Comparison of human (and other) genome browsers

*Terrence S. Furey**

Institute for Genome Sciences and Policy, Duke University, 101 Science Drive, Box 3382, Durham, NC 27708, USA
* *Correspondence to*: Tel: +1 919 668 4728; Fax: +1 919 668 0795; E-mail: terry.furey@duke.edu

Date received (in revised form): 28th October 2005

Abstract

The sequence of the human genome provides a scaffold on which numerous annotations, such the locations of genes, can be laid. Genome browsers have been created to allow the simultaneous display of multiple annotations within a graphical interface. In addition, they provide the ability to search for markers and sequences, to extract annotations for specific regions or for the whole genome and to act as a central starting point for genomic research. This review describes the basic functionality of genome browsers and compares three of them: the University of California Santa Cruz (UCSC) Genome Browser, the Ensembl Genome Browser and the NCBI MapViewer.

Keywords: *genome browsers, genome annotations, genome databases*

Introduction

Genome browsers allow researchers to navigate the genome in an analogous way to navigating the internet with Internet Explorer or Mozilla. As with the internet, the amount of available genomic data is overwhelming, and browsers aim to make these data accessible to all researchers. The number and variety of annotations has increased dramatically, enabling a detailed view of many aspects of the genome. Of course, one of the primary annotations is still the location and structure of genes, but even this is not straightforward, as many sources of information (sometimes conflicting) necessitate the creation of several gene-related annotations. These include the locations of mRNA and expressed sequence tag (EST) sequences deposited in the major sequence databases, curated gene sequence projects such as the Vertebrate Genome Annotation (VEGA),[1] RefSeq,[2] MGC[3] and ENSEMBL[4] and computational predictions such as GenScan[5] and Twinscan.[6]

There is a wide range of additional annotations. The locations of clones from bacterial artificial chromosome (BAC) and other clone libraries, sequence-tagged site (STS) markers from genetic maps[7–9] and estimated boundaries of cytogenetic bands[10] provide crucial mapping information. Alignments with genomic sequences from other species delineate regions of synteny and help to identify orthologous genes. Single nucleotide polymorphisms (SNPs) and other types of variation point to differences within a species. Locations of repetitive sequences, due both to retrotransposable elements and to simple repeats such as microsatellites, help to provide a more complete description of the genomic landscape. An incomplete listing of annotations is shown in Table 1. Browsers simultaneously display these annotations, allowing for the investigation and appreciation of the genomic context in which to consider a gene or region of interest.

Three browsers in particular, the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu),[11] the Ensembl Genome Browser (http://www.ensembl.org)[12] and the National Center for Biotechnology Information (NCBI) MapViewer (http://www.ncbi.nih.gov/mapview)[13] provide information portals for multiple genome sequences, including human. They share many common features, but differ in significant ways. The following presents an overview and comparison of these browsers.

Genome browser comparisons

Genome browsers can be described and compared with respect to presentation, content and functionality. Presentation refers to how the data are displayed in a graphical form and the overall structure of the website. Content refers to what data is accessible, such as particular genome sequences and annotations for a specific genome. Functionality refers to tools available for mining the genome sequence and annotations, such as sequence and text searches and data extraction.

The UCSC, Ensembl and NCBI genome browsers aim to present genomic data in a manner that will facilitate research, but they do so in different ways. Table 2 summarises some of these differences, and a more complete, yet still high-level, discussion of these is presented below.

Table 1. A sample of annotations found in one or more of the UCSC, Ensembl and NCBI genome browsers.

Type	Annotations
Mapping and sequence	Chromosome bands; GC percent; CpG Islands; restriction enzyme recognition sites; BAC and fosmid clones; STS markers from genetic, RH maps; Mitelman breakpoints
Genes, transcription and expression	RefSeq mRNAs; VEGA genes; Ensembl genes; UniGene; pseudogenes; retroposed genes; Non-coding RNA genes; tRNAs; mRNAs and ESTs; computational gene predictions; GNF Atlas expression values; Affymetrix microarray probes; DNase1 hypersensitive sites
Variation and repeats	SNPs from dbSNP, HapMap projects haplotypes; recombination rates and hotspots; segmental duplications; repetitive sequences (RepeatMasker); tandem repeats
Cross-species	Evolutionarily conserved regions; syntenic mappings to many organisms including chimp, mouse, rat, chicken, cow, dog, opossum, fish

Abbreviations: BAC, bacterial artificial chromosome; EST, expressed sequence tag; GNF, Genomics Institute of the Novartis Research Foundation; NCBI, National Center for Biotechnology Information; RH, Radiation hybrid; SNP, single nucleotide polymorphism; STS, sequence-tagged site; UCSC, University of California Santa Cruz; VEGA, Vertebrate Genome Annotation.

Presentation

UCSC features three types of browsers: a genome browser, a gene family browser (Gene Sorter) and a proteome browser. The genome browser is the most widely used and will be the focus of this discussion, although this in no way implies that the other two are not very valuable research tools. The primary web page of the genome browser consists of a graphic that displays annotations for some specified genomic region surrounded by navigational buttons and links to tools. The navigational buttons allow for zooming in and out or moving left or right along the genomic sequence. Within the graphic, annotations — also referred to as 'tracks' — are displayed horizontally, with the genome sequence running from left to right. The locations of specific elements within annotations are primarily indicated by boxes with lines sometimes connecting them to show relationships, such as in gene structures (boxes = exons, lines = introns). Arrows indicate forward or reverse strand, where applicable. The use of different colours and shading of boxes highlights the properties of certain annotations, such as confidence in the underlying data — as is the case in the Known Genes track — and quantitative traits, employed by the GC Percent track to indicate differing levels of content of guanine (G) and cytosine (C) base pairs. Clicking on an element within an annotation will bring up a separate 'details' web page with specific information about that element and links to other databases and resources such as GenBank[13] and SwissProt.[14] The amount of this additional information varies between annotations. Drop-down menus towards the bottom of the page, also accessible through a separate 'configuration' page, allow for the selection of annotations to display in the graphic (Table 1).

Ensembl structures its site around 'Views'. For humans, there are 22 Views that display different types of data and/or provide various functions. The primary View, analogous to the UCSC main browser page, is the ContigView. Within this View are three graphic displays that provide information at different resolutions for a region in the genome. The Overview graphic displays multiple megabases (Mbs), the Detailed view shows approximately 1 Mb and the Basepair view details about 100 bases. Similarly to UCSC, the genome is shown in a horizontal fashion with navigational buttons located within the Detailed view graphic. In the three graphics on this page, annotations are delineated by boxes, sometimes connected by lines and other times contained within a larger box. In the Detailed and Basepair views, the DNA contigs annotation divides the graphic with elements on the forward strand appearing above and on the reverse strand below. Clicking on an element in an annotation will cause a small pop-up window to appear with some basic information and possibly links to other Views within Ensembl or resources at other sites. For example, clicking on an Ensembl gene provides links to GeneView, TranscriptView and ProtView pages, which contain additional information about the gene or a region of the gene. Menus at the top of the Detailed view graphic provide the ability to select specific annotations for display.

The primary display of NCBI's MapViewer differs significantly from both UCSC and Ensembl by orienting the genome sequence in a vertical fashion. Again, boxes and lines indicate positions of elements in annotations, also referred to

Table 2. Feature comparison of the UCSC Genome Browser, Ensembl Genome Browser and NCBI MapViewer.

	UCSC	Ensembl	NCBI
Presentation	Genome in horizontal orientation Main page contains a single graphic displaying annotation ('tracks') Clicking on annotation element presents web page of detailed information and links to other resources	Genome in horizontal orientation Main ContigView page contains three graphics displaying annotations at different resolutions Clicking on annotation element presents box with links to other resources or Views with more detailed information	Genome in vertical orientation Annotations graphically presented in columns ('maps') Clicking on annotation elements or links in columns provides quick access to other, primarily NCBI, resources
Content	13 vertebrate, 15 invertebrate Many cross-species annotations including conservation across eight species ENCODE Project annotations	13 vertebrate, six invertebrate Heavy focus on gene annotations such as Ensembl genes and VEGA HapMap project-related Views	11 vertebrate, five invertebrate, one protozoan, 12 plant, eight fungi Annotations primarily from NCBI resources
Functionality	Text search, BLAT sequence search, isPCR primer search Advanced annotation extraction using Table Browser Ability to upload and view own annotations	Text search, BLAST and SSAHA sequence search, e-PCR primer search Advanced annotation extraction using BioMart Ability to upload and view own annotations Simultaneous view of syntenic regions	Text search, BLAST sequence search, e-PCR primer search Basic annotation extraction

Abbreviations: BLAT, BLAST-like alignment tool; ENCODE, ENCyclopedia Of DNA Elements; e-PCR, electronic polymerase chain reaction; NCBI, National Center for Biotechnology Information; SSAHA, Sequence search and alignment by hashing algorithm; UCSC, University of California Santa Cruz; VEGA, Vertebrate Genome Annotation.

as 'maps', which are presented in columns. The ability to navigate the genome is provided in a side bar to the left of the screen. Links within the annotations, as well as the LinkOut column, provide easy access to other relevant resources at the NCBI, such as Entrez Gene (formerly LocusLink),[15] Online Mendelian Inheritance in Man (OMIM)[16] and dbSNP.[17] A 'Maps & Options' button brings up a separate window, allowing one to select annotations to display.

Content

The NCBI provides access to the largest number of genome sequence assemblies, including 11 vertebrates, five invertebrates, one protozoan, eight plants and 12 fungi. Ensembl and UCSC are more heavily slanted towards the larger eucaryotic genomes, providing access to a similar set of 13 vertebrate genomes and six (Ensembl) or 15 (UCSC) invertebrates, and are devoid of the other classes of species.

Annotations available within the NCBI MapViewer primarily originate in the numerous databases available at the NCBI. The MapViewer, therefore, is very tightly integrated with these data sources, some of which — such as the Mitelman Breakpoint annotation — are not available at the other sites. UCSC and Ensembl also present annotations that originate from outside resources, such as the databases at NCBI, but supplement these with numerous additional annotations contributed by in-house or third-party researches.

The UCSC browser arguably contains the broadest set of annotations, especially in the area of cross-species comparisons. For example, the Conservation annotation, developed at and displayed only at UCSC, shows a measure of evolutionary conservation across eight vertebrate species, as determined by a phylogenetic hidden Markov model.[18] UCSC is also the official repository for, and displays data from, the ENCODE (Encyclopedia Of DNA Elements) project,[19] containing annotations ranging from histone modifications to regions of DNase 1 hypersensitivity.

The Ensembl browser contains the most extensive set of gene and transcription-related data, with 14 of its 22 Views primarily focused on the presentation of gene- or protein-related data. There is tight integration with gene data originating from both the Ensembl genes annotation[4] — a computationally generated evidence-based set that Ensembl produces — and the VEGA project[1] — a manual curation effort. The Ensembl browser also has the most extensive presentation of haplotype data, especially in their LDView, which was generated by the HapMap project.[20]

The underlying genomic sequence is exactly the same at all three sites, but analogous annotations may differ. For example, locations of mRNA and EST sequences require an alignment to the genome sequence. Their precise alignment may vary, based on the alignment program used and specific parameter settings within the program. The three sites do not employ the same alignment methods, resulting in slight differences, although they are in agreement for the vast majority of the time.

Functionality

There are many common functions that all three sites provide. Specific regions of interest can be quickly and easily displayed using keywords such as gene or marker names, exact base pair positions within chromosomes, or sequences via alignment programs like BLAST[21] (Ensembl and NCBI) or BLAST-like alignment tool (BLAT)[22] (UCSC). Locations of paired primer sequences can be obtained via electronic polymerase chain reaction (ePCR)[23] (NCBI and Ensembl) or isPCR (UCSC). Associated FTP sites allow for the download of complete genome sequences and annotations.

Annotation data can also be downloaded for particular regions. NCBI allows users to view annotations in a tabular format that can then be downloaded into a text file. Ensembl's BioMart[24] and the UCSC Table Browser[25] allow for both simple downloads of annotations and for quite complex datasets to be generated. These two tools also allow for the uploading of files of genomic regions or names of genes or markers for which annotation data, including the underlying sequence, can be obtained.

UCSC and Ensembl provide the ability for researchers to display their own annotation information within the browser. A simple text file denoting the base pair locations of annotation elements is uploaded and used to create a corresponding temporary annotation within the graphic, which is essentially only viewable by the originator. In this way, researchers can usefully view their own data within the context of all other available genomic data.

Ensembl provides the ability to view syntenic regions of two genomes simultaneously in their MultiContigView. The layout is similar to the ContigView described previously, but with the addition of data from two separate genomes being displayed in the Detailed view graphic, and a Navigational view replacing the Overview with a zoomed-out display of the regions being analysed in both genomes.

Last words

This overview of the UCSC, Ensembl and NCBI genome browsers is by no means complete and is not meant to recommend the use of one or the other of these sites. Users should explore the capabilities of each browser to determine the one they prefer. In the end, the browser that allows a researcher to be the most productive is the best.

The genome browsers reviewed here provide access to not only human genome sequence data, but also to annotations from an ever-growing set of species. Similar functionality for each genome assembly is provided for all species, although the range of annotations varies dramatically.

These are by no means the only genome-related browsers available, but they are among the most comprehensive. Similar browsers with more narrow foci, such as for a single organism, share many of the features and functions described above.

The quality of the publicly available data displayed in browsers is highly variable. Therefore, researchers must view this data as critically as any other. Appropriate experimentation is required as necessary to test the accuracy of any hypothesis generated using these data. Nevertheless, genome browsers offer a powerful research tool to be utilised by researchers worldwide.

References

1. Ashurst, J.L., Chen, C.K., Gilbert, J.G. *et al.* (2005), 'The vertebrate genome annotation (VEGA) database', *Nucleic Acids Res.* Vol. 33, pp. D459–D465, (Database issue).
2. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005), 'NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res.* Vol. 33, pp. D501–D504 (Database issue).
3. Gerhard, D.S., Wagner, L., Feingold, E.A. *et al.* (2004), 'The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC)', *Genome Res.* Vol. 14, pp. 2121–2127.
4. Curwen, V., Eyras, E., Andrews, T.D. *et al.* (2004), 'The Ensembl automatic gene annotation system', *Genome Res.* Vol. 14, pp. 942–950.
5. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.* Vol. 268, pp. 78–94.
6. Korf, I., Flicek, P., Duan, D. and Brent, M.R. (2001), 'Integrating genomic homology into gene structure prediction', *Bioinformatics* Vol. 17(Suppl 1), pp. S140–S148.
7. Kong, A., Gudbjartsson, D.F., Sainz, J. *et al.* (2002), 'A high-resolution recombination map of the human genome', *Nat. Genet.* Vol. 31, pp. 241–247.
8. Broman, K.W., Murray, J.C., Sheffield, V.C. *et al.* (1998), 'Comprehensive human genetic maps: Individual and sex-specific variation in recombination', *Am. J. Hum. Genet.* Vol. 63, pp. 861–869.
9. Dib, C., Faure, S., Fizames, C. *et al.* (1996), 'A comprehensive genetic map of the human genome based on 5,264 microsatellites', *Nature* Vol. 380, pp. 152–154.
10. Furey, T.S. and Haussler, D. (2003), 'Integration of the cytogenetic map with the draft human genome sequence', *Hum. Mol. Genet.* Vol. 12, pp. 1037–1044.

11. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al.* (2002), 'The human genome browser at UCSC', *Genome Res.* Vol. 12, pp. 996–1006.
12. Hubbard, T., Andrews, D., Caccamo, M. *et al.* (2005), 'Ensembl 2005', *Nucleic Acids Res.* Vol. 33, pp. D447–D453 (Database issue).
13. Wheeler, D.L., Barrett, T., Benson, D.A. *et al.* (2005), 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res.* Vol. 33, pp. D39–D45 (Database issue).
14. Bairoch, A., Apweiler, R., Wu, C.H. *et al.* (2005), 'The universal protein resource (UniProt)', *Nucleic Acids Res.* Vol. 33, pp. D154–D159 (Database issue).
15. Maglott, D., Ostell, J., Pruitt, K.D., *et al.* (2005), 'Entrez Gene: Gene-centered information at NCBI', *Nucleic Acids Res.* Vol. 33, pp. D54–D58 (Database issue).
16. Hamosh, A., Scott, A.F., Amberger, J.S. *et al.* (2005), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res.* Vol. 33, pp. D514–D517 (Database issue).
17. Sherry, S.T., Ward, M.H., Kholodov, M. *et al.* (2001), 'dbSNP: The NCBI database of genetic variation', *Nucleic Acids Res.* Vol. 29, pp. 308–311.
18. Siepel, A. and Haussler, D. (2004), 'Combining phylogenetic and hidden Markov models in biosequence analysis', *J. Comput. Biol.* Vol. 11, pp. 413–428.
19. ENCODE Project Consortium (2004), 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science* Vol. 306, pp. 636–640.
20. The International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
21. Altschul, S.F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.* Vol. 215, pp. 403–410.
22. Kent, W.J. (2002), 'BLAT — The BLAST-like alignment tool', *Genome Res.* Vol. 12, pp. 656–664.
23. Schuler, G.D. (1997), 'Sequence mapping by electronic PCR', *Genome Res.* Vol. 7, pp. 541–550.
24. Durinck, S., Moreau, Y., Kasprzyk, A. *et al.* (2005), 'BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis', *Bioinformatics* Vol. 21, pp. 3439–3440.
25. Karolchik, D., Hinrichs, A.S., Furey, T.S. *et al.* (2004), 'The UCSC Table Browser data retrieval tool', *Nucleic Acids Res.* Vol. 32, pp. D493–D496 (Database issue).