BMC Bioinformatics

BioMed Central

Software

Open Access

mlstdbNet – distributed multi-locus sequence typing (MLST) databases

Keith A Jolley*[1], Man-Suen Chan[2] and Martin CJ Maiden[1]

Address: [1]The Peter Medawar Building for Pathogen Research and Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3SY, UK and [2]Department of Paediatrics, Institute for Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, UK

Email: Keith A Jolley* - keith.jolley@medawar.ox.ac.uk; Man-Suen Chan - man-suen.chan@paediatrics.ox.ac.uk;
Martin CJ Maiden - martin.maiden@zoo.ox.ac.uk

* Corresponding author

Published: 01 July 2004

BMC Bioinformatics 2004, **5**:86 doi:10.1186/1471-2105-5-86

Received: 26 May 2004
Accepted: 01 July 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/86

Abstract

Background: Multi-locus sequence typing (MLST) is a method of typing that facilitates the discrimination of microbial isolates by comparing the sequences of housekeeping gene fragments. The mlstdbNet software enables the implementation of distributed web-accessible MLST databases that can be linked widely over the Internet.

Results: The software enables multiple isolate databases to query a single profiles database that contains allelic profile and sequence definitions. This separation enables isolate databases to be established by individual laboratories, each customised to the needs of the particular project and with appropriate access restrictions, while maintaining the benefits of a single definitive source of profile and sequence information. Databases are described by an XML file that is parsed by a Perl CGI script. The software offers a large number of ways to query the databases and to further break down and export the results generated. Additional features can be enabled by installing third-party (freely available) tools.

Conclusion: Development of a distributed structure for MLST databases offers scalability and flexibility, allowing participating centres to maintain ownership of their own data, without introducing duplication and data integrity issues.

Background

Multi-Locus Sequence Typing (MLST) is a method for characterising microbial isolates by means of sequencing internal fragments of housekeeping genes [1]. It was designed primarily for global epidemiology and surveillance [2] and has the advantages that data are highly reproducible and can be shared over the Internet without the need for exchanging live cultures. Fragments of approximately 500 bp length of usually between six and eight loci are sequenced, with each unique fragment sequence assigned an allele marker. Each allelic combination, or profile, is then assigned a sequence type (ST) number.

Following the introduction of MLST, database software was developed [3] that worked well for the small datasets initially produced. This was used to run the MLST websites that act as central repositories for sequence and profile definitions, as well as providing information on submitted isolates. It became apparent, however, that the original monolithic design had problems of scalability and data redundancy, while also requiring all isolate data to be submitted to a central database – which is often inappropriate in the public health setting. To overcome these issues, a new distributed database structure was required that enabled alleles and profiles to be defined centrally while allowing individual laboratories to host and control their own isolate databases. Here we describe mlstdbNet, a package that implements a network of web-accessible MLST databases that can be linked widely over the Internet. The central profiles database can be queried directly or via network protocols by client databases or other software.

Implementation

The premise behind the design of mlstdbNet was that isolate-specific information should be separated from allelic profiles and nucleotide sequences that may be shared by multiple isolates. By storing the profile and sequence data in its own database, any number of isolate databases can be constructed, each of which can interact with this profiles database but whose structure is not constrained by it. Consequently isolate databases can be set up for individual projects or populations of bacteria, or by individual organisations, with access controls set and fields included appropriately, while maintaining the benefits of having a single definitive source for sequence type and allele definitions.

The mlstdbNet package uses parts of an early implementation of MLST database software [3] and runs on Linux systems using the PostgreSQL database and Apache web server. The core software is written in Perl as a single, mod_perl compatible, CGI web script and requires, at a minimum, the CGI, DBI and XML::Parser::PerlSAX (part of libxml-perl) Perl modules. The functionality of the software can be enhanced by installing other modules and third party packages, such as EMBOSS [4] and Bioperl [5], which are used for sequence alignments, generating allele files in multiple formats and interacting with the PubMed database. Editing a single configuration file can enable the functionality offered by these external programs, but they are not required for the basic operation. Databases are specified in XML files that can be generated using 'dbConfig' [6]. The dbConfig program is written in Java (J2RE version 1.4 or later required) and offers an easy to use graphical interface to aid database design. As well as generating the database description XML file that is parsed by the web scripts, dbConfig also generates a SQL file that can be used to create the database. In addition, dbConfig will also create the configuration files for WDBI, an interface that allows the database to be curated over the web (written by Jeff Rowe, available from the mlstdbNet website [7]).

Remote connections from isolate to profile databases can be configured simply by adding the profile database host and port number to the XML description file, and configuring the profiles host to allow such connections. The software also allows the databases to be run on a separate machine from the web server, for improved speed and scalability. The HTML produced by the script uses cascading style sheets (CSS) to enable the look-and-feel to be modified easily and to minimise the size of the generated pages for fast response times (see figures 1, 2, 3, 4 for screenshots).

In addition to the script that dynamically generates the web pages, two other scripts can be run on a nightly basis. The first of these enables searching by reference by checking any database on the system for reference fields, and if it finds that a PubMed id entered in that field has not been seen before, it will download the citation from PubMed and store it in a local database. This local reference database is then used to create a searchable list of cited papers that can be selected to display information on isolates described in the paper. The other script generates an HTML page of summary statistics for the whole database.

Results and Discussion

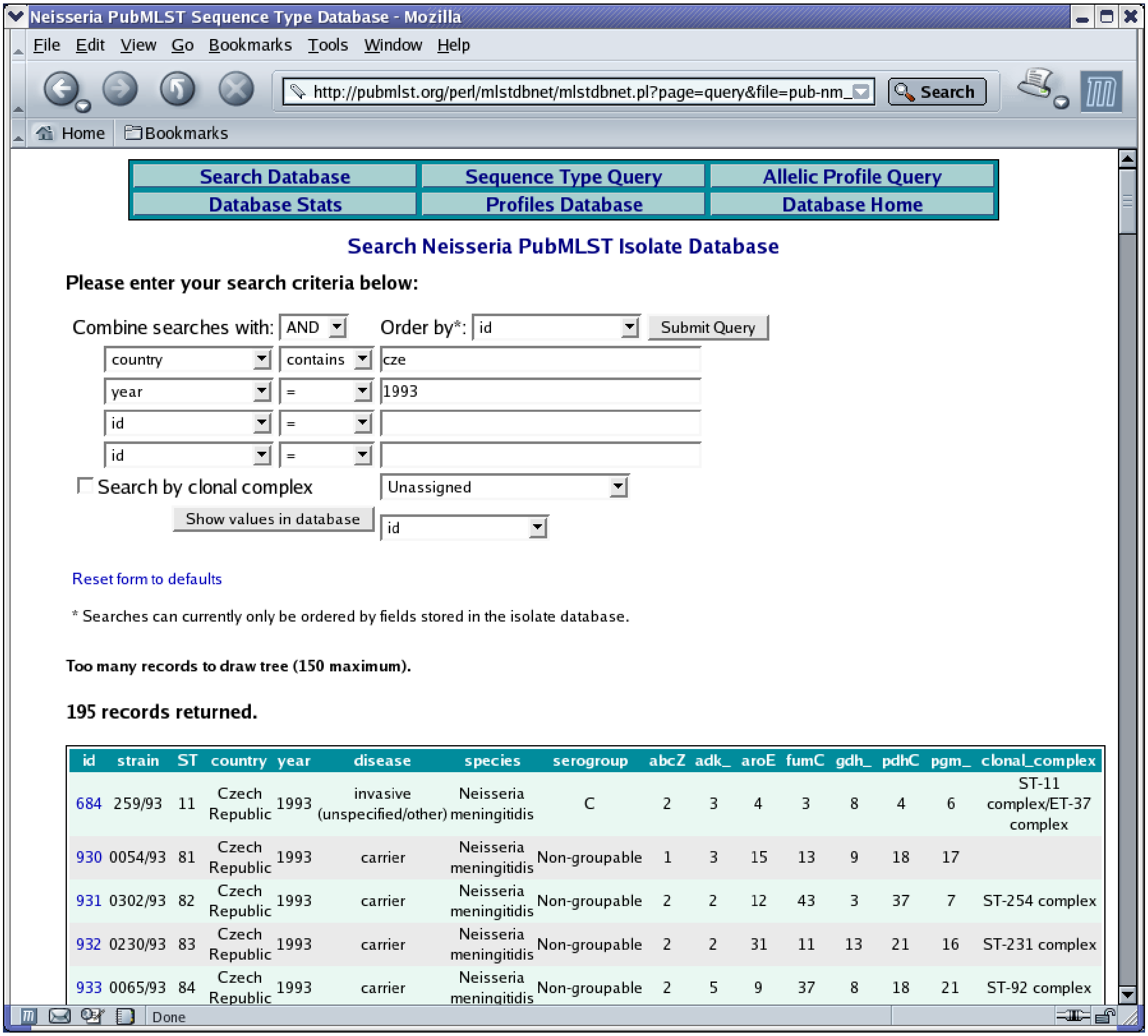
Distributed databases using this software

Examples of public isolate databases using mlstdbNet are the 'PubMLST' databases for *Neisseria* [1], *Campylobacter* [8], *Helicobacter pylori* and *Bacillus cereus* found on the PubMLST website [9]. These databases encourage submissions and describe the reported diversity of the organisms, but do not necessarily represent their natural populations. At least twelve other, mostly private, project- or organisation-specific isolate databases for these organisms have been established recently, all of which are clients of the central profile databases.

Queries to mlstdbNet isolate databases can also pull in data from other compatible sources, such as antigen databases (e.g. the *Neisseria* PorA variable region database [10]) and PubMed (Figure 5). Editing a single line in the XML configuration for the database can make connections to these external sources available. By including a PubMed id in a reference field of the database, the software can return information for isolates grouped by the publication they are described in, following a query by cited reference or author.

Novel features

Apart from the distributed structure, other novel features include the graphical breakdown of datasets, including displays of allele frequency and polymorphic sites. Following a search, datasets can be exported by E-mail with a choice of included fields and sequences from isolates can be concatenated for use in external packages. An important choice available is the ability to set up an isolate database to store either ST or allelic profile information. The latter choice enables the database to be used with partial



Search Neisseria PubMLST Isolate Database

Please enter your search criteria below:

Combine searches with: AND Order by*: id

country contains cze

year = 1993

id =

id =

□ Search by clonal complex Unassigned

id

* Searches can currently only be ordered by fields stored in the isolate database.

Too many records to draw tree (150 maximum).

195 records returned.

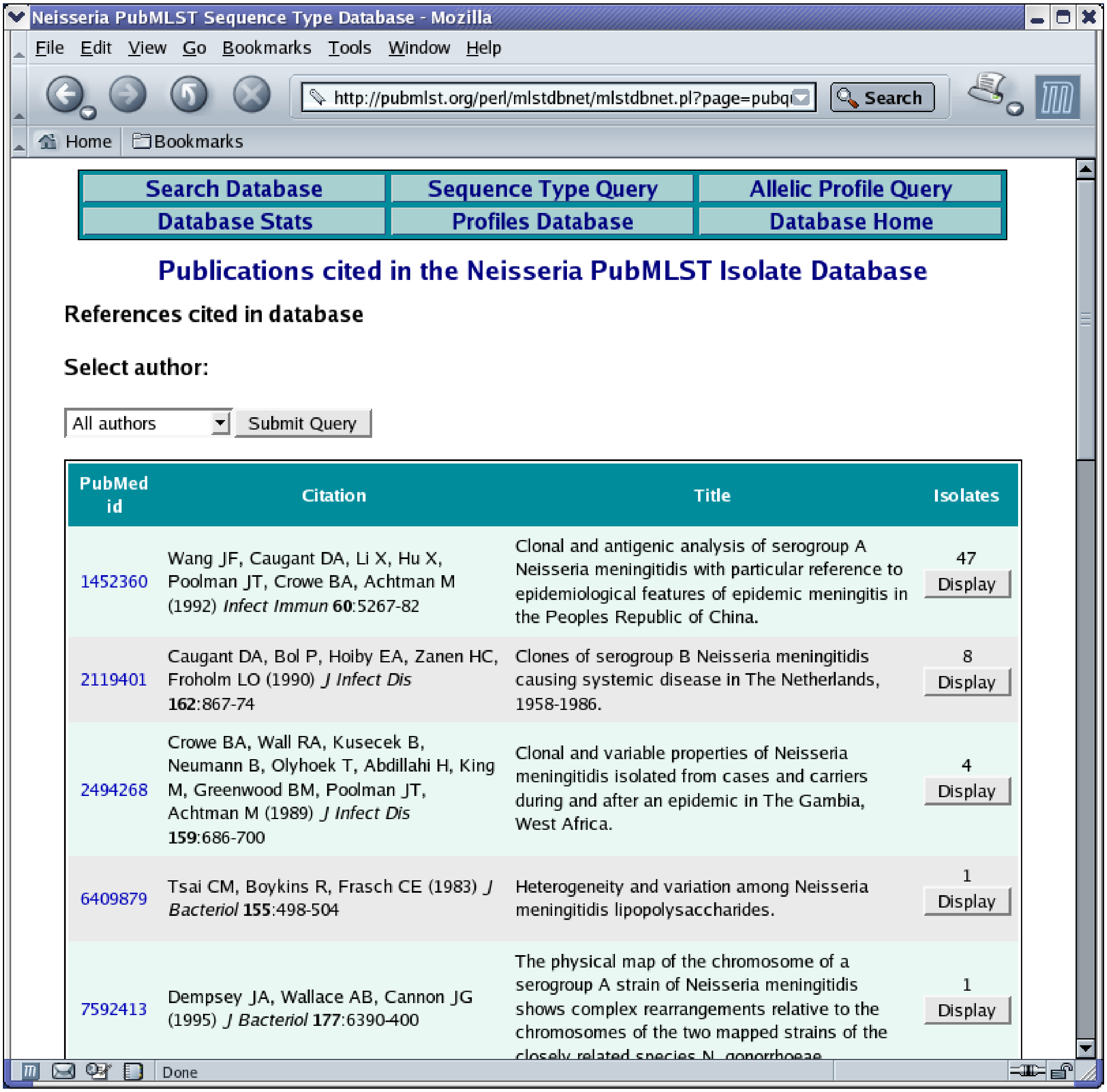
id	strain	ST	country	year	disease	species	serogroup	abcZ	adk_	aroE	fumC	gdh_	pdhC	pgm_	clonal_complex
684	259/93	11	Czech Republic	1993	invasive (unspecified/other)	Neisseria meningitidis	C	2	3	4	3	8	4	6	ST-11 complex/ET-37 complex
930	0054/93	81	Czech Republic	1993	carrier	Neisseria meningitidis	Non-groupable	1	3	15	13	9	18	17	
931	0302/93	82	Czech Republic	1993	carrier	Neisseria meningitidis	Non-groupable	2	2	12	43	3	37	7	ST-254 complex
932	0230/93	83	Czech Republic	1993	carrier	Neisseria meningitidis	Non-groupable	2	2	31	11	13	21	16	ST-231 complex
933	0065/93	84	Czech Republic	1993	carrier	Neisseria meningitidis	Non-groupable	2	5	9	37	8	18	21	ST-92 complex

Figure 1
Screenshot: Isolate query. Isolate databases can be queried by searching against any field or combination of fields.

profiles so that data can be entered as soon as sequencing results are obtained, with ST and clonal complex information retrieved once the profile has been completed. Isolate information can be retrieved by exact or partial matches to any field including those stored in the profiles database. Further, the profiles database can be queried in many ways, including batch methods for profiles and sequences.

Conclusions

This software represents a number of important enhancements over previous systems for storing and searching MLST data. Aside from the benefits to scalability offered by the distributed structure, it enables organisations to maintain ownership and control of their own data while still benefiting from centralised assignments of allele sequences and profiles, ensuring data integrity and consistency. In many cases, data confidentiality is required for research or legislative reasons. The central


Neisseria PubMLST Sequence Type Database - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://pubmlst.org/perl/mlstdbnet/mlstdbnet.pl?page=pubq
Search
Home
Bookmarks
Search Database
Sequence Type Query
Allelic Profile Query
Database Stats
Profiles Database
Database Home
Done

Publications cited in the Neisseria PubMLST Isolate Database

References cited in database

Select author:

All authors ▾ Submit Query

PubMed id	Citation	Title	Isolates
1452360	Wang JF, Caugant DA, Li X, Hu X, Poolman JT, Crowe BA, Achtman M (1992) *Infect Immun* **60**:5267-82	Clonal and antigenic analysis of serogroup A Neisseria meningitidis with particular reference to epidemiological features of epidemic meningitis in the Peoples Republic of China.	47 Display
2119401	Caugant DA, Bol P, Hoiby EA, Zanen HC, Froholm LO (1990) *J Infect Dis* **162**:867-74	Clones of serogroup B Neisseria meningitidis causing systemic disease in The Netherlands, 1958-1986.	8 Display
2494268	Crowe BA, Wall RA, Kusecek B, Neumann B, Olyhoek T, Abdillahi H, King M, Greenwood BM, Poolman JT, Achtman M (1989) *J Infect Dis* **159**:686-700	Clonal and variable properties of Neisseria meningitidis isolated from cases and carriers during and after an epidemic in The Gambia, West Africa.	4 Display
6409879	Tsai CM, Boykins R, Frasch CE (1983) *J Bacteriol* **155**:498-504	Heterogeneity and variation among Neisseria meningitidis lipopolysaccharides.	1 Display
7592413	Dempsey JA, Wallace AB, Cannon JG (1995) *J Bacteriol* **177**:6390-400	The physical map of the chromosome of a serogroup A strain of Neisseria meningitidis shows complex rearrangements relative to the chromosomes of the two mapped strains of the closely related species N. gonorrhoeae	1 Display

Figure 2
Screenshot: Isolate search by cited publication. A list of publications cited within isolate records can be displayed and a narrower search performed by searching for individual authors. All isolates described in a particular paper can be displayed by selecting the 'Display' button from this list.

profiles database can be readily mirrored to other sites as it contains no confidential data.

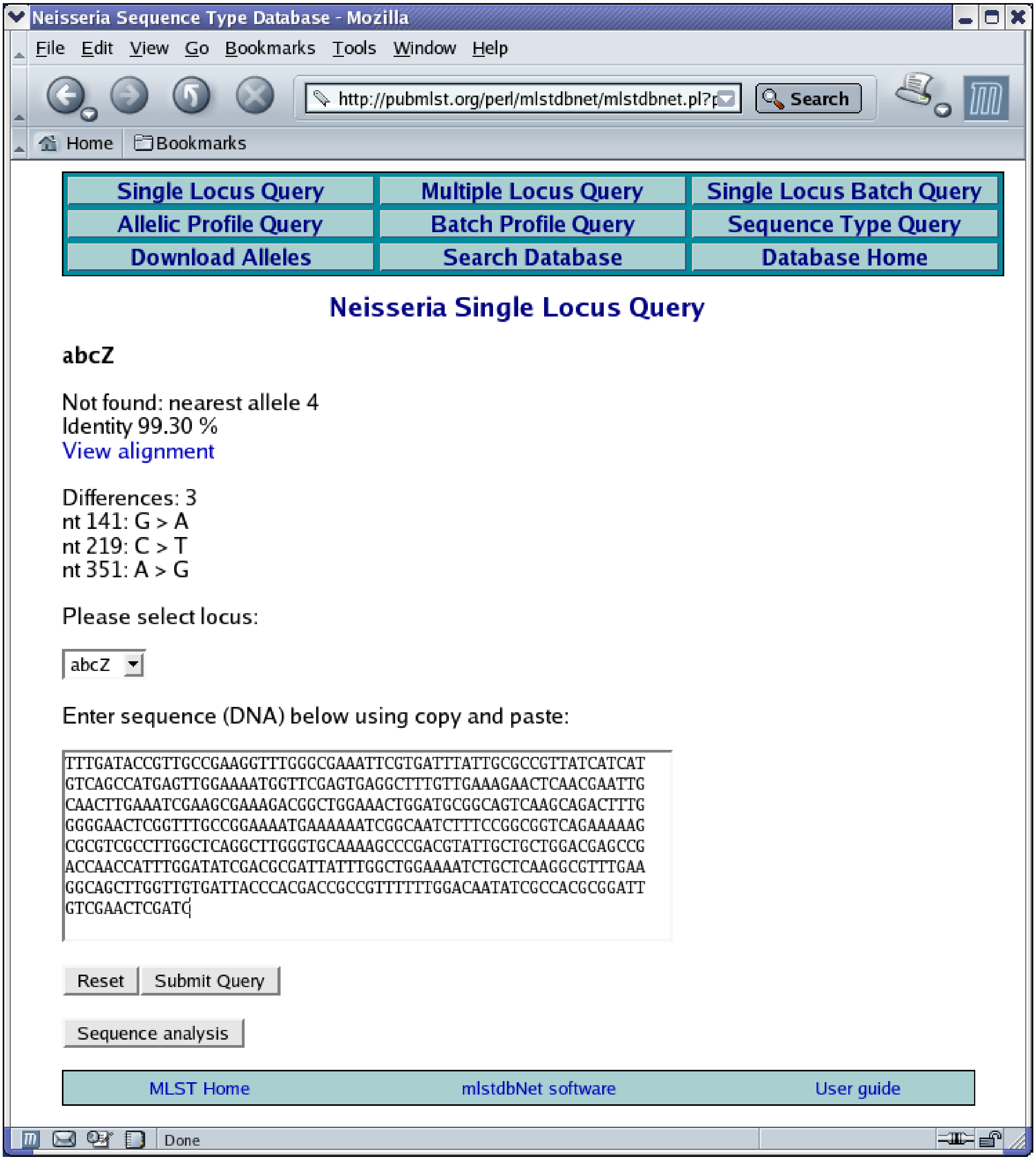
Availability and requirements

Project name: mlstdbNet

Project home page: http://pubmlst.org/software/database/mlstdbnet/

Operating system: Linux

Programming language: Perl



Neisseria Single Locus Query

abcZ

Not found: nearest allele 4
Identity 99.30 %
View alignment

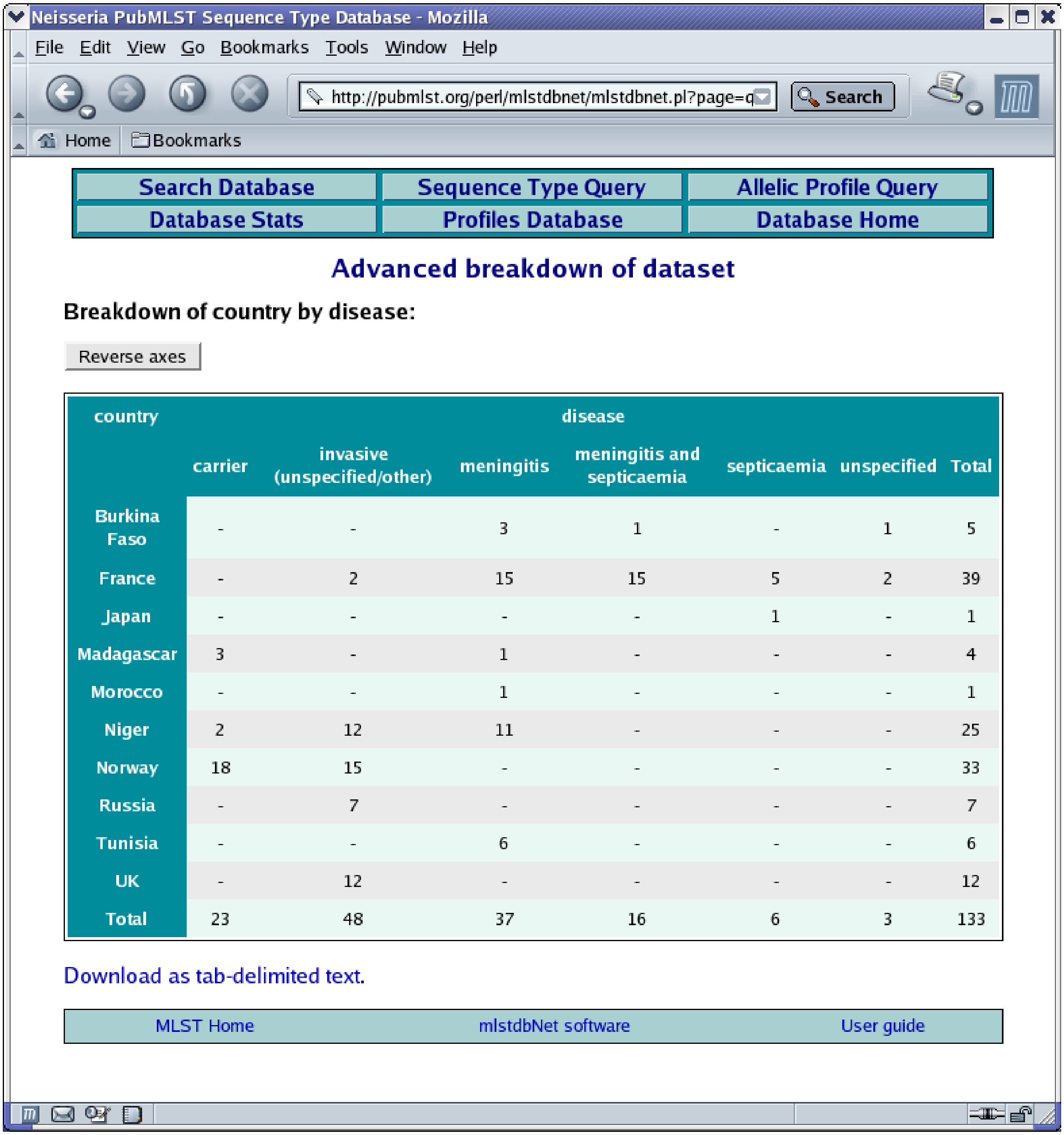
Differences: 3
nt 141: G > A
nt 219: C > T
nt 351: A > G

Please select locus: abcZ

Enter sequence (DNA) below using copy and paste:

```
TTTGATACCGTTGCCGAAGGTTTGGGCGAAATTCGTGATTTATTGCGCCGTTATCATCAT
GTCAGCCATGAGTTGGAAAATGGTTCGAGTGAGGCTTTGTTGAAAGAACTCAACGAATTG
CAACTTGAAATCGAAGCGAAAGACGGCTGGAAACTGGATGCGGCAGTCAAGCAGACTTTG
GGGGAACTCGGTTTGCCGGAAAATGAAAAAATCGGCAATCTTTCCGGCGGTCAGAAAAAG
CGCGTCGCCTTGGCTCAGGCTTGGGTGCAAAAGCCCGACGTATTGCTGCTGGACGAGCCG
ACCAACCATTTGGATATCGACGCGATTATTTGGCTGGAAAATCTGCTCAAGGCGTTTGAA
GGCAGCTTGGTTGTGATTACCCACGACCGCCGTTTTTTGGACAATATCGCCACGCGGATT
GTCGAACTCGATC
```

Figure 3
Screenshot: Allele query. Allele comparison will identify known alleles or determine the nearest allele with the nucleotide differences shown. An alignment of the query sequence to the nearest allele will be offered if EMBOSS is installed.



http://pubmlst.org/perl/mlstdbnet/mlstdbnet.pl?page=q

Advanced breakdown of dataset

Breakdown of country by disease:

country	disease						
	carrier	invasive (unspecified/other)	meningitis	meningitis and septicaemia	septicaemia	unspecified	Total
Burkina Faso	-	-	3	1	-	1	5
France	-	2	15	15	5	2	39
Japan	-	-	-	-	1	-	1
Madagascar	3	-	1	-	-	-	4
Morocco	-	-	1	-	-	-	1
Niger	2	12	11	-	-	-	25
Norway	18	15	-	-	-	-	33
Russia	-	7	-	-	-	-	7
Tunisia	-	-	6	-	-	-	6
UK	-	12	-	-	-	-	12
Total	23	48	37	16	6	3	133

Download as tab-delimited text.

Figure 4
Screenshot: Advanced breakdown. Following a database query, the displayed dataset can be analysed further, including breaking down one field against another and displaying frequencies of unique combinations of selected fields.

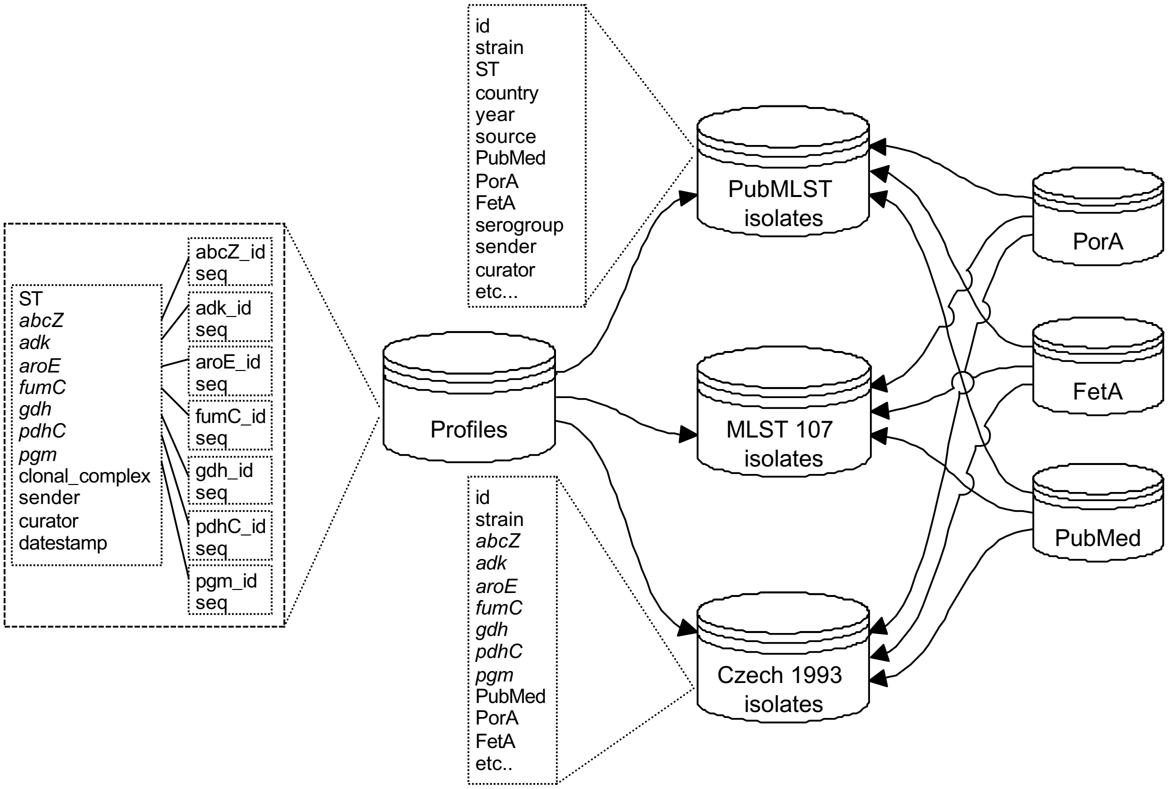


Figure 5

The distributed structure of the *Neisseria* MLST and related databases. The profiles database contains all the allele sequences and allelic profiles (sequence types) and can be queried via network connections or directly through the web. The PubMLST isolate database encourages general submissions and represents the known diversity of strains. The MLST 107 and Czech 1993 isolate databases are also available through the *Neisseria* MLST website and contain reference [1] and project [11, 12] sets of isolate data (see site for further details). These isolate databases also retrieve information from non-MLST databases including those for PorA and FetA antigen typing and PubMed. Other private isolate databases (not shown) also make use of the profiles database.

Other requirements: Apache; PostgreSQL; CGI, DBI and XML::Parser::perlSAX Perl modules

License: GNU GPL

Any restrictions to use by non-academics: none

Authors' contributions

KAJ carried out the main programming work. MSC developed the first generation MLST database software, parts of which have been used in this implementation. MM conceived the software development and participated in its design.

Additional material

Additional File 1

Distribution archive of software (version 1.1.5). This file contains the software.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-86-S1.gz]

Acknowledgements

The European Union (QLK2-CT-2001-01436) and the Wellcome Trust funded development.

References

1. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95:**3140-3145.
2. Urwin R, Maiden MC: **Multi-locus sequence typing: a tool for global epidemiology.** *Trends Microbiol* 2003, **11:**479-487.
3. Chan MS, Maiden MC, Spratt BG: **Database-driven multi locus sequence typing (MLST) of bacterial pathogens.** *Bioinformatics* 2001, **17:**1077-1083.
4. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
5. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
6. **dbConfig Home Page** [http://pubmlst.org/software/database/dbconfig/]
7. **mlstdbNet Home Page** [http://pubmlst.org/software/database/mlstdbnet/]
8. Dingle KE, Colles FM, Wareing DRA, Ure R, Fox AJ, Bolton FJ, Bootsma HJ, Willems RJL, Urwin R, Maiden MCJ: **Multilocus sequence typing system for Campylobacter jejuni.** *J Clin Microbiol* 2001, **39:**14-23.
9. Jolley KA: **PubMLST website - Publicly-accessible MLST databases and software.** [http://pubmlst.org].
10. Russell JE, Jolley KA, Feavers IM, Maiden MC, Suker JS: **PorA variable regions of Neisseria meningitidis.** *Emerg Infect Dis* 2004, **10:**674-678.
11. Jolley KA, Kalmusova J, Feil EJ, Gupta S, Musilek M, Kriz P, Maiden MC: **Carried Meningococci in the Czech Republic: a Diverse Recombining Population.** *J Clin Microbiol* 2002, **40:**3549-3550.
12. Jolley KA, Kalmusova J, Feil EJ, Gupta S, Musilek M, Kriz P, Maiden MC: **Carried meningococci in the Czech Republic: a diverse recombining population.** *J Clin Microbiol* 2000, **38:**4492-4498.