Visually integrating genomic data in the UCSC Genome Browser

Angie S. Hinrichs, Fan Hsu, Robert M. Kuhn, Brian J. Raney, Galt Barber, Ann S. Zweig, Greg Roe, David Haussler[1] and W. Jim Kent
Center for Biomolecular Science and Engineering and [1]Howard Hughes Medical Institute, University of California Santa Cruz (UCSC)

Landscape of a Chromosome

The UCSC Genome Browser can display genomic data at any resolution from an entire chromosome down to a single base. Here is a bird's-eye view of human chromosome 6, featuring several phenotype and variation data tracks.

In addition, a small "custom track" has been uploaded (see "Custom Tracks", right) for illustrative purposes: several deletions from a fictional patient are displayed in purple, for comparison against known chromosomal abnormalities and regions associated with disease.

chromosome bands

disease gene annotations

uploaded custom track

large deletions, amplifications, regions associated with disease

structural variants observed in populations

Custom Tracks: View Your Data

Custom tracks are user-supplied data that can be uploaded to the Genome Browser. Any data that maps with genomic coordinates can be displayed down to base-pair resolution. Multiple file formats allow simple block display, gene structures, sequence alignments, and plots of values mapped to coordinates.

This simple example custom track appears in the first three sections on the left:

```
  track name=pretendPatient description="======> Custom Track: Deletions
observed in Imaginary Patient <======" color=200,0,100 visibility=pack
  chr6    65000000    65100000    smallDeletion
  chr6    127000000   132500000   bigDeletion
  chr6    132910000   132930000   smallDeletion
```

The last section on the left contains three custom tracks using 1000 Genomes Project files served by FTP from NCBI: two BAM alignment files and one VCF file with genotypes.

Homing in on Disease Genes

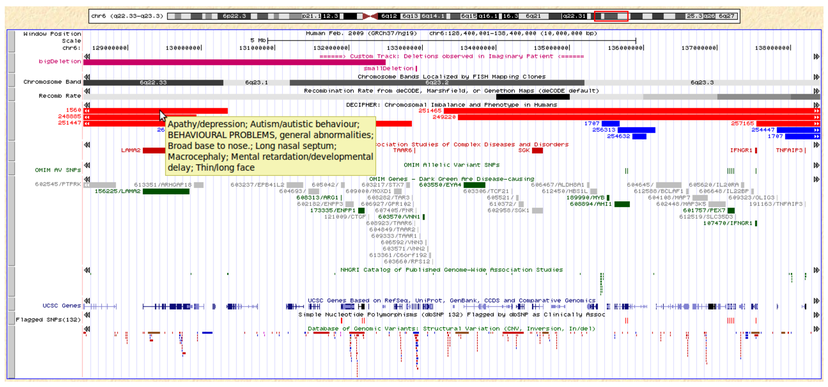
In this 10 million base region of chromosome 6, we can expand some of our tracks to see more details about their annotations. Hovering the mouse/pointer over items in some tracks provides a pop-up with more information. For example, chromosomal abnormalities in the DECIPHER track are labeled by patient number, and the pop-up displays patient phenotypes.

mouse over for more info

disease gene annotations

GWAS catalog

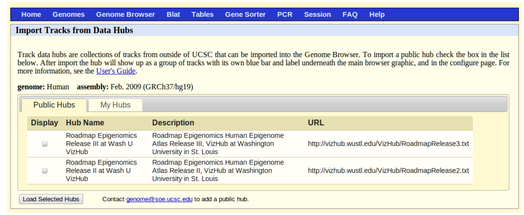
dbSNP "clinically associated"



Data Hubs: Organize Hundreds of CTs

Data Hubs are web-accessible directories of genomic data files that can be viewed on the Genome Browser alongside the native annotation tracks. Creating a Data Hub allows a project's tracks to be organized into composite and super-tracks, so that collections of tracks that share an attribute such as cell line or ChIP factor can be selected and configured as a group.

Data hubs can be public (listed in the Genome Browser) or unlisted (users must paste in the URL of the data hub).



Regulation, Conservation, Variation

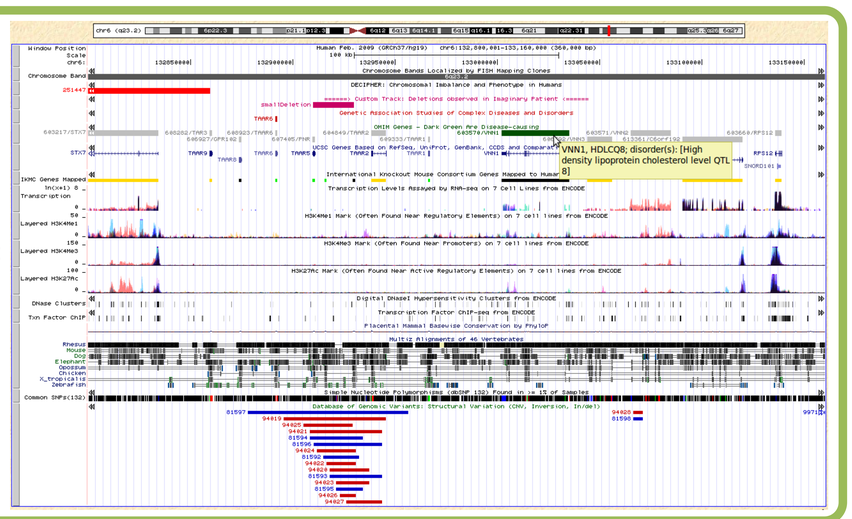
Zooming in further to view two gene clusters (TAARs, VNNs) in the neighborhood of our smaller fictional deletion, we can observe transcription levels and histone modifications identified by the ENCODE project, and cross-species nucleotide-level alignments used to infer conservation.

orthologous mouse knockouts (IKMC)

ENCODE project integrated regulation

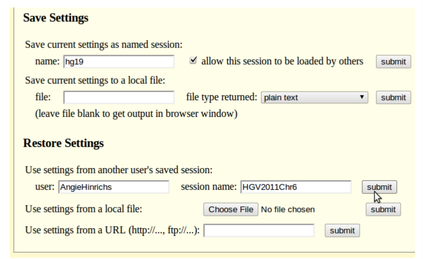
multi-species alignments and conservation

common short variants



Sharing Sessions

Once you have configured the Genome Browser to show exactly the regions and data tracks that make your point (including Custom Tracks, if any), save it as a session to share with colleagues. Load the session using the interface shown here or send a URL and a single click opens the session.



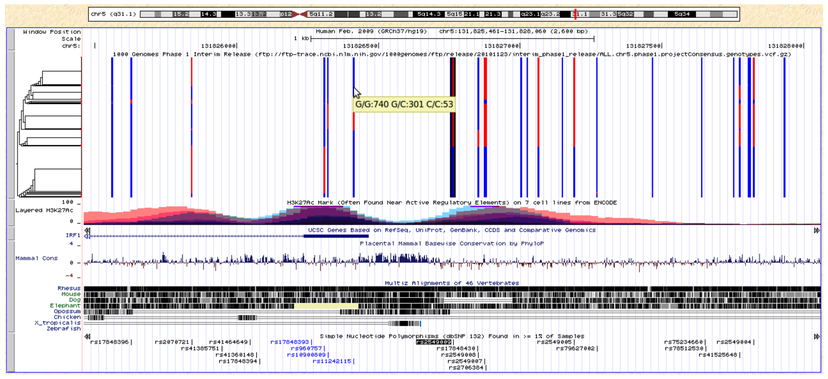
Haplotypes Near Regulatory SNP

Searching for a SNP such as rs2549009 (which partially predicts transcriptional activity of IRF1, possibly in conjunction with an unidentified variant*) causes the SNP ID to be highlighted; here, we can compare phased genotypes observed in 1,094 samples by the 1000 Genomes Project.

* Functional relevance of the IRF-1 promoter polymorphism rs2549009 on transcriptional activity in a native genomic environment.
Mertens J, Ramadori G, Mihm S. Hum Mol Genet. 2010 Dec 1;19(23):4587-94.

custom track: VCF genotypes from 1000 Genomes Project

SNP highlighted from search



Other UCSC Tools

Table Browser filters, intersects, combines, reformats Genome Browser data.

Genome Graphs plots signals (e.g. your GWAS data) along all chromosomes.

Gene Sorter finds similar genes by similarity of sequence, GO terms, PFAM and more.

Proteome Browser displays protein properties along peptide sequence.

Blat swiftly aligns submitted nucleotide or protein sequences to the reference genome.

isPCR, based on Blat, finds amplicons given primer sequences.

VisiGene provides a Google Maps-like interface to high-resolution mouse and frog *in situ* images.

Downloads: text files for entire database contents, repeat-masked assembly sequence, source files for selected tracks e.g. cross-species alignments and conservation.

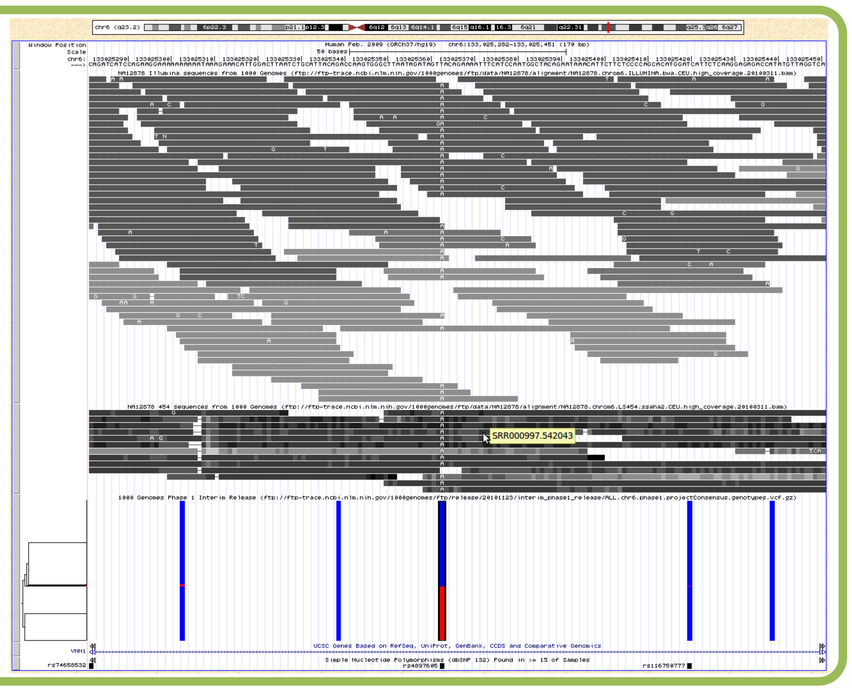
Resequencing: Reads to variants

Next-generation sequencing reads that have been aligned to the genome can be colored in several ways: gray scale alignment quality (1000 Genomes Illumina data, right), gray scale base qualities (1000 Genomes 454 data, right), or red/blue by strand (e.g. for RNA-seq reads, not shown).

When viewing a sufficiently small region, the reference genome bases are displayed at the top of the image; on aligned reads, bases that differ from the reference genome are highlighted. At right is an example of an apparent homozygous non-reference allele in 1000 Genomes Project's high-coverage sequencing of NA12878, also represented in the phased genotypes below.

custom tracks: BAM alignments from 1000 Genomes Project

custom track: VCF genotypes from 1000 Genomes Project



More Information

Click "**Help**" link (upper right) to get tool-specific help pages.

Search for answers to questions:
http://genome.ucsc.edu/contacts.html

Or **email** your question to the actively monitored public list:
genome@soe.ucsc.edu

OpenHelix provides free training material:
http://www.openhelix.com/downloads/ucsc/ucsc_home.shtml
and also offers training seminars (some free or discounted).

http://genomewiki.ucsc.edu/index.php/HGV2011

Acknowledgements

This work was funded by National Human Genome Research Institute (2 P41 HG002371-06 to UCSC Center for Genomic Science, 3 P41 HG002371-06S1 ENCODE supplement to UCSC Center for Genomic Science); National Cancer Institute (Contract No. N01-CO-12400 for Mammalian Gene Collection).
We would like to acknowledge the work of the UCSC Genome Bioinformatics technical staff (*http://genome.ucsc.edu/staff.html*), our many collaborators, and our users for their feedback and support.

Reference

The UCSC Genome Browser Database: update 2011. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. Nucleic Acids Res. 2011 Jan;39(Database issue):D876-82.

HGV 2011. Berkeley, CA. Sep. 8-10, 2011