Overmars *et al. BMC Genomics* 2013, **14**:209
http://www.biomedcentral.com/1471-2164/14/209

BMC Genomics

SOFTWARE Open Access

MGcV: the microbial genomic context viewer for comparative genome analysis

Lex Overmars[1,2,3*], Robert Kerkhoven[1], Roland J Siezen[1,3,4] and Christof Francke[1,2,3,5]

Abstract

Background: Conserved gene context is used in many types of comparative genome analyses. It is used to provide leads on gene function, to guide the discovery of regulatory sequences, but also to aid in the reconstruction of metabolic networks. We present the Microbial Genomic context Viewer (MGcV), an interactive, web-based application tailored to strengthen the practice of manual comparative genome context analysis for bacteria.

Results: MGcV is a versatile, easy-to-use tool that renders a visualization of the genomic context of any set of selected genes, genes within a phylogenetic tree, genomic segments, or regulatory elements. It is tailored to facilitate laborious tasks such as the interactive annotation of gene function, the discovery of regulatory elements, or the sequence-based reconstruction of gene regulatory networks. We illustrate that MGcV can be used in gene function annotation by visually integrating information on prokaryotic genes, like their annotation as available from NCBI with other annotation data such as Pfam domains, sub-cellular location predictions and gene-sequence characteristics such as GC content. We also illustrate the usefulness of the interactive features that allow the graphical selection of genes to facilitate data gathering (e.g. upstream regions, ID's or annotation), in the analysis and reconstruction of transcription regulation. Moreover, putative regulatory elements and their corresponding scores or data from RNA-seq and microarray experiments can be uploaded, visualized and interpreted in (ranked-) comparative context maps. The ranked maps allow the interpretation of predicted regulatory elements and experimental data in light of each other.

Conclusion: MGcV advances the manual comparative analysis of genes and regulatory elements by providing fast and flexible integration of gene related data combined with straightforward data retrieval. MGcV is available at http://mgcv.cmbi.ru.nl.

Keywords: Comparative genomics, Function annotation, Genome visualization, Network reconstruction, Regulatory element, Transcription regulation

Background

The number of sequenced prokaryotic genomes keeps expanding at a rapid pace. As a result, much of the function annotation of genes and other sequence elements relies increasingly on automated pipelines. Despite this tendency, human interference remains indispensable to translate genomic data correctly to biological meaning. Gene context and its evolutionary conservation is one of the genomic properties that can greatly aid the related (manual) genome analyses. The gene context provides many clues concerning function and biological role of a gene in a prokaryote [1,2]. Gene context data thus benefits the reconstruction of the metabolic network [3-5]. Moreover, conserved gene context can also be applied to guide the identification of regulatory elements and therewith the reconstruction of the transcription regulatory network (e.g. [6-9]).

From a practical point of view, a comprehensive visualization of genomics data and information on function facilitates the process of data integration, and thereby reduces the time needed for interpretation. There are several ways to achieve this goal, as reflected by the variety in genome browsers and annotation platforms that have been developed. Conventional genome

* Correspondence: L.Overmars@cmbi.ru.nl
[1] Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Geert Grooteplein Zuid 26-28, 6525GA, Nijmegen, The Netherlands
[2] TI Food and Nutrition, P.O. Box 5576700AN, Wageningen, The Netherlands
Full list of author information is available at the end of the article

browsers include for instance UCSC genome browser [10], Artemis [11] and GBrowse [12]. This type of genome browser is characterized by a generic, highly configurable setup (i.e. typically, users can upload their genomes in genbank- and/or gff3-format) and display genomic data in separate 'tracks'. On the other hand, resources such as IMG [13], Microscope [14], MicrobesOnline [15] and the SEED [16] serve as annotation platforms by providing the user genomic data, analysis tools and visualization options. In 2004 we introduced the Microbial Genome Viewer [17]. This web-based genome viewer allowed users to explore bacterial genomes in linear maps and create a genome-wide visualization of data in circular maps. Yet, other tools have a more specific focus. For instance, BAGET allows users to retrieve the gene-context for a single gene [18], whereas GeConT 2 allows users to visualize the genomic context of query genes [19]. Some tools specifically address conservation of gene order between orthologous genes, also denoted as "synteny". For instance, GeneclusterViz [20], GCView [21], PSAT [22] and Absynte [23] provide a local gene context comparison based on blast (–like) similarity searches.

In the public domain, various resources provide organism specific reconstructions of particular regulons through the integration of genome sequence data and stored motifs. Examples of these are PEPPER [24], RegulonDB [25], RegTransBase [26], PRODORIC [27], RegPrecise [28], ProdoNet [29], FITBAR [30], RegAnalyst [31] and MicrobesOnline [15]. Most of these resources enable automated predictions of regulatory sites based on stored motifs collected from literature. Some resources also in addition allow for *de novo* motif discovery, using tools such as MEME [32], Tmod [33] and GIMSAN [34], which were developed to identify significantly overrepresented sequence motifs.

The versatility of the above resources comes at the cost of some flexibility and speed. We have therefore developed the web-application MGcV, which aims specifically to serve as an integrative visual interface to speed up a manual genome analysis. MGcV is a light-weight and flexible viewer that provides: i) a comparative view of the genomic context for query genome segments, like genes, sets of genes, or (user defined-) gene trees; ii) the integration of information on gene function enriched with additional annotation data such Pfam domains and sub-cellular location-predictions within a single 'track'; iii) the possibility to visually select genes and extract diverse gene-linked information, like upstream regions, protein sequence or function annotation; and iv) the possibility to upload and integrate experimental data and user-defined regulatory elements in adaptable views. MGcV thus enables the exploitation of gene context information in the annotation of gene function, the analysis of the evolutionary conservation of that context, the recovery of associated regulatory elements and the ranked comparative view of the identified elements in combination with microarray- or RNA-seq data. Hereby MGcV provides a visual heart to the manual sequence-based analysis of gene-function and gene-regulation in bacteria.

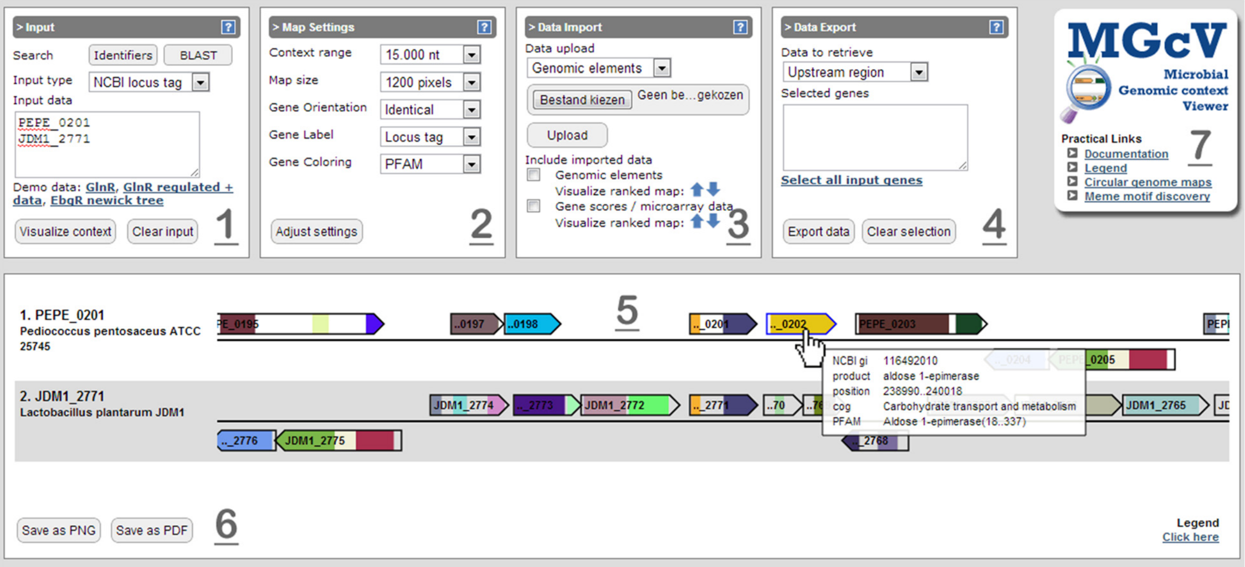
Methods

Data resources

The genome and protein sequences, the associated gene identifiers and function annotations (e.g. trivial names, COG categories, protein names) of all publicly available bacterial genomes are obtained from the FTP server of NCBI RefSeq [35,36]. Uniprot accessions mapped to NCBI GI-codes are retrieved from the Uniprot FTP server [37,38]. Pfam domains are obtained from the FTP server of EBI [39,40]. Gene-sequence characteristics like GC-content are calculated using in-house scripts. Subcellular location predictions are obtained from the PSORTdb website [41,42]. The data is updated on a weekly basis and stored in a local MySQL database to enable fast access. The microarray data that are used to illustrate the capabilities of MGcV in the second case study were taken from [43].

Implementation

MGcV is a web-application developed using a combination of python, javascript and SVG (Scalable Vector Graphics). We implemented MGcV as a single page application; the front-end makes server side calls through Jquery and AJAX and receives the response from the server. The interface consists of four boxes (see Figure 1). From left to right, these include: i) data input; ii) map settings; iii) data import and; iv) data export. The user can provide four types of input, NCBI RefSeq GI's, NCBI locus tags, genomic positions (NCBI Refseq genome accession *tab* start *tab* stop), or a Newick-tree (the leaf labels must contain NCBI-GI code). In addition, the user can search for specific genes by providing for instance gene product or gene names or by performing a BLAST search. Query data supplied by the user is parsed and mapped to the corresponding gene context data using python scripts. Uploaded (phylogenetic) trees are processed with Newick utilities [44], which is also used to create a visual representation of the tree in SVG format. For the COG (NCBI) and Pfam annotations in the gene context maps a color scheme was assigned by designating a unique color to each identifier. In a similar way colors were assigned to the different protein location predictions. The applied color schemes can be found in the legend. Gene-associated quantitative data (e.g. GC-content) are converted in a red-to-green gradient which is projected on top of the genes. Likewise, gene-associated quantitative data uploaded by the user



Search: Identifiers | BLAST
Input type: NCBI locus tag
Input data:
PEPE_0201
JDM1_2771

Context range: 15.000 nt
Map size: 1200 pixels
Gene Orientation: Identical
Gene Label: Locus tag
Gene Coloring: PFAM

Data upload: Genomic elements
Bestand kiezen Geen be...gekozen
Include imported data
☐ Genomic elements
Visualize ranked map:
☐ Gene scores / microarray data
Visualize ranked map:

Data to retrieve: Upstream region
Selected genes:

1. PEPE_0201
Pediococcus pentosaceus ATCC 25745

2. JDM1_2771
Lactobacillus plantarum JDM1

NCBI gi 116492010
product aldose 1-epimerase
position 238990..240018
cog Carbohydrate transport and metabolism
PFAM Aldose 1-epimerase(18..337)

Figure 1 The interface of MGcV. The interface comprises seven separate modules. 1) Input: Copy/paste the input containing regions of interest and select the appropriate input-type. The context is visualized by clicking "Visualize context". Input-types include lists of identifiers such NCBI GI-codes or NCBI locus tags, genomic regions, or phylogenetic trees in newick format. Identifiers can also be searched for using the buttons "Identifier" and "BLAST". 2) In the map settings, the appearance of the comparative context map can be altered. Various settings can be changed, which include the scale, the orientation, the gene-label and the data on display (e.g. COG coloring, PFAM, protein location). To apply the changes the "Adjust settings"-button has to be clicked. 3) Data can be added via data import. Users can upload the position of genomic elements (e.g. regulatory elements) and gene-associated quantitative data (e.g. microarray data). Subsequently, the user can include the data in the map or generate a (ranked) comparative context map based on the uploaded data (blue arrows). 4) Sequence and functional data like upstream regions, gene sequences, cog categories, annotations can be obtained for visually selected genes via data export. Genes can be included or excluded by clicking. 5) The functional data related to a particular gene context is displayed in a single lane in the comparative genome map, thereby allowing for a direct comparison of the data between various strains or species. The maps are interactive; hovering the mouse over the map will display additional information (e.g. NCBI GI, gene product, COG) and clicking genes will tag them for data export. 6) The map can be converted to PNG- or PDF- format. 7) The interface is linked to a tutorial, to a color legend, to the circular viewer of the original MGV, and also directly to the MEME website.

(e.g. microarray- or RNAseq- data) are converted in a red-to-green gradient. These data are then projected in a horizontal bar below the genes to allow the visual integration with annotation data and regulatory element predictions. Generated maps can be downloaded in SVG, PNG or PDF format. The conversion of SVG to PNG and PDF is done using "Batik Rasterizer" [45]. The interface and interactive maps allow the user to interact with the data. Map interactivity is achieved by ECMAscript; linked information on genes and other sequence elements can be inspected by mouse-over, whereas a mouse-click can be used to select genes for subsequent analysis and data retrieval. MGcV is operable in modern browsers like Firefox, Chrome and Internet Explorer, where for all browsers the latest version is recommended.

Results

Interface and functionality

Function annotation

The appropriate annotation of encoded function is essential for the correct interpretation of genomics data. The annotation process is initiated by the selection of genes and/or regions of interest. The flexible set-up of MGcV allows to generate an initial comparative context map simply by uploading a single identifier or a list of identifiers, like derived from a BLAST search, suffices to generate an initial comparative context map in MGcV. The uploaded identifiers may include NCBI gi-codes (RefSeq [36]), NCBI locus tags or genomic locations (designated by a RefSeq genome accession and position). In case the user does not have a list of gene identifiers, genes and their corresponding identifiers can be obtained via the built-in gene-search (input-box option "Identifiers"). In addition, a BLAST search can be performed to find proteins similar to a given protein sequence. The BLAST hits can be selected and used as input for MGcV. We have also implemented the possibility to upload and visualize any (phylogenetic) gene tree. The combined view of gene phylogeny and the gene context allows a quick evaluation of the potential for similarity in molecular function and biological role between the selected genes. The labeling of the genes (i.e. by trivial name, by locus tag, or by NCBI GI-code), and similarly, the coloring of the genes (i.e. by COG category

[46], by GC%, by sub cellular location [47] or by Pfam domain [40]) enhances the evaluation process. In addition, the genomic range of the maps can be altered and an identical orientation of the genes of interest can be enforced for purposes of presentation. The added value of MGcV in the manual function annotation is illustrated in more detail below (first case study).

Identification and comparison of regulatory elements

The starting point for a sequence-based reconstruction of transcription regulation is the identification of genes whose upstream region might contain a regulatory element, like a transcription factor (TF) binding site (e.g. [6,7,9,48]). We and others have shown that the identification of specific TF binding sites is particularly successful in the case of conserved gene context (e.g. [8,49,50]). We experienced that the ability to select upstream regions on basis of a visual representation of that context considerably speeds up the analysis and therefore have implemented this upstream region selection in MGcV. Moreover, we have added a “data import” option to allow the visualization of the predicted location of regulatory elements together with microarray or RNA-seq data. In this way, the location prediction of regulatory elements and the experimental data can be interpreted more easily in light of each other. In addition, the view can be ranked according to similarity score (for binding site predictions) or expression ratio (for microarray or RNA seq data). In fact, such a ranked view of expression data and gene context is also extremely useful in the interpretation of transcriptome experiments. The new features are illustrated below in the second case study.

Data export

An important aspect of data integration in comparative genome analyses is the combination of sequence and, sequence and function identifiers. Collecting these identifiers for a selected set of genes can be time-consuming, especially when the information linked to the genes found associated on the genome has to be included. We have added a “data export” option in MGcV to accommodate the rapid and comprehensive collection of gene-related data. The user can graphically select genes of interest by mouse-click, where the selected genes are highlighted and included in the “data export”-box. Subsequently, the data to be retrieved can be selected. These include for example upstream DNA sequences, protein sequences or function-related data like for instance: length, protein function, COG category or Pfam domains. The export option can be used without actually using the context view to, for instance, collect quickly the protein sequence or Uniprot accession codes for a set of gene IDs.

Case studies that illustrate the practical application of MGcV in manual comparative genome analysis

The main difference between MGcV and other resources is that MGvC is aimed to provide a platform to visually integrate one’s own data (i.e. data generated externally using other tools or obtained through experimentation) with annotation data and practical export options that enable further (external) analysis. Other resources, like for instance MicrobesOnline [15], in principle aim to offer a platform that is inclusive, i.e. that includes both calculation and visualization. Below we describe the results of two different manual comparative genome analysis using MGcV. In these two examples we highlight the flexible functionality of MGcV by visualizing the gene context and the associated functional information for a set of homologs that are present in a phylogenetic tree and by the visual integration of microarray data and de novo predictions of putative binding sites.

Case study 1: beta-galactosidases and associated regulators in Lactobacillus plantarum

The study of the *lac* operon and its control in *Escherichia coli* has set the paradigm of bacterial transcription regulation. The associated regulator in *E. coli* was named LacI. Most bacteria contain multiple homologs of this transcription factor family. In lactic acid bacteria, the *lac* operon is often associated with LacI-family regulators that form a separate clade within the family (e.g. EbgR in *E. coli*) [8]. *L. plantarum* WCFS1 has three regulators that belong to this clade: LacR (ortholog in *B. subtilis* [51]), GalR (ortholog in *S. thermophilus* [52]) and RafR [53]. To find functional equivalents in other Lactobacilli, the protein sequences of homologs were collected using a BLAST search. The recovered protein sequences were aligned and a neighbor-joining tree was constructed. To determine the degree of conservation the tree was used as input for MGcV. As shown in Figure 2A, integration with gene-context enhances both the interpretation of the tree and the identification of orthologs. Based on the integrated visualization of the phylogenetic tree and genomic context we can easily distinguish three different clusters. One of the first things that can be done on basis of the integrated view is a specification of annotation information as present in the NCBI database for orthologous genes that share context. The genes *JDM1_2771* (*galR1*) and *JDM_2780* (*galR2*) can be easily re-annotated to *galR* and *rafR*, respectively, on basis of the specific annotation that is available for *L. plantarum WCFS1*. Also the functional equivalency between genes can be evaluated, like for the gene *Lreu_1775*, which is the only regulator of e*bgR*-type in the *Lactobacillus reuterii* genome. Based on the tree and the fact that the gene has a similar gene context as *JDM1_2771* (*galR*) and *LVIS_1901* and not as *JDM_2780*

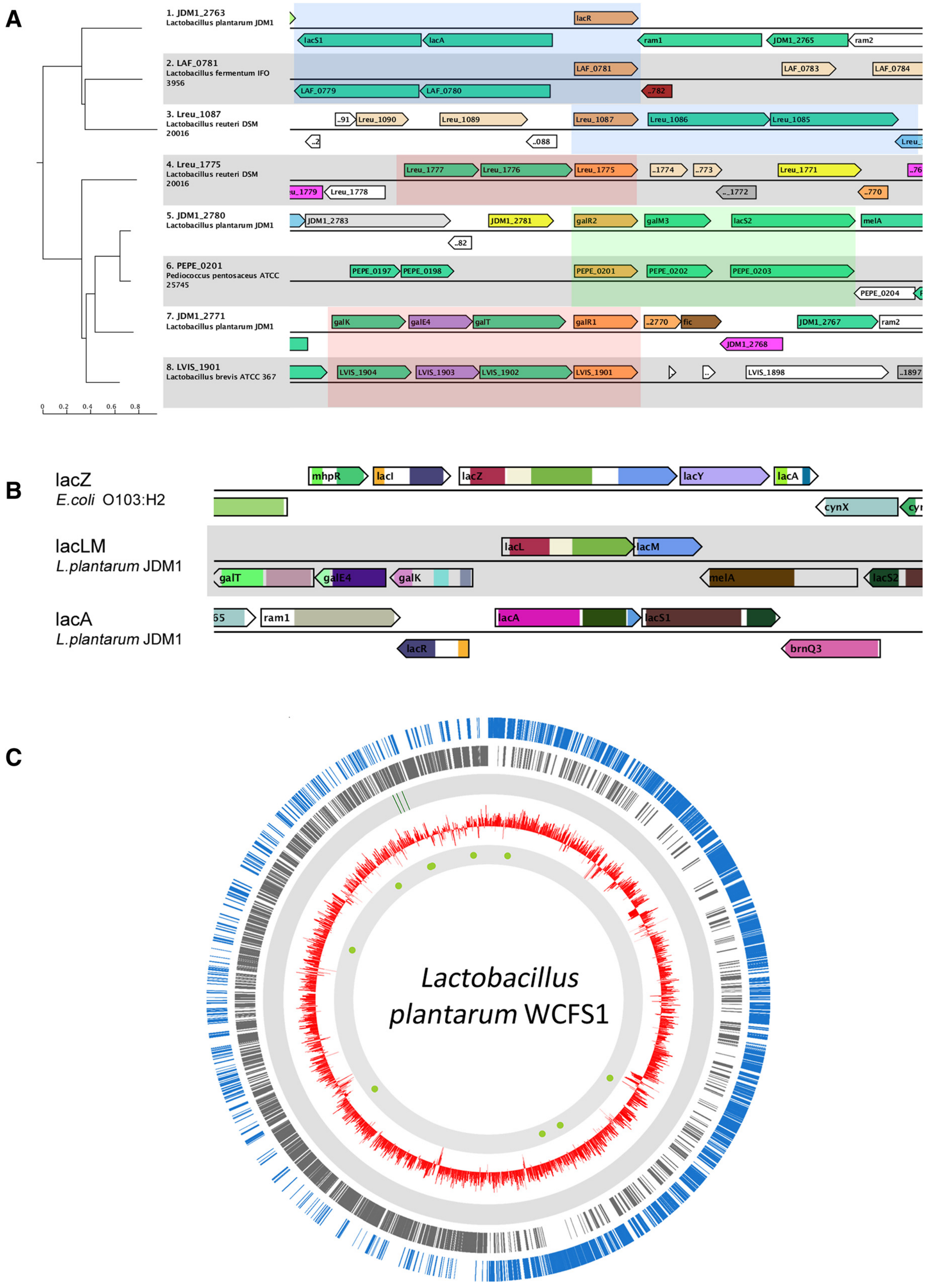


Figure 2 (See legend on next page.)

(See figure on previous page.)
Figure 2 EbgR-like transcription factors in *L. plantarum* and other lactobacilli. A) MGcV visualization of a phylogenetic tree of EbgR-type regulators in some Lactobacilli. To simplify, the tree was pruned (species: *Lactobacillus plantarum JDM1* and *WCFS1*, *Lactobacillus fermentum IFO3956*, *Lactobacillus reuteri DSM20016*, *Lactobacillus brevis ATCC 367* and *Pediococcus pentosaceus*). The context range was set to 10.000 nucleotides, genes were colored by COG-class and trivial names were used to label genes. The visual combination of the phylogenetic tree and genomic context allows to distinguish three groups; lacR-, rafR- and galR-like sequences (designated in blue, green and red, respectively). **B**) Comparative context map of the beta-galactosidase encoding genes *lacZ* (*E. coli*), *lacLM* (*L. plantarum*) and *lacA* (*L. plantarum*). Pfam domains are used for gene-coloring and trivial names are used to label genes. In *L. plantarum lacLM* and *lacA* are both annotated to encode a protein with the same name (beta-galactosidase). Yet, from this comparative view it becomes clear that *lacLM* and *lacA* must have a different evolutionary origin. Although lacLM is encoded by two genes, the domain structure appears identical to the single *lacZ* gene of *E. coli*. **C**) A circular genome map of *L. plantarum* in which the ORFs on the plus strand (blue), on the minus stand (grey), the locations of regulator encoding genes *lacR*, *rafR* and *galR* (green), the GC% (red) and putative binding sites (similarity to motif >90% [8]; represented by the green dots) are included.

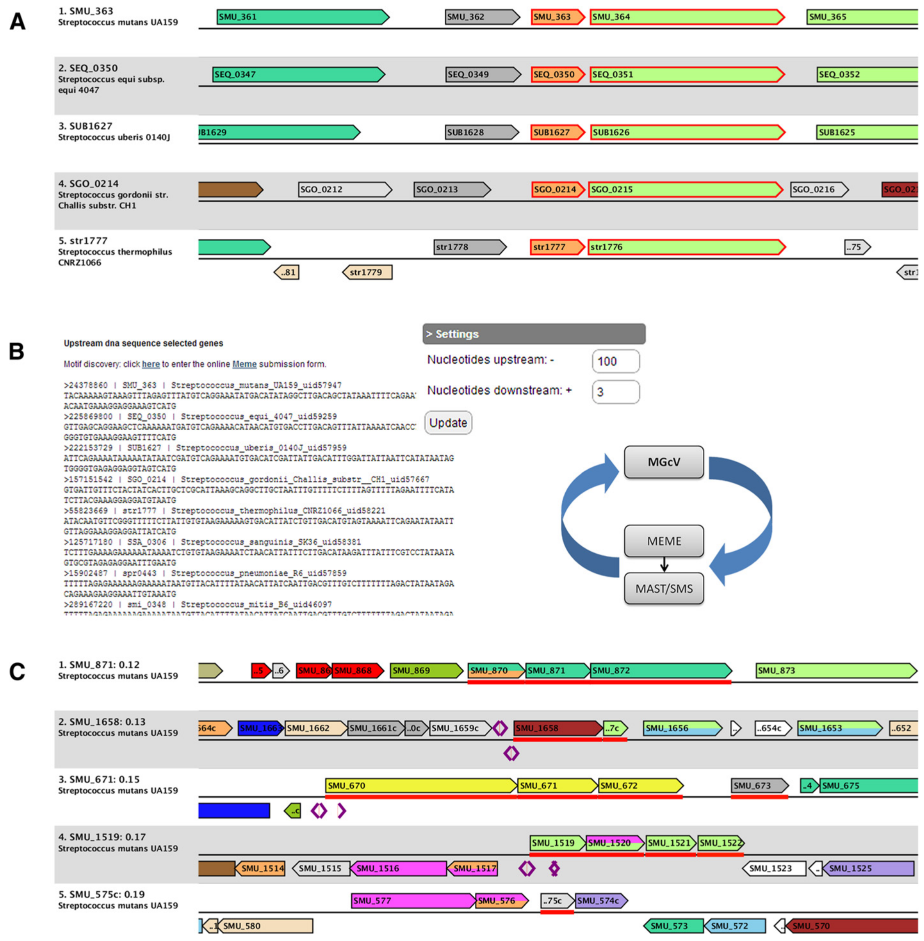
(*rafR*) and *PEPE_0201*, it can be annotated as *galR* with the expected inducer galactose.

The production of galacto-oligosaccharides using microbial beta-galactosidases is currently well-studied in the field of functional foods [54]. In *Escherichia coli* a gene encoding beta-galactosidase: *lacZ*, was described first by Joshua Lederberg in 1948 [55]. It took 25 years before a second beta-galactosidase encoding gene was described [56], which was designated *ebgA* from evolved beta-galactosidase. The discovery resulted in the classic study (designation by [57]) of molecular evolution (review in [58]). The Pfam and COG classification (Figure 2B) comply with the assertion that both genes have evolved from a common ancestor. In many lactobacilli a third closely-related variant is found, *lacLM*. In some Lactobacilli (e.g. *L. delbrueckii* and *L. salivarius*) the protein is encoded by a single gene. However, in most Lactobacilli the protein is encoded by two neighboring genes (probably the result of gene fission) and the active protein is a heterodimer [59]. It is the LacLM protein that is mostly exploited in biotechnological applications [60,61]. Like *E. coli*, various Lactobacilli have a second beta-galactosidase encoding gene, *lacA*. However, this gene has a completely different evolutionary origin and thus represents a functional analog. This conclusion can also easily be derived from the (pfam-) annotation information that is available in MGcV (Figure 2B).

We have maintained the circular viewer of the original MGV in which we constructed a circular genome map of *L. plantarum* (Figure 2C). In this map we included the locations of regulator-encoding genes *lacR*, *rafR* and *galR*, the GC-percentage and putative binding sites (similarity to motif >90% [8]). The genomic segment containing *lacR, rafR* and *galR* is flanking a region with a decreased GC-percentage, which was suggested to represent a lifestyle adaptation region in which many genes are acquired by horizontal gene transfer [62].

Case study 2: Reconstruction of GlnR-mediated regulation in Streptococcus mutans

Recently, we have published a comparative genomics study on the transcription factor GlnR [63]. GlnR is one of the four major transcription factors involved in the control of central nitrogen metabolism in *Bacillus subtilis*. A BLAST search was performed to retrieve GlnR orthologs from all sequenced Streptococcal genomes and the gene context for the resulting list was displayed in MGcV (see Figure 3A). We observed a clear conservation of the *glnRA* operon and its genomic context in all Streptococcaceae. MGcV was then used to collect selected upstream regions (Figure 3B). These were analyzed using MEME (via the available link; [32]) to search for a motif representing the GlnR-binding site. The motif was then refined and used to identify and score putative binding sites on all Streptococcal genomes (via e.g. MAST [32] or a Similar Motif Search [64]). Subsequently, the resulting list of putative sites with their corresponding similarity scores was uploaded to MGcV. The view was ranked according to similarity score and the binding site predictions could be evaluated in light of their position relative to the genes. Then, the consistency of the predictions with microarray data was checked visually in MGcV. Chen and colleagues constructed a GlnR gene knockout in *Streptococcus mutans* for which they performed a microarray experiment [43]. These data were retrieved and uploaded and the view was ranked according to microarray ratios (Figure 3C). The view makes immediately clear that the predicted binding-sites are consistent with the microarray data. In addition, the view shows that the operon showing the strongest response (consisting of SMU_870, SMU_871 and SMU_872) is not preceded by a putative binding-site and therefore probably is regulated indirectly. In fact, this operon encodes a PTS system for which no functional relation with nitrogen is described. Interestingly, many of the high-scoring putative binding sites are followed by a binding site in the N-terminus encoding part of the gene (Figure 3C: SMU_671 and SMU_1519), suggesting that this might be a particularity of the regulatory mechanism. Finally, the interactive map provides a convenient overview to determine a possible score



Upstream dna sequence selected genes

Motif discovery: click here to enter the online Meme submission form.

```
>24378860 | SMU_363 | Streptococcus_mutans_UA159_uid57947
TACAAAAAGTAAAGTTTAGAGTTTATGTCAGGAAATATGACATATAGGCTTGACAGCTATAAATTTTCAGAA
ACAATGAAAGGAGGAAAGTCATG
>225869800 | SEQ_0350 | Streptococcus_equi_4047_uid59259
GTTGAGCAGGAAGCTCAAAAAATGATGTCAGAAAACATAACATGTGACCTTGACAGTTTATTAAAATCAACC
GGGTGTGAAAGGAAGTTTTCATG
>222153729 | SUB1627 | Streptococcus_uberis_0140J_uid57959
ATTCAGAAAATAAAAATATAATCGATGTCAGAAAATGTGACATCGATTATTGACATTTGGATTATTAATTCATATAATAG
TGGGGTGAGAGGAGGTAGTCATG
>157151542 | SGO_0214 | Streptococcus_gordonii_Challis_substr__CH1_uid57667
GTGATTGTTTCTACTATCACTTGCTCGCATTAAAGCAGGCTTGCTAATTTGTTTTTCTTTTAGTTTTTAGAATTTTCATA
TCTTACGAAAGGAGGATGTAATG
>55823669 | str1777 | Streptococcus_thermophilus_CNRZ1066_uid58221
ATACAATGTTCGGGTTTTTCTTATTGTGTAAGAAAAAGTGACATTATCTGTTGACATGTAGTAAAATTCAGAATATAATT
GTTAGGAAAGGAGGATTATCATG
>125717180 | SSA_0306 | Streptococcus_sanguinis_SK36_uid58381
TCTTTGAAAAGAAAAAATAAAATCTGTGTAAGAAAATCTAACATTATTTCTTGACATAAGATTTATTTCGTCCTATAATA
GTGCGTAGAGAGGAATTTGAATG
>15902487 | spr0443 | Streptococcus_pneumoniae_R6_uid57859
TTTTTAGAGAAAAAAGAAAAATAATGTTACATTTTATAACATTATCAATTGACGTTTGTCTTTTTTTAGACTATAATAGA
CAGAAAGAAGGAAATTGTAAATG
>289167220 | smi_0348 | Streptococcus_mitis_B6_uid46097
TTTTTAGAGAAAAAAGAAAAATAATGTTACATTTTATAACATTATCAATTGACGTTTGTCTTTTTTTAGACTATAATAGA
```

Figure 3 Application of MGcV in the reconstruction of GlnR-mediated regulation in *Streptococcus mutans*. A) Comparative context map of GlnR homologs obtained via a BLAST search in all sequenced Streptococci (only five species are shown). The GlnRA operon and its direct genomic context is clearly conserved in the Streptococci. The map was used to graphically select those genes that could be preceded by a binding site. **B**) The "upstream region" option of the "Data-export"-box was used to obtain upstream regions of the selected genes. Subsequently, the available link to MEME was used to search for possible overrepresented motifs. The results were refined and a motif defined [63], which was then used to search and score putative binding sites (e.g. using MAST or SMS). **C**) A comparative context map ranked on expression ratios (low-to-high) of a GlnR mutant, visualized in conjunction with predicted GlnR binding sites. To exemplify, the figure is limited to the top 5 of down-regulated genes. In this map, gene expression ratios are represented in a colored bar (red-to-green gradient; red is down-regulated) at the baseline and putative binding sites are designated by purple arrows (direction representing the strand). Both the microarray data and putative binding sites and their corresponding binding sites were uploaded using the "Data import"-box. The resulting map allows the analysis of the putative GlnR binding sites in light of the expression data of the GlnR mutant. Most of the top down regulated genes (SMU_1658, SMU_671 and SMU,1519 in Figure 3C) indeed are preceded by a putative binding site.

threshold for both the predictions as well as the expression data.

Conclusion

Gene-context conservation is an important genomic property to exploit in genome analyses. Nine years ago we developed a Microbial Genome Viewer [17] to support our efforts in the gene annotation and metabolic reconstruction of the lactic acid bacterium *Lactobacillus plantarum WCFS1* [65,66]. Over the years we have experienced the need for additional functionality and more flexibility to enhance the work on the curation of function annotation and on the reconstruction of transcription regulatory networks. While maintaining the functionality, we have changed the complete setup and developed a new interface to create an adaptable interactive Microbial Genome context Viewer with high speed and versatile functionality to aid small-scale analyses. Both the input and output options of MGcV provide many practical features. The interactive maps allow users to graphically select sets of genes for data retrieval and subsequent analyses. Moreover, the maps provide a single integrated view of the data. The maps are made available in SVG, PNG and PDF format and are hereby suited to use as illustrations in publications, posters and presentations. The MGcV features that constitute its value to the manual analysis of genome sequence include: i) its light-weight and flexible interface; ii) the possibility to a) select multiple genes in the maps and extract gene-related data for these; and b) extract selected upstream regions to be used for further analysis; iii) the visual integration of a user-defined phylogenetic tree and the related gene context; and iv) the visual integration and ranking of microarray data or regulatory element predictions in the context of gene organization. Regarding the regulatory elements, any list of positions linked to a quantitative score can be uploaded, ranked and viewed. Possible applications of MGcV include: annotation refinement, function prediction on basis of a (phylogenetic) tree and conserved gene context, the sequence-based reconstruction of gene regulatory networks, and microarray/RNA-seq data analysis. We have presented two case studies to illustrate the practical applications of MGcV. Altogether, MGcV provides a flexible platform to exploit publicly available genomic data in small scale genome analysis in a fast and convenient manner.

Availability and requirements

Project name: MGcV
Project home page: http://mgcv.cmbi.ru.nl
Operating system(s): Platform independent
Programming language: Python/SVG/Javascript
Other requirements: Internet browser supporting SVG (Scalable Vector Graphics)
License: None required.
Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

LO and CF designed and coordinated the project. LO developed the web-application and wrote the manuscript. RK contributed to the development of the web-server scripts. RS and CF revised the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). We thank Marieke Bart, Tom Groot Kormelink, Lennart Backus and Mark de Been for their contributions to the project.

Author details

[1]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Geert Grooteplein Zuid 26-28, 6525GA, Nijmegen, The Netherlands. [2]TI Food and Nutrition, P.O. Box 5576700AN, Wageningen, The Netherlands. [3]Netherlands Bioinformatics Centre, P.O. Box 91016500HB, Nijmegen, The Netherlands. [4]Microbial Bioinformatics, Eikelakkers 2, 6711TE, Ede, The Netherlands. [5]CFLSc, Prinsenhof 12, 5616TE, Eindhoven, The Netherlands.

Received: 22 November 2012 Accepted: 22 March 2013
Published: 1 April 2013

References

1. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11**(3):356–372.
2. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10**(3):366–370.
3. Francke C, Siezen RJ, Teusink B: **Reconstructing the metabolic network of a bacterium from its genome.** *Trends Microbiol* 2005, **13**(11):550–558.
4. Park JM, Kim TY, Lee SY: **Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses.** *Proc Natl Acad Sci USA* 2010, **107**(33):14931–14936.
5. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nat Protoc* 2010, **5**(1):93–121.
6. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003, **2**(2):11.
7. Rodionov DA: **Comparative genomic reconstruction of transcriptional regulatory networks in bacteria.** *Chem Rev* 2007, **107**(8):3467–3497.
8. Francke C, Kerkhoven R, Wels M, Siezen RJ: **A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in Lactobacillus plantarum WCFS1.** *BMC Genomics* 2008, **9**:145.
9. Neph S, Tompa M: **MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes.** *Nucleic Acids Res* 2006, **34**(Web Server issue): W366–W368.
10. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, *et al*: **The UCSC Genome Browser database: update 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D876–D882.
11. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**(10):944–945.
12. Podicheti R, Gollapudi R, Dong Q: **WebGBrowse–a web server for GBrowse.** *Bioinformatics* 2009, **25**(12):1550–1551.
13. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al*: **IMG: the integrated microbial genomes database and comparative analysis system.** *Nucleic Acids Res* 2012, **40**(Database issue):D115–D122.
14. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C, *et al*: **MicroScope: a platform for microbial genome annotation and comparative genomics.** *Database (Oxford)* 2009, **2009**:bap021.
15. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, *et al*: **MicrobesOnline: an integrated portal for comparative and functional genomics.** *Nucleic Acids Res* 2010, **38**(Database issue):D396–D400.
16. 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.
17. Kerkhoven R, van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ: **Visualization for genomics: the microbial genome viewer.** *Bioinformatics* 2004, **20**(11):1812–1814.
18. Oberto J: **BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence.** *Bioinformatics* 2008, **24**(3):424–425.
19. Martinez-Guerrero CE, Ciria R, Abreu-Goodger C, Moreno-Hagelsieb G, Merino E: **GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W176–W180.
20. Pejaver VR, An J, Rhee S, Bhan A, Choi JH, Liu B, Lee H, Brown PJ, Kysela D, Brun YV, *et al*: **GeneclusterViz: a tool for conserved gene cluster visualization, exploration and analysis.** *Bioinformatics* 2012, **28**(11):1527–1529.
21. Grin I, Linke D: **GCView: the genomic context viewer for protein homology searches.** *Nucleic Acids Res* 2011, **39**(Web Server issue): W353–W356.
22. Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ: **PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes.** *BMC Bioinformatics* 2008, **9**:170.
23. Despalins A, Marsit S, Oberto J: **Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters.** *Bioinformatics* 2011, **27**(20):2905–2906.

24. de Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J: **PePPER: a webserver for prediction of prokaryote promoter elements and regulons.** *BMC Genomics* 2012, **13:**299.
25. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A, *et al*: **RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).** *Nucleic Acids Res* 2011, **39**(Database issue):D98–D105.
26. Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I: **RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes.** *Nucleic Acids Res* 2007, **35**(Database issue):D407–D412.
27. Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Munch R: **PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes.** *Nucleic Acids Res* 2009, **37**(Database issue):D61–D65.
28. Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA: **RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes.** *Nucleic Acids Res* 2010, **38**(Database issue):D111–D118.
29. Klein J, Leupold S, Munch R, Pommerenke C, Johl T, Karst U, Jansch L, Jahn D, Retter I: **ProdoNet: identification and visualization of prokaryotic gene regulatory and metabolic networks.** *Nucleic Acids Res* 2008, 36(Web Server issue):W460–W464.
30. Oberto J: **FITBAR: a web tool for the robust prediction of prokaryotic regulons.** *BMC Bioinformatics* 2010, **11:**554.
31. Sharma D, Mohanty D, Surolia A: **RegAnalyst: a web interface for the analysis of regulatory motifs, networks and pathways.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W193–W201.
32. Bailey T, Boden M, Buske F, Frith M, Grant C, Clementi L, Ren J, Li W, Noble W: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W202–W208.
33. Sun H, Yuan Y, Wu Y, Liu H, Liu JS, Xie H: **Tmod: toolbox of motif discovery.** *Bioinformatics* 2010, **26**(3):405–407.
34. Ng P, Keich U: **GIMSAN: a Gibbs motif finder with significance analysis.** *Bioinformatics* 2008, **24**(19):2256–2257.
35. *NCBI RefSeq FTP server.* [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/].
36. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Res* 2012, **40**(Database issue):D130–D135.
37. *Uniprot FTP server.* [ftp://ftp.uniprot.org/].
38. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database (Oxford)* 2011, **2011:**bar009.
39. *EBI FTP server.* [ftp://ftp.ebi.ac.uk/pub/databases/Pfam/].
40. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(Database issue):D290–D301.
41. *PSORTdb.* [http://db.psort.org/].
42. Yu NY, Laird MR, Spencer C, Brinkman FS: **PSORTdb–an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea.** *Nucleic Acids Res* 2011, **39**(Database issue):D241–D244.
43. Chen PM, Chen YY, Yu SL, Sher S, Lai CH, Chia JS: **Role of GlnR in acid-mediated repression of genes encoding proteins involved in glutamine and glutamate metabolism in Streptococcus mutans.** *Appl Environ Microbiol* 2010, **76**(8):2478–2486.
44. Junier T, Zdobnov EM: **The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell.** *Bioinformatics* 2010, **26**(13):1669–1670.
45. *Batik Rasterizer.* [http://xmlgraphics.apache.org/batik/tools/rasterizer.html].
46. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
47. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, *et al*: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26**(13):1608–1615.
48. Prakash A, Tompa M: **Discovery of regulatory elements in vertebrates through comparative genomics.** *Nat Biotechnol* 2005, **23**(10):1249–1256.
49. de Been M, Bart MJ, Abee T, Siezen RJ, Francke C: **The identification of response regulator-specific binding sites reveals new roles of two-component systems in Bacillus cereus and closely related low-GC Gram-positives.** *Environ Microbiol* 2008, **10**(10):2796–2809.
50. Rodionov DA, Gelfand MS: **Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling.** *Trends Genet* 2005, **21**(7):385–389.
51. Daniel RA, Haiech J, Denizot F, Errington J: **Isolation and characterization of the lacA gene encoding beta-galactosidase in Bacillus subtilis and a regulator gene, lacR.** *J Bacteriol* 1997, **179**(17):5636–5638.
52. Ajdic D, Ferretti JJ: **Transcriptional regulation of the Streptococcus mutans gal operon by the GalR repressor.** *J Bacteriol* 1998, **180**(21):5727–5732.
53. Silvestroni A, Connes C, Sesma F, De Giori GS, Piard JC: **Characterization of the melA locus for alpha-galactosidase in Lactobacillus plantarum.** *Appl Environ Microbiol* 2002, **68**(11):5464–5471.
54. Park AR, Oh DK: **Galacto-oligosaccharide production using microbial beta-galactosidase: current state and perspectives.** *Appl Microbiol Biotechnol* 2010, **85**(5):1279–1286.
55. Lederberg J: **Gene control of beta-galactosidase in Escherichia coli.** *Genetics* 1948, **33**(6):617.
56. Campbell JH, Lengyel JA, Langridge J: **Evolution of a second gene for beta-galactosidase in Escherichia coli.** *Proc Natl Acad Sci USA* 1973, **70**(6):1841–1845.
57. Dean AM: **The future of molecular evolution.** *EMBO Rep* 2010, **11**(6):409.
58. Hall BG: **The EBG system of E. coli: origin and evolution of a novel beta-galactosidase for the metabolism of lactose.** *Genetica* 2003, **118**(2–3):143–156.
59. Nguyen TH, Splechtna B, Krasteva S, Kneifel W, Kulbe KD, Divne C, Haltrich D: **Characterization and molecular cloning of a heterodimeric beta-galactosidase from the probiotic strain Lactobacillus acidophilus R22.** *FEMS Microbiol Lett* 2007, **269**(1):136–144.
60. Nguyen TH, Splechtna B, Yamabhai M, Haltrich D, Peterbauer C: **Cloning and expression of the beta-galactosidase genes from Lactobacillus reuteri in Escherichia coli.** *J Biotechnol* 2007, **129**(4):581–591.
61. Liu GX, Kong J, Lu WW, Kong WT, Tian H, Tian XY, Huo GC: **Beta-Galactosidase with transgalactosylation activity from Lactobacillus fermentum K4.** *J Dairy Sci* 2011, **94**(12):5811–5820.
62. Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Tarchini R, Peters SA, Sandbrink HM, Fiers MW, *et al*: **Complete genome sequence of Lactobacillus plantarum WCFS1.** *Proc Natl Acad Sci USA* 2003, **100**(4):1990–1995.
63. Groot Kormelink T, Koenders E, Hagemeijer Y, Overmars L, Siezen RJ, de Vos WM, Francke C: **Comparative genome analysis of central nitrogen metabolism and its control by GlnR in the class Bacilli.** *BMC Genomics* 2012, **13:**191.
64. Francke C, Groot Kormelink T, Hagemeijer Y, Overmars L, Sluijter V, Moezelaar R, Siezen RJ: **Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior.** *BMC Genomics* 2011, **12:**385.
65. Teusink B, van Enckevort FH, Francke C, Wiersma A, Wegkamp A, Smid EJ, Siezen RJ: **In silico reconstruction of the metabolic pathways of Lactobacillus plantarum: comparing predictions of nutrient requirements with those from growth experiments.** *Appl Environ Microbiol* 2005, **71**(11):7253–7262.
66. Siezen RJ, Francke C, Renckens B, Boekhorst J, Wels M, Kleerebezem M, van Hijum SA: **Complete resequencing and reannotation of the Lactobacillus plantarum WCFS1 genome.** *J Bacteriol* 2012, **194**(1):195–196.

doi:10.1186/1471-2164-14-209
Cite this article as: Overmars *et al.*: MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics* 2013 **14**:209.