

The Genome Atlas Resource

Matloob Qureshi, Eva Rotenberg, Hans-Henrik Stærfeldt,
Lena Hansson, and David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology,
The Technical University of Denmark, 2800 Lyngby, Denmark

Abstract. The Genome Atlas is a resource for addressing the challenges of synchronising prokaryotic genomic sequence data from multiple public repositories. This resource can integrate bioinformatic analyses in various data format and quality. Existing open source tools have been used together with scripts and algorithms developed in a variety of programming languages at the Centre for Biological Sequence Analysis in order to create a three-tier software application for genome analysis. The results are made available via a web interface developed in Java, PHP and Perl CGI. User-configurable and dynamic views of Chromosomal maps are made possible through an updated GeneWiz browser (version 0.94) which uses Java to allow rapid zooming in and out of the atlases.

Keywords: Genome atlas, web interface, chromosomal maps, genome analysis.

1 Introduction

There are, at the time of writing, over 1200 completed Archaeal and Bacterial genome sequences available in the major public repositories of sequence data. However, for a number of reasons these repositories are not in complete synchronisation with each other [1]. Furthermore, the number of genomes published per year has been rising rapidly since the first genome published in 1995 [2] and the advent of “next generation” sequencing technologies which allow a bacterial genome to be sequenced, assembled and annotated in a day [3,4] serve to highlight the need for tools to assist with comparative genomic analysis.

Most bacterial genomes would be thousands of pages long, if viewed as text. Therefore there is a need to collate these projects and present them together in an integrated site, along with links to a graphical overview of the chromosomal sequences. We use asynchronous client-server communication to develop zoomable atlases, which can go from a full chromosomal view, down to the level of individual nucleotides, smoothly and quickly. Although there are other web-based services such as Entrez NCBI genomes [5] and EnSEMBL genomes [6], only the CBS Genome Atlas application focuses on collecting prokaryotic sequence data from multiple sources together with the results of detailed genomic and structural analyses in a user-configurable and dynamic manner [7, 8].

In this work, open source tools such as BioPerl and eHive [9,10] have been used together with Perl scripts and algorithms in other programming languages to develop a three-tier software application for genome analysis [11]. This resource is available

at <http://www.cbs.dtu.dk/services/GenomeAtlas/>. The structure of the rest of this paper is as follows: section 2 covers the three tier architecture of the genome atlas, section 3 describes the GeneWiz chromosome atlas browser, section 4 presents the genome atlas data model, Some examples of scientific uses of the web pages are described in section 5 and finally conclusions are made in the last section.

2 Three Tier Architecture

The Genome Atlas has been redeveloped as three-tier application. Multi-tier applications are concerned with partitioning components of an application into layers, each concerned with a different aspect of the system in order to facilitate flexibility and reuse. Multi-tier architectures are often used in client server applications [11]. The three tiers of Genome Atlas system are described diagrammatically in **Fig. 1**. The data tier is concerned with storage and access to the information used by the application. A data access API developed using BioPerl presents an interface to this information that can be accessed from the logic tier. This tier is involved with the coordination and processing of data and is where the analysis pipeline resides. The final tier of the application is the presentation tier where Genome Atlas data is presented via a web server using a combination of PHP, Perl web pages and the GeneWiz Java application.

2.1 The Data Tier

Completed prokaryotic genome sequence data are downloaded from the three main public sequence repositories, GenBank, EMBL and DDBJ. Although they regularly synchronise data amongst themselves, a small number of projects can be missing from one or more resource. Therefore the Genome Atlas database draws data from all three resources and also from projects referenced in the Genomes On-Line Database (GOLD). The MD5 sum of the DNA sequence is used to minimise redundancy of information stored in the Genome Atlas. Project accession numbers and identifiers are used to map data between the different sources. The GOLD database provides references to genome projects that are not yet in the main repositories and these are also downloaded where possible.

The DNA sequence data are stored as files in GenBank and EMBL format in a directory and file structure as described previously [7]. A MySQL database is used to collate all the information associated with a project, including the location of the all sequence files and the associated taxonomic data. The results of any computationally intensive analyses are stored in the database where appropriate or in as a set of files in a similar directory structure to the raw sequence data files. A data access layer written in Perl with BioPerl provides an integrated object-oriented interface to all the stored information. This can be accessed, by the analysis pipeline in the logic tier and by any other scripts written by users with accounts at the institute.

2.2 The Logic Tier

The logic tier consists of a set of bioinformatic applications written in a variety of programming languages including C, Perl, Python and shell scripts. The execution of these applications in sequence to create an analysis pipeline is managed by the eHive

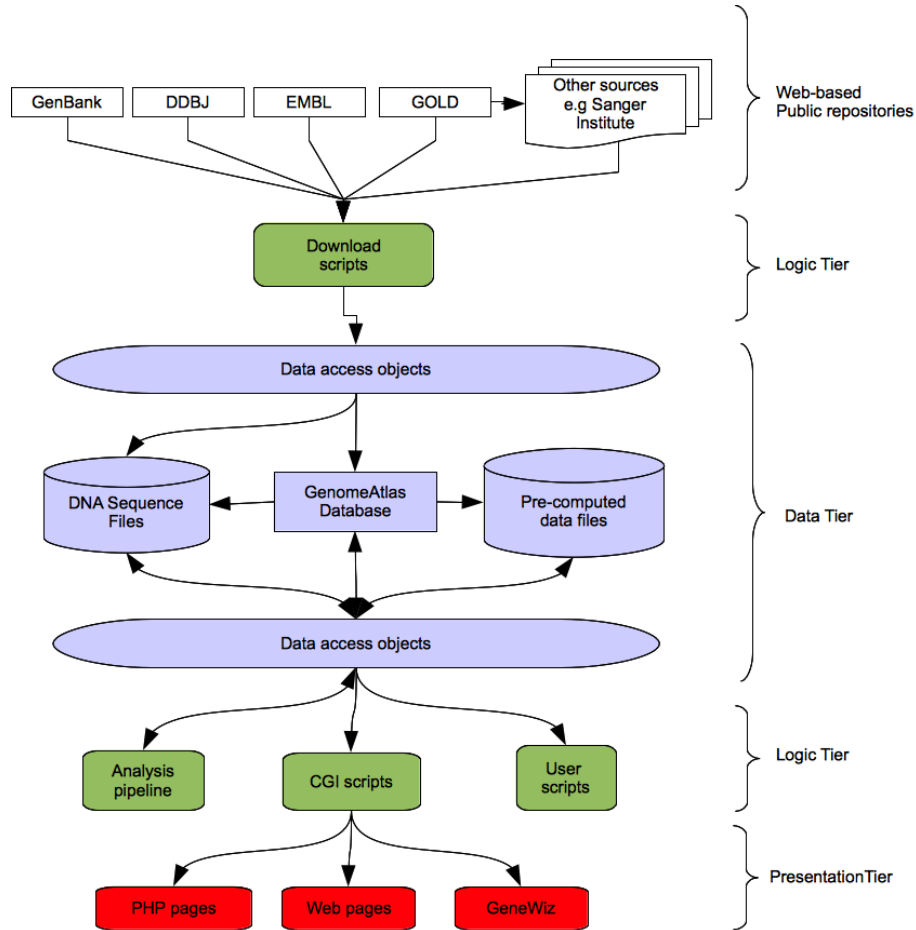


Fig. 1. Genome Atlas Application Diagram. The application is divided into three-tiers. The Data tier is concerned with the storage and retrieval of information. Data access objects are used to present a simple common application interface to the underlying data. The Logic tier performs more complex processing. The download scripts for acquisition of Sequence data from public repositories and the analysis pipeline reside in this tier. The third tier is the presentation tier, where data is reformatted and presented in web forms, static and dynamic chromosome atlases, and genome summaries.

workflow management system developed for the EnsEMBL project [10]. The executable of interest is called from a Perl module that controls the input, output and execution environment of the software. The module together with the pre- and post-conditions are stored in the eHive database. Thus the eHive system allows analyses to be grouped into a set of independent work packages which make efficient use of our computing cluster. It also manages the flow of data from one set of analyses to the next and provides a fault-tolerant system for handling situations when problems occur.

2.3 The Presentation Tier

The Genome Atlas web pages are written in PHP and are updated regularly. The basis for displaying data and analyses in the Genome Atlas is the chromosomal map [12]. Many kinds of chromosomal maps can be displayed together, so that the chromosome can be visualised together with genes, repeat sequences and structural data. Some of these are generated by the analysis pipeline and stored as vector graphics, compressed bit maps or binary files whilst others are created on request in the logic tier.

3 The GeneWiz Chromosome Atlas Browser

The GeneWiz browser provides a dynamic scalable and zoomable view of the chromosomal maps. It is written in Java, using the AWT and Swing libraries for the GUI components. The data can be taken from existing genome projects stored in Genome Atlas in which case the browser makes use of predefined settings and binary files. Alternatively, user defined data can be uploaded in a number of formats. The properties of the chromosome are calculated by the CBS computing cluster and displayed as tracks in the chromosome map display. Asynchronous requests are made to the server whilst the display is rendered so when the atlas is clicked or a section is selected there is a smooth zoom in or out from one level of magnification to another, all the way through to the individual bases in the sequence. Pre-calculated binary files are requested by the client and the server returns the data necessary for the display [8]. Some of the tracks, for example, global inverted repeats are displayed with a colour scheme based on global properties, whilst the colours and statistics for other tracks, such as GC skew, are recalculated according to the region of the chromosome being viewed. Details for open reading frames are displayed from the genome annotation when the pointer is moved over the CDS on the appropriate track. The GeneWiz browser also allows views and settings to be stored as a session in a user defined directory. Browsing may then be resumed at any time by loading the session file. The browser also exports data in fasta, support vector graphic, postscript or pdf format.

4 The Genome Atlas Data Model

The information stored in the Genome Atlas data tier is coordinated via a set of tables in a MySQL relational database. The data model for this database is shown in **Fig. 2**. The main table concerned with storing data for coordinating the Genome Atlas with external repositories is the Projects table. This table links an internal project identifier and version with data related to the Organism, such as NCBI organism taxonomy. It also references information about the project from public repositories stored in the Sources table. A project is linked to one or more DNA sequences through the Segments table. This table contains information about the actual location of the DNA sequence and other related data such as length and MD5 check-sum. The MD5 values are used to generate incremental version numbers whenever the DNA sequences referenced by Segments are updated. The results of a given bioinformatic analysis are

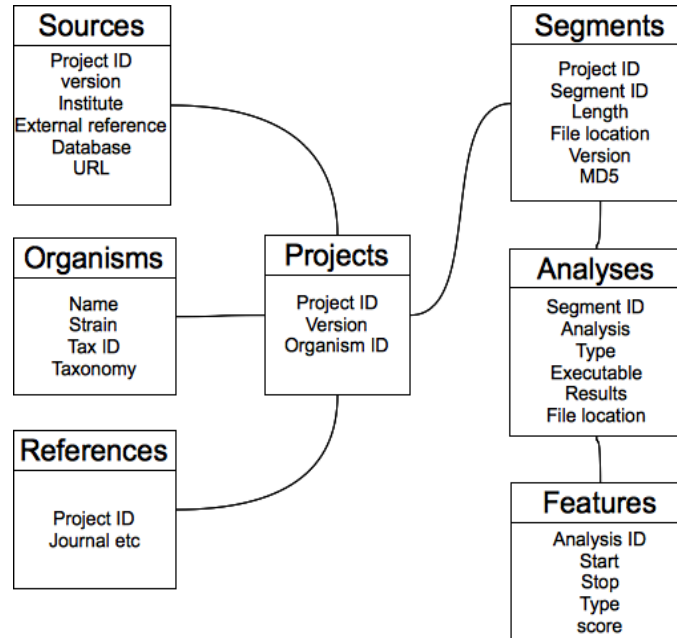


Fig. 2. Genome Atlas Database Data model. The main tables in the Genome Atlas are shown. The Projects table links to the Sources, Organisms and References tables. This provides a way of querying external meta-data linked to the project. The Segments table links the Projects to DNA Sequence files stored in a file system, whilst the Analyses table connects the DNA Sequence to the results produced by running the analysis pipeline on a project.

stored in the Analyses table. This table records the analysis module, the executable and version together with the results and any file locations or sequence features generated by the analysis pipeline. This allows the logic tier to select attributes and features based created by specific versions of analysis executables.

5 Using the Genome Atlas for Research

The Genome Atlas can be used to analyse the properties of genomes in order to test hypotheses. Here we present two short investigations conducted through the Genome Atlas application.

5.1 Correlation between Growth Rate and the Properties of Chromosomes

The Genome Atlas analysis pipeline includes analysis by RNAmmer [13], therefore we can display the number of predicted 5S, 16S and 23S ribosomal RNA genes in each genome. This can found on the General Genomes web page. The genomes table displayed here can be sorted on any column in ascending or descending order by clicking on the arrows underneath the column heading (**Fig. 3**).

1075 projects found

Available Tables

General Sigma Factors Two-Component systems Repeats Protein length AZ-DNA

| Row | Organism | Tax Group | NCBI Project ID | Replicons | Total Size (bp) | Number of genes | SS rRNA count | 16S rRNA count | 23S rRNA count | tRNA count | % AT |
|-----|----------------------------------------------------------------|-----------|-----------------|-----------|-----------------|-----------------|---------------|----------------|----------------|------------|------|
| 1 | <i>Photobacterium profundum</i> SS9 | BProt GV | 13128 | 3 | 6,403,280 | 5,480 | 19 | 15 | 15 | 169 | 58.3 |
| 2 | <i>Brevibacillus brevis</i> NBRC 100599 | BFirm BP | 29147 | 1 | 6,296,436 | 5,949 | 14 | 15 | 15 | 127 | 52.7 |
| 3 | <i>Clostridium beijerinckii</i> NCIMB 8052 | BFirm CC | 12637 | 1 | 6,000,632 | 5,020 | 15 | 14 | 14 | 94 | 70.1 |
| 4 | <i>Bacillus cereus</i> AH187 | BFirm BB | 17715 | 5 | 5,599,857 | 5,796 | 14 | 14 | 14 | 104 | 64.5 |
| 5 | <i>Bacillus cereus</i> B4264 | BFirm BB | 17731 | 1 | 5,419,038 | 5,408 | 14 | 14 | 14 | 108 | 64.7 |
| 6 | <i>Bacillus thuringiensis</i> str. Al Hakam | BFirm BB | 18255 | 2 | 5,313,030 | 4,798 | 14 | 14 | 14 | 104 | 64.6 |
| 7 | <i>Bacillus cereus</i> G38B102 | BFirm BB | 31307 | 2 | 5,449,908 | 5,621 | 14 | 14 | 14 | 105 | 64.7 |
| 8 | <i>Bacillus thuringiensis</i> BMB171 | BFirm BB | 43631 | 2 | 5,643,051 | 5,349 | 14 | 14 | 14 | 104 | 64.8 |
| 9 | <i>Bacillus cereus</i> ATCC 14579 | BFirm BB | 384 | 2 | 5,427,083 | 5,255 | 13 | 13 | 13 | 108 | 64.7 |
| 10 | <i>Bacillus cereus</i> E33L | BFirm BB | 12488 | 6 | 5,843,235 | 5,641 | 13 | 13 | 13 | 96 | 64.9 |
| 11 | <i>Bacillus cytotoxicus</i> NVH 391-98 | BFirm BB | 13624 | 2 | 4,094,159 | 3,844 | 13 | 13 | 13 | 106 | 64.1 |
| 12 | <i>Bacillus cereus</i> Q1 | BFirm BB | 16220 | 3 | 5,506,207 | 5,502 | 13 | 13 | 13 | 94 | 64.5 |
| 13 | <i>Bacillus cereus</i> G9842 | BFirm BB | 17733 | 3 | 5,736,823 | 5,857 | 13 | 13 | 13 | 98 | 65.0 |
| 14 | <i>Bacillus cereus</i> ATCC 10987 | BFirm BB | 74 | 2 | 5,432,652 | 5,844 | 12 | 12 | 12 | 98 | 64.5 |
| 15 | <i>Vibrio fischeri</i> ES114 | BProt GV | 12986 | 3 | 4,284,050 | 3,802 | 13 | 12 | 12 | 119 | 61.6 |
| 16 | <i>Bacillus cereus</i> AH90 | BFirm BB | 17711 | 4 | 5,688,834 | 5,810 | 12 | 12 | 12 | 96 | 64.7 |
| 17 | <i>Shewanella sediminis</i> HAW-EB3 | BProt GA | 18789 | 1 | 5,517,674 | 4,497 | 13 | 12 | 12 | 125 | 53.9 |
| 18 | <i>Paenibacillus</i> sp. JDR-2 | BFirm BP | 20399 | 1 | 7,184,930 | 6,213 | 11 | 12 | 12 | 88 | 49.7 |
| 19 | <i>Bacillus megaterium</i> QM B1551 | BFirm BB | 30165 | 8 | 5,523,192 | 5,629 | 13 | 12 | 12 | 139 | 62.1 |
| 20 | <i>Alivivitis salmonicida</i> LF1298 | BProt GV | 30703 | 6 | 4,655,860 | 4,284 | 13 | 12 | 12 | 104 | 61.0 |
| 21 | <i>Clostridium acetobutylicum</i> ATCC 824 | BFirm CC | 77 | 2 | 4,132,880 | 3,848 | 11 | 11 | 11 | 73 | 69.1 |
| 22 | <i>Clostridium difficile</i> 630 | BFirm CC | 78 | 2 | 4,298,133 | 3,787 | 10 | 11 | 11 | 87 | 70.9 |
| 23 | <i>Bacillus anthracis</i> str. Ames | BFirm BB | 309 | 1 | 5,227,293 | 5,311 | 11 | 11 | 11 | 95 | 64.6 |
| 24 | <i>Vibrio parahaemolyticus</i> HMD 2210633 | BProt GV | 380 | 2 | 5,165,770 | 4,832 | 12 | 11 | 11 | 126 | 54.6 |
| 25 | <i>Bacillus anthracis</i> str. Ames Ancestor | BFirm BB | 10784 | 3 | 5,503,926 | 5,617 | 11 | 11 | 11 | 95 | 64.8 |
| 26 | <i>Bacillus anthracis</i> str. Sterne | BFirm BB | 10878 | 1 | 5,228,863 | 5,287 | 11 | 11 | 11 | 95 | 64.6 |
| 27 | <i>Vibrio Harveyi</i> ATCC BAA-1116 | BProt GV | 18857 | 3 | 6,058,377 | 6,064 | 11 | 11 | 10 | 121 | 54.6 |
| 28 | <i>Clostridium botulinum</i> E3 str. Alaska F43 | BFirm CC | 28855 | 1 | 3,659,644 | 3,256 | 12 | 11 | 11 | 79 | 72.6 |
| 29 | <i>Clostridium botulinum</i> B str. Eklund 17B | BFirm CC | 28857 | 2 | 3,847,969 | 3,527 | 12 | 11 | 11 | 77 | 72.5 |
| 30 | <i>Bacillus anthracis</i> str. CDC 884 | BFirm BB | 31329 | 3 | 5,506,763 | 5,902 | 11 | 11 | 11 | 96 | 64.8 |
| 31 | <i>Bacillus anthracis</i> str. A0248 | BFirm BB | 33543 | 3 | 5,503,926 | 5,291 | 11 | 11 | 11 | 95 | 64.8 |
| 32 | <i>Vibrio</i> sp. Fx25 | BProt GV | 40507 | 2 | 5,089,025 | 4,518 | 12 | 11 | 11 | 124 | 55.1 |
| 33 | <i>Bacillus megaterium</i> DSM319 | BFirm BB | 42425 | 1 | 5,097,447 | 5,124 | 11 | 11 | 11 | 115 | 61.9 |
| 34 | <i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168 | BFirm BB | 75 | 1 | 4,214,630 | 4,106 | 10 | 10 | 10 | 86 | 56.6 |
| 35 | <i>Clostridium perfringens</i> str. 13 | BFirm CC | 78 | 2 | 3,085,740 | 2,723 | 10 | 10 | 9 | 96 | 71.5 |
| 36 | <i>Clostridium perfringens</i> SMI01 | BFirm CC | 12621 | 4 | 2,960,088 | 2,631 | 10 | 10 | 10 | 95 | 71.8 |
| 37 | <i>Alkaliphilus metallirediensis</i> QYMF | BFirm CC | 13008 | 1 | 4,929,568 | 4,625 | 11 | 10 | 10 | 106 | 63.2 |
| 38 | <i>Shewanella baltica</i> CSI-55 | BProt GA | 13385 | 6 | 5,342,896 | 4,489 | 11 | 10 | 10 | 117 | 53.8 |
| 39 | <i>Shewanella baltica</i> CSI-98 | BProt GA | 13388 | 4 | 5,547,544 | 4,688 | 11 | 10 | 10 | 104 | 53.8 |
| 40 | <i>Halobacterium modesticalium</i> Ioc1 | BFirm CC | 13427 | 1 | 3,075,407 | 3,000 | 10 | 10 | 10 | 109 | 43.0 |
| 41 | <i>Psychromonas ingrahamii</i> 37 | BProt GA | 16187 | 1 | 4,559,598 | 3,545 | 12 | 10 | 10 | 86 | 59.9 |
| 42 | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> ATCC 7966 | BProt GA | 16897 | 1 | 4,744,448 | 4,122 | 11 | 10 | 10 | 128 | 38.4 |
| 43 | <i>Clostridium novyi</i> NT | BFirm CC | 16820 | 1 | 2,847,720 | 2,225 | 10 | 10 | 10 | 81 | 71.1 |
| 44 | <i>Shewanella woodyi</i> ATCC 51908 | BProt GA | 17455 | 1 | 5,935,403 | 4,880 | 11 | 10 | 10 | 126 | 58.3 |

Fig. 3. The General Genome Atlas table here shows more than a thousand bacterial genomes sorted by number of 16S Ribosomal RNA in descending order

The first entry in the table is *Photobacterium profundum* (GenBank project ID 13128) which is a member of Vibrionales; known for their short doubling time [14]. The other group heavily represented at the top of the table are Firmacutes, notably Clostridia. The second entry, *Brevibacillus brevis* (Genbank project ID 29147), is a member of this group. It has been observed that the leading and lagging strand of bacterial chromosomes have differing “skews” of nucleotides present in one of the two stands. The A/T and G/C biases of short oligomers can be plotted along the chromosome to allow visualisation of any such biases [15]. Fig. 4 shows these plots of the oligomer bias towards the leading strand for a number of different organisms.

Fig. 4a shows the strand bias of *Photobacterium profundum* (a Gram negative Gammaproteobacteria); it can be seen that the Guanines are over-represented on the leading strand whilst the Adenines are on the lagging strand. In contrast, Fig. 4b shows *Brevibacillus brevis* (a Gram positive Firmacute). The green and the blue lines follow the same direction with respect to replication origin and terminus. The Adenine bias in Fig. 4a is much weaker and in the opposite direction. Thus, although both are highly streamlined, they show the opposite strand bias. This difference in the bias probably reflects a historical contingency as the *polC* gene encodes a proofreading subunit in DNA polymerase which is present in *Brevibacillus brevis* but absent from *Photobacterium profundum* [14].

Looking back at the table in Fig. 3, the third organism on the list, *Clostridium beijerinckii* (GenBank project ID 12637) has an AT content of 70%, which is considerably higher than many bacterial genomes. Many of the Firmicute genomes in this list

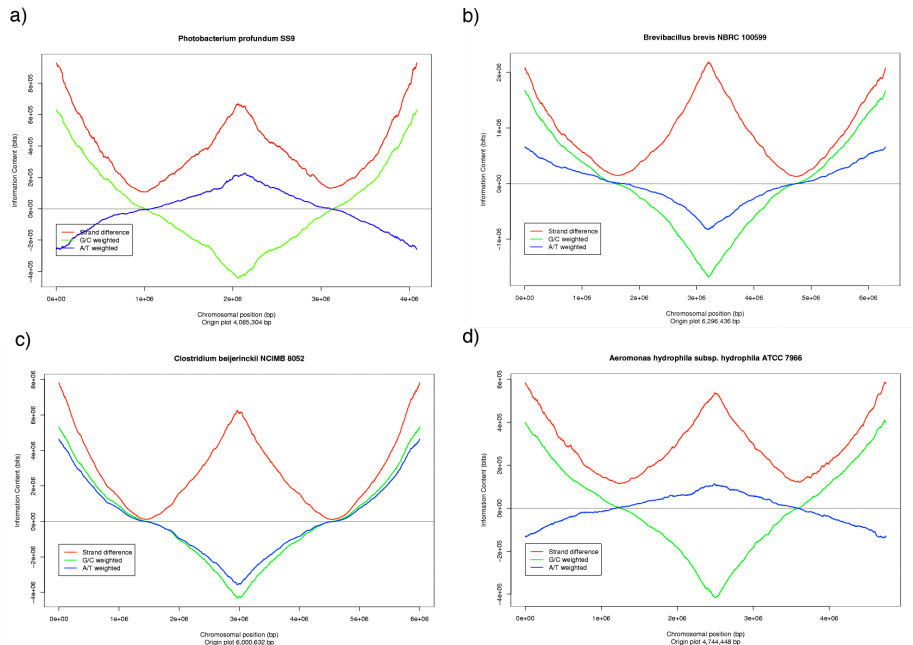


Fig. 4. Plots of the G/C skew, A/T skew on the leading strand and the strand difference of oligomers from the chromosome. (a) shows *Photobacterium profundum* SS9, (b) is *Brevibacillus brevis* NBRC 100599, (c) is *Clostridium beijerinckii* NCIMB 8052 and (d) is *Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966.

are AT rich. We can see a similar oligomer skew in **Fig. 4c**. For comparison, the skew diagram for a GC rich organism from the table in **Fig. 3** is also shown (*Aeromonas hydrophila*: GenBank project 16697). Notice there is still a strong strand bias, although, like *Photobacterium profundum*, the Guanines and Adenines are biased on the opposite strand (i.e the blue and green lines are opposite).

5.2 Zooming into Regions of Chromosomes

We chose to visualise the chromosome of *Brevibacillus brevis* (the second organism shown in **Fig. 3**) via the GeneWiz browser (**Fig. 5**). There are many repeat regions visible in the chromosome map (shown in the red and blue lanes in the inner circles). There are also some regions of high intensity in the 'position preference' lane (shown as dark green). These areas of the chromosome are likely to exclude chromatin proteins, and as such are often the location for highly expressed genes [16].

Fig. 5a Shows the whole chromosome zoomed out whilst **Fig. 5b**, shows a magnified view of the area near the top of the circle. This is in a region closer to the replication origin, containing genes coding for the beta/beta prime subunit of RNA polymerase, which are known to be highly expressed. At the bottom of the chromosome, close to the replication terminus, is another region, containing the *lrg* operon, which controls synthesis on an antibiotic peptide linear gramicidin. This operon can

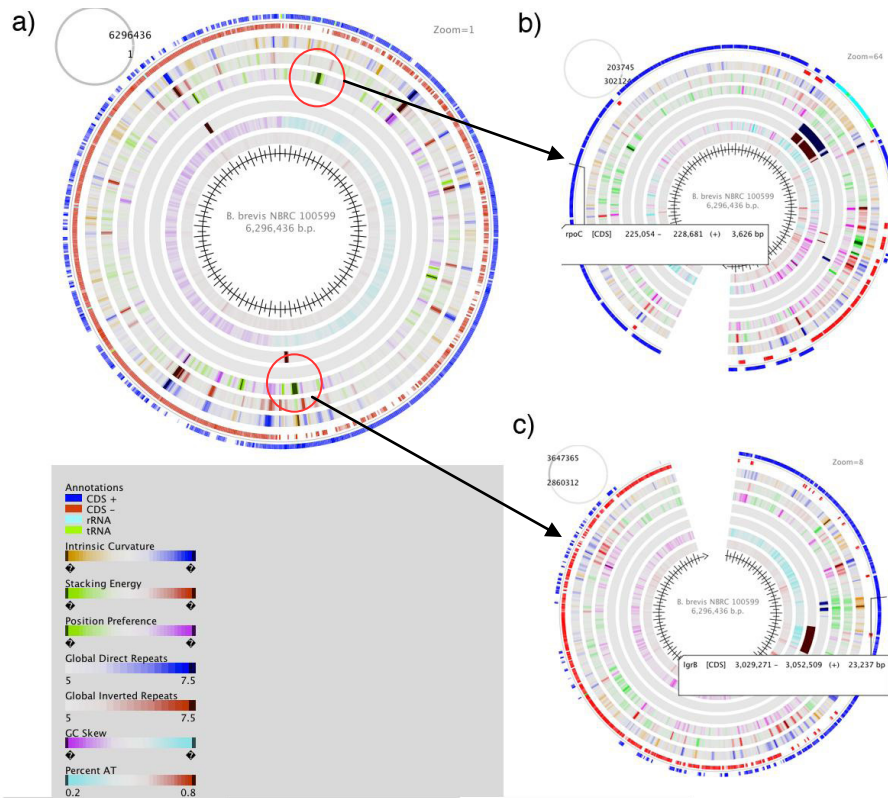


Fig. 5. Zoomable atlases. (a) shows the whole chromosome, and (b) and (c) a zoom of two regions.

also be highly expressed under the right conditions such as sporulation [17]. This ability to view the physical properties of the chromosome in order to identify regions likely to contain highly expressed genes followed by closer analysis by magnification of the region of interest represents a very useful tool for studying bacterial chromosomes.

6 Conclusion

We hope that the Genome Atlas application will result in the development of an application capable of scaling to provide a useful resource for analysis of completed prokaryotic genomes as the number of such projects continues to increase. The use of widely accepted software development patterns such as multi-tier architecture and of existing open source projects wherever possible should provide a robust platform for dealing with the explosion of various new types of high-throughput sequencing data [18], as well as emerging single-molecule methods or the so-called ‘third generation’ sequencing technologies [19].

We have also demonstrated how some of the unique features of the Genome Atlas database, such as the ability to view and sort by the number of predicted ribosomal RNA genes and the zoomable atlases can be used to assist biological investigations.

The only method of interaction with the Genome Atlas currently is a web browser, but future work on the Genome Atlas could include making the application available as a web service based on the Representational State Transfer (REST) architecture [20]. This would allow users to make use of resources provided by the application for via automated tools and incorporated in services of their own.

References

1. Lagesen, K., Ussery, D.W., Wassenaar, T.W.: The One Thousandth Genome - A Cautionary Tale. *Microbiology* 156, 603–608 (2010)
2. Fleischmann, R.D., et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Rd. Science* 269, 496–512 (1995)
3. Flicek, P., Birney, E.: Sense from sequence reads: methods for alignment and assembly. *Nature Methods* 6, S6–S12 (2009)
4. Reeves, G.A., Talavera, D., Thornton, J.M.: Genome and proteome annotation- organization, interpretation and integration. *J. R. Soc. Interface* 6, 129–147 (2009)
5. Wheeler, D.L., et al.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 36, D13–D21 (2008)
6. Kersey, P.J., et al.: Ensembl Genomes: Extending ensembl across the taxonomic space. *Nucleic Acids Research* 38, D563–D569 (2010)
7. Hallin, P.F., Ussery, D.W.: CBS Genome Atlas Database: A dynamic storage for bioinformatic results and sequence data. *Bioinformatics* 20, 3682–3686 (2004)
8. Hallin, P., Stærfeldt, H., Rotenberg, E., Binnewies, T., Benham, C., Ussery, D.: GeneWiz browser: An Interactive Tool for Visualizing Sequenced Chromosomes. *Standards in Genomic Sciences* 1(2), October 14 (2009), <http://standardsingenomics.org/index.php/sigen/article/view/sigs.28177> (Date accessed: August 26, 2010)
9. Stajich, J.E., et al.: The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12, 1611–1618 (2002)
10. Severin, J., Beal, K., Vilella, A., Fitzgerald, S., Schuster, M., Gordon, L., Ureta-Vidal, A., Flicek, P., Herrero, J.: eHive: An Artificial Intelligence workflow system for genomic analysis. *BMC Bioinformatics* 11, 240 (2010)
11. Ramirez, A.O.: Three-Tier Architecture. *Linux Journal* (75), July 01 (2000), <http://www.linuxjournal.com/article/3508>
12. Jensen, L.J., Carsten, F., Ussery, D.W.: Three Views of Microbial Genomes. *Research in Microbiology* 150, 773–777 (1999)
13. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W.: RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108 (2007)
14. Reen, F., Almagro-Moreno, S., Ussery, D., Boyd, E.: The genomic code: inferring Vibrionaceae niche specialization. *Nature Reviews: Microbiology* 4, 697–704 (2006)
15. Worning, P., Jensen, L.J., Hallin, P.F., Stearfeltd, H.H., Ussery, D.W.: Origin of Replication in Circular Prokaryotic Chromosomes. *Environmental Microbiology* 8, 353–361 (2006)

16. Willenbrock, H., Ussery, D.W.: Prediction of highly expressed genes in microbes based on chromatin accessibility. *BMC Mol. Biol.* 13, 8–11 (2007)
17. Nakai, T., Yamauchi, D., Kubota, K.: Enhancement of linear gramicidin expression from *Bacillus brevis* ATCC 8185 by casein peptide. *Biosci. Biotechnol. Biochem.* 69(4), 700–704 (2005)
18. Hawkins, R.D., Hon, G.C., Ren, B.: Next-generation genomics: an integrative approach. *Nature Reviews Genetics* 11, 476–486 (2010)
19. Munroe, D.J., Harris, T.J.: Third-generation sequencing fireworks at Marco Island”. *Nature Biotechnology* 28, 426–428 (2010)
20. Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology Association for Computing Machinery* 2(2), 115–150 (2002)