MLVA-NET – A STANDARDISED WEB DATABASE FOR BACTERIAL GENOTYPING AND SURVEILLANCE

G Guigon[1], J Cheval[1], R Cahuzac[2], S Brisse (sbrisse@pasteur.fr)[1]
1. Institut Pasteur, Genotyping of Pathogens and Public Health, Paris, France
2. Institut Pasteur, Genome Analysis and Integration, Paris, France

Background

Strain typing is an important aid to surveillance networks and outbreak investigations of infectious diseases [1]. MLVA (Multilocus VNTR Analysis, with VNTR standing for Variable Number of Tandem Repeats) has emerged as a highly discriminatory and widely applicable genotyping method that is now being applied for strain tracking in a growing number of bacterial pathogens [2,3]. The genomic loci containing tandem repeats are often maintained among strains of a bacterial species, while individual strains harbour different copy numbers that can be determined simply by PCR amplification. Similar to sequence-based methods such as Multilocus Sequence Typing (MLST), the MLVA method indexes genetic variation at well defined genomic loci and produces reproducible allelic profiles that can be coded in a simple digital format. Hence, they represent an attractive alternative to banding profile-based methods such as pulsed-field gel electrophoresis (PFGE), which requires dedicated efforts (e.g. http://www.cdc.gov/pulsenet) in order to produce fingerprinting data that are comparable across laboratories. Indeed, to be useful to surveillance networks and for global epidemiology, a genotyping method has to be technically accessible, reproducible and to yield easily portable data. In addition, electronic databases that are made accessible through the Internet can render exchange and comparison of data among laboratories very effective for local, national, and international surveillance.

Existing databases of MLST data accessible through web portals (http://www.pubmlst.org, http://www.mlst.net, http://www.pasteur.fr/mlst) represent a common language for strain typing that has proven extremely useful for collaborative research and global epidemiology of bacterial and fungal pathogens [4]. However, given the much faster evolutionary rate of tandem repeats compared to nucleotide sequences, MLVA markers provide much improved resolution compared to MLST, thus representing a subtyping tool that is especially useful for strain discrimination in genetically homogeneous pathogens, such as M. tuberculosis [5], Bacillus anthracis [6] or Salmonella enterica serotype Typhimurium [7]. Web-accessible MLVA databases are not yet widely used for international collaboration [8], but the development in this area is very active (http://mlva.u-psud.fr/, http://www.mlva.eu/, http://www.miru-vntrplus.org).

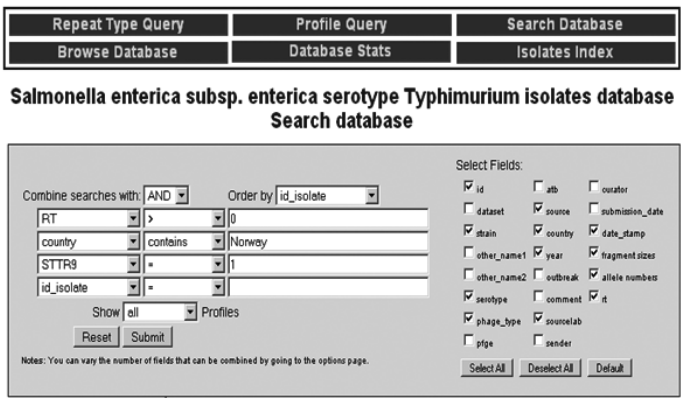
Description of MLVA database

We have developed MLVA-NET (http://www.pasteur.fr/mlva), a web-accessible database system dedicated to the comparison of MLVA genotyping profiles and to retrieval of relevant epidemiological information for the corresponding isolates. An unlimited number of organisms (species, subspecies, serovars or other categories) can be entered into the system. Curators, working through the internet, create and maintain one or several datasets (groups of isolates) for one or more organisms, Individuals who are in charge of data management for a collaborative network can request curator rights from the MLVA-NET administrator. There is no limit to the number of curators and datasets for a given organism.

The database contains two types of data – profiles and isolates – which are accessed through distinct links. Each allele at a given locus is assigned a so-called 'allele number'. When combined over all loci, these numbers make up a numeric code that defines a particular MLVA profile, or repeat type (RT). All MLVA profiles are immediately made public in order to provide the necessary common language for microbial strain typing. In contrast, the curator, in agreement with the person who supplied the data, can decide to keep private the epidemiological information related to isolates, such as isolate name, country or date of isolation. The decision to

FIGURE 1

Example of a MLVA-NET isolates query using the <Search database> menu



keep isolate information private or to make it public is made for entire datasets, not for individual isolates. Hence, the web pages showing information on isolates contain public datasets that are available for all external users, as well as private datasets that are accessible only for registered users through a password-protected identification step. Registered users can either have only reading access, or predefined curator rights that allow them to import or modify isolates.

An important principle of MLVA-NET is to store raw data, i.e. the length of PCR fragments, as determined on agarose gels or capillary electrophoresis. The fragment sizes are automatically translated by the system into allele numbers. Each allele is assigned to a bin, corresponding to PCR fragment lengths ranging between a lower and a higher bound. For each organism, different ways of defining bins ("coding methods") can co-exist according to the preferences of user networks. Therefore, our system retains maximal information (fragment lengths) while providing flexibility and adaptability in the way data can be analysed. For example, the discovery of incomplete repeats in some strains can be taken into account without having to rebuild the database. Because tandem repeats can evolve by stepwise loss or gain of a single repeat, it can be useful to take into account in phylogenetic analyses the difference in the number of repeats between strains. Therefore, a coding method can be defined so that allele numbers correspond to the repeat number, instead of arbitrary numbers (e.g., numbered successively as they are discovered).

The system accepts missing data, which is important given the fact that not all strains contain all possible VNTR loci. The same organism can be analysed by several methods, which can differ by the marker set (number and identity of loci), their order in the allelic profile, and by the definition of bins and alleles. Hence, for a given set of markers, data can be compared across datasets even if contributing laboratories have different preferences for bin definition or allele number assignment.

So far, the database is suitable only for haploid organisms.

Besides download and search functions that give access to the entire public contents of the database, a number of flexible query and comparison functions are available. Notably, they allow strains that have been newly genotyped by the user to be compared to the content of the database. The user can search for all RTs that are identical or similar to a query profile, or retrieve the profile corresponding to a particular RT. An advanced search function is available that allows combining queries with comparison operators (=, >, <, NOT, NOT contains, contains). The search form (Figure 1) allows (i) to enter search criteria in chosen fields, (ii) the way

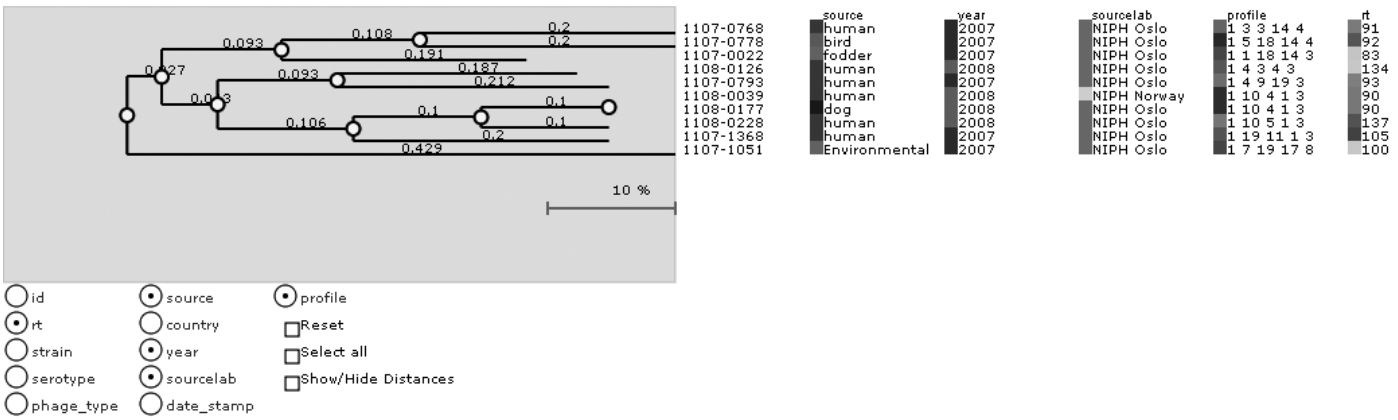
Figure 2

Example of a MLVA-NET results page for S. Typhimurium isolates from Norway with allele number 1 for marker STTR9

Isolates information									Fragment sizes					Allele numbers					RT
id	strain	serotype	phage_type	source	country	year	sourcelab	date_stamp	STTR9	STTR5	STTR6	STTR10	STTR3	STTR9	STTR5	STTR6	STTR10	STTR3	rt
189	1107-0022			fodder	Norway	2007	NIPH Oslo	2007-11-14	162	227	394	363.00	524.00	1	1	18	14	3	83
369	1107-0768			human	Norway	2007	NIPH Oslo	2007-11-22	162	239	300	362	549	1	3	3	14	4	91
377	1107-0778			bird	Norway	2007	NIPH Oslo	2007-11-22	162	252	394	362	550	1	5	18	14	4	92
380	1107-0793			human	Norway	2007	NIPH Oslo	2007-11-22	162	246	348	350.00	523	1	4	9	19	3	93
423	1107-1051			Environmental	Norway	2007	NIPH Oslo	2007-11-22	162	264	305	344	325	1	7	19	17	8	100
457	1107-1368			human	Norway	2007	NIPH Oslo	2007-11-22	162	306	359	356	523	1	19	11	1	3	105
599	1108-0039			human	Norway	2008	NIPH Norway	2008-01-21	162	300	318	356	524.00	1	10	4	1	3	90
603	1108-0126			human	Norway	2008	NIPH Oslo	2008-02-26	162	246	301	393	523	1	4	3	4	3	134
606	1108-0177			dog	Norway	2008	NIPH Oslo	2008-02-26	161	301	319	357	523	1	10	4	1	3	90
608	1108-0228			human	Norway	2008	NIPH Oslo	2008-02-26	162	300	325	356	524.00	1	10	5	1	3	137

Figure 3

Interactive phylogenetic tree on MLVA-NET



criteria are combined, (iii) the order of displayed results, and (iv) the category (complete, incomplete or all) of isolates' profiles that are searched. The buttons on the right panel allow selection of fields that will be displayed on the results page.

It is, for example, possible to search for all isolates from Norway that have allele number 1 for marker STTR9 (Figure 2). From this selection of isolates users can access analysis tools (diversity indices, phylogenetic trees, data export).

The browse database mode gives access to all entries and allows the user to retrieve information for selected fields of interest (e.g. the raw data can be hidden by un-checking the corresponding columns of the table).

Batch functions are available for comparison of large numbers of isolates at once. In the profiles interface, MLVA-NET can assign existing RTs to multiple query profiles from a spreadsheet, and assign allele numbers to raw fragment size data. A specified field can be chosen for ordering the query results. The user can choose to restrict queries to complete profiles (no missing locus information), incomplete profiles, or both.

On the isolates interface, a number of diversity indices can be calculated on the selected datasets and isolates. Unweighted Pair-Group Method with Arithmetic Averages (UPGMA) dendrograms and neighbour-joining phylogenetic tree functions are available to cluster isolates for efficient comparison in an epidemiological context. The resulting interactive graphs can be displayed with the user-defined isolate information attached (Figure 3). The tree can be exported as Newick format for analysis with other tree visualization tools.

Three layouts are possible: phenogram, circular, or radial; each layout includes a re-rooting option. Distances between profiles can be calculated using several evolutionary models such as the saltational (infinite alleles) model or by applying (on user-defined loci) the stepwise mutation model [9,10] which considers alleles with similar repeat numbers as being more likely to be closely related.

Finally, a curator interface allows curators to manage their datasets: insert new isolates one by one or in batch, change or create a new coding method, and change the status (public or private) of datasets. This provides a convenient way for collaborative networks to make datasets public at a chosen date (e.g. once the data have been published).

Conclusion

MLVA-NET, the Institut Pasteur's MultiLocus VNTR Analysis database and web interface system, should help considerably in establishing a common language on microbial strain typing based on MLVA data for large numbers of pathogens. The database structure was tailored to allow distinct access rights to separate datasets. In contrast to alternative MLVA databases, MLVA-NET incorporates raw size data, which extends the possibilities for comparisons across public datasets from distinct networks. Of note, sizing data may vary slightly across distinct experimental platforms, and it is therefore crucial for curators to ensure that size data are normalised before they are entered into the MLVA-NET database.

Our data export functions render it possible to compare MLVA-NET data with data stored in other systems. However, discussions are in progress with the administrators of other MLVA databases to improve harmonisation and avoid redundancy of datasets. The user-friendly design of MLVA-NET was inspired by mlstdbNet [11], a system for MLST databases that used with a large success at pubmlst.org and www.pasteur.fr/mlst. As this design clearly separates profiles on the one hand and isolates on the other hand, the requirement for a common language is ensured by the immediate availability of profiles, even though information on isolates can be kept private for security or confidentiality reasons.

Epidemiological surveillance networks and collaborative networks of microbiologists interested in population biology should benefit from MLVA-NET. It is hoped that this system will contribute to a standardisation of MLVA, allowing the exchange of knowledge on the geographic and temporal distribution of strain types for epidemiology and evolutionary purposes.

Acknowledgements

We acknowledge Sandrine Rousseau for help in database design and Bjorn-Arne Lindstedt for helpful discussions. Financial support was provided by Institut Pasteur and the Institut de Veille Sanitaire (Saint Maurice, France).

References

1. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13(Suppl 3):1-46.
2. van Belkum A, Scherer S, van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev. 1998;62(2):275-93.
3. Lindstedt BA. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis. 2005;26(13):2567-82.
4. Aanensen DM, Spratt BG (2005) The multilocus sequence typing network: mlst.net. Nucleic Acids Res 33: W728-733.
5. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol. 2006;44(12):4498-510.
6. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, et al. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. J Bacteriol. 2000;182(10):2928-36.
7. Lindstedt BA, Heir E, Gjernes E, Kapperud G. DNA fingerprinting of Salmonella enterica subsp. enterica serovar Typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci. J Clin Microbiol. 2003;41(4):1469-79.
8. Grissa I, Bouchon P, Pourcel C, Vergnaud G. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie. 2007;Jul 28 [Epub ahead of print].
9. Kimura M, Ohta T. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A. 1978;75(6):2868-72.
10. Dettman JR, Taylor JW. Mutation and evolution of microsatellite loci in Neurospora. Genetics. 2004;168(3):1231-48.
11. Jolley KA, Chan MS, Maiden MC. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics. 2004;5:86.

This article was published on 8 May 2008.

Citation style for this article: Guigon G, Cheval J, Cahuzac R, Brisse S. MLVA-NET – a standardised web database for bacterial genotyping and surveillance. Euro Surveill. 2008;13(19):pii=18863. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18863