Key Features of the UCSC Genome Site

Heather Trumbower and Jennifer Jackson
UCSC Genome Bioinformatics Group
heather@soe.ucsc.edu

Abstract

The UCSC Genome site has been contributing to the public genomics community for over 5 years. We wish to highlight the key features, both new and old.

1. Introduction

The UCSC Genome site is *http://genome.ucsc.edu*. This site is a free, publically available web server that is a highly organized and curated repository for numerous genomic data sets. Tools are provided which support the ability to interactively view, query and collate these data. The fundamental data are genomic assemblies; annotations are published from both UCSC researchers and external collaborators. Key features are the Genome Browser[1,2], BLAT[3], the Gene Sorter[4], the Table Browser[5], the liftOver Utility, the Conservation tracks, and support for custom annotations. All source code and data annotations are available for download from *http://hgdownload.cse.ucsc.edu*. Restrictions exist for a limited number of data sets and for commercial licensing.

2. Data sets and sources

The UCSC Genome site houses assemblies for vertebrate, insect and other species. The vertebrate assemblies include human, chimp, rhesus, dog, cow, mouse, rat, opossum, chicken and frog. Three species of fish assemblies are also available: zebrafish, tetraodon and fugu. The insect assemblies include 6 fly species as well as honeybee and mosquito. Additional assemblies include the nematodes C. elegans and C. briggsae and the yeast S. cerevisiae. These assemblies are produced by numerous sequencing centers and are fully credited at http://genome.ucsc.edu/cgi-bin/hgGateway and http://genome.ucsc.edu/credits.html.

3. Annotations

Annotations are presented by the Genome Browser. Primary categories of annotations are mapping and sequencing, genes and gene predictions, mRNA and ESTs, expression and regulation, comparative genomics, and variation and repeats. Mapping annotations include STS markers and BAC End pairs, and a newly added restriction enzyme track. Gene annotations include RefSeq, Ensembl and Twinscan, the UCSC Known Genes collection, and the new CCDS collection (human-only; in collaboration with EBI and NCBI). mRNA and EST data is from Genbank, aligned by BLAT on a daily basis. A unique alternative-splicing visualization is available for the July 2003 human assembly. Expression data is from collaborators such as Affymetrix, GNF and NCI. Comparative genomics features the chained and netted pairwise alignments, multiple species alignments, a Most Conserved filter, cross-species protein alignments, and Ecores. Variations include presentation of the dbSNP alignments and segmental duplications from the Eichler lab. Repeats are presented from RepeatMasker and TRF.

4. Genome Browser and custom tracks

The Genome Browser is available at http://genome.ucsc.edu/cgi-bin/hgTracks. The browser can be positioned to a chromosomal location, or a location can be chosen based on the results of a search by identifier. Each track has a choice of display density; generally 4 modes are available: dense, squish, pack or full. Details are provided for each annotation displayed, including extended details for Known Genes. A new configuration page supports the ability to apply visibility settings by annotation category. Each annotation includes a description of the methods used and most have additional filtering options.

Site visitors can add their own annotations to their session via a procedure described at

http://genome.ucsc.edu/goldenPath/help/customTrack.html. Custom tracks can be in GFF, GTF, PSL, BED or WIG format. The data can be pasted directly, uploaded from a file or redirected from a URL. The data can be shared with others by adding the hgsid value to the URL. The hgsid value is the session ID stored on the UCSC servers and available from http://genome.ucsc.edu/cgi-bin/cartDump.

5. BLAT, In-Silico PCR and Short Match

UCSC continues to support BLAT servers for all assemblies, allowing for fast and thorough alignment of nucleotide or protein sequences. A recently added extension is the support for primer validation for some assemblies at http://genome.ucsc.edu/cgi-bin/hgPcr. All assemblies also offer a Short Match track in the Genome Browser which displays all matches of a nucleotide sequence 2 to 30 bases in length.

6. Gene Sorter

The Gene Sorter is a gene-centric view of annotation relationships including protein-level homology, GO classification, and Pfam domain structures. Genes of interest can be grouped into clusters and filtered based on criteria such as expression profiles and/or tissue specificity. Results include links to the genomic positional view and to tools allowing customized sequence and text download options. http://www.genome.ucsc.edu/cgi-bin/hgNear

7. Table browser

The Table Browser has recently been redesigned to support annotation track based queries. The Table Browser allows query construction using filters and intersections, with output options of BED, GTF, custom track, all or some fields, or sequence. Filters can include nested linked tables; intersections can accommodate overlaps.

8. Lift over

The liftOver utility effectively maps one genome to another, allowing rapid identification of regions of interest between successive assemblies of the same species or between two distinct species. Both the unix-based and web-based versions of the liftOver utility accept a range of user-specified mapping parameters. http://genome.ucsc.edu/cgi-bin/hgLiftOver.

9. Mailing list and FAQs

The UCSC Genome Bioinformatics group hosts a mailing list genome@soe.ucsc.edu for member discussion and individual questions regarding use of the Genome site. In addition, a set of FAQs and answers is available within the site at http://www.genome.ucsc.edu/FAQ.

10. Conservation track

The Conservation track, based on a phylogenetic hidden Markov model (phastCons) constructed using the multiz [6] algorithm, shows a measure of evolutionary conservation between multiple species against a target genome. Both overall and paired conservation is scored and displayed in manner that can be customized to highlight specific aspects of the modeled relationships.

11. References

[1] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, "The Human Genome Browser at UCSC", *Genome Research*, CSHL Press, Vol. 12, 2002, pp. 996-1006.

[2] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler and W.J. Kent, "The UCSC Genome Browser Database", *Nucleic Acids Research,* Oxford University Press, Vol. 31, No. 1, 2003, pp. 51-54.

[3] W.J. Kent, "BLAT: The BLAST-like Alignment Tool", *Genome Research,* CSHL Press, Vol. 12, No. 4, 2002, pp. 656-664.

[4] W.J. Kent, F. Hsu, D. Karolchik, R.M. Kuhn, H. Clawson, H. Trumbower, D. Haussler, "Exploring relationships and mining data with the UCSC Gene Sorter", *Genome Research*, CSHL Press, Vol. 15, 2005, pp. 737-741

[5] D. Karolchik, A. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler and W.J. Kent, "The UCSC Table Browser data retrieval tool", *Nucleic Acids Research,* Oxford University Press, Vol. 32, Database issue, 2004, pp. D493-D496.

[6] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. (2004). Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner. Genome Res. 14(4):708-15.

.