http://microbesgenotyping.i2bc.paris-saclay.fr

Legionella pneumophila genotypes database deduced from available whole genome sequence data

Gilles Vergnaud, David Christiany, Christine Pourcel

University Paris-Saclay, Gif-sur-Yvette, France

Genotypes databases are central

Genotypes databases used to be assembled from published data directly obtained by strain typing. Currently, because relevant strains are increasingly being whole genome sequenced (WGS), it becomes easier and more robust to extract genotype data from WGS. We apply this approach to *L. pneumophila* and create an *in silico* database with 646 entries and two genotyping assays, MLVA and SBT.

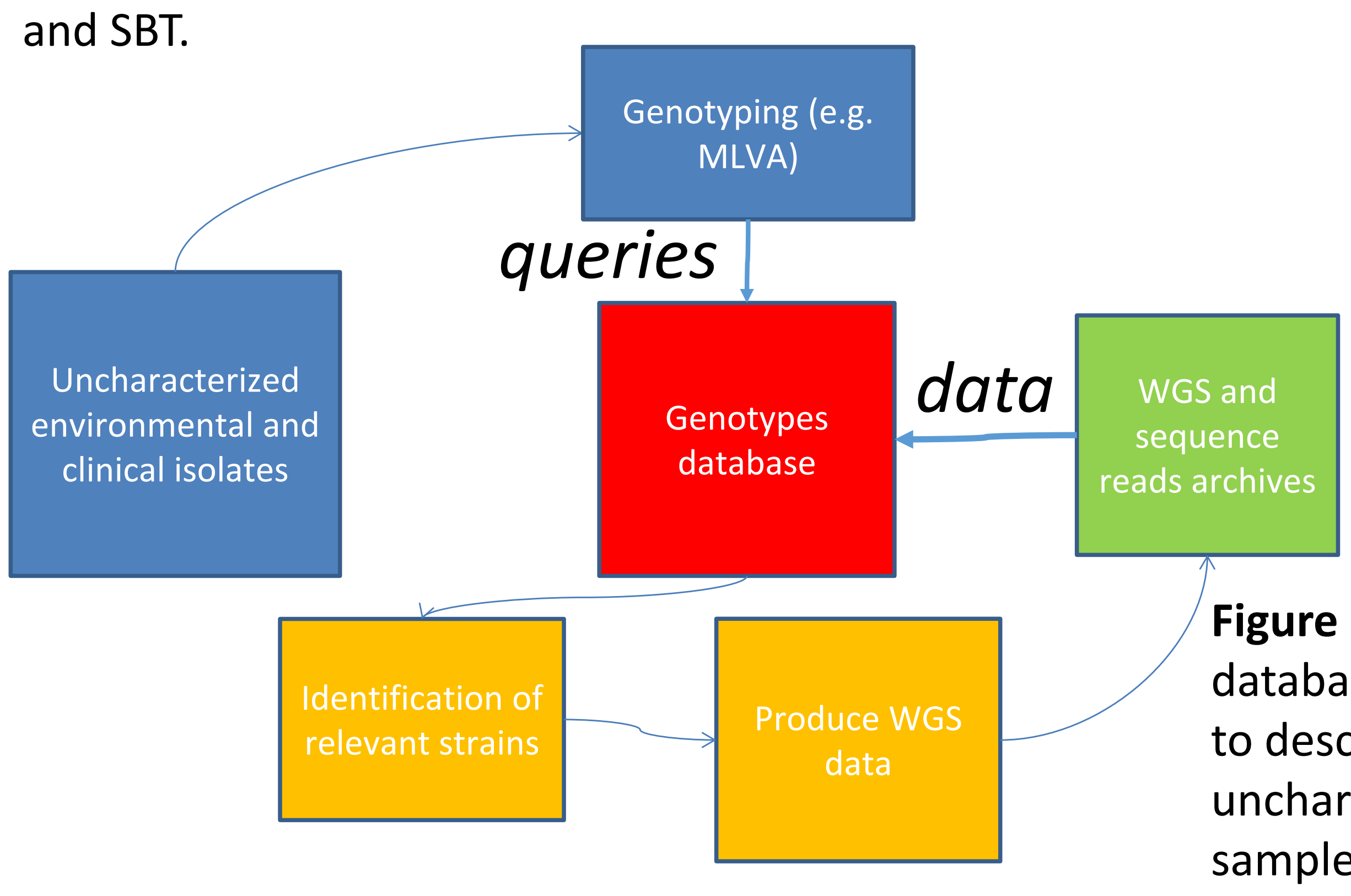

Genotyping (e.g. MLVA)
queries
Uncharacterized environmental and clinical isolates
Genotypes database
data
WGS and sequence reads archives
Identification of relevant strains
Produce WGS data

Figure 1: genotypes databases are central to describe uncharacterized samples but also WGS archives.

Genotypes for *L. pneumophila:* SBT and MLVA

The central genotype database in **Figure 1** contains genotypes from assays appropriate for uncharacterized isolates. Multiple Loci VNTR (Variable Number of Tandem Repeats) Analysis (MLVA) has proved applicable as a first line assay for investigations including hundreds of strains. MLVA allows the production of a code corresponding to the number of repeats at each VNTR investigated. MLVA can be applied directly on environmental samples (ref. 1).

Fourteen VNTR loci have been described in the *Legionella pneumophila* genome (ref. 2, 3, 6). Three selections of loci have been proposed for routine typing MLVA8, MLVA10 and MLVA12, but additional combinations or subsets can be used to fit local epidemiology.

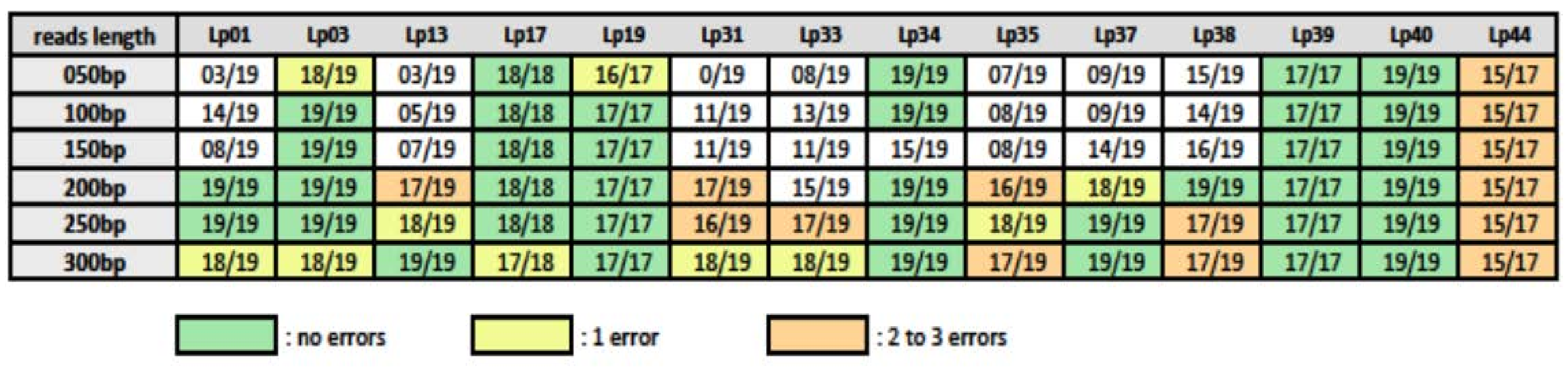
SBT is deduced from the sequence of seven loci. More than 2000 STs (genotypes) have been defined, see

http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php

In silico analysis of available sequence data

Two types of data sets are available for *in silico* typing, "complete genomes" and sequence reads archives (SRA). Nineteen complete *L. pneumophila* genomes are publicly available compared to more than 700 SRA files. SBT and MLVA codes of complete genomes were deduced *in silico* using the BioNumerics (Applied-Maths) version 7.61 tools and an *in house* python script respectively.

In order to estimate the read length necessary for a correct reconstruction of tandem repeats length from read archives, the complete genomes were used to produce reads data using artificial fastq (ref. 8). The reads were assembled using SPAdes version 3.9 (reference 9). **Table 1** indicates that **reads longer than 200 bp can be used to reconstruct all VNTRs** with a reasonable success rate, whereas only some VNTRs can be confidently reconstructed with shorter reads.



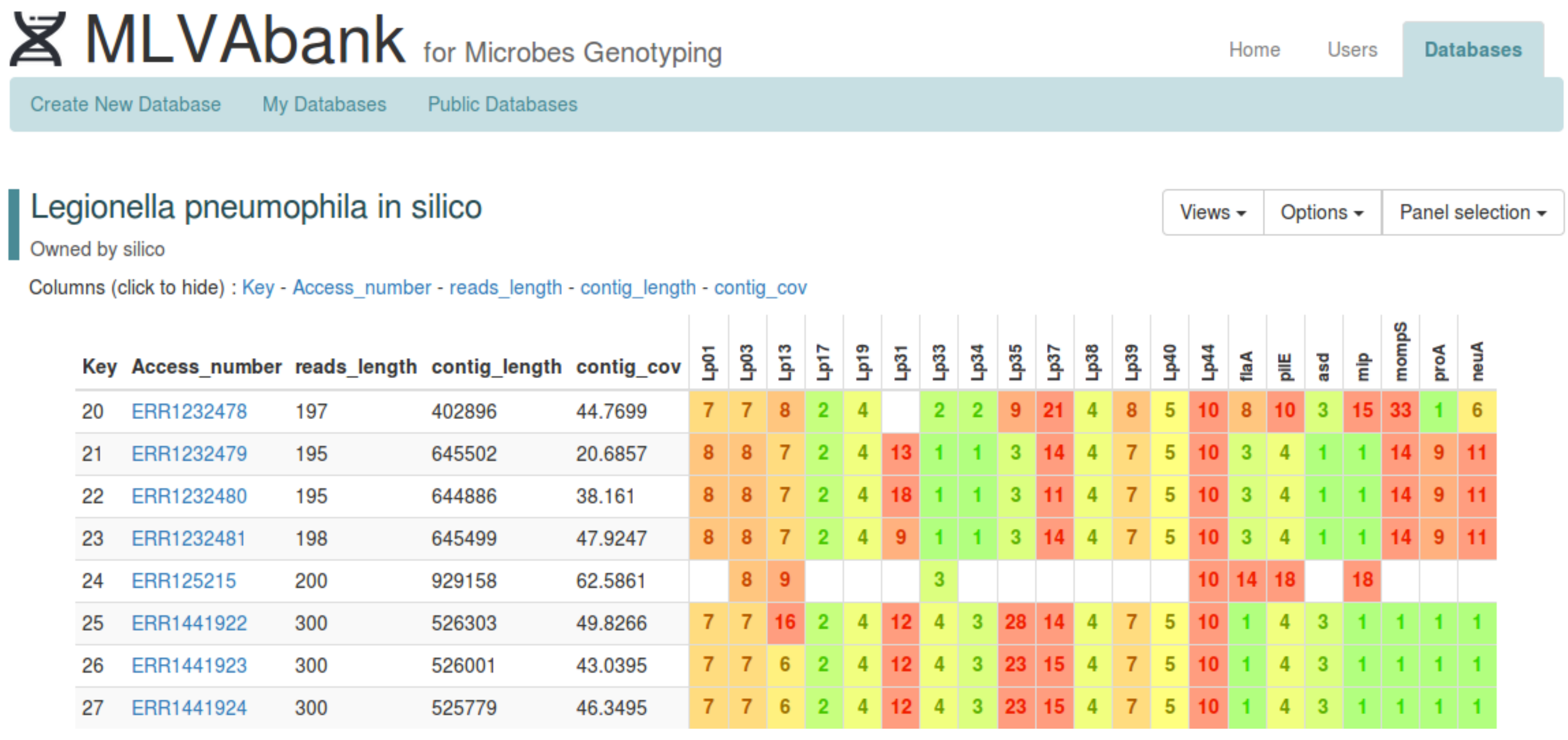
reads length	Lp01	Lp03	Lp13	Lp17	Lp19	Lp31	Lp33	Lp34	Lp35	Lp37	Lp38	Lp39	Lp40	Lp44
050bp	03/19	18/19	03/19	18/18	16/17	0/19	08/19	19/19	07/19	09/19	15/19	17/17	19/19	15/17
100bp	14/19	19/19	05/19	18/18	17/17	11/19	13/19	19/19	08/19	09/19	14/19	17/17	19/19	15/17
150bp	08/19	19/19	07/19	18/18	17/17	11/19	11/19	15/19	08/19	14/19	16/19	17/17	19/19	15/17
200bp	19/19	19/19	17/19	18/18	17/17	17/19	15/19	19/19	16/19	18/19	19/19	17/17	19/19	15/17
250bp	19/19	19/19	18/19	18/18	17/17	16/19	17/19	19/19	18/19	19/19	17/19	17/17	19/19	15/17
300bp	18/19	18/19	19/19	17/18	17/17	18/19	18/19	19/19	17/19	19/19	17/19	17/17	19/19	15/17

🟩 : no errors 🟨 : 1 error 🟧 : 2 to 3 errors

Table 1: efficiency of VNTR assembly according to read length. Reads with length ranging from 50bp up to 300 bp were simulated. For each locus, the number of correct and incorrect reconstructions is indicated.

Creating the *Legionella pneumophila in silico* database

All reads archives deposited in the European Nucleotide Archive before july 2016 (more than 750) were downloaded and assembled with SPAdes 3.9. Some datasets did not assemble correctly (largest contig smaller than 100 kb) and were discarded. The SBT and MLVA were deduced when appropriate according to **Table 1**. The resulting data was used to create the *Legionella pneumophila in silico* database (646 entries in august 2016 release).



Legionella pneumophila in silico

Owned by silico

Columns (click to hide) : Key - Access_number - reads_length - contig_length - contig_cov

Key	Access_number	reads_length	contig_length	contig_cov	Lp01	Lp03	Lp13	Lp17	Lp19	Lp31	Lp33	Lp34	Lp35	Lp37	Lp38	Lp39	Lp40	Lp44	flaA	pilE	asd	mip	mompS	proA	neuA
20	ERR1232478	197	402896	44.7699	7	7	8	2	4		2	2	9	21	4	8	5	10	8	10	3	15	33	1	6
21	ERR1232479	195	645502	20.6857	8	8	7	2	4	13	1	1	3	14	4	7	5	10	3	4	1	1	14	9	11
22	ERR1232480	195	644886	38.161	8	8	7	2	4	18	1	1	3	11	4	7	5	10	3	4	1	1	14	9	11
23	ERR1232481	198	645499	47.9247	8	8	7	2	4	9	1	1	3	14	4	7	5	10	3	4	1	1	14	9	11
24	ERR125215	200	929158	62.5861		8	9				3							10	14	18		18			
25	ERR1441922	300	526303	49.8266	7	7	16	2	4	12	4	3	28	14	4	7	5	10	1	4	3	1	1	1	1
26	ERR1441923	300	526001	43.0395	7	7	6	2	4	12	4	3	23	15	4	7	5	10	1	4	3	1	1	1	1
27	ERR1441924	300	525779	46.3495	7	7	6	2	4	12	4	3	23	15	4	7	5	10	1	4	3	1	1	1	1

Figure 2: a view of the *in silico* database

Using the *Legionella pneumophila in silico* database

The database can be queried, i.e. a genotype (MLVA, SBT) can be compared to the stored profiles. In addition, the whole database can be downloaded in order to run local analyses with preferred software. **Figure 3** shows one such analysis .

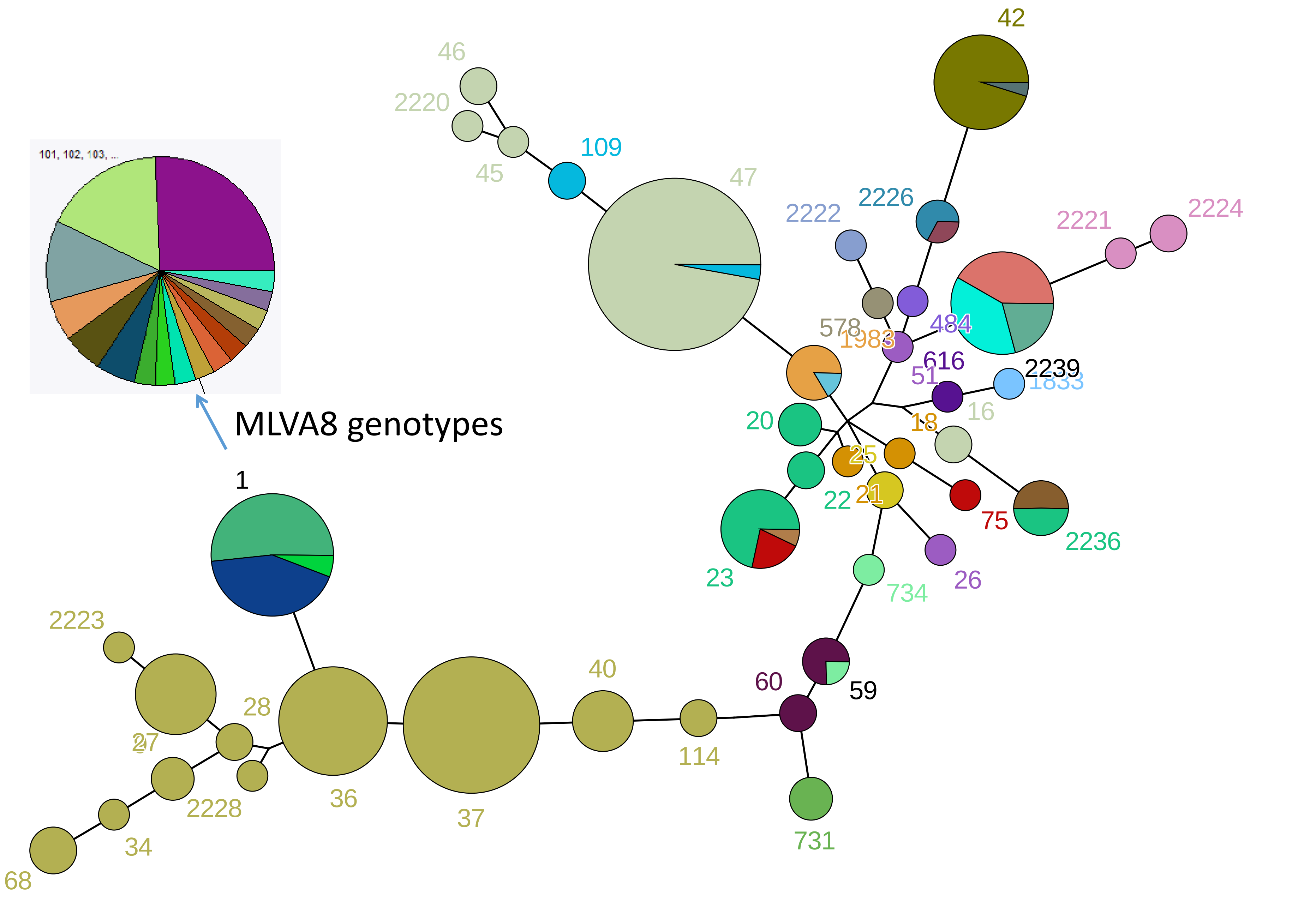


Figure 3: Minimum spanning Tree based upon SBT data, colored according to MLVA8. A full SBT and MLVA8 dataset is available for 326 entries. SBT and MLVA8 resolves 43 and 105 genotypes respectively. The color code reflects MLVA clonal complexes clustering (31 clusters). ST numbers are indicated. ST1 is divided into 3 MLVA CC or 16 genotypes.

Conclusions

MLVA data can be deduced from the majority of *L. pneumophila* reads archives. Genotypes databases can consequently be produced from sequence reads archives as well as complete genomes. Such databases which can include any relevant genotyping assay will allow the classification of new strains. Subsequently, relevant strains can be identified for whole genome sequencing. Once genotypes have been gathered, private or public genotypes databases can be created in a few minutes using http://microbesgenotyping.i2bc.paris-saclay.fr.

Acknowledgments

We thank Drs Baharak Afshar and Georgia Kapatai, Public Health England, for the provision of SBT alleles and profiles data from the Legionella Sequence Based Typing genotypes database at http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php.

Bibliography

1. **Spatial distribution of *Legionella pneumophila* MLVA-genotypes in a drinking water system.** Rodríguez-Martínez S, Sharaby Y, Pecellín M, Brettar I, Höfle M, Halpern M. **Water Res. 2015 Jun 15;77:119-32. PMID:25864003**

2. **High-throughput typing method to identify a non-outbreak-involved *Legionella pneumophila* strain colonizing the entire water supply system in the town of Rennes, France.** Sobral D, Le Cann P, Gerard A, Jarraud S, Lebeau B, Loisy-Hamon F, Vergnaud G, Pourcel C. **Appl Environ Microbiol. 2011 Oct;77(19):6899-907. PMID:21821761**

3. **Investigation of the population structure of *Legionella pneumophila* by analysis of tandem repeat copy number and internal sequence variation.** Visca P, D'Arezzo S, Ramisse F, Gelfand Y, Benson G, Vergnaud G, Fry NK, Pourcel C.
Microbiology. 2011 Sep;157(Pt 9):2582-94. PMID:21622529

4. **High-resolution in situ genotyping of *Legionella pneumophila* populations in drinking water by multiple-locus variable-number tandem-repeat analysis using environmental DNA.** Kahlisch L, Henne K, Draheim J, Brettar I, Höfle MG. **Appl Environ Microbiol. 2010 Sep;76(18):6186-95. PMID:20656879**

5. **Multiple-locus variable-number tandem repeat analysis of *Legionella pneumophila* using multi-colored capillary electrophoresis.** Nederbragt AJ, Balasingham A, Sirevåg R, Utkilen H, Jakobsen KS, Anderson-Glenna MJ. **J Microbiol Methods. 2008 May;73(2):111-7. PMID:18374436**

6. **Identification of variable-number tandem-repeat (VNTR) sequences in *Legionella pneumophila* and development of an optimized multiple-locus VNTR analysis typing scheme.** Pourcel C, Visca P, Afshar B, D'Arezzo S, Vergnaud G, Fry NK. **J Clin Microbiol. 2007 Apr;45(4):1190-9. PMID:17251393**

7. **Characterization of a tandem repeat polymorphism in *Legionella pneumophila* and its use for genotyping.** Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C. **J Clin Microbiol. 2003 May;41(5):1819-26. PMID:12734211**

8. **Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines.** Frampton M1, Houlston R. **PLoS One. 2012;7(11):e49110. PMID:23152858**

9. **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. **J Comput Biol. 2012 May;19(5):455-77. PMID:22506599**