Briefings in Bioinformatics Advance Access published July 3, 2012

doi:10.1093/bib/bbs029

A brief introduction to web-based genome browsers

Jun Wang, Lei Kong, Ge Gao and Jingchu Luo

Submitted: 24th January 2012; Received (in revised form): 17th May 2012

Abstract

Genome browser provides a graphical interface for users to browse, search, retrieve and analyze genomic sequence and annotation data. Web-based genome browsers can be classified into general genome browsers with multiple species and species-specific genome browsers. In this review, we attempt to give an overview for the main functions and features of web-based genome browsers, covering data visualization, retrieval, analysis and customization. To give a brief introduction to the multiple-species genome browser, we describe the user interface and main functions of the Ensembl and UCSC genome browsers using the human alpha-globin gene cluster as an example. We further use the MSU and the Rice-Map genome browsers to show some special features of species-specific genome browser, taking a rice transcription factor gene *OsSPL14* as an example.

Keywords: *genome browsers; genomic databases; visualization; data retrieval and analysis; customization*

INTRODUCTION

The initial sequence generated by the human genome project [1] together with the draft genome sequence of several model organisms, including the house mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), nematode (*Caenorhabditis elegan*), baker's yeast (*Saccharomyces cerevisiae*), Gram-negative bacterium (*Escherichia coli*) and thale cress (*Arabidopsis thaliana*), completed at the beginning of this millennium create a paradigm shift within biological research, as predicted by Gilbert in the early 1990s [2]. With the rapid development of next-generation sequencing technologies, hundreds of eukaryotic and thousands of prokaryotic genomes have been sequenced (http://www.genomesonline.org/). All the sequence data as well as the annotations generated through most completed or ongoing genome projects are collected in the genome databases and are publicly available through web portals such as the NCBI genome portal (http://www.ncbi.nlm.nih.gov/genome/) and the EBI genome database website (http://www.ebi.ac.uk/Databases/genomes.html). Great efforts were made in the construction of genome databases such as the human genome database (GDB) [3] and the *E. coli* genome database (Colibri) [4] in the early 1990s to enable subsequent quantities of data to be accommodated.

By systematic integration of genome sequences together with annotations generated through much heterogeneous data, genome browser provides a unique platform for molecular biologists to browse, search, retrieve and analyze these genomic data efficiently and conveniently. With a graphical interface, genome browser helps users to extract and summarize information intuitively from huge amount of raw data. Furthermore, different types of annotations from multiple sources can be integrated into one genome browser, helping users to analyze data

Corresponding author: Jingchu Luo, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences and Center for Bioinformatics, Peking University, Beijing 100871, P.R. China. Tel: +86 10 6275 7281; Fax: +86 10 6275 9001; E-mail: luojc@pku.edu.cn

JunWang is a PhD student in the College of Life Sciences at Peking University. Her research interests include new-generation genome browser development, heterogeneous biological data integration and bioinformatics platform construction.

Lei Kong is an Engineer in the College of Life Sciences at Peking University. His research interests include bioinformatics infrastructure construction and bioinformatics software development.

Ge Gao is an Assistant Professor in the College of Life Sciences at Peking University. His research interests include developing bioinformatics tools for mining and integrating heterogeneous biological data, with special focus on comparative genomics.

Jingchu Luo is a Professor in the College of Life Sciences at Peking University. His main research focus is bioinformatics platform development, bioinformatics database construction and comparative plant genomics.

across different data providers effectively. With a uniform interface, users can navigate the whole genome using the same genomics coordinate system, and make comparative analysis across different lineages such as primates, mammalians, vertebrates and plants.

In general, genome browser can be divided into web-based browsers and stand-alone applications. In this review, we focus on web-based genome browsers which are useful in promoting biological research due to their data quality, flexible accessibility and high performance. First, dedicated organizations often collect and integrate high-quality annotation data into web-based genome browsers, providing plentiful up-to-date information for the community. Second, users can access them anywhere with a standard web browser, avoiding any additional effort of setting up local environment for application installation and data preparation. Third, web-based genome browsers are usually installed on high performance servers and can support more complex and larger scale data types and applications. Here, we attempt to give a brief introduction to the main web-based genome browsers and the underlying frameworks.

Web-based genome browsers

Currently, there are two types of web-based genome browsers. The first type is the multiple-species genome browsers implemented in, among others, the UCSC genome database [5], the Ensembl project [6], the NCBI Map viewer website [7], the Phytozome and Gramene platforms [8]. These genome browsers integrate sequence and annotations for dozens of organisms and further promote cross-species comparative analysis. Most of them contain abundant annotations, covering gene model, transcript evidence, expression profiles, regulatory data, genomic conversation, etc. Each set of pre-computed annotation data is called a track in genome browsers. The essence of a genome browser is to pile up multiple tracks under the same genomic coordinate along the *Y*-axis, thus users could easily examine the consistency or difference of the annotation data and make their judgments of the functions or other features of the genomic region. Table 1 lists several mainstream web-based genome browsers, including Ensembl, the UCSC genome browser and the NCBI Map Viewer, which are accessed by a large number of users worldwide.

The other type is the species-specific genome browsers which mainly focus on one model organism and may have more annotations for a particular species. Powered by the Generic Model Organism Database (GMOD) project (http://gmod.org/), dozens of open-source software tools are collected for creating and managing genome biological databases, and the GBrowse framework [9] is one of the most popular tools in the GMOD project. Currently, most of these species-specific genome browsers are implemented based on the GBrowse framework, such as MGI, FlyBase, WormBase, SGD and TAIR (Table 2).

Table 1: List of main web-based general genome browsers with multiple species

Name	URL	Description
Ensembl	http://www.ensembl.org/	Major species with completed genome sequences providing lineage-specific web portals for vertebrates, metazoa, plants, fungi, protists and bacteria.
UCSC	http://genome.ucsc.edu/cgi-bin/hgGateway	Major species with completed genome sequences including vertebrates, deuterostomes, insects and nematodes. No plant species.
Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/	Major species with completed genome sequences including vertebrates, invertebrates, protozoa, plants and fungi, as well as dozens of uncompleted plant genomes.
Phytozome	http://www.phytozome.net/cgi-bin/gbrowse/	Major plant species with completed and ongoing genome sequences including monocots, dicots, fern, moss and green algae, with VISTA alignments.
Gramene	http://www.gramene.org/genome.browser/	Major plant species with completed genome sequences including monocots, dicots, fern, moss and green algae, and the short arm of chromosome 3 of several wild rice species.
VISTA	http://pipeline.lbl.gov/cgi-bin/gateway2/	Whole genome alignment presentation, including vertebrates, insects, nematodes, deuterostomes, plants, fungi, alga, annelids, stramenopiles and metazoa.
Genome Projector	http://www.g-language.org/g3/	Several hundreds of bacteria genomes with circular or linear maps.
Annmap	http://annmap.picr.man.ac.uk/	A genome browser that includes mappings between genomic features and Affymetrix microarrays for human, mouse, rat and yeast.

Table 2: List of some web-based species-specific genome browsers

Name	URL	Species
Animals		
MGI	http://gbrowse.informatics.jax.org/cgi-bin/gbrowse/	*Mus musculus* (Mouse)
RGD	http://rgd.mcw.edu/fgb2/gbrowse/	*Rattus norvegicus* (Rat)
Xenbase	http://www.xenbase.org/fgb2/gbrowse/	*Xenopus tropicalis* (Frog)
ZFIN	http://zfin.org/cgi-perl/gbrowse/	*Danio rareo* (Zebrafish)
Flybase	http://flybase.org/cgi-bin/gbrowse/	*Drosophila* (Fruit fly)
BeetleBase	http://beetlebase.org/cgi-bin/gbrowse/	*Tribolium Castaneum* (Beetle)
AphidBase	http://isyip.genouest.org/cgi-bin/gb2/gbrowse/	*Acyrthosiphon pisum* (Aphid)
wFleaBase	http://wfleabase.org/gbrowse/	*Daphnia* (Water flea)
Wormbase	http://www.wormbase.org/db/gb2/gbrowse/	*Caenorhaditus elegans* (Worm)
Plants		
TAIR	http://www.arabidopsis.org/browse/	*Arabidopsis thaliana* (Wall cress)
BRAD	http://brassicadb.org/cgi-bin/gbrowse/	*Brassica rapa* (Brassica)
SGN	http://solgenomics.net/gbrowse/bin/gbrowse/	*Solanum pimpinellifolium* (Tomato)
Popgenie	http://www.popgenie.org/tool/gbrowse/	*Populus trichocarpa* (Populus)
Rice Genome	http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/	*Oryza sativa* japonica (Rice)
Rice-Map	http://www.ricemap.org/	*Oryza sativa* japonica/indica (Rice)
MaizeDB	http://gbrowse.maizegdb.org/	*Zea mays* (Maize)
Microbes		
dictyBase	http://dictybase.org/db/cgi-bin/ggb/gbrowse/	*Dictyostelium discoideum* (Dictyostelid)
SGD	http://browse.yeastgenome.org/	*Saccharomyces cerevisiae* (Yeast)
ParameciumDB	http://paramecium.cgm.cnrs-gif.fr/cgi-bin/gbrowse2/	*Paramecium tetraurelia*

Genome browser frameworks

The creation of large sequencing data for various species increases the demand for building genome browsers to help researchers view and analyze these data in a more intuitive way. However, building a web-based genome browser from scratch is both time and labor consuming, while well-designed genome browser frameworks could be useful in this aspect. As listed in Table 3, there are some web-based genome browser frameworks for users to install locally, and to configure and customize their own annotation data efficiently. In addition to providing rich resources, Ensembl and the UCSC systems are also released as software packages for local installation. GBrowse is the most popular genome browser framework [9] and has been widely used in model organism projects for data visualization. The new genome browsers, JBrowse [10], ABrowse [11] and Anno-J [12], support Google-map like navigation, and LookSeq [13] is designed for raw sequencing reads presentation. Furthermore, there are some synteny genome browsers for comparative visualization of several species, with separated genome coordinates.

The underlying technology of genome browsers

Traditional genome browsers often employ classical web technologies based on synchronized transferring and static web pages. By using the AJAX-based web technology, modern genome browsers could bring better user experience to scientists [14]. Currently, most traditional genome browsers have actively employed this new web technology, but they still cannot provide as interactive browsing experience as the newly designed ones. For example, GBrowse supports smooth browsing within a limited region by preloading larger pictures (version 2.20 onwards) (http://cpan.uwinnipeg.ca/htdocs/GBrowse/Changes.html), while new AJAX-based browsers such as ABrowse [9], Anno-J [12] and JBrowse [10] support smooth dragging along the whole genome.

Genome browser can be divided into two categories based on whether the image is rendered on the server side or on the client side. Server-side rendering browsers such as UCSC, Ensembl, GBrowse and ABrowse extract the requested data from the back-end databases and render them into pictures on the server, and then send the pictures to the client web browsers. Client-side rendering browsers such as Anno-J and JBrowse send the requested data to client web browsers directly and draw the pictures dynamically in client web browsers. Client-side rendering reduces the server burden and the network data flow by distributing computing tasks to client sides. On the other hand, the pictures produced by server side could bring users much richer details of

Table 3: List of web-based genome browser frameworks

Name	URL	Description
General viewer		
Ensembl	http://www.ensembl.org/	An extensible software architecture with powerful API support.
UCSC	http://genome.ucsc.edu/	A freely downloadable package for local browser installation.
GBrowse	http://gmod.org/wiki/GBrowse	The most popular genome viewer for multiple genome annotation visualization, especially for model organism database projects.
JBrowse	http://jbrowse.org/	A JavaScript-based genome browser, providing Google-map like new browsing experience.
ABrowse	http://www.abrowse.org/	A new-generation customizable genome browser framework.
Anno-J	http://www.annoj.org/	An interactive application designed for visualizing genome annotation data and deep sequencing data.
SViewer	http://www.ncbi.nlm.nih.gov/projects/sviewer/	NCBI's new Sequence Viewer.
Synteny viewer		
GBrowse.syn	http://gmod.org/wiki/GBrowse.syn	A GBrowse-based synteny browser designed to display multiple genomes, with a central reference species compared with two or more additional species.
SynBrowse	http://www.synbrowse.org/	A generic sequence comparison tool for visualizing genome alignments both within and between species.
SynView	http://www.eupathdb.org/apps/SynView/	A light-weight, interactive and highly customizable comparative genomic visualization tool based on GBrowse.
Sybil	http://sybil.sourceforge.net/	A web-based software package for comparative genomics.
Next generation sequencing data viewer		
LookSeq	http://www.sanger.ac.uk/resources/software/lookseq/	A web-based application for alignment visualization, browsing and analysis of genome sequence data including short reads generated by next-generation sequencing.

the annotation data since current web browsers support limited drawing functions.

FUNCTIONATILIES AND FEATURES

Visualization

Taking the advantage of high-throughput sequencing technology and high-performance computing resource, immense volumes of genomic annotation data are being made available. The principal function of the genome browser is to aggregate different types of annotation data together and integrate them into an abstract graphical view. Annotations are organized under a uniform genome coordinate with the chromosome as the *X*-axis and various types of data being displayed along the *Y*-axis.

A web-based genome browser often provides a centralized or a set of databases to store different types of annotation data obtained from several organizations. The challenge for the general genome browsers is how to display this information properly for different genomic scales. Massive amounts of information need to be incorporated into the picture when a large genomic area is requested, which could overburden the server and the network. And too many heavy and complicated details also disturb the user attention. The UCSC genome browser tries to solve the problem by providing multiple views for a track [5]. The dense view of some tracks could be displayed to hide complicated details when zooming out to a large area of a chromosome, so that the user has a broad picture of the selected chromosome region.

It is useful to have an overview of a large area of the chromosome and look into several small regions for details simultaneously. Putting paralogous genes together in one page could promote comparative analysis greatly. NCBI sequence viewer supports users to view different regions inside the same chromosome, providing flexible multiple-panel-based navigating approach with different color cursors indicating the corresponding genomic locations. ABrowse supports users to visualize multiple genome regions of different chromosomes/genomes in separated in-page windows [11], while the separated windows are not fully operable as the main browsing canvas.

Displaying genome alignments within or between species helps users to make comparative studies such as finding the conserved or fast involved elements among several genomes. Examples of this type of browsers are the Generic Synteny Browser [15],

Sybil (http://sybil.sourceforge.net/), SynBrowse [16], SynView [17] and VISTA [18]. Although some of the general genome browsers could also display genomes alignments, the alignments could only be organized under the coordinates of one of the genomes. Most of the synteny browsers could do it in a better way by arranging each segment of the alignment under the coordinate of its own genome while piling them together. This type of genome browsers often focus on sequence alignment of DNA or protein sequences rather than other types of annotation data.

The NGS raw data viewer aims to provide graphic views for the short read sequences aligned to the genome and helps users to find genome structure variations. LookSeq provides a simple graphical representation for paired sequence reads, which helps to reveal potential insertions and deletions [13]. Users can have a view of high-depth original reads besides general annotation results, and manually manipulate them to assimilate information at different levels of resolution.

Data retrieval and analysis

In addition to graphical data navigation, data retrieval and analysis are useful features for a genome browser. Most of the existing genome browsers support search functions to locate genomic regions by coordinates, sequences or keywords. Some genome browsers employ a system to retrieve bulk data. For example, the UCSC system offers Table Browser to retrieve specified datasets [19], while the Ensembl, Gramene and ABrowse projects employ the BioMart system [20, 21] for making large data queries.

To facilitate further data analysis, multiple data access approaches are supported for analysis tools to retrieve data from the genome browsers. The Galaxy genome browser Trackster supports analysis by integrating tools in the same platform, connecting data manipulation with visualization tightly. The users can view the data in the genome browser seamlessly and further filter the visualized data on-the-fly, which helps to refine the results conveniently and efficiently. Some genome browsers offer plug-in mechanism, supporting third-party development and integration of tools based on the open interface, so that various tools could be added and launched easily within one platform [9]. Furthermore, it is also valuable for genome browsers to have close integration with external applications to perform the sequence and annotation data analysis transparently. The UCSC genome browser [22] supports users to submit selected data directly to Galaxy [23] and GREAT [24] platforms, while ABrowse supports data submission to Galaxy and WebLab [25] for data analysis.

In addition to human-oriented interfaces, machine-oriented data retrieval is becoming even more essential for large-scale data analysis [26]. Currently, web service has been widely used for exchanging structured information through networks among various data resources [27]. ABrowse supports native standard SOAP-based web service for underlying data access [11], which is also supported in BioMart [20, 21] and employed by the Ensembl and Gramene projects. In addition, BioDAS [28] has been widely used in some genome browsers for data exchange [5, 9, 29] of distributed platforms.

Customization

It is much easier to build a genome browser based on a framework. Most of the frameworks have configuration files for users to customize local data. Currently, it is easy for users to integrate annotation into general genome browsers with several popular data formats, such as GFF, BED, SAM and WIG. However, if the data format is not compatible with the genome browser, it is difficult for average users to convert data formats to meet the system requirements. Some browser frameworks such as GBrowse and ABrowse provide plug-in mechanisms or API to extend new data types [9, 11].

The genome browser is fast becoming a collaboration platform for researchers to share discoveries and to exchange knowledge [14], promoting remote cooperation among a group of scientists. Most genome browsers provide a facility for end-users to upload, create and share their own annotation data, providing a collaborative platform. The user annotation comments can also be attached for selected items on-the-fly [11, 29] and shared with specified users and groups [29], or the whole research community [11]. Besides, users can save any important analysis status as bookmarks [9, 11, 29] or sessions [30], and share them efficiently among several researchers.

To summarize, we list the main functions including visualization, data retrieval and analysis, and customization for different genome browsers in Table 4. Currently, there are many applications built on these platforms. For example, the Epigenome Roadmap Genome Browser [31] and Cistrome [32] are built based on the UCSC system; Gramene [8] is built based on Ensembl; Rice-Map [33] is built based on

Table 4: Main functions of the mainstream genome browsers.

Features	UCSC	Ensembl	SViewer	ABrowse	GBrowse	JBrowse	Anno-J
Visualization							
Annotation navigation	Page-based browsing, enabling dragging	Page-based browsing	Map-like browsing within an entry	Map-like browsing along whole genome	Map-like browsing within a limited region	Map-like browsing along whole genome	Map-like browsing along whole genome
Multiple in-page windows	/	/	√	√	/	/	/
Data retrieval and analysis							
Query system	Table Browser	BioMart	/	BioMart	/	/	/
User-oriented analysis	Direct data submission	/	Integrated tools	Direct data submission	Plugin tools	/	/
Machine-oriented interface	Through BioDAS	Through BioMart	/	SOAP-based Web Service/Through BioMart	Through BioDAS	Through Amazon	/
Customization							
Upload user tracks	√	√	√	√	√	√	/
User-contributed contents	Session-based data restoring	Personal annotation, bookmark and group mechanism	/	Comments for tracks/entries and landmarks	/	/	/

ABrowse; Flybase [34] and Wormbase [35] are built based on GBrowse; and the Arabidopsis epigenome map [12] is built based on Anno-J. These applications provide valuable resources for biologists with the support of general genome browser platforms.

EXAMPLES

As described above, genome browser is a useful tool in computational molecular biology, especially in genetic, genomic and evolutionary researches. In the following section, we introduce the main functions of general genome browsers using the Ensembl and UCSC genome browsers as examples. We further describe the features of species-specific genome browsers based on the MSU and Rice-Map genome browsers.

General genome browsers

Hemoglobin is the key molecule used to transport oxygen in vertebrates. Human adult hemoglobins are encoded by the alpha-globin and beta-globin gene clusters. Taking the human alpha-globin gene cluster as an example, we describe the features of the Ensembl and UCSC genome browsers (Figure 1).

Data visualization

The Ensembl genome browser (Release 67) provides the same user interface for each organism. By choosing 'Human' as the target organism and searching the keyword 'alpha hemoglobin', then following the links in the search results page, users can enter the chromosome region 16:226,679-227,521 where the HBA1 gene encoding alpha hemoglobin is located. By expanding the region to chromosome 16:200,001-235,000, user can have a view of the alpha-globin gene cluster with default setting (Figure 1a). The main body of the interface contains two panels. The left panel lists the main menu for location-based displays at different levels from whole genome, chromosome summary to region overview and region in detail. And links to comparative genomics, genetic variation as well as sequence markers are also provided. The main panel is arranged in three sections from top to bottom, providing different scales for users to analyze the genome.

In addition to the location view, Ensembl provides separate pages to display various types of information, organized in a tabbed structure. In the gene page, different transcripts for HBA1 are organized in a table, and the summary of this gene is described below, with related transcripts highlighted in light green. Furthermore, users can jump to the transcript

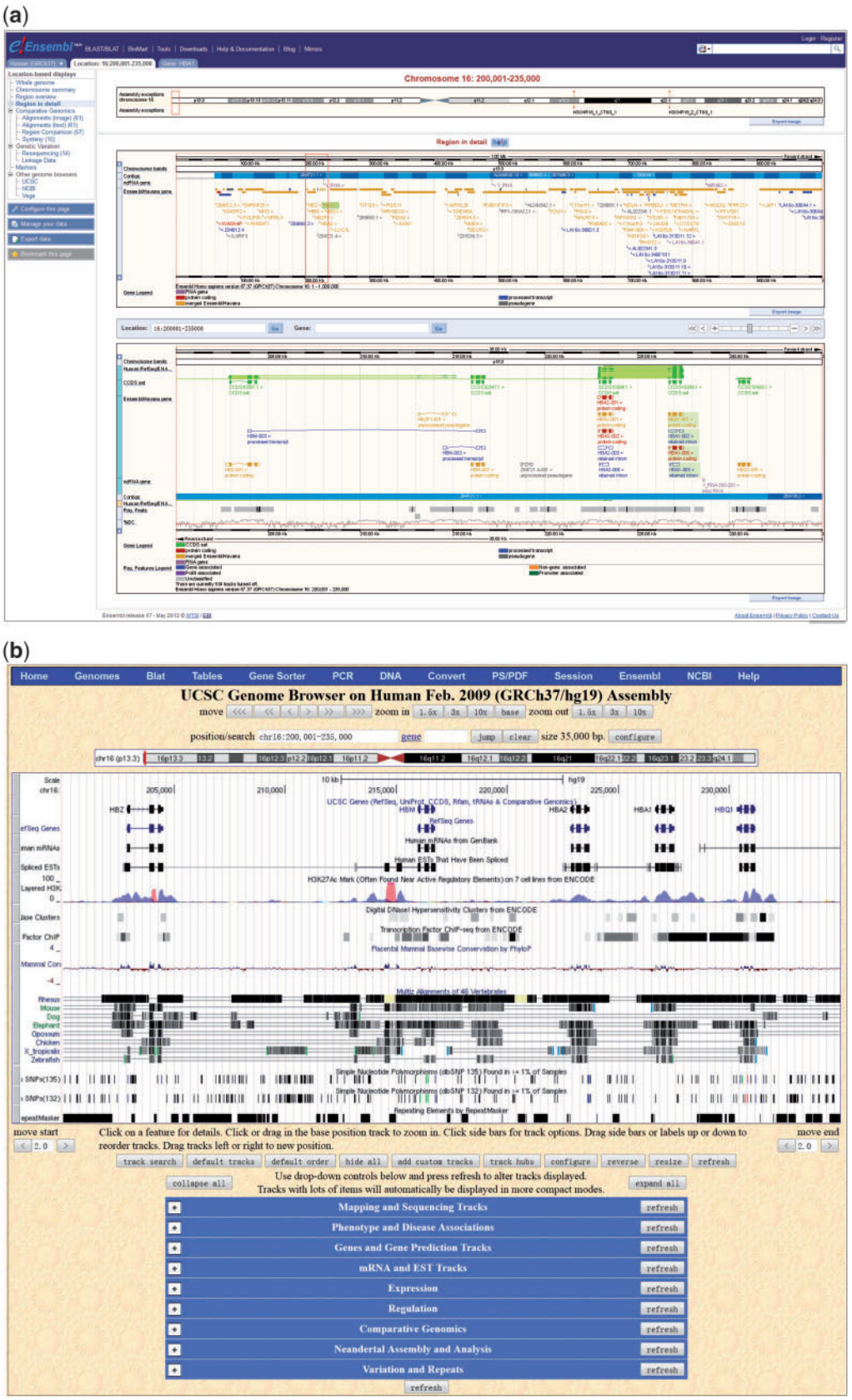

Figure 1: The screenshot of the user interface of the Ensembl and UCSC genome browsers for the alpha-globin gene cluster located in the forward strand of chromosome 16, consisting of five members (HBZ, HBM, HBA2, HBA1 and HBQ1) and two pseudogenes. (**a**) The user interface of the Ensembl genome browser with default setting

(continued)

page by clicking the individual transcript. The detailed transcript descriptions are shown in the middle panel. In addition, other information such as regulation and variation can also be displayed in different pages for users to investigate with different views, and cross-links are also listed in the left panel for specific pages. Detailed description of the user interface can be found in the Ensembl online tutorial (http://www.ensembl.org/info/website/tutorials/).

Besides the general functionalities as described above, the main panel has some special features designed for end-users to access the annotations conveniently. Clicking on the graphical item in the main panel will show the detailed information of this annotation unit. Moving the mouse over the track name at the left side of each track will show several logos. Detailed track description will be shown by moving the mouse over the information logo, and the display style of each track can be changed by moving the mouse to the configuration logo. Furthermore, users can click and drag the column bar beside the track name to re-order the track location freely for customized visualization.

The interface of the UCSC genome browser (v267, hg19) has some common features with that of the Ensembl genome browser, including the overall layout for graphical display of chromosome coordinates, the search boxes to make location or text queries and the layout of exon–intron structure of the genes. However, there are some special features implemented by the UCSC genome browser (Figure 1b).

The default tracks displayed are different from that of Ensembl. The annotation of protein-coding genes and their transcripts include the UCSC genes, the RefSeq genes, the Human mRNAs and the spliced ESTs. For example, the UCSC and RefSeq annotation for the human alpha-globin gene cluster including HBZ, HBM, HBA2, HBA1 and HBQ1 are displayed together with their mRNA and spliced EST transcript evidence. Besides, the histone modification H3K27Ac mark on seven cell lines, the digital DNaseI hypersensitivity clusters and the transcription factor ChIP-Seq assay generated by the ENCODE project are also displayed as default tracks. The blue peaks show that all these genes have H3K27Ac histone mark enrichment in the K562 cell line, while the two red peaks under the HBZ and HBM genes show the enrichment in the GM12878 cell line. Furthermore, multiple alignments of 46 vertebrate species are provided to measure the evolutionary conservation of this region. And most genes have high conservation value among mammals, while the intergenic regions are less conserved. SNP variants from dbSNP, repeating elements generated by RepeatMasker are also provided as default tracks.

The user experience of UCSC genome browser is different from that of the Ensembl system. Users can freely drag the graphical canvas upstream or downstream, and right-click functions are provided for the whole canvas for quick configuration. In contrast to the Ensembl system, the column bar is linked to track description and configuration page, and the tracks can be re-ordered by clicking and dragging the track name.

Data retrieval and analysis

In addition to data browsing, data analysis is also supported by the Ensembl platform. A BioMart-powered query system is provided for bulk data retrieval, and advanced users can use standard Mart API to retrieve bulk data from the backend database directly. Furthermore, sequence similarity searches with BLAST/BLAT and other tools such as assembly converter, ID history converter, region report and variant effect predictor are supported for users to process data conveniently.

Figure 1: Continued

of the annotation tracks, showing the alpha-globin gene cluster. The graphical annotations are displayed in the main body divided in three sections from top to bottom. The top section shows the location of this alpha-globin gene cluser on chromosome 16, where the selected region is highlighted by a red box. The middle section shows the chromosome band, the contigs and the protein-coding genes annotated by the Ensembl/Havana project. The bottom section shows more annotations in this 35 kb region. (**b**) The main user interface of the UCSC genome browser showing the default tracks in default order for the human alpha-globin gene cluster. The top graphical window shows the annotations, and the text window below supports users to make configuration simultaneously in the same page. User can choose four styles to view different tracks, i.e. the dense, squish, pack and full styles. Some of the groups and tracks are similar to that of Ensembl, e.g. the annotations for genes, mRNAs and ESTs, while some are different, e.g. the group of phenotype and disease associations.

As for the UCSC system, some other useful analysis functions are also supported besides data navigation. Different from the BioMart retrieval system employed by the Ensembl platform, the UCSC platform offers the Table Browser system for bulk data retrieval, supporting operations of intersection and correlation for dataset processing. And the retrieved result can be directly sent to external platforms for further analysis with a single click. In addition to the sequence similarity search using BLAT, the UCSC system provides some online tools for users to perform quick data analysis, such as In Silico PCR, Gene Sorter, VisiGene and several handy utilities.

Customization

In the left panel of the Ensembl browser page, three buttons 'Configue this page', 'Manage your data' and 'Export data' for customization can be invoked to add or remove annotation tracks, and to upload and analyze users' own data (http://www.ensembl.org/info/website/upload/). Users can also add annotations freely for specified items as user comments. To support quick link to previous browsing status, users can bookmark the layout of the current browser for future study. To promote knowledge sharing, group mechanism is supported in the Ensembl system for collaboration among several colleagues [29].

To facilitate data customization, the UCSC system allows users to save settings as sessions for restoring and sharing through a wiki system. And users can also upload tracks for personal data visualization and analysis with the precomputed annotations. Moreover, the UCSC system provides the Track Hubs functionality on the home page, offering a sharing mechanism for large custom datasets from other individuals and labs [36].

Species-specific genome browsers

As listed in Table 2, dozens of species-specific genome browsers are available online. Here, we introduce the two species-specific genome browsers dedicated to the rice genome, the MSU rice genome browser [37] and the Rice-Map genome browser [33] using the rice transcription factor *OsSPL14* as an example (Figure 2). Recently, *OsSPL14* has been reported as a member of the rice SBP-like gene family and is essential for rice grain productivity. Regulation of *OsSPL14* by *OsmiR156* defines ideal plant architecture in rice, promoting panicle branching and enhanced grain yield [38, 39].

Data visualization

In the MSU rice genome browser (Release 7), users can search the *OsSPL14* gene by specifying 'SPL14' in the search box. As shown in Figure 2a, *OsSPL14* is encoded by the transcript LOC_Os08g39890.1 with three exons, and located in the reverse strand of chromosome 8, from 25 274 449 to 25 278 696. Powered by the GBrowse platform, the MSU rice genome browser provides annotation views with different scales, including chromosome overview, regional view and detailed view. The large-scale view provides a broader picture for users to inspect the upstream and downstream annotation conveniently. In the detailed annotation canvas, more than 82 annotation tracks are provided, covering gene model, transcript evidence, expression profiling, sequence alignment, genetic marker, SNP, RNA-Seq coverage and other genomic features. In addition to the basic gene model information, users may inspect this gene in different development stages through various RNA-Seq expression data. According to the integrated expression data, this gene has a highly expressed signal in the pre-emergence inflourescence stage, pistil stage and embryo-25 days after pollination stage, which is consistent with SBP-box gene function reported in the literature (Figure 2a).

In the Rice-Map genome browser (v1.0), different annotation tracks are organized in a map-like visualization canvas, with the name of opened tracks listed in the right panel. Currently, 81 *japonica* tracks and 82 *indica* tracks have been compiled and loaded into Rice-Map. Besides basic gene annotation, there are rich annotation for cross genome alignments and conservation values, offering important clues to investigate this gene in other plants. From the conservation data, users can find that the sequence of this gene is highly conserved among rice and other grasses including the wild grass (*Brachypodium distachyon*), maize (*Zea mays*) and sorghum (*Sorghum bicolor*), especially in the exon regions (Figure 2b). Furthermore, users can use the 'magic wand' in the navigation panel (left-top corner) to select an interesting region by mouse, and inspect it in the main canvas or in a new sub-window. After clicking the annotation item, detailed descriptions for track and entry are provided in the right panel, along with the graphical view.

Data retrieval and analysis

The sequence and annotation data can be dumped from the MSU genome browser database through

Figure 2: The screenshot of the user interface of the MSU and Rice-Map genome browsers for the rice transcription factor *OsSPL14*. (**a**) The user interface of the MSU rice genome browser. The chromosome overview is displayed at the top, the regional view is shown at the middle and the bottom section is the detailed view for four annotation tracks including gene model, expression data for pre-emergence inflorescence, pistil and embryo—25 days after pollination stages. (**b**) The Rice-Map genome browser. In the middle canvas, different annotation tracks are listed, including the MSU gene model, and three comparative genome alignments from VISTA. The detailed information for individual entries is shown in the right panel, interpreting the data resource, entry location, sequence and function etc.

the online user interface. In the Rice-Map platform, a BioMart-powered system 'Rice Mart' is provided for users to retrieve large datasets, with a standard interface. The retrieved dataset, as well as the genomic, CDS and protein sequence information of each entry can also be submitted to external bioinformatic platforms, i.e. WebLab and Galaxy for further analysis.

Customization

The MSU genome browser supports users wishing to view their own data in the genome browser, alongside the precomputed annotations. By uploading the local data file or specifying the URL of a remote annotation file, users can easily integrate their own data into the online platform.

In the Rice-Map platform, users can add user annotation by clicking and dragging a genomic region freely, providing a more flexible approach for average users to add annotation on-the-fly. Furthermore, Rice-Map provides a multi-functional user space for users to save contributed data online, including track evaluation, entry comments and browsing landmarks. Moreover, users can also

set privilege for the contributed items, and the private items can be taken as personal online notebook, while the public items can be shared among the whole community.

CHALLENGES AND PERSPECTIVES

The web-based genome browsers have many challenges such as how to visualize different data types produced by the NGS sequencing instruments. For the annotation visualization, FlyBase tries to provide a 3D view for different development stages of RNA expression data in one track, providing an overall view of differential gene expression patterns over and above the traditionally separated tracks. As for the circular genomes, presenting them as circular maps is more natural than linear views, and Genome Projector provides new circular views for more than 400 bacterial genomes [40]. For the original reads presentation, LookSeq can provide a reads view for small genomes currently but performs poorly with large-scale genomes. Based on the HTML5 language, semantic tags and web storage are introduced into the client browser, which helps users to improve their browsing experience efficiently. With the launch of the large-scale sequencing projects such as 1000 Genomes, visualizing individual sequencing data becomes a big challenge. It is necessary to develop new approaches to visualize several individual genomes together under a coordinate-free system to figure out insertions and deletions conveniently.

Since personalized functions are becoming more and more crucial to promote scientific research, some novel features are being added to genome browsers, improving user participation, collaboration and communication. The WebApollo project (http://gmod.org/wiki/WebApollo/) is now aiming to develop an online collaborative annotation editor, allowing users to interactively create and edit genomic annotations in a web-based graphical environment. The management and transmission of large-scale data are becoming a great challenge. Some genome browsers set the data storage lifespan and limit the data size for custom tracks [29, 30]. High-speed network facilities and powerful servers are needed to solve such problems in the future. Furthermore, a distributed system is needed to exchange data across different platforms and resources to achieve a collaborative society for biologists [27], such as the P2P data transmission approach, enabling user participation in data sharing.

Though web service is a becoming a standard protocol for data exchange and application communication, the problem of how to define the data exchange format and the application interface is still unsolved. Putting the genome browser and the bioinformatics application platform on a cloud environment and allowing them to share the same storage could be a possible solution to avoid heavy data transmission. Due to the demand of access speed and large-scale data integration, web-based genome browsers are gradually moving to cluster servers or cloud environments. UCSC provides powerful hardware to offer a high-speed browsing experience [41], while Ensembl and JBrowse are actively using Amazon web services to improve the online service [10, 42]. In the future, more and more cloud technologies would provide high performance for the end users.

In addition, genome browsers may take advantage of the latest hardware technology to offer a richer presentation of data. For example, GenomePad is a novel mobile-based genome browser, supporting users to navigate and share genome information conveniently on mobile phones (http://research.oicr.on.ca/genomepad/). Genome wowser (http://www.research.chop.edu/programs/cbmi/index.php/genome-wowser-support.html) is an iPad-enabled human genome browser, providing an intuitive and portable presentation of the popular UCSC genome browser. Moreover, JBrowse is also planning to use mobile devices for sequencing application (http://jbrowse.org/?p=173). Tablet PCs could be another choice for intuitive and convenient genome browsing (http://www.microsoft.com/surface/).

In conclusion, with the development of new web technologies, genome browser is becoming a collaboration platform for researchers to share data and to exchange knowledge [14], promoting collaboration among scientists in different locations.

Key Points

- Genome browser provides a graphical interface for users to browse, search, retrieve and analyze genomic sequence and annotation data.
- Web-based genome browsers promote biology research greatly for their flexible accessibility, data quality and high performance.
- The main functions and features of web-based genome browsers include data visualization, retrieval, analysis and customization.

- The Ensembl and UCSC genome browser are the two most popular general genome browsers. We use them to introduce the main functions of the multiple-species genome browser, taking the human alpha-globin gene cluster as an example.
- There are many species-specific genome browsers. We use the MSU and Rice-Map genome browsers to show some special features of species-specific genome browser, taking the rice transcription factor gene *OsSPL14* as an example.

FUNDING

This work was supported by National Science and Technology Infrastructure Program (2009FY120100), National Key Basic Research Program (2011CB A01102) and the Natural Science Foundation of China (1071160 and 31171242).

References

1. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**: 860–921.
2. Gilbert W. Towards a paradigm shift in biology. *Nature* 1991;**349**:99.
3. Cuticchia AJ, Fasman KH, Kingsbury DT, *et al*. The GDB human genome data base anno 1993. *Nucleic Acids Res.* 1993;**21**:3003–6.
4. Medigue C, Viari A, Henaut A, *et al*. Colibri: a functional data base for the Escherichia coli genome. *Microbiol Rev* 1993;**57**:623–54.
5. Karolchik D, Baertsch R, Diekhans M, *et al*. The UCSC genome browser database. *Nucleic Acids Res* 2003;**31**:51–4.
6. Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. *Nucleic Acids Res* 2002;**30**:38–41.
7. Wolfsberg TG. Using the NCBI Map Viewer to browse genomic sequence data. *Curr Protoc Hum Genet* 2011; Chapter 18:Unit18 15.
8. Ware D, Jaiswal P, Ni J, *et al*. Gramene: a resource for comparative grass genomics. *Nucleic Acids Res* 2002;**30**: 103–5.
9. Stein LD, Mungall C, Shu S, *et al*. The generic genome browser: a building block for a model organism system database. *Genome Res* 2002;**12**:1599–610.
10. Skinner ME, Uzilov AV, Stein LD, *et al*. JBrowse: a next-generation genome browser. *Genome Res* 2009;**19**:1630–8.
11. Kong L, Wang J, Zhao S, *et al*. ABrowse - a customizable next-generation genome browser framework. *BMC Bioinformatics* 2012;**13**:2.
12. Lister R, O'Malley RC, Tonti-Filippini J, *et al*. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36.
13. Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res* 2009;**19**: 2125–32.
14. Nielsen CB, Cantor M, Dubchak I, *et al*. Visualizing genomes: techniques and challenges. *Nat Methods* 2010;**7**:S5–15.
15. McKay S, Vergara I, Stajich J. Using the generic synteny browser (GBrowse_syn). *Curr Protoc Bioinformatics* 2010; Chapter 9:Unit 9.12.
16. Pan X, Stein L, Brendel V. SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics* 2005;**21**: 3461–8.
17. Wang H, Su Y, Mackey AJ, *et al*. SynView: a GBrowse-compatible approach to visualizing comparative genome data. *Bioinformatics* 2006;**22**:2308–9.
18. Frazer KA, Pachter L, Poliakov A, *et al*. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 2004;**32**:W273–9.
19. Karolchik D, Hinrichs AS, Furey TS, *et al*. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004; **32**:D493–6.
20. Smedley D, Haider S, Ballester B, *et al*. BioMart–biological queries made easy. *BMC Genomics* 2009;**10**:22.
21. Haider S, Ballester B, Smedley D, *et al*. BioMart Central Portal–unified access to biological data. *Nucleic Acids Res* 2009;**37**:W23–7.
22. Fujita PA, Rhead B, Zweig AS, *et al*. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2010;**39**: D876–82.
23. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;**11**:R86.
24. McLean CY, Bristor D, Hiller M, *et al*. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;**28**:495–501.
25. Liu X, Wu J, Wang J, *et al*. WebLab: a data-centric, knowledge-sharing bioinformatic platform. *Nucleic Acids Res* 2009;**37**:W33–9.
26. Sen TZ, Harper LC, Schaeffer ML, *et al*. Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database* 2010; doi:10.1093/database/baq007.
27. Stein L. Creating a bioinformatics nation. *Nature* 2002;**417**: 119–20.
28. Dowell RD, Jokerst RM, Day A, *et al*. The distributed annotation system. *BMC Bioinformatics* 2001;**2**:7.
29. Flicek P, Aken BL, Beal K, *et al*. Ensembl 2008. *Nucleic Acids Res* 2008;**36**:D707–14.
30. Karolchik D, Kuhn RM, Baertsch R, *et al*. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 2008;**36**:D773–9.
31. Zhou X, Maricque B, Xie M, *et al*. The human epigenome browser at Washington University. *Nat Methods* 2011;**8**: 989–90.
32. Liu T, Ortiz JA, Taing L, *et al*. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* 2011;**12**:R83.
33. Wang J, Kong L, Zhao S, *et al*. Rice-Map: a new-generation rice genome browser. *BMC Genomics* 2011;**12**:165.
34. Drysdale R. FlyBase: a database for the Drosophila research community. *Methods Mol Biol* 2008;**420**:45–59.
35. Harris TW, Lee R, Schwarz E, *et al*. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res* 2003;**31**:133–7.

36. Zweig AS, Karolchik D, Kuhn RM, *et al*. UCSC genome browser tutorial. *Genomics* 2008;**92**:75–84.

37. Ouyang S, Zhu W, Hamilton J, *et al*. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* 2007;**35**:D883–7.

38. Jiao Y, Wang Y, Xue D, *et al*. Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat Genet* 2010;**42**:541–4.

39. Miura K, Ikeda M, Matsubara A, *et al*. OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nat Genet* 2010;**42**:545–9.

40. Arakawa K, Tamaki S, Kono N, *et al*. Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics* 2009;**10**:31.

41. Dreszer TR, Karolchik D, Zweig AS, *et al*. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 2011;**40**:D918–23.

42. Flicek P, Aken BL, Ballester B, *et al*. Ensembl's 10th year. *Nucleic Acids Res* 2009;**38**:D557–D562.