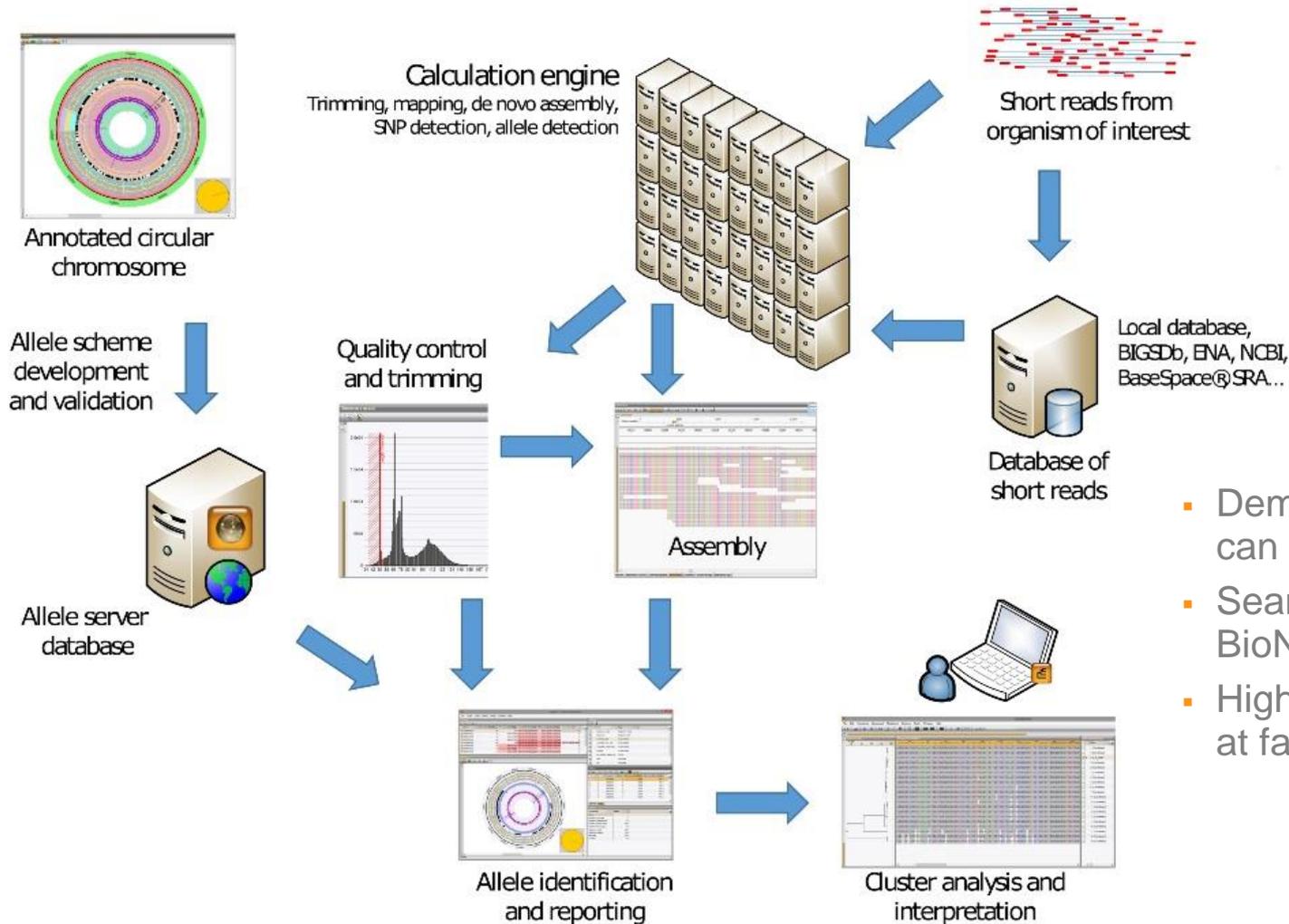
Local WGS tools using the Power Assembler

Point & click application for mapping and assembly of WGS data

- Build new or modify predefined project templates
- Easy access to all actions and parameters
- Save and share analysis templates
- Both mapping and de novo assembly
- Extensive logging and reporting
- Batch analysis from entry selection



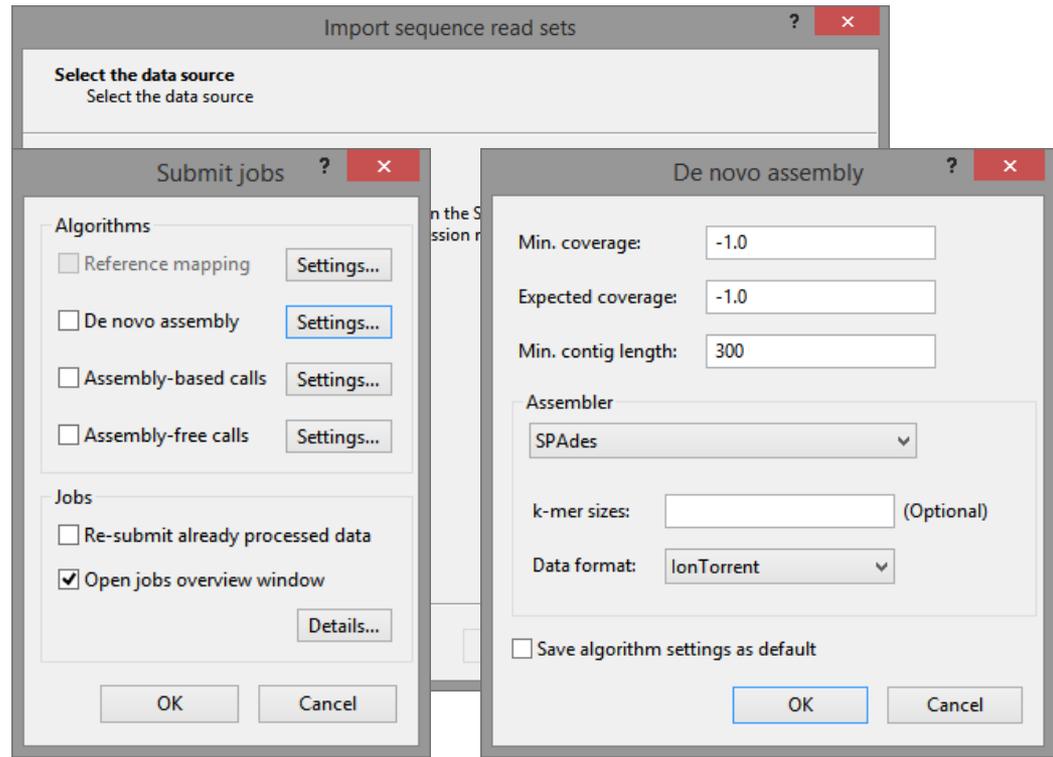
WGS tools integrated on the calculation engine



- Demanding calculations can be performed on CE
- Seamless integration with BioNumerics
- High throughput analyses at fast turnaround times

wgMLST for cluster detection

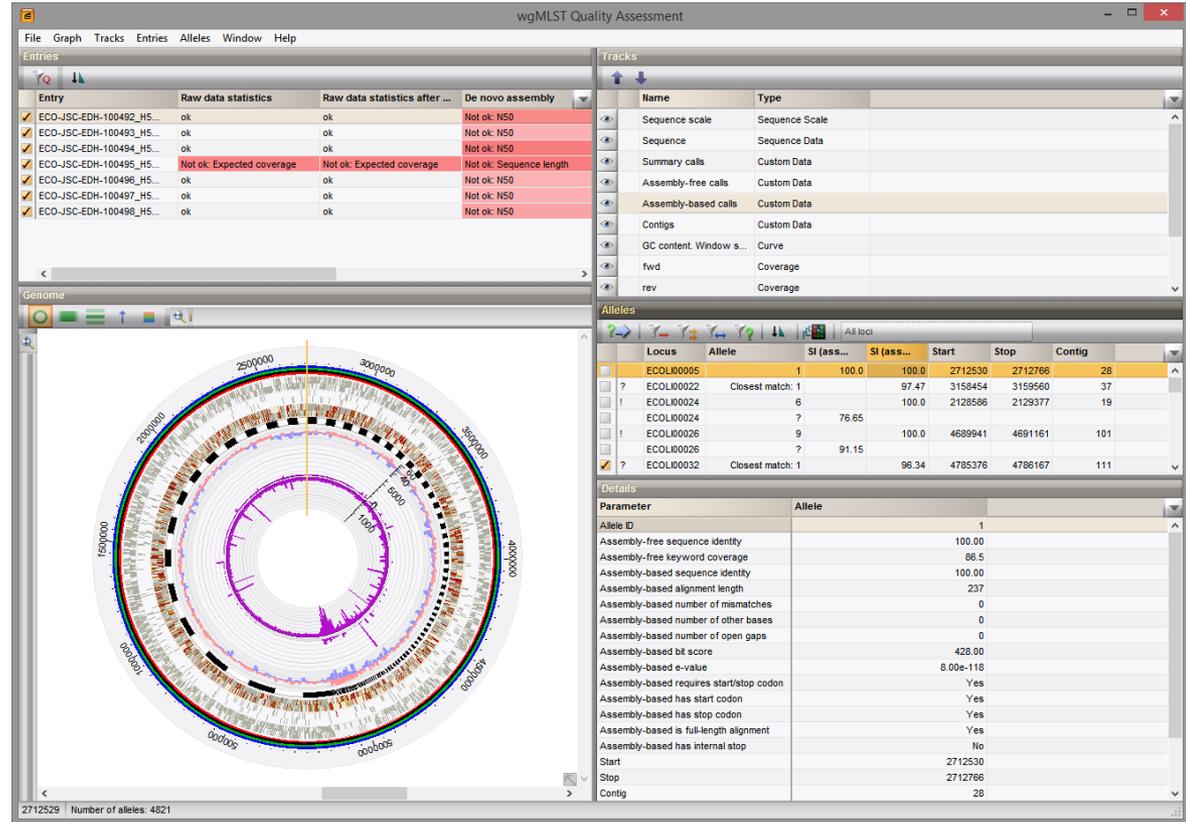
- 1 Automated import from sequence read sets
- 2 Batch job processing on the calculation engine
- 3 Quality assessment of imported results
- 4 Automated submission of new alleles
- 5 Automated assignments and sample reporting
- 6 Calculate population modelling networks



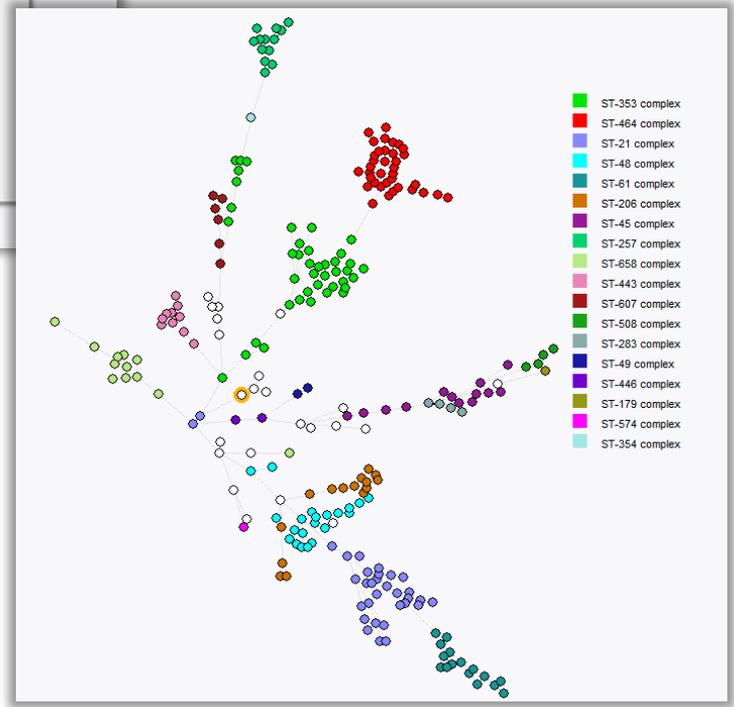
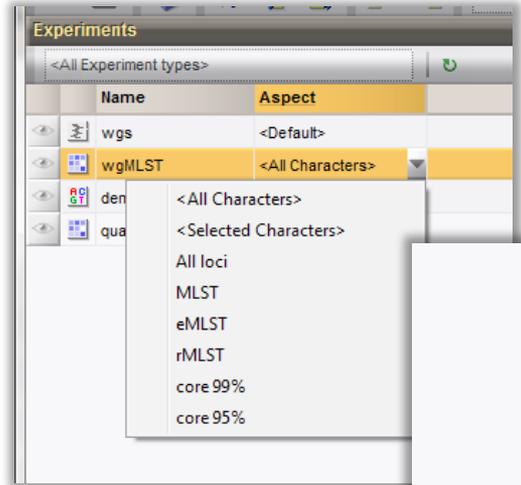
The screenshot displays three overlapping dialog boxes from a software application:

- Import sequence read sets:** A dialog box with the title "Import sequence read sets" and a subtitle "Select the data source". It contains a single button labeled "Select the data source".
- Submit jobs:** A dialog box with the title "Submit jobs". It features a section for "Algorithms" with four options:
 - Reference mapping (Settings...)
 - De novo assembly (Settings...)
 - Assembly-based calls (Settings...)
 - Assembly-free calls (Settings...)
 Below this is a "Jobs" section with:
 - Re-submit already processed data
 - Open jobs overview window
 At the bottom are "OK" and "Cancel" buttons, and a "Details..." button.
- De novo assembly:** A dialog box with the title "De novo assembly". It contains several input fields:
 - Min. coverage: -1.0
 - Expected coverage: -1.0
 - Min. contig length: 300
 - Assembler: SPAdes (dropdown menu)
 - k-mer sizes: (Optional)
 - Data format: IonTorrent (dropdown menu)
 At the bottom, there is a checkbox for "Save algorithm settings as default" and "OK" and "Cancel" buttons.

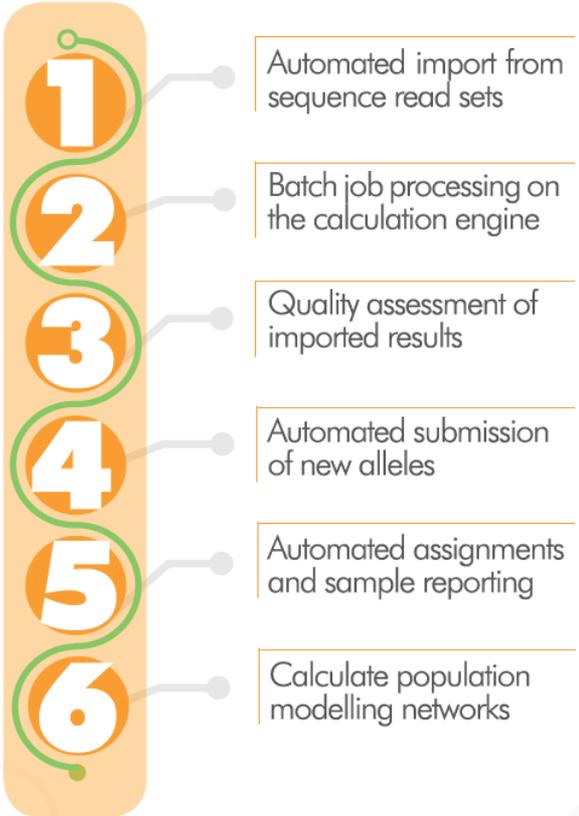
- 1 Automated import from sequence read sets
- 2 Batch job processing on the calculation engine
- 3 Quality assessment of imported results
- 4 Automated submission of new alleles
- 5 Automated assignments and sample reporting
- 6 Calculate population modelling networks



- 1 Automated import from sequence read sets
- 2 Batch job processing on the calculation engine
- 3 Quality assessment of imported results
- 4 Automated submission of new alleles
- 5 Automated assignments and sample reporting
- 6 Calculate population modelling networks



wgMLST for cluster detection



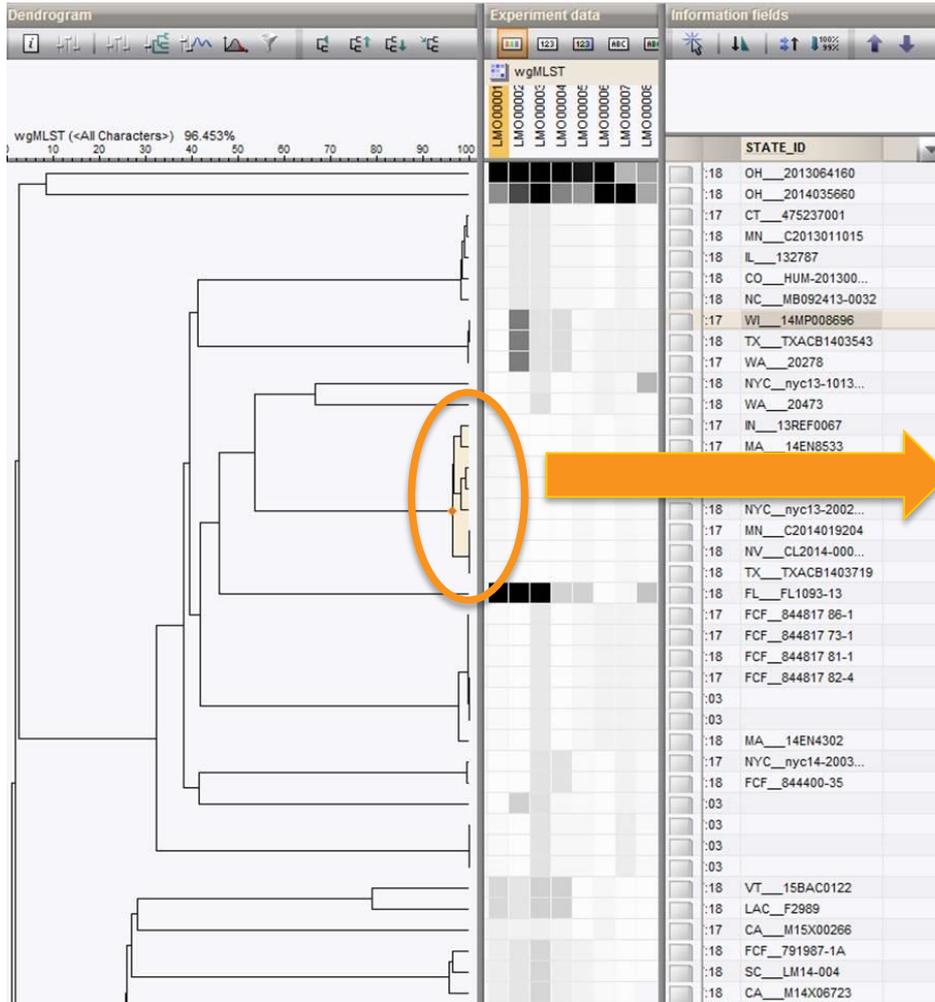
BioNumerics wgMLST schemes

Released

- *Staphylococcus aureus*
- *Listeria monocytogenes*

In preparation:

- *Campylobacter*
- *E. coli / Shigella*
- *Salmonella enteritidis*
- *Mycobacterium tuberculosis*
- *Pseudomonas aeruginosa*
- *Legionella pneumophila*
- *Bacillus anthracis/subtilis/cereus*
- *Neisseria gonorrhoeae*
- *Clostridium difficile*
- ...





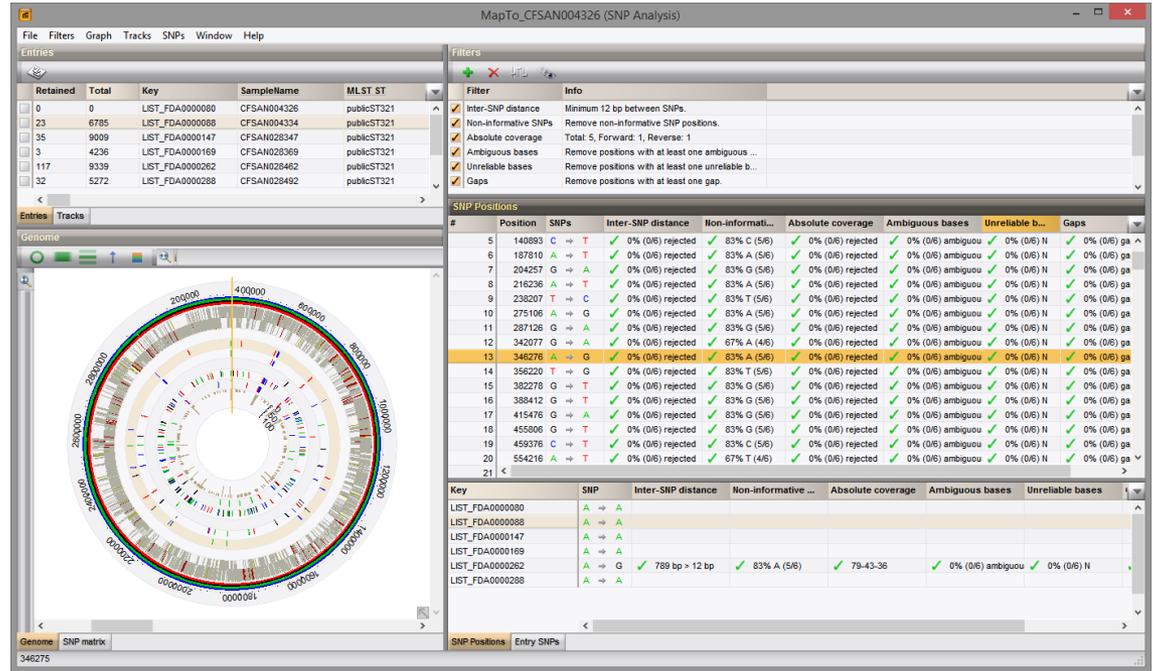
1 Automated import from sequence read sets

2 Choice of reference sequence

3 Mapping against the reference sequence

4 wgSNP assessment and filtering

5 Calculate population modelling networks



Add SNP filters

- Coverage
 - Absolute coverage
 - Relative coverage
- Base quality
 - Unreliable bases
 - Ambiguous bases
 - Gaps
- Abundance
 - Non-informative SNPs
 - Singleton SNPs
 - Majority SNPs
- Position
 - Position mask
 - Inter-SNP distance
- Mutation types

Minimum coverage
 Forward: 1
 Reverse: 1
 Total: 5

Position retention threshold
 Maximum frequency: 100 %

Description: Per experiment, remove SNPs with insufficient (absolute) coverage. This filter is ignored if no coverage information is present.

OK Cancel

Pairwise alignment

- Fast algorithm
- Standard

Multiple alignment

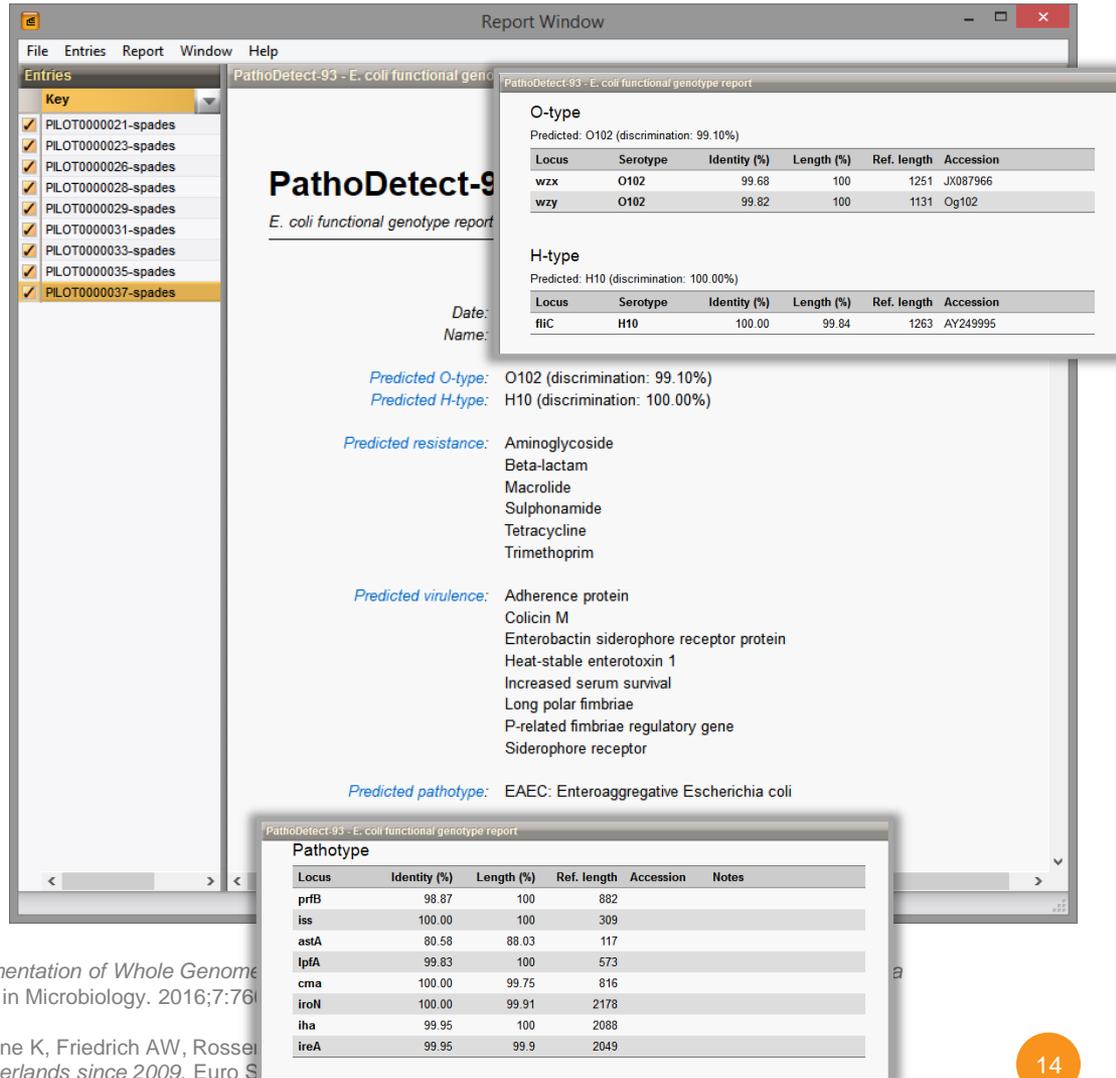
- Multiple alignment based
- SNP-based
- Categorical

SNP Filtering

Template: Basic filtering

ATTAAGGATCAAGCGCTT

- Epidemiological metadata
isolation date, isolation place, linked patient information
- Typing data
MLST, coreMLST, wgMLST typing
- Pathotype detection
- O/H serotyping
- Resistance and virulence detection
- Plasmids and prophage detection
- In-silico PCR



Report Window

File Entries Report Window Help

PathoDetect-93 - E. coli functional genotype report

PathoDetect-93
E. coli functional genotype report

Date: _____
Name: _____

Predicted O-type: O102 (discrimination: 99.10%)
Predicted H-type: H10 (discrimination: 100.00%)

Predicted resistance: Aminoglycoside
Beta-lactam
Macrolide
Sulphonamide
Tetracycline
Trimethoprim

Predicted virulence: Adherence protein
Colicin M
Enterobactin siderophore receptor protein
Heat-stable enterotoxin 1
Increased serum survival
Long polar fimbriae
P-related fimbriae regulatory gene
Siderophore receptor

Predicted pathotype: EAEC: Enterocaggregative Escherichia coli

Pathotype

Locus	Identity (%)	Length (%)	Ref. length	Accession	Notes
prfB	98.87	100	882		
iss	100.00	100	309		
astA	80.58	88.03	117		
lpfA	99.83	100	573		
cma	100.00	99.75	816		
iroN	100.00	99.91	2178		
iha	99.95	100	2088		
ireA	99.95	99.9	2049		

O-type
Predicted: O102 (discrimination: 99.10%)

Locus	Serotype	Identity (%)	Length (%)	Ref. length	Accession
wzx	O102	99.68	100	1251	JX087966
wzy	O102	99.82	100	1131	Og102

H-type
Predicted: H10 (discrimination: 100.00%)

Locus	Serotype	Identity (%)	Length (%)	Ref. length	Accession
flfC	H10	100.00	99.84	1263	AY249995

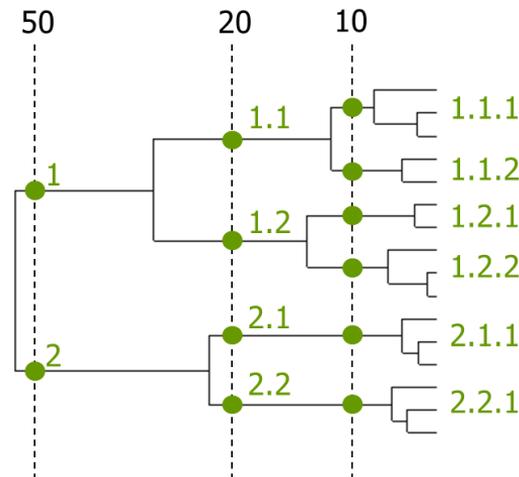
Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. *Implementation of Whole Genome Toxin-Producing Escherichia coli (STEC) in the United States*. *Frontiers in Microbiology*. 2016;7:761-770.

Kluytmans-van den Bergh MF, Huizinga P, Bonten MJ, Bos M, De Bruyne K, Friedrich AW, Rosser J, et al. *Enterobacteriaceae in retail chicken meat but not in humans in the Netherlands since 2009*. *Euro Surveill*. 2016;21:9.30149.

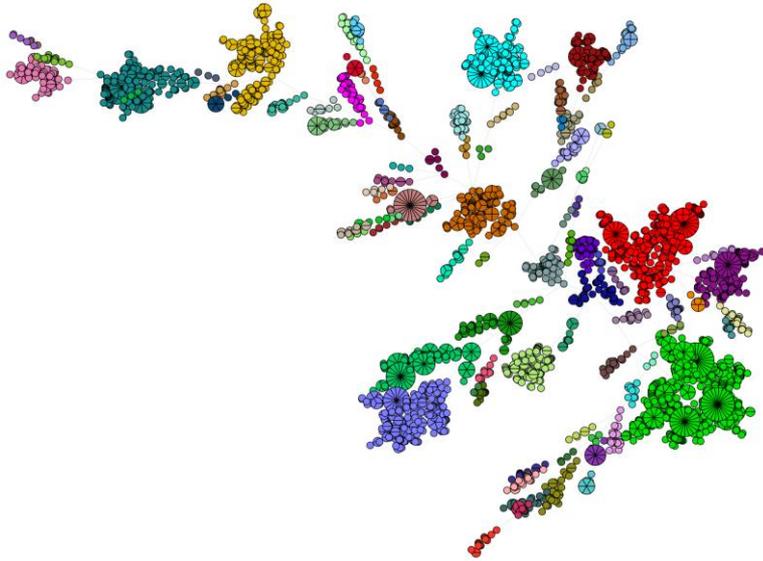
Strain nomenclature principles

- **The ST idea, i.e. 100% identity rule won't work**
 Almost every strain has some unique mutations, therefore would have a unique ST

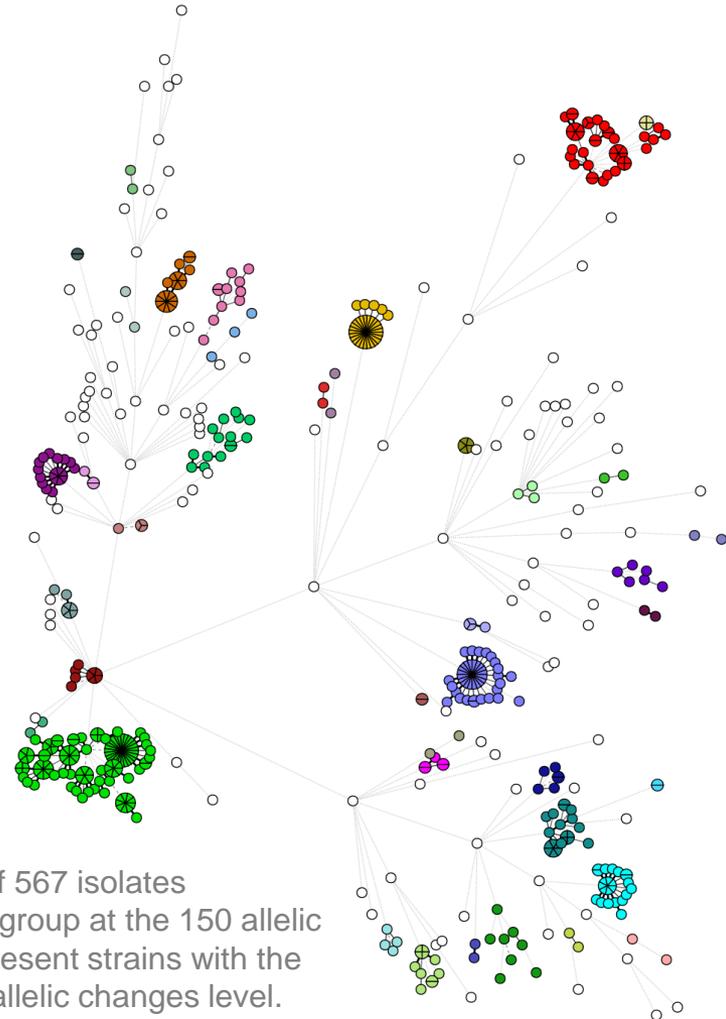
- **Hierarchical strain nomenclature**
 - Reflects the distance between strains
 - Based on single linkage clustering at different distance cutoffs (e.g. using cgMLST)
 - The cutoff for each level has been optimized and validated to minimize the error rate of assignments



Strain nomenclature applied on Listeria



Minimum spanning tree of 3652 isolates
 Colors represent strains with the same name at the 150 allelic changes level.
 Sporadic groups (<3 isolates) have been removed



Minimum spanning tree of 567 isolates
 All isolates in the biggest group at the 150 allelic changes level. Colors represent strains with the same name up to the 11 allelic changes level.
 Sporadic isolates indicated in white.



A B I O M É R I E U X C O M P A N Y

BioNumerics [GET A FREE TRIAL](#)

Katrien_DeBruyne@applied-maths.com