wgMLST

Whole genome multi-locus sequence typing

Hannes Pouseele

Applied Maths

Introduction

Goals

Dataset

Credits

Disclaimer

Goals

- The wgMLST analysis strategy
- Why you should like the wgMLST approach
- An automated pipeline for wgMLST

- Analysis of *Listeria monocytogenes* surveillance data using wgMLST

Introduction

Goal

Dataset

Credits

Disclaimer

Dataset: real-time Listeria surveillance data

- 349 samples (356 sequencer runs) of Listeria monocytogenes
- Clinical and environmental samples

- Data set contains both sporadic samples and outbreak samples

- In particular, the data set contains data from the Crave Brothers Cheese outbreak.

Introduction

Goal

Dataset

Credits

Disclaimer

Credits

- Data generation: CDC
- Sequencing platforms: Illumina (GAII, MiSeq)
- Data analysis: BioNumerics® 7.5 (Applied Maths)

Introduction

Goal

Dataset

Credits

Disclaimer

Disclaimer

For the last 7 years, I was affiliated with the following organization:

- Organization:

  Applied Maths NV, Keistraat 120, B-9830 Sint-Martens-Latem, Belgium

- Relationship:

  employee (salary)

wgMLST

- Starting point : a set of loci

  A locus is determined by

  - a set of variants for this locus, and
  - possibly, a set of validations for this locus (*in silico* PCR primers)

- For each sample, determine for each locus whether
  - the locus is present, and
  - if present, which variant (“allele”) of the locus is present

- Two procedures: Identify alleles
  - from (de novo) assembled genomes
  - directly from the reads

Allele identification

Two procedures:

- Assembly-based:

  Identify alleles from (de novo) assembled genomes using BLAST

  ✓Computationally intensive

  ✓Multiple contigs lead to loci that are missing from the assembly

  ✓Undefined behavior for multi-copy loci

  ✓Required for extrinsic validation

- Assembly-free:

  Identify alleles directly from the reads

  ✓Computationally less intensive

  ✓Missing loci are missing in the reads, not in the assembly

  ✓Exhaustive procedure for multi-copy loci

  ✓No extrinsic validation possible

Strong points

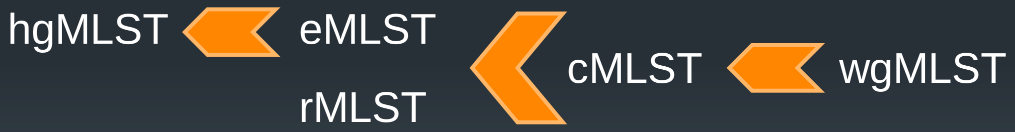
✓Based on the concept of allelic variation, not only point mutations

Different estimate of evolutionary distance for events such as

- recombination
- simultaneous close-range mutations

are counted as one event, which might biologically more relevant

✓Fixed set of loci leading to typing schemes on different levels



✓Close to functional analysis

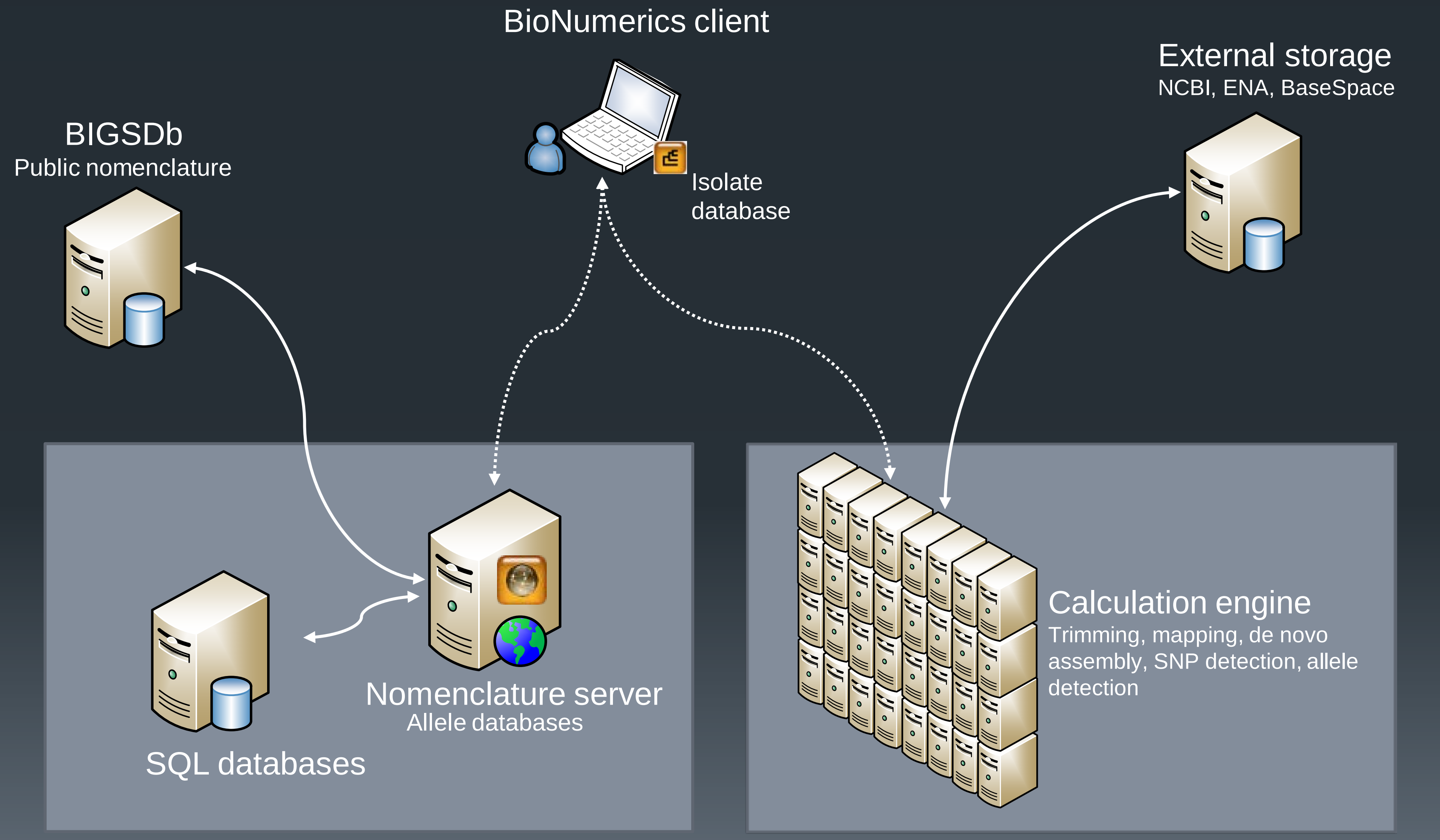
✓Naturally incorporates multiple references

✓Can be translated to SNPs within loci

Drawbacks

- ✓Does not consider the “complete” genome
  - By default, intergenic regions only
- ✓Requires curation
  - Since the goal is to have a universal set of loci and alleles that can be communicated about, consistent naming is required.

wgMLST pipeline



wgMLST pipeline

Sample side: lightweight sample database

- No heavy lifting in storage or in calculations
- Metadata remains local
- Total automation: “back to biology”
  - synchronized with NCBI BioProject
  - jobs are posted and retrieved automatically

wgMLST pipeline

Curation side: automated curation tools

- Alleles are automatically named (independent of public nomenclature)

  validation criteria: deviation from closest known allele, check for start/stop codon, …

- Automated sequence type assignment

- Automatic synchronization with public nomenclature

- Annotation of alleles, loci, sequence types, typing schemes … used in sample reporting.

wgMLST pipeline demo

- Complete setup running at the Applied Maths booth

- Sample and allele databases available for
  - Campylobacter
  - Listeria monocytogenes
  - Mycobacterium tuberculosis
  - Salmonella

- New schemes need about a day to develop

Lmo wgMLST scheme

- 5 Reference sequences representing 4 lineages

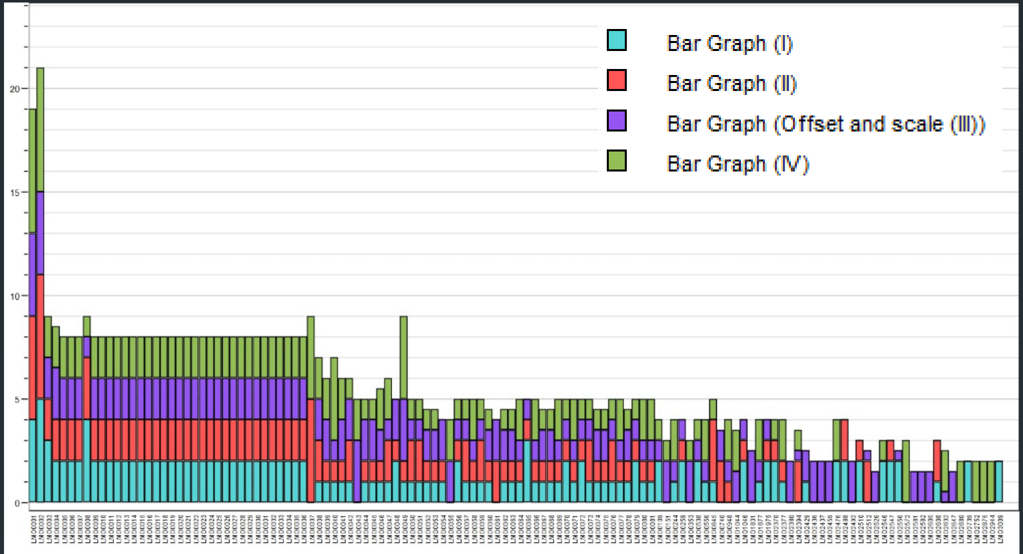
  FM242711, NC_003210, FM211688, CP001175, NZ_CM001469

- CDS's of these 5 sequences used to obtain homologous groups (Deng2010) and thus a set of well-defined loci.
  - Sequence identity > 50%
  - E-value < 0.00001

Lmo wgMLST scheme

Number of loci:	3436	
Number of core loci: (note: core = present in all lineages)	2345	68.25%
Number of lineage-specific loci: (I: 134, II: 283, III: 156, IV: 76)	649	18.9%
Number of multi-allelic loci: (I: 56, II: 53, III: 74, IV: 66)	126	3.7%

Lmo wgMLST scheme



- up to 6 copies of the same locus per sample
- some loci tend to be multi-copy in all lineages
- some loci are lineage-specific and multi-copy

PCR validations (20 bp length): largely lineage-specific

- Unique set of PCR primers: 779 loci (22.7%)
- Preceding PCR primer: 1147 loci (33.4%)
- Succeeding PCR primer: 976 loci (28.4%)

Acknowledgements

- CDC team:

  Peter Gerner-Smidt, Lee Katz, Steven Stroika, Kelly Jackson, …

- Applied Maths team:

  Johan Goris, Bruno Pot, Katrien De Bruyne, Koen Janssens
  Jeroen Van Goey, Jan Stout, Dolf Michielsen, Sander Valcke,
  Kenny Knecht