BMC Bioinformatics

SOFTWARE Open Access

BIGSdb: Scalable analysis of bacterial genome variation at the population level

Keith A Jolley*, Martin CJ Maiden

Abstract

Background: The opportunities for bacterial population genomics that are being realised by the application of parallel nucleotide sequencing require novel bioinformatics platforms. These must be capable of the storage, retrieval, and analysis of linked phenotypic and genotypic information in an accessible, scalable and computationally efficient manner.

Results: The Bacterial Isolate Genome Sequence Database (BIGSDB) is a scalable, open source, web-accessible database system that meets these needs, enabling phenotype and sequence data, which can range from a single sequence read to whole genome data, to be efficiently linked for a limitless number of bacterial specimens. The system builds on the widely used mlstdbNet software, developed for the storage and distribution of multilocus sequence typing (MLST) data, and incorporates the capacity to define and identify any number of loci and genetic variants at those loci within the stored nucleotide sequences. These loci can be further organised into 'schemes' for isolate characterisation or for evolutionary or functional analyses. Isolates and loci can be indexed by multiple names and any number of alternative schemes can be accommodated, enabling cross-referencing of different studies and approaches. LIMS functionality of the software enables linkage to and organisation of laboratory samples. The data are easily linked to external databases and fine-grained authentication of access permits multiple users to participate in community annotation by setting up or contributing to different schemes within the database. Some of the applications of BIGSDB are illustrated with the genera *Neisseria* and *Streptococcus*.
The BIGSDB source code and documentation are available at http://pubmlst.org/software/database/bigsdb/.

Conclusions: Genomic data can be used to characterise bacterial isolates in many different ways but it can also be efficiently exploited for evolutionary or functional studies. BIGSDB represents a freely available resource that will assist the broader community in the elucidation of the structure and function of bacteria by means of a population genomics approach.

Background

Parallel sequencing technology, sometimes referred to as 'next-generation' sequencing, makes possible the rapid determination of large numbers of complete bacterial genome sequences at a low cost. This is leading increasingly to its use in population studies including epidemiological investigations, a trend that will accelerate with the continual introduction of technical advances such as single molecule sequencing [1]. The possibility of comparing any number of gene targets among multiple, disparate, isolates, allows the assembled data resource to be used to address a wide range of research questions concerning bacterial evolution, ecology and pathogenicity. Harnessing this resource will enable a diversity of information to be efficiently exploited in functional studies.

Investigations of bacterial population biology and epidemiology have utilised whole genome data but, until now, its application has been limited to largely clonal organisms or closely related isolates [2-8]. In order to facilitate wider bacterial population genomic research, there is a need to link whole genome data to the population sample data, including detailed provenance, clinical information and phenotype as appropriate, allowing integrated studies irrespective of the diversity of the isolates. One method that has been widely used to achieve this for both population studies and especially epidemiological investigation is multilocus sequence typing (MLST)

* Correspondence: keith.jolley@zoo.ox.ac.uk
Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK

[9,10]. MLST indexes the sequences of representative housekeeping gene fragments, approximately 500 bp in length usually from seven loci, with each unique allele assigned an arbitrary integer identifier. Unique combinations of the alleles at each locus, allelic profiles, are identified by a sequence type (ST) number with definitions stored in authoritative online databases overseen by a curator for each species or group of species. Housekeeping gene profiles may be combined with sequence data from more rapidly-evolving genes, such as those encoding surface antigens, where higher resolution is required to address questions concerning antigenic variation or antibiotic resistance [11,12]. The current generation of sequence typing databases [13-15] have been highly successful at linking isolate provenance to sequence data, contributing significantly to the widespread adoption of MLST and antigen sequence typing. Indexing population genomic data provided by parallel sequencing technologies, however, provides challenges requiring the development of new methodologies and informatics solutions.

We have extended the proven concept of MLST to genome scale data, where particular combinations of loci can be analysed depending on the question being addressed. As with conventional MLST, each unique sequence at each locus can be assigned an allele number, allowing the range of analyses developed for MLST, or for other techniques utilizing numerical profiles such as multiple locus VNTR analysis (MLVA) [16], to be directly applied to whole genomes, while still providing full access to the underlying sequence data when required. This approach provides the isolate- or specimen-centric view required for epidemiology, ecology and population analysis, with the isolate provenance and phenotype linked to any quantity of sequence data, from individual dye-terminator sequencing reads, through partial genomes consisting of multiple unassembled contigs to complete genome sequences. It also enables routine typing applications for epidemiological studies to use the same methodology as required for population analysis, which has been one of the factors in the success and widespread adoption of MLST. While analogous to MLST, the application of the concept goes far beyond typing, allowing detailed investigation of population diversity in bacterial systems. Here we describe a software platform for population genomics that has been designed and developed to exploit this concept.

Implementation

BIGSDB is written in Perl for installation on UNIX/Linux systems. It utilizes the PostgreSQL database and Apache web server software running under mod_perl [17] to avoid the performance penalty of Perl interpreter start-up times. Sequence handling routines are provided by the BIOPERL library [18] and EMBOSS suite of programs [19]. Client-side Javascript makes use of the JQUERY library [20]. Built-in authentication is based on Perl/Javascript MD5 secure user authentication [21], with Javascript MD5 code written by Paul Johnston [22]. Sequence homology matching uses BLAST [23] with a default word size of 15 and identity of 70% for DNA sequences (values are configurable by the user).

Configuration

Global configuration settings are stored in a text file. Settings provide the locations of helper applications like BLAST and EMBOSS and the names of the preference and authentication databases. Logging utilizes the Log::Log4perl module, enabling fine-grained control of error and status logging by modifying a configuration file, setting the overall log level for discrete components of the system.

Individual databases are configured with a XML file that describes any isolate provenance fields including default display properties (that can be over-ridden by user preferences), sample table for use as a LIMS database, connection information, web paths, authentication type in use and enabled plug-ins.

Sequence definition databases

As well as storing isolate information, BIGSDB is able to host separate sequence definition databases so that new allele sequences can be defined for any locus and made available on the Internet. Users are able to paste in and query sequences against all known alleles from a particular locus or against all loci. The nearest matches are displayed along with nucleotide differences and the start and end positions within the sequence are identified, allowing exact polymorphisms to be identified and checked rapidly and efficiently without the user having to trim their sequence to match the defined allele.

Sequence definition databases can also define schemes so can, for instance, make MLST profile data available.

User customization

Individual users are able to customize the query interface. The fields and loci that are returned within main results tables following a query can be selected so that only the results of interest are shown without leading to the table width going beyond the confines of the page. Drop-down list boxes that filter search results based on particular provenance criteria or publications can be added to the standard query interface. All settings are stored in a separate preference database linked to a unique identifier stored as a HTTP cookie so that these are remembered between sessions.

Data export

Isolate data, along with all defined allele identifiers, can be exported for a subset of the data returned from a

query or for the whole database. Isolate data are exported in tab-delimited text format, suitable for importing into spreadsheets or easy parsing by automated scripts. Concatenated sequences can be exported for isolates in FASTA or extended-FASTA format suitable for use in third-party phylogenetic packages.

Plug-in architecture

The software employs a plug-in architecture, allowing additional features and analysis packages to be added by third-parties without modification of the core code. Various attributes are defined for each plug-in which specify whether it works with isolate or sequence definition databases, where in the interface it should be available, e.g. main index page and/or following a query to utilise a returned dataset, and a feature category, e.g. export, breakdown, or analysis, allowing tools to be grouped by function.

Most software functionality that can be considered optional to the core requirements of the database form part of plug-ins, easing maintenance of the main code and allowing installations to be tailored to individual requirements. Since plug-ins are self-contained units, they can be distributed under different licenses to the main software package.

Authentication and access control

There are three types of user in BIGSDB: i) 'users' can view data but never modify it; ii) 'curators' can add and modify data with specific permissions enabling particular roles to be defined and controlled; and iii) 'admins' have full control over the database structure, data and curator permissions.

The software can be configured with either built-in authentication or controlled by apache. Built-in authentication uses client-side Javascript to hash passwords together with session identifiers so that passwords are not transmitted in clear text over the network. Controlling authentication within the program also allows users to change their own passwords from the web interface. Using apache authentication allows any supported external authentication scheme to be employed.

Isolate databases can be configured to be public, where either everybody on a public website or all authenticated users can view all records. Alternatively, controls can be configured allowing read and write access of individual isolate records to specific users or user groups. By default, new records are viewable by everybody and writable by the curator who adds them, but access can be controlled easily within the curator interface in either a single isolate manner or batch mode. Curation access to individual loci in sequence definition databases can be set so that curators are allowed to define allele sequences for certain loci only. This allows a single definitions database to be serviced by a community of curators, expert in particular areas of the organism's biology.

Results

Design philosophy

The Bacterial Isolate Genome Sequence Database (BIGSDB) is an informatics system that can hold provenance and phenotype information on an unlimited number of isolates, along with nucleotide sequence data (Figure 1). These sequence data can be of any size scale ranging from individual dye-terminator sequencing reads through partial assemblies generated from parallel sequencing technologies to complete assembled genomes. These are stored within a 'sequence bin' linked to the isolate records. A reference sequence for each locus can be defined, or alternatively they can be linked to external databases that hold allele sequence definitions. This enables the positions of loci within individual sequences stored in the bin to be determined automatically using any algorithm, at present BLAST[23], and the sequences extracted along with flanking regions if required. If an external sequence definition database, containing allele sequences and their identifiers, has been defined it can automatically be queried to determine the allele number for each locus (Figure 2). BIGSDB facilitates the construction of these definition databases. Provided the locus is fully encompassed within a single contig, the length of individual sequences is unimportant for extracting allele data, avoiding many of the problems associated with assembly of short-read data.

Genetic loci can be grouped into schemes with membership defined by any criteria and with each unique combination of alleles associated with a primary key (a field that uniquely defines this combination) and any number of other fields. One example of this is the standard seven locus MLST scheme where each allelic profile is defined by a ST number, the primary key in this case. Since the ST can also define membership of a clonal complex, an epidemiologically related grouping of STs, this can be included as an additional field in the scheme. The flexibility of BIGSDB enables loci to belong to any number of schemes, allowing multiple strain nomenclatures to be cross-referenced or for schemes based on specific aspects of the biology of the organism to be employed, such as particular biochemical pathways, surface components coded by antigen genes, antibiotic resistance or members of macromolecular complexes.

Allele assignment and locus tagging

As well as automatic allele assignment using BLAST, allele numbers can be assigned manually, enabling existing isolate datasets to be imported where these designations are already determined. Competing allele designations,

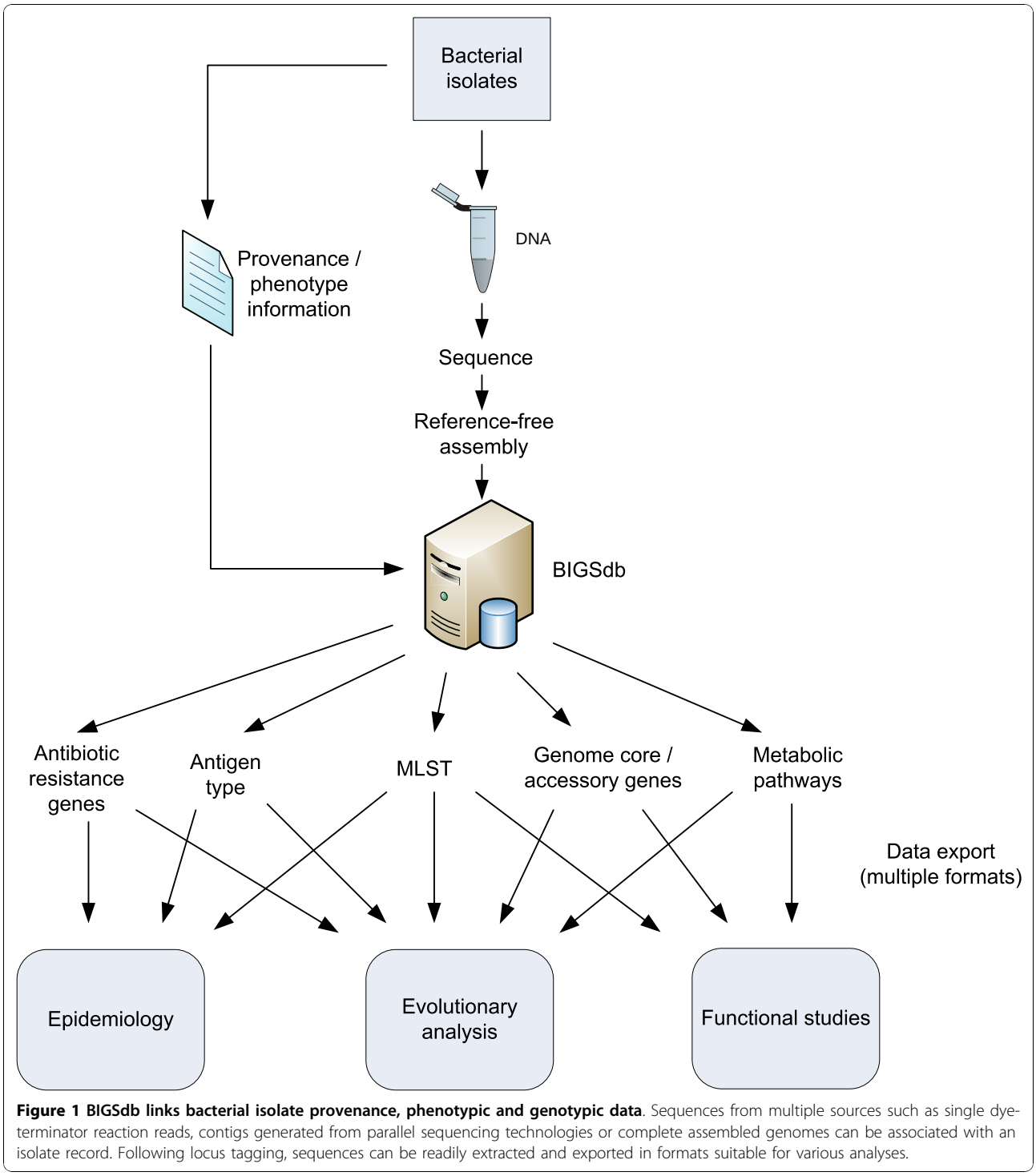

Figure 1 BIGSdb links bacterial isolate provenance, phenotypic and genotypic data. Sequences from multiple sources such as single dye-terminator reaction reads, contigs generated from parallel sequencing technologies or complete assembled genomes can be associated with an isolate record. Following locus tagging, sequences can be readily extracted and exported in formats suitable for various analyses.

identified by different users or with different analysis methods, can be handled with the first determined designation used for analysis, but the presence of conflicting data highlighted within the user interface. Designations can be promoted or demoted from the active or pending state by the curator; with a fine-grained permission system allowing specific curator roles (see access control in Methods). A full history of changes made to an isolate record is logged, so that it is possible to track which curator made a change and when.

Loci are usually defined by nucleotide sequences, but BIGSDB will also handle loci that are defined by the

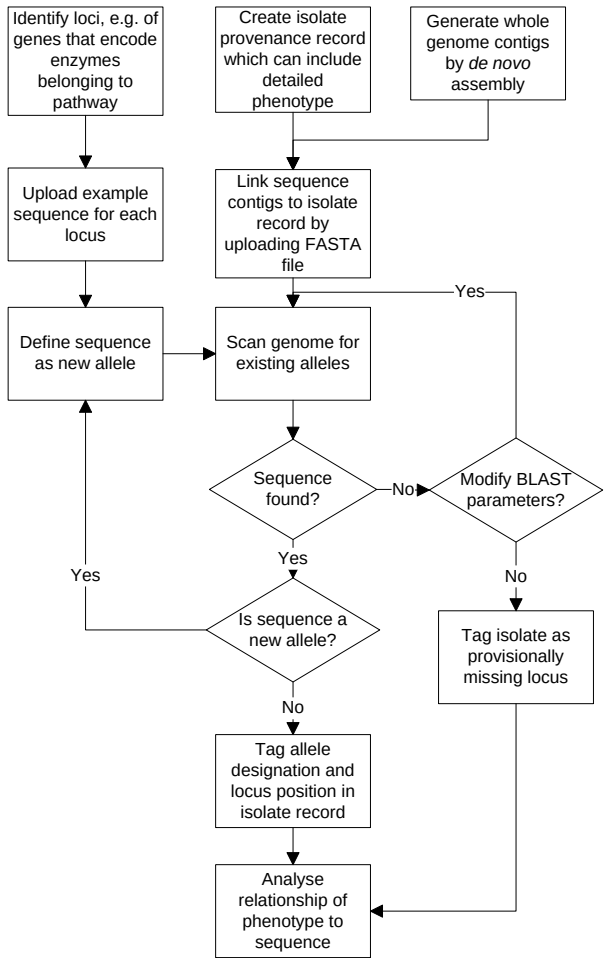

Figure 2 Iterative process of analysing loci. Since detailed phenotypic information can be included with the isolate record, the correlation between enzyme sequence diversity and phenotype can be examined using the integrated analysis tools.

translated sequence peptide, commonly used, for example, to define variable regions of antigens important in typing or vaccine development. Irrespective of the locus type, the sequence definition database can be queried using either nucleotide or peptide sequences, with the query type recognized automatically and the appropriate BLAST algorithm called. Alleles can be named using either simple integers, or by a text string, with the format naming constrained by a regular expression defined in the locus table if required. Coding sequence definitions from existing published annotations can be retrieved automatically by entering a Genbank or EMBL accession number, assisting the process of setting up new loci for analysis.

Isolate and locus aliases

A common issue with population datasets is that of alternative nomenclature of isolates, with many samples having multiple names having been stored in different laboratories or collections. BIGSDB allows isolates to have any number of names by storing aliases in a linked table. These aliases are treated in the same manner as the primary name, and will be found by searches against the 'isolate' field. In a similar manner, loci can also have aliases, all of which are accessible within the interface. Loci that are members of schemes defined within external databases are not constrained to the names used in those databases, ensuring that data organisation within BIGSDB is not impacted by external naming constraints.

Data export and analysis

Sequence and provenance data can be exported from the database in multiple formats. Where genome data are represented by single or a small number of large contigs, export in EMBL format with the locus information included provides a method of consistent feature annotation, allowing newly defined loci to be applied rapidly to any number of existing genomes. Other formats include extended multi-FASTA where data for each locus are grouped in aligned blocks, facilitating whole genome phylogenetic analysis with applications such as ClonalFrame [24].

Datasets can be further analysed by provenance or allele content using various breakdown tools that determine value frequencies or that breakdown one field against another, allowing analyses such as clonal complex against serogroup.

Laboratory Information Management System (LIMS)

BIGSDB can form the basis of a laboratory information management system (LIMS). An optional samples table can be defined containing any choice of fields which can, for example, describe sample type and freezer location. Each isolate record can then be associated with any number of samples which are displayed within the detailed isolate information pages, where the records can be updated by users with appropriate privileges.

Demonstration 1: PubMLST and reference gene-based analysis of partially assembled genome data

We have installed BIGSDB on PubMLST.org and converted the *Neisseria* MLST databases [25] to use the system in place of the previously used MLSTDBNET[13] software, the functionality of which is generically incorporated. The MLST databases for *Neisseria* are the largest of all such databases, containing provenance and genetic data for over 17,000 isolates and over 8,000 STs, providing a valid test of scalability and performance for population level data. The isolate database has been linked to existing antigen typing and antibiotic resistance gene databases [26], enabling automated allele assignment and sequence tagging of typing antigens and

genes encoding candidate vaccine proteins. We have further populated this database with publicly-available *Neisseria* species genome data and performed automated tagging and assignment of alleles for loci with existing definition databases. Some of these are represented by single contigs of approximately 2.2 Mbp, but there are also samples with unfinished genomes consisting of multiple smaller contigs. Finally, we have deposited contigs for isolate OX9932088, a ST-41/44 isolate collected during the UK meningococcal carriage study [27], generated from Illumina Solexa reads, to demonstrate the ease of analysis of such data using the gene reference approach [Neisseria PubMLST:14923]. Velvet assembly [28] of the Solexa runs yielded 295 contigs of >100 bp length, which were uploaded to the PubMLST BIGSDB database. Automated sequence tagging determined the full strain designation, B: P1.21,16: F1-5: ST-1415 (cc41/44), incorporating MLST and the PorA and FetA antigen types. Additionally, the peptide sequence for Factor H binding protein, a principal component of two meningococcal recombinant protein vaccines undergoing clinical trials, was identified as variant 19 [29]. Allele sequences for *penA* [30] and *rpoB* (definitions incorporated into PubMLST) were identified as *penA*-1 and *rpoB*-18 indicating the isolate has intermediate susceptibility to penicillin and high susceptibility to rifampicin respectively. Isolate records for which genome data are available can be readily extracted from the database by selecting 'whole genome' in the project filter of the query interface.

Demonstration 2: Relationships within the genus *Streptococcus*

Since BIGSDB can define multiple schemes for a dataset, it can be used to cross-reference typing methods. A database containing 43 published streptococci genomes was constructed [31] and had loci defined for all streptococcal MLST schemes (*S. agalactiae* [32], *S. oralis* [33], *S. pneumoniae* [34], *S. pyogenes* [35], *S. suis* [36], *S. uberis* [37] and *S. zooepidemicus* [38]). Sequences and ST definitions from these schemes were imported into a unified definitions database, and the genomes tagged with all loci found. In addition, unique alleles from the streptococcal MLSA database [39], whose loci were chosen to be present across the viridans streptococci, were imported and assigned allele numbers. Genome isolates were then tagged against these loci as well.

Further loci were defined based on a sample of the coding sequences extracted from the annotated *S. equi* genome [40]. The BIGSDB genome comparator tool was used to identify loci found in all 43 genomes using BLAST with a 70% identity cut-off and a word size of 8. Because the search used nucleotide sequences it would be expected to only find the more conserved classes of protein-coding genes [41]. Seventy-seven coding loci, consisting largely of genes whose products are involved in translation were identified. Sequences for the MLSA scheme and the 77 trans-genus loci were then extracted as two separate datasets from the database as aligned sequences in multi-FASTA format using the BIGSDB export functionality. ClonalFrame trees were generated from these sequence data (Figure 3). The trees from the MLSA loci and from the 77 loci identified without *a priori* knowledge produced highly similar species clustering. The only major differences between the two trees was that the branch points for the *S. equi*/*S. zooepidemicus* and the *S. uberis* branches were positioned nearer to the *S. pyogenes* cluster in the MLSA tree.

Discussion

Flexible storage of population-scale bacterial genome data

There is a dichotomy in the approach to data handling and analysis between the researchers that have been involved in generating complete annotated genome sequences and those engaged in large scale bacterial population studies. To date, the former have worked with relatively few isolates as exemplars of their species, while the latter have collected and analysed datasets that may include hundreds or thousands of isolates with less complete genome sampling. Many of the genomes hosted in online databases [42-44] have detailed and comprehensive annotation, but what these databases have in common is that for any particular species they contain very few genomes (Figure 4) and the analysis is geared specifically to the attributes of the sequence itself rather than of the isolate from which it was derived. For example, data cannot be analysed based on host clinical outcome, geographical location or any number of attributes that are important in evolutionary or epidemiological studies.

Conversely, population biologists and bacterial epidemiologists collect and analyse many more isolates, but until now, have sampled the genome by sequencing relatively few genes. Methods of sampling a genome with manageable units of data that are epidemiologically meaningful will continue to be essential for characterisation of isolates and disease management and with the decreasing costs of parallel sequencing it is likely, in the near future, to be more efficient and cost-effective to obtain MLST and antigen profiles by sequencing the whole genome rather than by a gene-by-gene approach using dideoxy chain termination methods. MLST and antigen sequencing have proved sufficient for routine typing, and they have been used successfully for association studies that link host, disease or sample site to

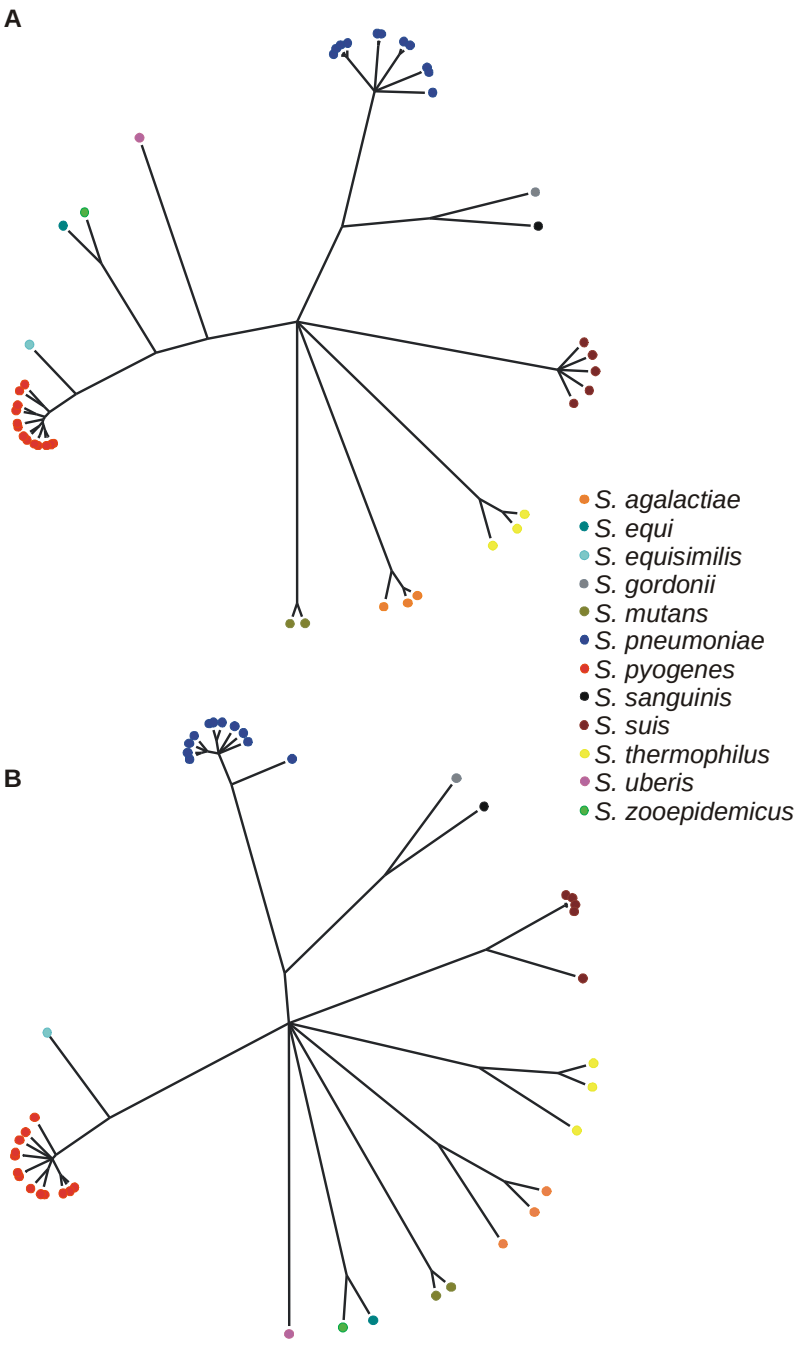

Figure 3 The Genome Comparator plugin can identify loci shared among genomes. ClonalFrame trees were generated from 43 Streptococcal genome sequences using A) seven MLSA gene fragment loci and B) 77 complete genes found to be present throughout the genus identified by BIGSDB. Aligned sequences were exported from the database and 50% consensus trees generated from six independent runs with 50 k iterations, 50 k burn-in iterations, and a thinning interval of 100.

genetic type [45-49]. The information available from the complete genome, however, will facilitate much higher resolution association analyses. The BIGSDB platform has been designed to store such population-scale bacterial genome datasets with no constraint on the number of samples or loci that can be analysed.

Reference gene-based analysis of genome data

As the number of new genomes from individual species or genera increases, a problem emerges as to how they can be compared. Existing tools [50-52] do not scale well to the challenge of handling multiple genomes at the scale that will be required in the near future.

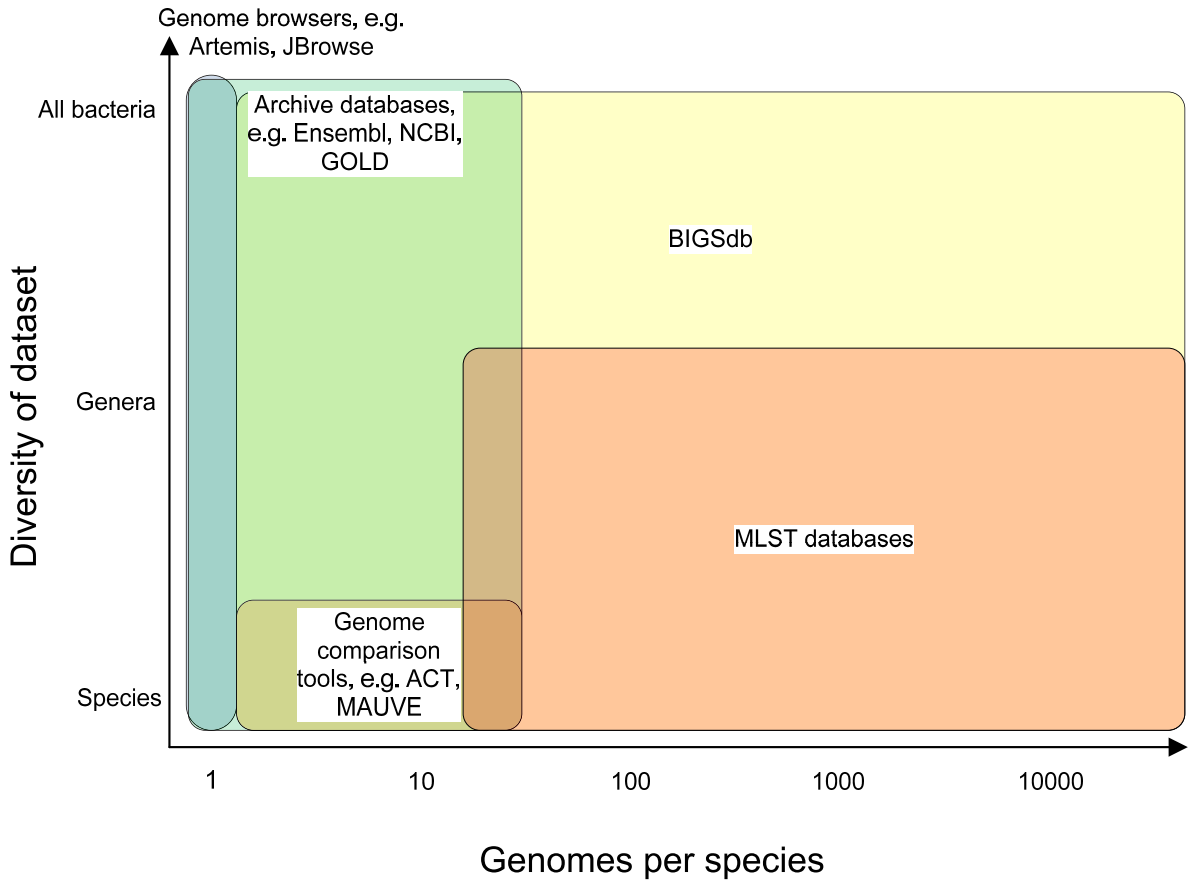

Figure 4 The BIGSdb database platform is highly scalable. The system can be used to analyse a few isolates up to many thousands, each with full genome data attached. This compares favourably with existing database and analysis resources.

When more than a handful of comparisons are to be made, one method is to use single nucleotide polymorphisms (SNPs) to generate phylogenetic trees, identifying informative SNP markers to differentiate particular nodes [4,53-56]. Ongoing SNP discovery resulting in continual tree generation makes this approach cumbersome for large ongoing studies, and the effects of recombination on population structure may make any resulting tree misleading. A recent study used short-read Solexa data to investigate hospital transmission and intercontinental spread of MRSA ST-239 [5]. This study identified thousands of high-quality SNPs, allowing a phylogeny to be determined, but this approach depends on mapping the sequences to a high-quality reference genome of a related isolate. Such a method will not be generically applicable to large scale population studies with diverse genotypes.

Adopting a reference gene-based, as opposed to a reference genome-based, approach to analysis of whole or partial genomes avoids many of the problems associated with determining how related strains are to each other by existing genomic methods. As genes are generally the unit of selection, treating them as discrete units of analysis is valid and analogous to MLST where each unique allele at a locus is defined with an allele number. Instead of a seven locus typing scheme the technique can be applied to each of the coding sequences in the genome. Identifying gene-length regions of DNA by comparison to a database of reference sequences is conceptually straightforward, computationally simple, and highly accurate. The method scales well, without the need for reanalysis every time new variants are discovered or new isolates added to a dataset. This is a particular advantage for population datasets that can include many thousands of isolates.

The method can utilize the array of analysis tools developed or applicable for analysing MLST data [24,57-60] with the loci chosen to provide the level of discrimination required or to focus on particular aspects of the biology of the organism. A further advantage is that the method is additive to and fully compatible with existing sequence typing schemes, so that new isolates with genome data can easily be categorized and compared to the large isolate datasets that have been accumulated over many years and geographical locations.

Intra-genus analysis of species groups

With the exception of those defining rRNA variation, most existing sequence typing and characterisation databases host data for one or a few related microbiological species. BIGSDB allows loci to belong to groups which could be related to their level of diversity or by their presence in specific organisms so it can be specified that, for example, a *N. meningitidis* isolate should be scanned against *Neisseria*-specific loci and against more widely applicable loci such as genes for DNA and RNA metabolism or protein folding which are found widely throughout the bacterial domain [61]. Such groupings allow a single database to hold information for a disparate range of species, with characterisation ranging from the genus to a finetype, as appropriate, depending on their similarity and the research question being addressed. The trans-genus *Streptococcus* database demonstrates how multiple MLST schemes can be accommodated within a single database, allowing such schemes to be cross-referenced and sequences exported for phylogenetic analysis. From the analysis of 43 Streptococcal genomes, 77 loci were found in all sequences using a nucleotide BLAST search with the *S. equi* genome [40] as a reference. The identification of these genes, without prior knowledge, allowed a ClonalFrame tree to be constructed showing the individual species as distinct clusters, very similar to the tree constructed from MLSA locus data. Many more loci would be expected to be found throughout the genus if the BLAST search for each locus used an expanding database of all known alleles rather than the single reference genome sequence, as used for the MLSA scheme incorporated in the same database. Alternatively, loci can be defined based on the translated protein sequences, enabling protein BLAST searches to be used. The relative insensitivity of nucleotide BLAST searches compared to those for proteins [41], however, is unlikely to be an issue for isolates belonging to the same or closely related species.

Conclusions

Recent advances in sequencing technology have removed the collection of these data as a limiting step in the study of bacterial populations. It is now possible to undertake whole genome studies on multiple isolates and such limits of cost and speed as remain are likely to be breached in the very near future. The study of bacterial populations now faces the challenge of exploiting this rich source of inference, potentially from whole genomes of thousands of isolates, and to do this it will be necessary to collect structured representative samples of populations and to link precise provenance and phenotype information with the sequence data. The success of the MLST approach to bacterial isolate characterisation was greatly facilitated by the accessibility of the data via the Internet, which enabled community participation in the collection and analysis of the data. MLST also provided a hierarchical and structured approach to population analysis, as well as linking the sequence type to relevant phenotype and provenance information. BIGSDB replicates and extends this paradigm by enabling whole genomes, or fragments of them, to be archived and the data to be organised and interpreted by any number of schemes, which can comprise any number of loci. Using BIGSDB, genomic data can be used to characterise isolates in many different ways but it can also be efficiently exploited for evolutionary or functional studies. Permitting indexing of loci on a functional basis, by treating loci or groups of loci as independent units of analysis, opens the way for genome annotation to become a community-based process [62,63]. BIGSDB represents a freely available resource that will assist the broader community in the elucidation of the structure and function of the bacteria by means of a population genomics approach.

Availability and requirements

Project name: BIGSDB

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

Operating systems: Linux/UNIX

Programming language: Perl, Javascript

Other requirements: Apache, PostgreSQL

License: GNU GPL

Any restrictions to use by non-academics: none

Abbreviations

LIMS: Laboratory Information Management System; MLST: Multi-Locus Sequence Typing; MLVA: Multi-Locus VNTR Analysis; MRSA: Methicillin-Resistant *Staphylococcus auerus*; SNP: Single Nucleotide Polymorphism; ST: Sequence Type; VNTR: Variable Number Tandem Repeats

Acknowledgements

This work was funded by The Wellcome Trust. MCJM is a Wellcome Trust Senior Research Fellow. We are grateful to Holly Bratcher (University of Oxford) and Stephen Bentley (Sanger Institute) for allowing us to deposit Illumina Solexa contigs for isolate OX9932088 in the PubMLST database.

Authors' contributions

MM and KJ conceived the design concept and wrote the manuscript. KJ developed the software. Both authors read and approved the manuscript.

Received: 20 May 2010 Accepted: 10 December 2010
Published: 10 December 2010

References

1. Pettersson E, Lundeberg J, Ahmadian A: **Generations of sequencing technologies.** *Genomics* 2009, **93(2)**:105-111.
2. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, *et al*: **Evolutionary history of *Salmonella typhi*.** *Science* 2006, **314(5803)**:1301-1304.
3. Baker S, Holt K, van de Vosse E, Roumagnac P, Whitehead S, King E, Ewels P, Keniry A, Weill FX, Lightfoot D, *et al*: **High-throughput genotyping of *Salmonella enterica* serovar Typhi allowing geographical assignment of haplotypes and pathotypes within an urban District of Jakarta, Indonesia.** *J Clin Microbiol* 2008, **46(5)**:1741-1746.
4. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, *et al*: **High-throughput sequencing provides**

insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet* 2008, **40(8)**:987-993.
5. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, *et al*: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327(5964)**:469-474.
6. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, Pallen MJ: **High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak.** *J Hosp Infect* 2010, **75(1)**:37-41.
7. Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C: **High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak.** *BMC Genomics* 2010, **11**:120.
8. Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ, *et al*: **Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics.** *Proc Natl Acad Sci USA* 2010, **107(9)**:4371-4376.
9. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, *et al*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95(6)**:3140-3145.
10. Maiden MC: **Multilocus Sequence Typing of Bacteria.** *Annu Rev Microbiol* 2006, **60**:561-588.
11. Jolley KA, Brehony C, Maiden MC: **Molecular typing of meningococci: recommendations for target choice and nomenclature.** *FEMS Microbiol Rev* 2007, **31(1)**:89-96.
12. Dingle KE, McCarthy ND, Cody AJ, Peto TE, Maiden MC: **Extended sequence typing of Campylobacter spp., United Kingdom.** *Emerg Infect Dis* 2008, **14(10)**:1620-1622.
13. Jolley KA, Chan MS, Maiden MC: **mlstdbNet - distributed multi-locus sequence typing (MLST) databases.** *BMC Bioinformatics* 2004, **5(1)**:86.
14. Jolley KA, Maiden MC: **AgdbNet - antigen sequence database software for bacterial typing.** *BMC Bioinformatics* 2006, **7**:314.
15. Aanensen DM, Spratt BG: **The multilocus sequence typing network: mlst.net.** *Nucleic Acids Res* 2005, , **33 Web Server:** W728-733.
16. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*.** *J Bacteriol* 2000, **182(10)**:2928-2936.
17. **mod_perl home page.** [http://perl.apache.org/].
18. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12(10)**:1611-1618.
19. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6)**:276-277.
20. **jQuery home page.** [http://jquery.com/].
21. **Perl/Javascript MD5 secure user authentication home page.** [http://perl-md5-login.sourceforge.net/].
22. **Paul Johnston's home page.** [http://pajhome.org.uk/].
23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
24. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175(3)**:1251-1266.
25. ***Neisseria* MLST website.** [http://pubmlst.org/neisseria/].
26. **Neisseria.org meningococcal typing website.** [http://neisseria.org/nm/typing/].
27. Maiden MC, Stuart JM, Group UMC: **Carriage of serogroup C meningococci 1 year after meningococcal C conjugate polysaccharide vaccination.** *Lancet* 2002, **359(9320)**:1829-1831.
28. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18(5)**:821-829.
29. Brehony C, Wilson DJ, Maiden MC: **Variation of the factor H-binding protein of *Neisseria meningitidis*.** *Microbiology* 2009, **155**:4155-4169.
30. Taha MK, Vazquez JA, Hong E, Bennett DE, Bertrand S, Bukovski S, Cafferkey MT, Carion F, Christensen JJ, Diggle M, *et al*: **Target gene sequencing to characterize the penicillin G susceptibility of *Neisseria meningitidis*.** *Antimicrob Agents Chemother* 2007, **51(8)**:2784-2792.
31. ***Streptococci* genomes demonstration website.** [http://pubmlst.org/streptococci/].
32. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, Glaser P, Rusniok C, Crook DW, Harding RM, *et al*: **Multilocus Sequence Typing System for Group B Streptococcus.** *J Clin Microbiol* 2003, **41(6)**:2530-2536.
33. Do T, Jolley KA, Maiden MC, Gilbert SC, Clark D, Wade WG, Beighton D: **Population structure of *Streptococcus oralis*.** *Microbiology* 2009, **155(Pt 8)**:2593-2602.
34. Enright MC, Spratt BG: **A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease.** *Microbiology* 1998, **144(11)**:3049-3060.
35. Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE: **Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between emm type and clone.** *Infect Immun* 2001, **69(4)**:2416-2427.
36. King SJ, Leigh JA, Heath PJ, Luque I, Tarradas C, Dowson CG, Whatmore AM: **Development of a multilocus sequence typing scheme for the pig pathogen *Streptococcus suis*: identification of virulent clones and potential capsular serotype exchange.** *J Clin Microbiol* 2002, **40(10)**:3671-3680.
37. Coffey TJ, Pullinger GD, Urwin R, Jolley KA, Wilson SM, Maiden MC, Leigh JA: **First insights into the evolution of *Streptococcus uberis*: a multilocus sequence typing scheme that enables investigation of its population biology.** *Appl Environ Microbiol* 2006, **72(2)**:1420-1428.
38. Webb K, Jolley KA, Mitchell Z, Robinson C, Newton JR, Maiden MC, Waller A: **Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group.** *Microbiology* 2008, **154(Pt 10)**:3016-3024.
39. Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG: **Assigning strains to bacterial species via the internet.** *BMC Biology* 2009, **7**:3.
40. Holden MT, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, Jourdan T, Bason NC, Holroyd NE, Mungall K, *et al*: **Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens.** *PLoS Pathogens* 2009, **5(3)**: e1000346.
41. Pearson WR: **Effective protein sequence comparison.** *Meth Enzymol* 1996, **266**:227-258.
42. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser.** *Genome Res* 2004, **14(5)**:951-955.
43. Davidsen T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, *et al*: **The comprehensive microbial resource.** *Nucleic Acids Res* 2009, , **38 Database:** D340-345.
44. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2009, , **38 Database:** D346-354.
45. Nielsen LN, Sheppard SK, McCarthy ND, Maiden MC, Ingmer H, Krogfelt KA: **MLST clustering of Campylobacter jejuni isolates from patients with gastroenteritis, reactive arthritis and Guillain-Barre syndrome.** *J Appl Microbiol* 2009.
46. Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC, *et al*: **Campylobacter genotyping to determine the source of human infection.** *Clin Infect Dis* 2009, **48(8)**:1072-1078.
47. Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, Strachan NJ, Ogden ID, Maiden MC, Forbes KJ: **Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6.** *Int J Food Microbiol* 2009, **134(1-2)**:96-103.
48. Baldwin A, Loughlin M, Caubilla-Barron J, Kucerova E, Manning G, Dowson C, Forsythe S: **Multilocus sequence typing of *Cronobacter sakazakii* and *Cronobacter malonaticus* reveals stable clonal structures with clinical significance which do not correlate with biotypes.** *BMC Microbiology* 2009, **9**:223.
49. Arvand M, Feil EJ, Giladi M, Boulouis HJ, Viezens J: **Multi-locus sequence typing of Bartonella henselae isolates from three continents reveals hypervirulent and feline-associated clones.** *PLoS One* 2007, **2(12)**:e1346.
50. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21(16)**:3422-3423.
51. Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA: **Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database.** *Bioinformatics* 2008, **24(23)**:2672-2676.

52. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14(7)**:1394-1403.
53. Pandya GA, Holmes MH, Petersen JM, Pradhan S, Karamycheva SA, Wolcott MJ, Molins C, Jones M, Schriefer ME, Fleischmann RD, *et al*: **Whole genome single nucleotide polymorphism based phylogeny of *Francisella tularensis* and its application to the development of a strain typing assay.** *BMC Microbiology* 2009, **9**:213.
54. Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PS, Roberto FF, Hnath J, Brettin T, Keim P: **Whole-genome-based phylogeny and divergence of the genus *Brucella*.** *J Bacteriol* 2009, **191(8)**:2864-2870.
55. Pearson T, Okinaka RT, Foster JT, Keim P: **Phylogenetic understanding of clonal populations in an era of whole genome sequencing.** *Infect Genet Evol* 2009, **9(5)**:1010-1019.
56. Kennedy AD, Otto M, Braughton KR, Whitney AR, Chen L, Mathema B, Mediavilla JR, Byrne KA, Parkins LD, Tenover FC, *et al*: **Epidemic community-associated methicillin-resistant *Staphylococcus aureus*: recent clonal expansion and diversification.** *Proc Natl Acad Sci USA* 2008, **105(4)**:1327-1332.
57. Jolley KA, Feil EJ, Chan MS, Maiden MC: **Sequence type analysis and recombinational tests (START).** *Bioinformatics* 2001, **17(12)**:1230-1231.
58. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG: **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data.** *J Bacteriol* 2004, **186(5)**:1518-1530.
59. Francisco AP, Bugalho M, Ramirez M, Carrico JA: **Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach.** *BMC Bioinformatics* 2009, **10**:152.
60. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155(2)**:945-959.
61. Gil R, Silva FJ, Pereto J, Moya A: **Determination of the core of a minimal bacterial gene set.** *Microbiol Mol Biol Rev* 2004, **68(3)**:518-537.
62. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, *et al*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33(17)**:5691-5702.
63. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, *et al*: **The integrated microbial genomes system: an expanding comparative analysis resource.** *Nucleic Acids Res* 2010, , **38 Database:** D382-390.

doi:10.1186/1471-2105-11-595
Cite this article as: Jolley and Maiden: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010 **11**:595.