*D494–D498* *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 14 October 2008*
*doi:10.1093/nar/gkn674*

# Strepto-DB, a database for comparative genomics of group A (GAS) and B (GBS) streptococci, implemented with the novel database platform ‘Open Genome Resource’ (OGeR)

**Johannes Klein, Richard Münch, Ilona Biegler, Isam Haddad, Ida Retter* and Dieter Jahn**

Institute for Microbiology, Technische Universität Braunschweig, Spielmannstrasse 7,
38106 Braunschweig, Germany

Received August 15, 2008; Revised September 19, 2008; Accepted September 22, 2008

## ABSTRACT

**Streptococci are the causative agent of many human infectious diseases including bacterial pneumonia and meningitis. Here, we present Strepto-DB, a database for the comparative genome analysis of group A (GAS) and group B (GBS) streptococci. The known genomes of various GAS and GBS contain a large fraction of distributed genes that were found absent in other strains or serotypes of the same species. Strepto-DB identifies the homologous proteins deduced from the genomes of interest. It allows for the elucidation of the GAS and GBS core- and pan-genomes via genome-wide comparisons. Moreover, an intergenic region analysis tool provides alignments and predictions for transcription factor binding sites in the non-coding sequences. An interactive genome browser visualizes functional annotations. Strepto-DB (http://oger.tu-bs.de/strepto_db) was created by the use of OGeR, the Open Genome Resource for comparative analysis of prokaryotic genomes. OGeR is a newly developed open source database and tool platform for the web-based storage, distribution, visualization and comparison of prokaryotic genome data. The system automatically creates the dedicated relational database and web interface and imports an arbitrary number of genomes derived from standardized genome files. OGeR can be downloaded at http://oger.tu-bs.de.**

## INTRODUCTION

The development of cost-efficient DNA sequencing methods has caused an explosion of prokaryotic genome sequencing projects (1,2). The exploration of new genome sequences is strongly supported by the availability of related genomes that can be used as templates. Correspondingly, strain-specific properties can be traced back to differences in the genomes of compared strains. The comparison of the gene composition of several bacterial genomes from different strains of the same species revealed that only a fraction of genes is shared among the analyzed strains. This so-called core-genome is complemented by a fraction of distributed genes that are only present in some strains and absent in others (3). The supra- or pan-genome of a species is defined as the core-genome plus all distributed genes. It became clear that the pathogenicity of certain bacteria strongly depends on the fraction of distributed genes in the genome (4).

Due to the medical impact of pathogenic *Streptococcus pyogenes* (GAS) and *Streptococcus agalactiae* (GBS) infections, several genome projects focused on the elucidation of serotypic variants of these Gram-positive bacteria (5). Several streptococci genomes are available at the NMPDR (6), a database that focuses on microbial pathogens. Moreover, comprehensive databases provide comparative analysis features for prokaryotic genomes. These include MicrobesOnline (7), IMG (8) and GenoList (9), amongst others. However, for *S. agalactiae* it was predicted that the available reservoir of distributed genes is so large that new genes will be discovered even after hundreds of elucidated genomes (10). Therefore, comparative analyses of GAS and GBS genomes require the incorporation of all available sequence data.

For sequencing projects usually confidential data handling is required prior publishing of the results. For this purpose, local data storage and analysis is essential. Several software tools have recently been published that offer local solutions for comparative analysis of prokaryotic genome data. PSAT (11) is a web tool that visualizes the conservation of gene order among a given set of organisms.

*To whom correspondence should be addressed. Tel: +49 531 391 5810; Fax: +49 531 391 5854; Email: i.retter@tu-bs.de

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Although PSAT supplies a very useful overview about the relatedness of different genomes, it does not offer the query functions of a typical genome database, i.e. direct gene and protein queries with detailed information on obtained results. These features are provided for example by JCoast (12), a tool for the comparative analysis of prokaryotic genomes that is based on GenDB (13). However, JCoast is a local solution that does not support the distribution of data by a web server.

For this reason, we have developed Open Genome Resource (OGeR) as a generic web-accessible database and bioinformatics tool platform for the storage, visualization and comparative analysis of prokaryotic genome data. OGeR is suited to supply convenient assistance for reading and interpretation of genome files for biologists. The system is very flexible as it supports the import of an arbitrary selection of prokaryotic genome DNA sequence flat files. After the initial installation, the system is automatically generated, so that the update to new genome releases is very simple. Thus, OGeR can aid annotation and controlled data distribution in sequencing projects that depend on confidential data handling.

In this article, the functionalities of Strepto-DB are introduced as an example of application for OGeR. The database Strepto-DB provides an up-to-date resource for all GAS and GBS genomes that are currently publicly available, including unfinished WGS sequences. It supplies a convenient platform for the (pan-)genome analysis and interpretation of GAS and GBS. Strepto-DB was developed as part of the ERA-NET PathoGenoMics project that conducts a comprehensive comparative molecular analysis of GAS and GBS pathogenesis (http://www.pathogenomics-era.net).

## FEATURES OF STREPTO-DB

### Data content, exploration and visualization

The current Strepto-DB release 8.8 provides access to 13 GAS genomes, 8 GBS genomes and 7 plasmids. These comprise 41804 protein coding genes, including 902 'unique' genes for which no orthologs in any of the other strains could be detected (Table 1). To visualize the respective sizes of pan-genomes and core-genomes, Venn diagrams are provided as Supplementary Data.

The query options of the Strepto-DB web interface are summarized in Table 2. The database can be searched by gene and protein names, gene ontology (GO) and other functional annotation terms. Sequences can be searched either as strings and regular expressions or by BLAST. A genome viewer provides a scalable overview over the locus of the genes of interest on the chromosome. For each gene, Strepto-DB provides a gene and a corresponding protein entry that comprise functional annotation including GO terms and EC numbers, respectively. Furthermore, links to external data resources are provided. These include EMBL-Bank (14), UniProt (15), Integr8 (16), ExPASy (17), NCBI Gene and Protein (18), KEGG (19), BRENDA (20) and PRODORIC (21). For gene entries, the genomic context is visualized as a map in an interactive genome browser that centers on a gene when selected by a mouse click. The selected gene is marked in red. Below this genome map, the genome browser displays a frame plot of the GC content. The genome browser also displays the DNA sequence of the referring genome section with coding regions in color. At the bottom of the gene entry, the Genomic Data field provides the gene sequence in various formats and the option for download in FASTA format.

**Table 1.** Protein content of Strepto-DB

| Strain | Total no. of proteins | No. of proteins for which no orthologs were found |
|---|---|---|
| *S. agalactiae* 18RS21* | 2146 | 44 |
| *S. agalactiae* 2603V/R | 2123 | 81 |
| *S. agalactiae* 515* | 2275 | 74 |
| *S. agalactiae* A909 | 1996 | 11 |
| *S. agalactiae* CJB111* | 2197 | 73 |
| *S. agalactiae* COH1* | 2376 | 63 |
| *S. agalactiae* H36B* | 2376 | 101 |
| *S. agalactiae* NEM316 | 2094 | 144 |
| *S. pyogenes* M1 GAS | 1697 | 5 |
| *S. pyogenes* M49 591* | 1172 | 96 |
| *S. pyogenes* MGAS10270 | 1987 | 18 |
| *S. pyogenes* MGAS10394 | 1886 | 20 |
| *S. pyogenes* MGAS10750 | 1979 | 41 |
| *S. pyogenes* MGAS2096 | 1898 | 34 |
| *S. pyogenes* MGAS315 | 1865 | 5 |
| *S. pyogenes* MGAS5005 | 1851 | 3 |
| *S. pyogenes* MGAS6180 | 1890 | 13 |
| *S. pyogenes* MGAS8232 | 1844 | 14 |
| *S. pyogenes* MGAS9429 | 1877 | 33 |
| *S. pyogenes* SSI-1 | 1860 | 19 |
| *S. pyogenes* str. Manfredo | 1746 | 10 |
| All | 41804 | 902 |

*The sequence of this strain was not completely finished. The remaining gaps might contain protein coding genes, and other genes might be redundantly annotated within several overlapping contigs.

### Search for homologous proteins and intergenic region analysis

Strepto-DB allows for the alignment of both coding and non-coding DNA sequences within the *Streptococcus* genomes of interest. Homologous proteins were precalculated by reciprocal BLAST searches. The proteome comparison query supplies an overview about the conservation of proteins between different strains. After the selection of a reference genome and one or more comparison genomes, this query returns lists of those proteins that are conserved between the selected strains. In addition, each protein entry provides a list of homologous proteins. On demand, the identified homologs are aligned with the MUSCLE alignment tool (22) and displayed with the Jalview visualization software (23). Furthermore, the genomic context of the various homologous genes can be displayed as genome maps. As an example, Figure 1 shows the genomic context of the *cylE* gene for β-hemolytic/cytolytic activity (24). The *cylE* gene is present in all sequenced strains of *S. agalactiae* but was found absent in all *S. pyogenes* strains. The genome map shows differences in the annotation of the region of the *cyl* operon in *S. agalactiae* COH1, CJB111 and NEM316.

**Table 2.** Query options of Strepto-DB

| Query name | Query action |
|---|---|
| BLAST | Perform a sequence comparison<br>–with the blastp program against the Strepto-DB protein sequences<br>–with the blastn program against the Strepto-DB gene sequences<br>–with the blastn program against the Strepto-DB chromosome, contig and plasmid sequences |
| Genes/Proteins | Search for genes and proteins by name, locus tag, GO term, keyword, EC number or database identifier |
| Genome Viewer | Browse a selected genome with the Circular Genome Viewer (CGView) |
| Intergenic Region Analysis | Search for a gene and its homologs to select an intergenic region for analysis with MUSCLE, Virtual Footprint and MEME |
| Proteome Comparison | Select a reference genome and one or more comparison genomes to view homologous proteins within the selected genomes |
| Sequences | Search for sequences or regular expressions within a selected chromosome, plasmid or contig |

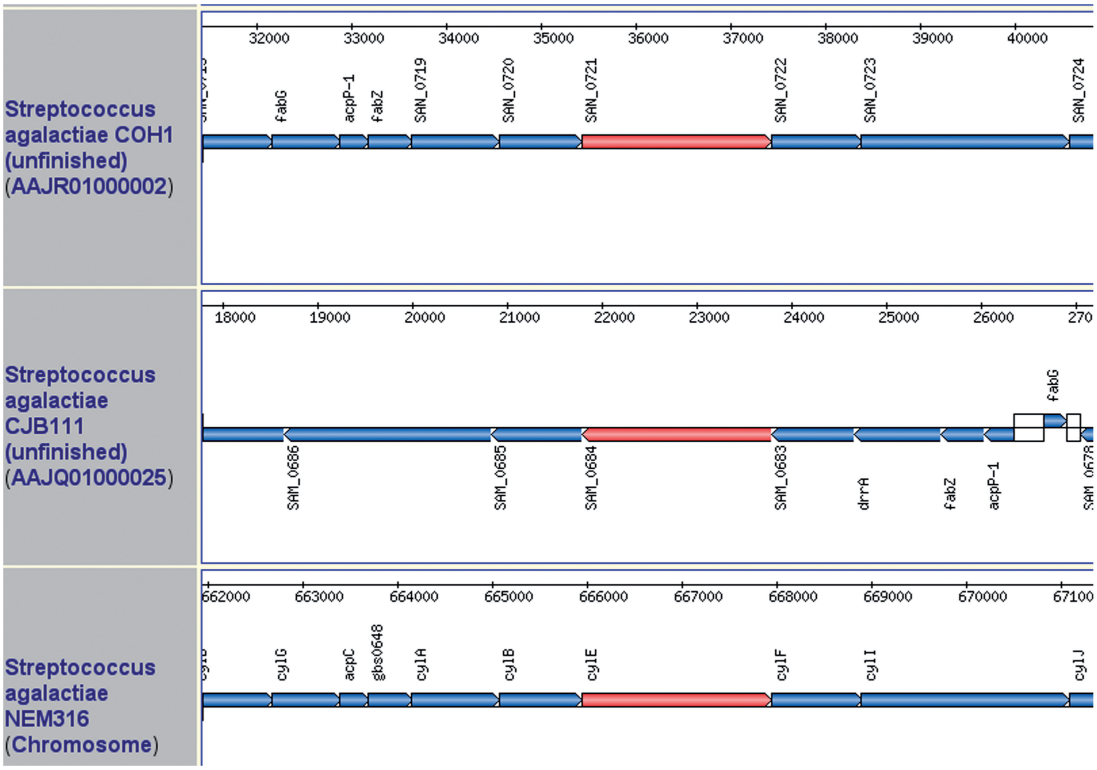

**Figure 1.** Map of homologous protein coding genes. The screenshot shows the *cyl* operon in the GBS strains COH1, CJB111 and NEM316. The gene *cylE* is marked in red. To view homologs of the adjacent genes, the reference gene can be changed by a mouse click.

In the intergenic regions, conserved DNA sequence motifs can function as regulator binding sites. Thus, an analysis of the intergenic DNA sequences might reveal information on the regulation of the respective downstream genes. Strepto-DB provides an intergenic region analysis that is composed of three tools: first, a BLAST search that aligns the intergenic region DNA sequence of choice with the intergenic regions of the referring homologous genes of other *Streptococcus* strains. This similarity search can be started by a mouse click on the region of interest on the homologs' genome map. Second, selected intergenic regions can be analyzed for conserved sequence motifs with the MEME motif discovery tool (25). Third, each intergenic region entry includes a link to the Virtual Footprint analysis tool (21). Virtual Footprint uses position weight matrices from the PRODORIC database to predict transcription factor binding sites within the promoter region of a gene. Taken together, these methods provide very useful supplementary evidence for potential regulator binding sites, generating hypotheses for experimental verification.

## THE 'OPEN GENOME RESOURCE' (OGeR) PLATFORM FOR THE COMPARATIVE ANALYSIS OF PROKARYOTIC GENOMES

OGeR is generically applicable for the storage and comparison of related prokaryotic genomes. As one example, Strepto-DB was set up and is maintained with OGeR

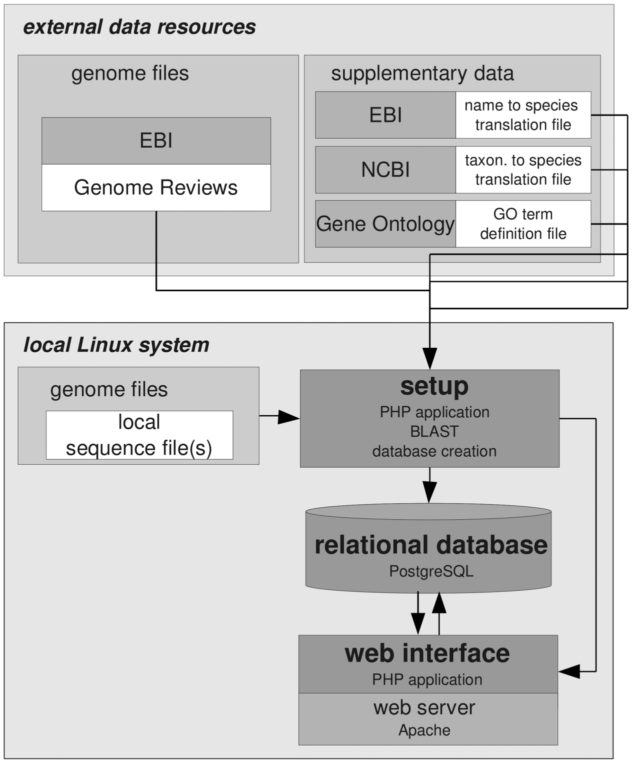

**Figure 2.** Setup and components of the OGeR system.

and therefore provides an example for its functionalities. Thus, the Strepto-DB database and all features of the web interface were automatically compiled.

### System architecture

OGeR consists of three components, a relational database, a setup that processes input data and imports them into the database and a web interface that queries the database (Figure 2). By default, genome sequences are downloaded from the EBI Genome Reviews database. GenBank files and other local data sources can also be loaded. Additional Supplementary Data is automatically downloaded from the Gene Ontology, EBI and NCBI websites. The setup creates the database schema and processes sequences and other input files. This procedure includes the extraction of gene and protein annotations and the detection of homologous proteins. Finally, sequence data and corresponding annotations are imported into the database. The web interface presents data stored in the database and performs multiple alignments on demand. It links to various external databases, provided the referring database identifiers were included in the input files.

### Implementation and local installation

OGeR is implemented as a PHP application that uses an Apache web server and operates on a PostgreSQL database. Local installation requires a Linux operating system and the installation of the corresponding PHP and Apache software packages. For the creation of a new OGeR-based database, the OGeR setup procedure requests the required information and imports the desired genomes into the system. Data download is performed by the wget program. Local genome sequences can be imported in EMBL or GenBank format. Subsequently, homologous proteins are determined by an all-against-all BLAST search (26) of the proteins that are annotated in the imported flat files. As the BLAST search follows a quadratic time complexity, this step limits the number of genomes that can be imported in a reasonable amount of time on a given computing hardware. The BLAST results are evaluated to detect homologous proteins. Thereby, ‘homology’ is defined as a double reciprocal BLAST hit with a given maximal E-value. For Strepto-DB, an E-value cutoff of 1*e-5 was chosen. Finally, the setup finishes with the creation of a new web interface for the database. A detailed installation instruction facilitates the installation and setup procedure.

The OGeR web interface uses CGView (27) for the genome viewer. Multiple alignments are performed with MUSCLE (22) and depicted with Jalview (23). As CGView and JalView are implemented as Java applets, the client web browser requires Java installation. However, multiple alignments can alternatively be shown in a simple view that does not depend on Java.

## CONCLUDING REMARKS

We have implemented a simple integrated database and bioinformatics platform named OGeR for the comparative analysis of related genomes. This platform was subsequently employed for comparative genomic analyses of 21 *Streptococcus* genomes with establishment of the Strepto-DB platform. Conserved and distributed genes were deduced for the analyzed strains and used for core- and pan-genome prediction.

## ACKNOWLEDGEMENTS

We would like to thank Bernd Hoppe for excellent technical assistance and financial management.

## FUNDING

German Bundesministerium für Bildung und Forschung (ERA-NET grant 0313936C to J.K. and R.M.). and Deutsche Forschungsgemeinschaft (Sonderforschungsbreich 578 to I.B. and R.M.). Funding for open access charges: Deutsche Forschungsgemeinschaft (Sonderforschungsbreich 578).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–479.
2. Medini,D., Serruto,D., Parkhill,J., Relman,D.A., Donati,C., Moxon,R., Falkow,S. and Rappuoli,R. (2008) Microbiology in the post-genomic era. *Nat. Rev. Micro.*, **6**, 419–430.

3. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
4. Ehrlich,G.D., Hiller,N.L. and Hu,F.Z. (2008) What makes pathogens pathogenic. *Genome Biol.*, **9**, 225.
5. Lefébure,T. and Stanhope,M.J. (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.*, **8**, R71.
6. McNeil,L.K., Reich,C., Aziz,R.K., Bartels,D., Cohoon,M., Disz,T., Edwards,R.A., Gerdes,S., Hwang,K., Kubal,M. *et al.* (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.*, **35**, D347–353.
7. Alm,E.J., Huang,K.H., Price,M.N., Koche,R.P., Keller,K., Dubchak,I.L. and Arkin,A.P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.
8. Markowitz,V.M., Szeto,E., Palaniappan,K., Grechkin,Y., Chu,K., Chen,I.M.A., Dubchak,I., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–533.
9. Lechat,P., Hummel,L., Rousseau,S. and Moszer,I. (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res.*, **36**, D469–474.
10. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.
11. Fong,C., Rohmer,L., Radey,M., Wasnick,M. and Brittnacher,M. (2008) PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics*, **9**, 170.
12. Richter, M., Lombardot, T., Kostadinov, I., Kottmann, R., Duhaime, M., Peplies, J. and Glockner, F. (2008) JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics*, **9**, 177.
13. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al.* (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
14. Cochrane,G., Akhtar,R., Aldebert,P., Althorpe,N., Baldwin,A., Bates,K., Bhattacharyya,S., Bonfield,J., Bower,L., Browne,P. *et al.* (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **36**, D5–12.
15. UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–195.
16. Mulder,N.J., Kersey,P., Pruess,M. and Apweiler,R. (2008) In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol.*, **38**, 165–177.
17. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
18. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–21.
19. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–484.
20. Barthelmes,J., Ebeling,C., Chang,A., Schomburg,I. and Schomburg,D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.*, **35**, D511–514.
21. Münch,R., Hiller,K., Grote,A., Scheer,M., Klein,J., Schobert,M. and Jahn,D. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, **21**, 4187–4189.
22. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
23. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
24. Tettelin,H., Masignani,V., Cieslewicz,M.J., Eisen,J.A., Peterson,S., Wessels,M.R., Paulsen,I.T., Nelson,K.E., Margarit,I., Read,T.D. *et al.* (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc. Natl Acad. Sci. USA*, **99**, 12391–12396.
25. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–373.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Grant,J.R. and Stothard,P. (2008) The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res.*, **36**, W181–184.