Genome browsing 'for the rest of us'

Software and resources for visualizing genomic data

by David P. Leader
(Glasgow, UK)

Anyone whose sporting prowess has given him the opportunity to examine X-rays of his own broken bones will know the perverse fascination that such pictures of one's 'physical inner-self' can provide. As biochemists, our fascination is less likely to be with our bones than with our genomes, and this article considers some of the ways in which we can survey our 'genetic inner-selves' on a desktop computer, in the office or in the home. The article takes the standpoint that, in addition to any particular genetic 'bone' (broken or otherwise) that we may wish to view, we are also interested in traversing the whole genomic 'skeleton'. Those of a squeamish nature, who think that genome browsing should be left to the scientific equivalent of the surgeon, need read no further.

Initially, systems for visualizing genomes were developed to serve the needs of those research workers actually assembling genomes from experimental data. The most venerable of these is AceDB[1], developed for the *Caenorhabditis elegans* sequencing project. Such systems were geared to representing and referencing the clones from which the sequence had been derived, to identifying genes in the raw sequence, and to annotating them. It is no criticism, therefore, to say that their features do not necessarily coincide with those required by third parties wishing to 'browse' the genome after it has been assembled. Of course, new features can always be added to software, but there is another aspect of these systems that acts as a barrier to 'lay' use. This is the fact that they are integrated with a database (the 'DB' in AceDB) which requires specialist expertise to use. Two alternatives to this sort of system are considered here. The first is desktop software specifically designed to allow the average biological scientist to browse completed genomes, details of which are held in simple ('flat') files; the second involves web pages which present the genomic information held in complex databases.

Key words: AceDB, BugView, Ensembl, GenBank®, gene cluster, genome, synteny.

BugView

The genome visualization software that we shall consider, BugView[2] (Figures 1 and 2), does not require the user to deal with a specialized database, as its initial input is individual annotated files that can be downloaded from GenBank®. Such files are available, not only for the scores of bacterial genomes for which the software was originally designed, but also for the individual chromosomes of eukaryotic genomes such as yeasts, *C. elegans*, *Drosophila* and *Arabidopsis*.

A disadvantage of this solution to the 'database problem' is that the user is restricted to the information provided in these files, and must rely on that information being correct and current. The software is also less well suited to browsing those large mammalian and other chromosomes which, because of unsequenced simple repetitive regions, are split into several GenBank® files.

What aspects of a GenBank® file are represented in BugView's visualization? The initial view is restricted to the genes themselves, their directionality, whether they overlap, and, where relevant, intron–exon structure, including alternative splicing. Users can scroll and zoom smoothly, allowing easy navigation of a genome.

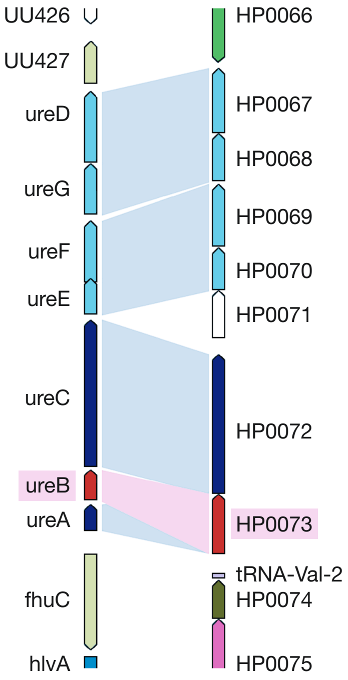


Figure 1. Comparison of genes of the urease cluster in *Ureaplasma urealyticum* (left) and *Helicobacter pylori* (right) using BugView

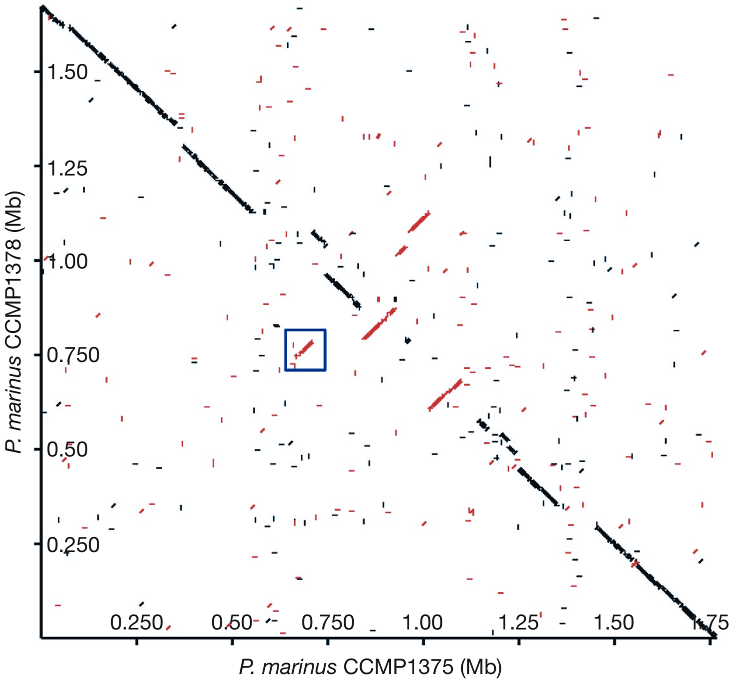


Figure 2. BugView matrix comparison of two species of *Prochlorococcus marinus*. The red lines are homologous regions that have been inverted with respect to one another. The blue rectangle is a cursor selection from which the user may transfer to a view of the same area aligned as in Figure 1.

Double-clicking on a particular gene of interest invokes a small pop-up window from which the DNA and protein sequences, and other gene information, can be accessed without leaving the overall map view. Users may also make and store their own annotations using this window.

When browsing genomes, an important aid to perception is to have a visual indication of the functional category of the product of a gene, and this is usually provided by colour-coding. Such representation depends, of course, on annotation of the genes, and although BugView allows one to make such annotations oneself, most users will prefer to have the annotating done for them. The genes of many genomes have been functionally classified in the database of Clusters of Orthologous Groups of proteins (COGs)[3]. If this information is available for a genome, it is included in the '.ptt' file, to be found in the same directory as the corresponding '.gbk' file on the GenBank® FTP site[4]. The COGs from '.ptt' files can be imported directly into BugView, giving one an instant perspective of, for example, gene clusters. The colour-coding also enhances the usefulness of some of BugView's alternative genomic views, such as the circular view, popular for bacterial chromosomes. Even users with imperfect colour vision can take advantage of the functional categorization, as it is possible to search for genes by category — as well as by the more usual criteria — and then 'step through' a set of found genes.

Comparing genomes in BugView

A common reason for wishing to view a genome is to see how similar or different it is from another genome. Indeed, BugView was written to compare pathogenic and non-pathogenic strains of *Streptococcus pneumoniae.* The facilities allowing such comparisons do not limit the user to viewing different genomes or chromosomes next to one another, but include an in-built sequence-comparison program that allows one to assign gene pairs for two genomes of interest. Such pairs are represented by 'bands' between the genes, although if the genes are far apart on their respective genomes, the bands will be too oblique for the pairs to be seen. One needs to juxtapose a gene pair (or group of pairs) of interest by sliding the two genomes relative to one another, and, if necessary, inverting their relative orientation. This facility is relatively easy to provide in software such as BugView, but is difficult to implement satisfactorily in the web-based interfaces considered below.

Although comparing whole genomes is most useful if two species have not diverged too far, in more distantly related species it can still be valuable to compare individual clusters of related genes. We illustrate this with results from a biochemistry undergraduate unit in which students use BugView and other Internet and bioinformatics resources to identify and explore biochemical features of an assigned bacterium of medical or economic interest. Figure 1 shows the cluster of urease-related genes in *Heliobacter pylori* and *Ureaplasma urealyticum.* The fascinating biochemistry underlying the different roles of urease in these pathogens is an effective argument that a knowledge of metabolism is needed to make sense of the results of genome projects. Looking at genes has limited value if you do not understand what they do.

Browsing genomes on the web

As already mentioned, chromosomes such as those of mammals have unsequenced simple repetitive regions, and are split into several, often very large, GenBank® files. For individual scientists, the only realistic way to browse these is using the web interfaces provided by organizations that are able to integrate and serve the data from their databases. The web interfaces that will be considered here are those provided by the two largest bioinformatics organizations: NCBI, with its Map Viewer; and EBI-Sanger with Ensembl.

The experience of using such web interfaces to browse genomes is generally inferior to that which desktop software can provide, because there is a much slower response each time a user tries to scroll or zoom. The reason for this is that these sites have implemented their browsing facilities solely in HTML (the web-page description language) so that each change of view involves generation of a new web page. Such web pages cannot be displayed until the necessary bitmap graphic files have been transferred across the Internet from the remote server to the user's machine. (Of course, to browse genomes in desktop software, one must transfer much larger data files across the Internet. However, this, and the initial time it takes to load the files into computer memory, are perceived by the user as predictable preliminaries — an acceptable price to pay for subsequent smooth and near-instantaneous scrolling and zooming.)

Given this limitation, web-based facilities are oriented more to providing information about particular genes than to genome browsing (they call themselves map viewers, rather than browsers). Their strength is the links that they provide to the wide range of information held in their databases.

NCBI Map Viewer

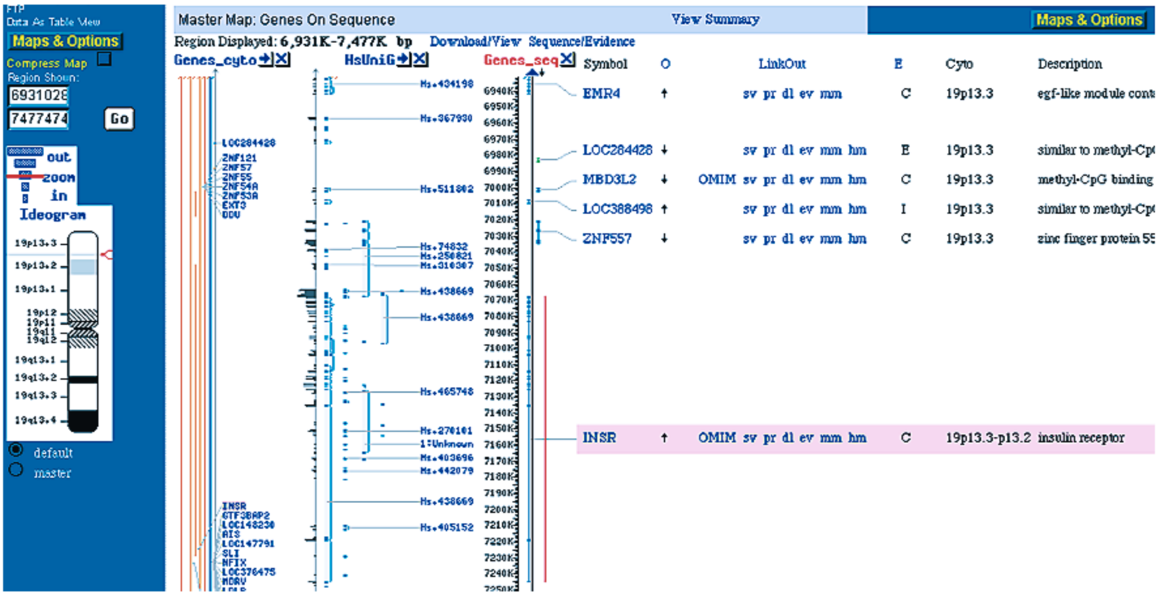
In the NCBI Map Viewer[5], one is initially presented with pictorial representations of the chromosomes of an organism. If one then searches by name for a particular gene, hits are 'marked' on the chromosome, and clicking on a particular 'hit' generates a new view of the gene within a relatively wide genomic context. A cartoon representation of the chromosome of interest — with the current display range indicated — is always in view, and the textbook vertical representation reinforces the metaphor of browsing a chromosome.

The default display of pictorial information in Map Viewer (Figure 3) is relatively sparse, with the graphic display limited to features related to gene identification, although there are text links to other pages. This sparseness is to be commended — there is sufficient information to be useful, but no more than is initially manageable. However potential users should be aware of the range of additional features that can be added from the 'Maps Options' pop-up window.

There are currently about two dozen genomes available for browsing through the NCBI Map Viewer.

EMBL-EBI/Sanger Ensembl

There are only about half as many genomes available for browsing with Ensembl[6], but those available are furnished with equally abundant linked data, which is presented in a variety of views (MapView, ContigView, GeneView, etc).



Symbol	O	LinkOut	E	Cyto	Description
EMR4	↑	sv pr dl ev mm	C	19p13.3	egf-like module cont
LOC284428	↓	sv pr dl ev mm hm	E	19p13.3	similar to methyl-Cp
MBD3L2	↓	OMIM sv pr dl ev mm hm	C	19p13.3	methyl-CpG binding
LOC388498	↑	sv pr dl ev mm hm	I	19p13.3	similar to methyl-Cp
ZNF557	↓	sv pr dl ev mm hm	C	19p13.3	zinc finger protein 55
INSR	↑	OMIM sv pr dl ev mm hm	C	19p13.3-p13.2	insulin receptor

Figure 3. NCBI Map Viewer, showing a view of human chromosome 19 with the insulin-receptor gene highlighted

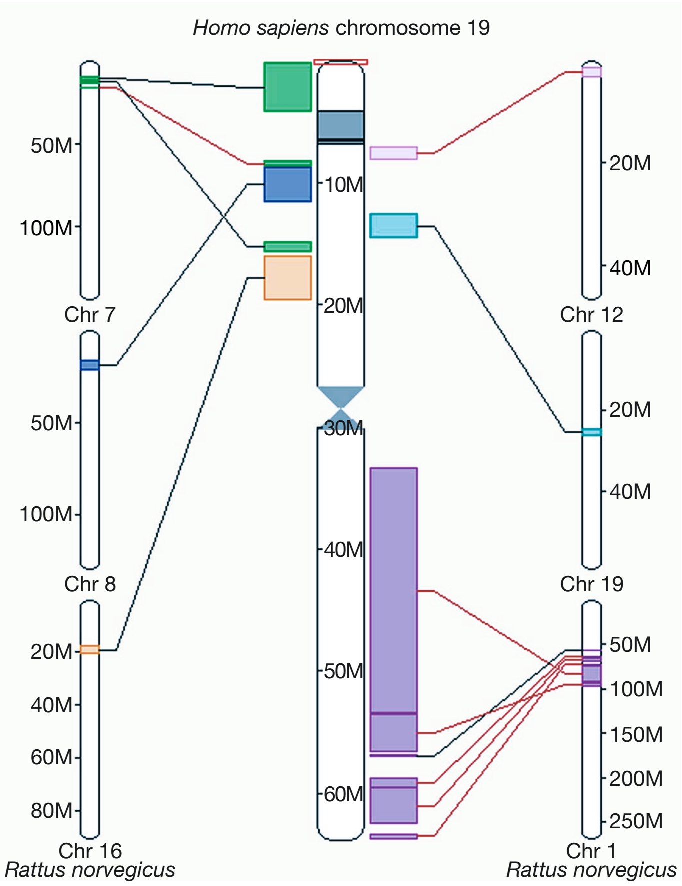


Figure 4. Ensembl SyntenyViewer view of human chromosome 19, showing syntenic regions in different rat chromosomes

Initial entry is again from cartoon chromosomes, and clicking on one of these invokes an enlarged view (MapView) containing several features, including a dense histogram of genes. If one clicks at a particular position one is transferred to ContigView, which corresponds roughly to NCBI's Map Viewer. ContigView shows a set of different views of a gene, from broad chromosome context to fine nucleotide detail; and these views are in separate horizontal boxes, one below the other. Compared with NCBI's Map Viewer, the initial ContigView may appear overloaded with information — in the default view, half of the information is generally off-screen. However, this approach does make users aware of what features are available, and they can reclaim 'screen real estate' by collapsing any boxes in which they are not interested.

In contrast to NCBI's facility, an initial text search for the name of a gene of interest in Ensembl leads to the TextView, not the ContigView. To reach the latter, it appears necessary to proceed to GeneView and then click within 'Genomic Location'. This can be rather irritating, as can the 'bugginess' of the DHTML pop-ups when using Mac versions of Internet Explorer 5.*x*.

Comparing genomes on the web

A feature of both the Ensembl and NCBI sites is presentation (albeit static) of conserved synteny — groups of genes on a chromosome that are also grouped in another organism. In Ensembl, this is displayed in the SyntenyView shown in Figure 4. In order to explore synteny properly, it is necessary to be able to browse the genomes of both organisms being compared, and to slide the chromosomes relative to one another so as to juxtapose syntenic regions. This requires sophisticated software[7], and, wisely, neither Ensembl nor NCBI have tried to implement synteny browsing through serving a succession of HTML pages.

Asking for more

It is difficult to praise too highly the service that NCBI and EBI-Sanger — and their dedicated staff — have provided in making the results of the public genome sequencing projects available to individual scientists. In this context, it may seem churlish to criticize the current facilities for browsing and comparing large mammalian genomes — in our defence, we would plea that we wish to see these facilities made even better. We can imagine two reasons that these organizations implemented genome browsing by serving a succession of HTML pages. First, this provides consistency to the whole genomic resource, an important prerequisite for ease of use; and second, the low technical requirement for viewing HTML allows the resources to be as widely accessible as possible. These are valid arguments, and we accept that any more sophisticated browsing facility would need to be an adjunct to current facilities, rather than a replacement for them.

One approach to providing a more sophisticated alternative would be to use a Java applet — a small program running within an HTML page. This need impose only minimal additional demands on the user's software, and one such applet, RGD Mapview[8], on the Rat Genome Database site, allows comparison of rat, mouse and human chromosomes. An alternative approach[9] involves scalable vector graphics (SVG). In SVG, information about objects to be displayed is in the form of simple mathematical descriptions, which occupy much less memory than the bitmap images supported by HTML. Resizing (scaling) or repositioning an object only involves changing the values of mathematical variables, in contrast with generating a new bitmap graphic and transferring it from the server. Unfortunately, SVG and the context in which it operates (the document object model) are currently supported only patchily, and implemented inconsistently, by different web browsers.

Finally, we would emphasize that providing a satisfactory web interface for browsing large genomes does pose both technical and financial problems. However, the effort of overcoming these problems will be worthwhile if it eventually provides 'human genome browsing for the rest of us'.

References

1. http://www.acedb.org/
2. http://www.gla.ac.uk/~dpl1n/BugView/
3. http://www.ncbi.nlm.nih.gov/COG/
4. ftp://ftp.ncbi.nih.gov/genomes/
5. http://www.ncbi.nlm.nih.gov/mapview/
6. http://www.ensembl.org/
7. http://www.gla.ac.uk/~dpl1n/SyntenyVista/
8. http://www.rgd.mcw.edu/VCMAP/mapview.shtml
9. http://www.gla.ac.uk/~dpl1n/SVGBrowser/

David Leader is at IBLS in the University of Glasgow, where for many years he studied the protein kinases of DNA viruses, and the phosphorylation of ribosomal proteins. His baptism into genomics was during the historic (prehistoric?) 1980s, but he marked the millennium by giving up experimental work on actin gene sequences and herpesvirus genomes to focus on bioinformatics. He is a Deputy Director of the Glasgow University Bioinformatics Research Centre, and is particularly interested in visualizing genomic data. He aspires to design biological software and systems that will (rather than can) be used by 'the man, or woman, at the bench'.

email: d.leader@bio.gla.ac.uk