

UiO : **Department of Biosciences**
University of Oslo

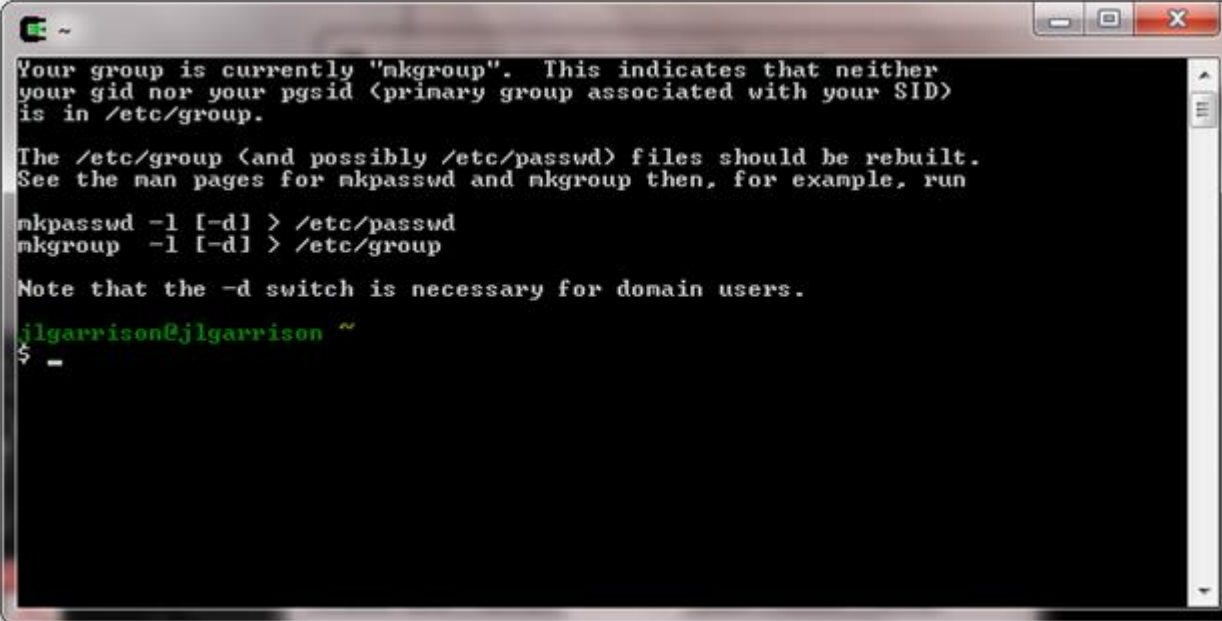
MBV4410/9410 Fall 2016

Bioinformatics for Molecular Biology



Use Unix locally

Windows: Cygwin (<https://www.cygwin.com/>)



```
~  
Your group is currently "nkgroup". This indicates that neither  
your gid nor your pgsid (primary group associated with your SID)  
is in /etc/group.  
  
The /etc/group (and possibly /etc/passwd) files should be rebuilt.  
See the man pages for mkpasswd and mkgroup then, for example, run  
  
mkpasswd -l [-d] > /etc/passwd  
mkgroup -l [-d] > /etc/group  
  
Note that the -d switch is necessary for domain users.  
jlarrison@jlarrison ~  
$ -
```

Use Unix locally

- Mac: Terminal, iTerm2 (<https://www.iterm2.com/>)

The screenshot shows a Mac Terminal window titled "2. Default (Vim)". The terminal output is as follows:

```
Processes: 174 total, 4 running, 170 sleeping, 620 threads
Load Avg: 0.74, 0.61, 0.59  CPU usage: 5.35% user, 7.58% sys, 87.5% idle
SharedLibs: 5196K resident, 6732K data, 0B linkedit,
MemRegions: 30149 total, 1630M resident, 47M private, 635M shared.
PhysMem: 953M wired, 2108M active, 1002M inactive, 4064M used, 36M free.
VM: 396G vsize, 1042M framework vsize, 1277547(0) pageins, 226575(0) pageouts.
Networks: packets: 538782/496M in, 340809/116M out.
Disks: 804097/126 read, 1354778/286 written.

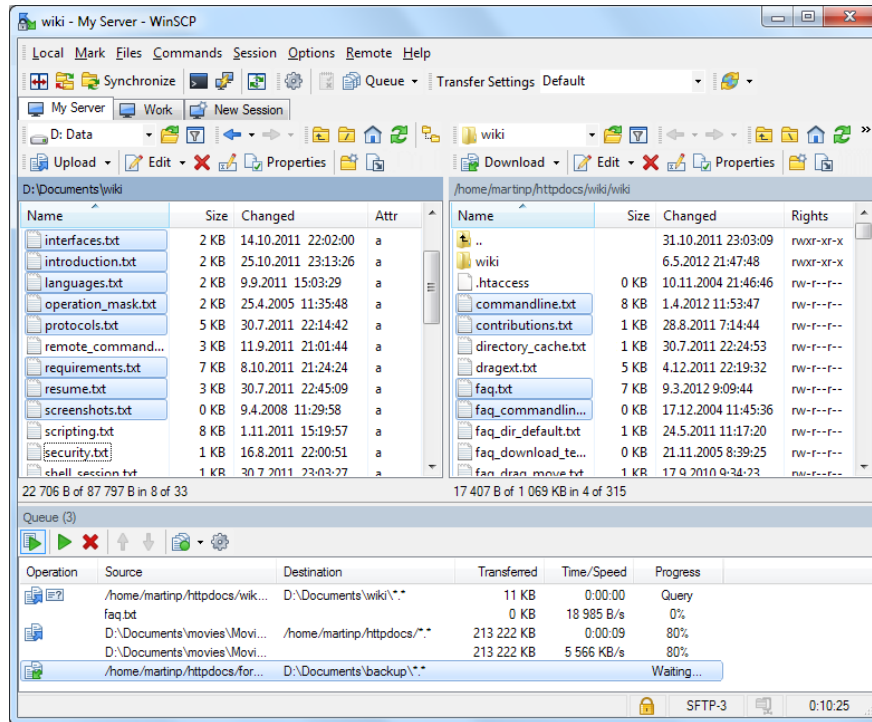
PID  COMMAND   %CPU  TIME    #TH  #VQ  #PORT #MREG RPRVT  RSHRD  RSIZE  VPRVT
90830- Google Chrom 0.0  00:03.05  6    1   100   215  9672K  49M   31M   34M
39058- screencaptur 0.0  00:00.03  2    1   41-   87-   476K-  13M  2948K- 12M-
39055- quicklookd  0.0  00:00.27  7    3   84-   84-  2728K- 7388K  6632K- 543M-
39052- top          10.3 00:02.02 1/1  0    26    33  1468K+ 264K  2044K+ 17M
```

Below the process list, the terminal shows the source code of the xterm program, starting with a license notice:

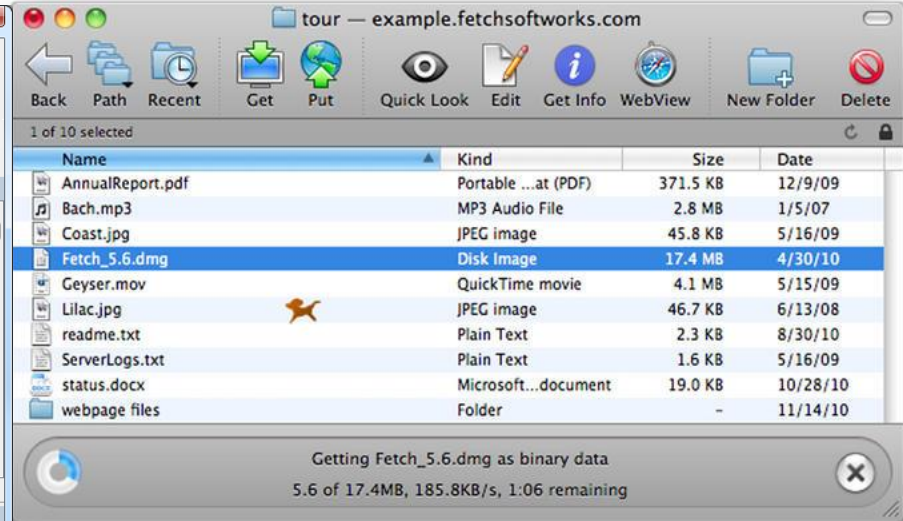
```
/* -*- mode:objc -*-
// $Id: PseudoTerminal.m,v 1.437 2009-02-06 15:07:23 delx Exp $
//
/*
** PseudoTerminal.m
**
** Copyright (c) 2002, 2003
**
** Author: Fabian, Ujwal S. Settler
** Initial code by Kiichi Kusama
**
** Project: iTerm
**
** Description: Session and window controller for iTerm.
**
** This program is free software; you can redistribute it and/or modify
** it under the terms of the GNU General Public License as published by
** the Free Software Foundation; either version 2 of the License, or
** (at your option) any later version.
**
** This program is distributed in the hope that it will be useful,
** but WITHOUT ANY WARRANTY; without even the implied warranty of
** MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
** GNU General Public License for more details.
**
** You should have received a copy of the GNU General Public License
** along with this program; if not, write to the Free Software
** Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
**/
```

The source code continues with various includes and function definitions, including a table of character sets and a large block of code for handling window and session management.

WinSCP



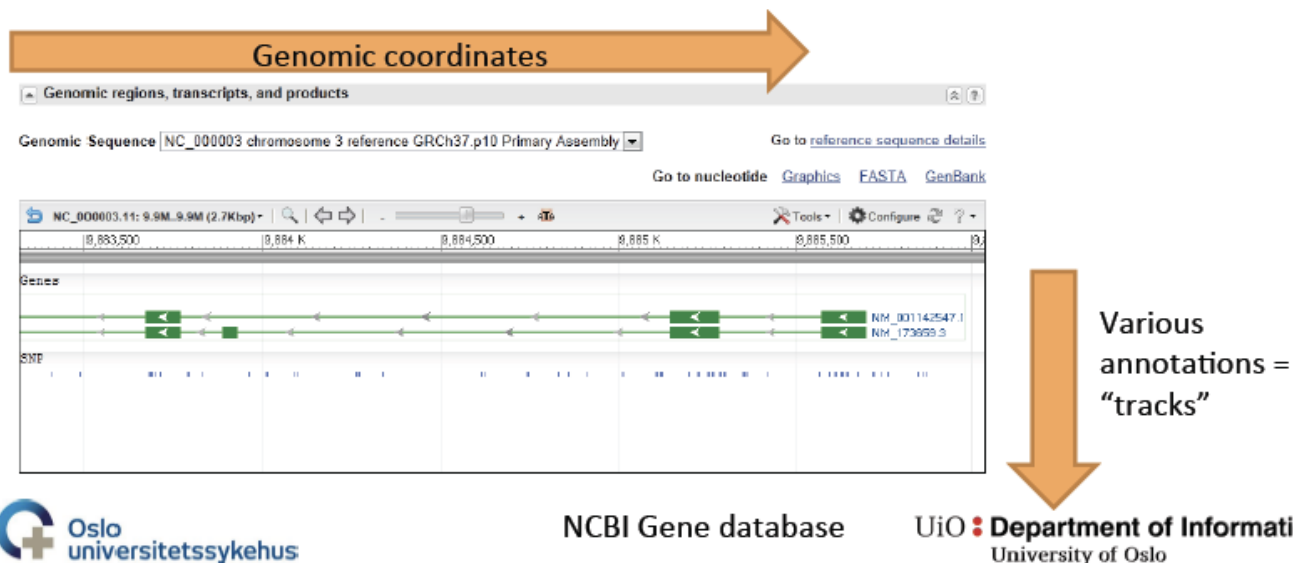
Fetch



Ensembl genome browser and database

Genome browsers

- Graphical interface for genomic data
- Shows information from biological databases mapped onto genomic sequence



Ensembl Genome Browser

- Joint project between EMBL-EBI and the Wellcome Trust Sanger Institute
- Central resource for studying genomes of vertebrates
 - Mainly chordates, but some few extra (*e.g. C. elegans* and *S. cerevisiae*)
 - Updated several times a year with new genome assemblies and new species
 - Annotations of genomes (*e.g.* genes and their splice variant, SNPs) added by the Ensembl pipeline
 - Automatic gene prediction (with or without experimental evidence) & some curator input

Ensembl Genome Browser

Search: for

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

- Human (GRCh37)
- Mouse (GRCm38)
- Zebrafish (Zv9)

★ [Log in to customize this list](#)

All genomes

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre](#) and [Ensembl Genomes](#)

What's New in Release 73 (September 2013)

- New search engine using Solr
- New species: Duck (*Anas platyrhynchos*) and Flycatcher (*Ficedula albicollis*)
- Updated patches for the human assembly (GRCh37.p12)

[Full details of this release](#)

[More release news on our blog →](#)

Latest blog posts

- 01 Nov 2013: [Upgrade of public MySQL](#)
- 01 Nov 2013: [Retirement of archive](#)
- 29 Oct 2013: [Take our API course](#)

[Go to Ensembl blog →](#)

Did you know...?

If you're using the VEP, use the [cache](#). Our Predictor tool also has an [interface](#).

Ensembl is a joint project between EMBL, EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements page](#) includes a list of additional current and previous funding bodies.

Ensembl release 73 - September 2013 © [WTS](#) / [EBI](#)

[About Ensembl](#) | [Privacy Policy](#) | [Contact Us](#)

Permanent link - [View in archive site](#)

<http://www.ensembl.org>

Excellent resource for exploring vertebrate species where the genome has been sequenced

Ensembl Genome Browser

The screenshot shows the Ensembl Genome Browser interface with a grid of species icons and names. Two large blue arrows point towards the center: one labeled 'Oslo' and another labeled 'New'. Below the grid, the text 'Currently approx. 80 species' is displayed.

Species listed (from top-left to bottom-right):

- Aardvark (review - assembly only)
- Alpaca
- Anole lizard
- Armadillo (review - new assembly, GRCv3.0)
- Baboon (review - assembly only)
- Budgerigar (review - assembly only)
- Bushbaby
- Ciona intestinalis
- Ciona savignyi
- Caenorhabditis elegans
- Cat
- Cave fish (review - assembly only)
- Chicken
- Chimpanzee (review - new assembly)
- Gibbon
- Gorilla
- Guinea Pig
- Hedgehog
- Horse
- Human
- Hyrax
- Kangaroo rat
- Lamprey
- Lesser hedgehog tenrec
- Macaque
- Marmoset
- Medaka
- Megabat
- Platypus
- Prairie vole (review - assembly only)
- Rabbit
- Chinese hamster
- Chinese softshell turtle
- Cod
- Coelacanth
- Cow
- Dog
- Dolphin (review - new assembly, T10v14)
- Duck
- Elephant
- Ferret
- Flycatcher
- Fruitfly
- Fugu
- Microbat
- Mouse
- Mouse Lemur
- Naked mole-rat (review - assembly only)
- Olive Baboon (review - assembly only)
- Opossum
- Orangutan
- Painted Turtle (review - assembly only)
- Panda
- Pig
- Pig FPC map (review - assembly only)
- Pika (review - new assembly, Oct2013)
- Platyfish
- Tasmanian devil
- Tetraodon
- Tilapia
- Tree Shrew
- Turkey
- Vervet monkey (review - assembly only)
- Wallaby
- Xenopus
- Zebra Finch
- Zebrafish

Currently approx. 80 species

EnsemblGenomes

The screenshot shows the EnsemblGenomes website. The header includes the logo, navigation links (About, Species, Working with communities, FAQs), a search bar, and a taxonomic filter (Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates). The main content area is divided into two columns. The left column, titled 'Ensembl Genomes: Extending Ensembl across the taxonomic space.', features a list of updates: 'Assembly Mapping' (with a diagram of genomic assembly), 'Rice genome updated', 'Improved resources for wheat genomes', 'Four aquatic metazoan genomes', and 'Ensembl Genomes REST Service'. Below this list is a paragraph about the development and funding of the project, accompanied by logos for EMBL-EBI and the Norwegian Research Council. The right column contains a section for the 'Agricultural-Omics Training Course' with a registration link, a 'What's New in Release 20 (September 2013)' section detailing updates to Bacteria, Fungi, and Metazoa, and a 'Have a question?' section with a link to frequently asked questions.

Ensembl Genomes: Extending Ensembl across the taxonomic space.

- Assembly Mapping**
For genomes where Ensembl Genomes has provided older assembly versions in the past, assembly mappings are now available. These can be accessed using the Post API or REST service, or via the assembly converter in the web interface. This is available for plants, metazoa, fungi and protists.
- Rice genome updated**
- Improved resources for wheat genomes**
- Four aquatic metazoan genomes**
- Ensembl Genomes REST Service**

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualization of genomic data. For details of our funding please click [here](#).

EMBL-EBI

Agricultural-Omics Training Course

Register now for an upcoming EBI training course in Agricultural-Omics. For more details, please go to <http://www.ebi.ac.uk/training/course/agricultural-omics>
posted 2013-10-03

What's New in Release 20 (September 2013)

The twentieth release of Ensembl Genomes features updates to version 73 of the Ensembl software across all divisions, and a number of new genomes added bringing the total number of genomes to 9225 ([full list](#)). Detailed notes can be found [here](#). See the individual homepages for more information.

Ensembl Bacteria

Ensembl Bacteria has been updated to include the latest versions of 9 069 genomes (8 842 eubacteria and 247 archaea) from the INSDC archives. Cross-references to Rhea and MetaCyc have also been added, as have Enzyme Commission classifications. In addition, data from RegulonDB have been used to add operon and other regulatory features to *E. coli* K-12 MG1655.

Ensembl Fungi

Two new plant pathogen genomes, *Microbotryum violaceum* and *Blumeria graminis*. Cross-references to PHI-base were added for plant pathogens.

Ensembl Metazoa

Three metazoan species have updated assemblies and gene models in release 20 of Ensembl Metazoa: the pea aphid, the western honey bee, and the purple sea urchin. The variation data for *Anopheles gambiae* has been updated to include ~7.5 million additional variants.

Ensembl Plants

The first assembly of the bread wheat genome, *Triticum aestivum*, from the IWGSC has been added in this release. In addition we have loaded the latest assembly for *Oryza sativa* from IRGSP.

Have a question?

Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, [contact us](#).

- Bacteria, protists, fungi, plants and other metazoa

Ensembl Genome Browser

very brief demo

The screenshot shows the Ensembl Genome Browser homepage. At the top, there's a navigation bar with links like BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar is on the right. Below the navigation bar, there's a search section with a dropdown menu for 'All species' and a text input field. The main content area is divided into several sections: 'Browse a Genome' with a description of the project and a list of popular genomes (Human, Mouse, Zebrafish); 'What's New in Release 73 (September 2013)' with a list of updates; 'Latest blog posts' with a list of recent posts; and a 'Did you know...' section with a link to the VEP tool. The footer contains information about the project's funding and a link to the 'About Ensembl' page.

<http://www.ensembl.org>

Explore in
exercise!

UCSC Genome Browser

UCSC Genome Browser

- Developed and maintained at the University of California, Santa Cruz (UCSC)
- Interactive website
- Access to genome sequence data from
 - Human genome
 - Latest assembly (GRCh38), the 2nd latest (GRCh37), but also earlier versions
 - Mouse, rat, and approx. 50 other mammals
 - Chicken, turkey, budgerigar, reptiles, frogs, and fishes
 - Insects, nematodes, *S. cerevisiae* and more
 - In total 91 species in 2014

UCSC Genome Browser

BRIEFINGS IN BIOINFORMATICS, VOL 14, NO 2, 144–161
Advance Access published on 20 August 2012

doi:10.1093/bib/bbs038

The UCSC genome browser and associated tools

Robert M. Kuhn, David Haussler and W James Kent

Submitted: 8th February 2012; Received (in revised form): 9th June 2012

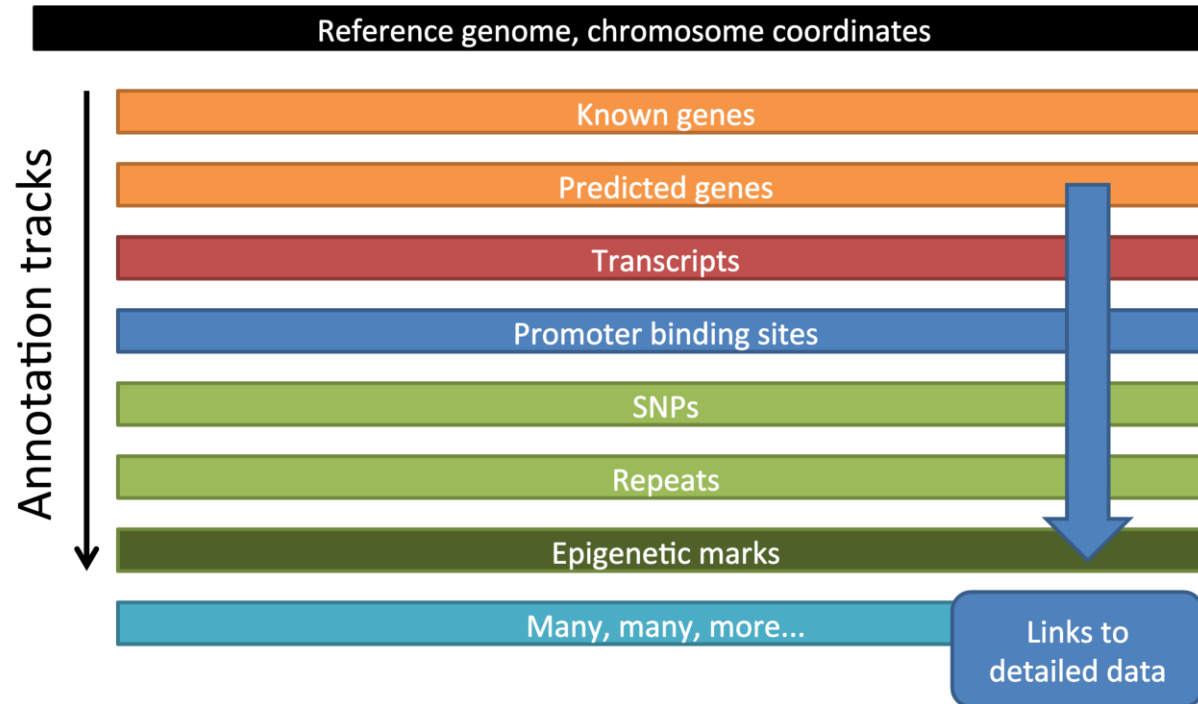
<http://genome.ucsc.edu>

Abstract

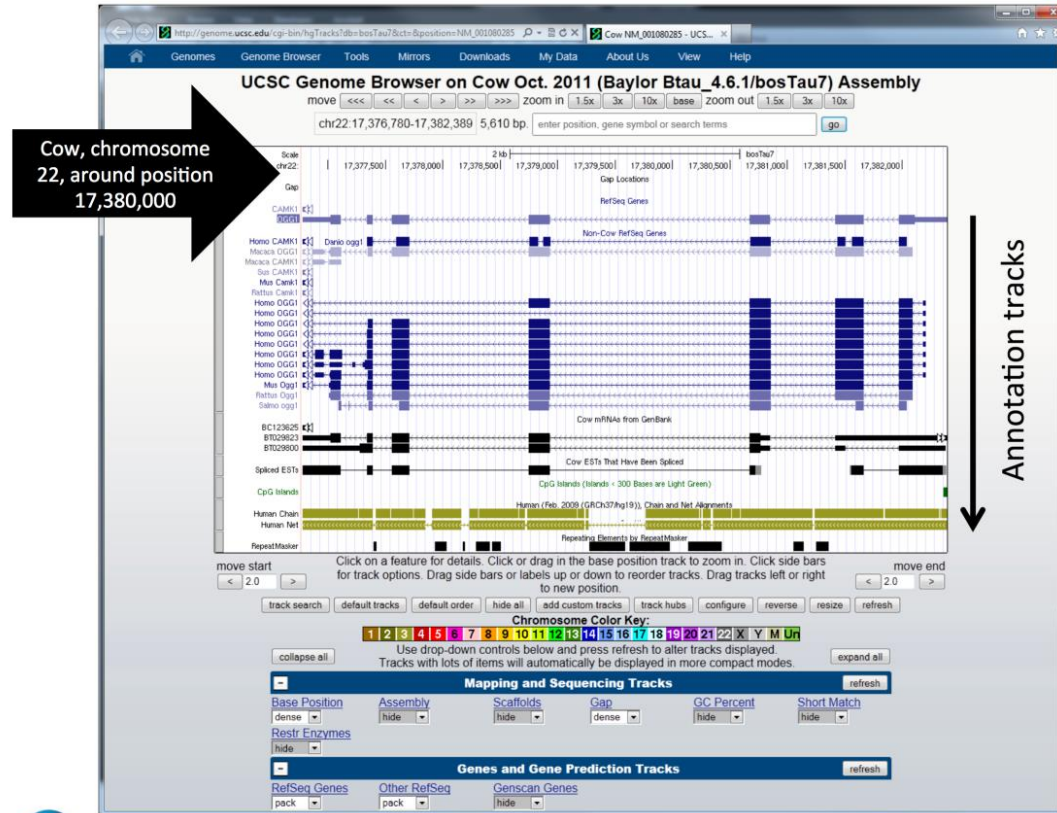
The UCSC Genome Browser (<http://genome.ucsc.edu>) is a graphical viewer for genomic data now in its 13th year. Since the early days of the Human Genome Project, it has presented an integrated view of genomic data of many kinds. Now home to assemblies for 58 organisms, the Browser presents visualization of annotations mapped to genomic coordinates. The ability to juxtapose annotations of many types facilitates inquiry-driven data mining. Gene predictions, mRNA alignments, epigenomic data from the ENCODE project, conservation scores from vertebrate whole-genome alignments and variation data may be viewed at any scale from a single base to an entire chromosome. The Browser also includes many other widely used tools, including BLAT, which is useful for alignments from high-throughput sequencing experiments. Private data uploaded as Custom Tracks and Data Hubs in many formats may be displayed alongside the rich compendium of precomputed data in the UCSC database. The Table Browser is a full-featured graphical interface, which allows querying, filtering and intersection of data tables. The Saved Session feature allows users to store and share customized views, enhancing the utility of the system for organizing multiple trains of thought. Binary Alignment/Map (BAM), Variant Call Format and the Personal Genome Single Nucleotide Polymorphisms (SNPs) data formats are useful for visualizing a large sequencing experiment (whole-genome or whole-exome), where the differences between the data set and the reference assembly may be displayed graphically. Support for high-throughput sequencing extends to compact, indexed data formats, such as BAM, bigBed and bigWig, allowing rapid visualization of large datasets from RNA-seq and ChIP-seq experiments via local hosting.

Kuhn *et al.* Brief. Bioinform. **14**, 144 (2012)

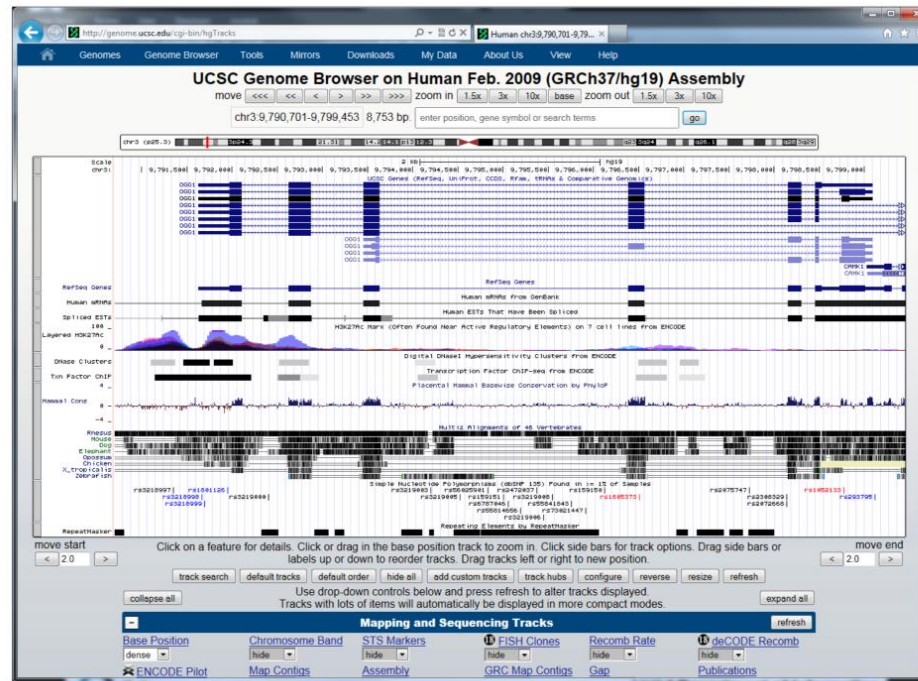
UCSC Genome Browser



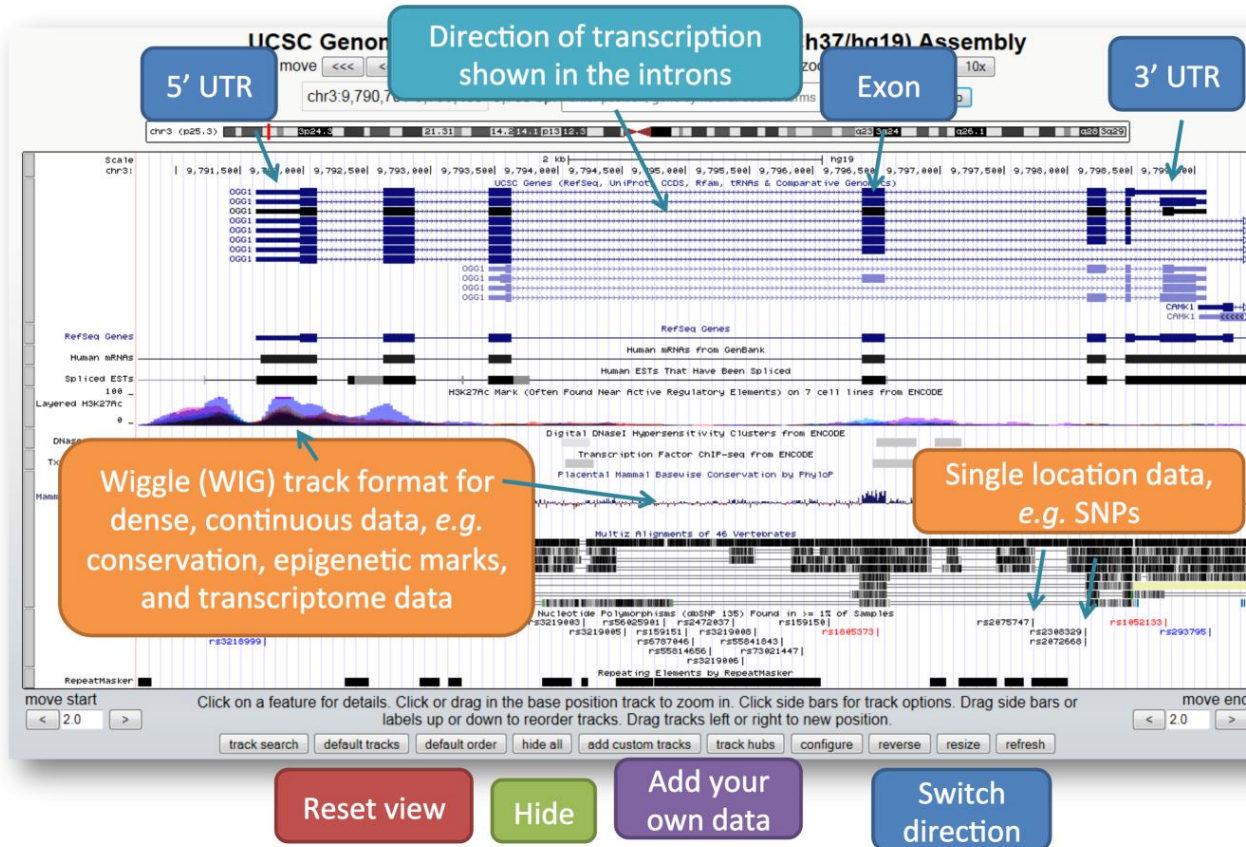
UCSC Genome Browser



UCSC Genome Browser brief demo



Different kinds of data



Help to investigate correct splicing?

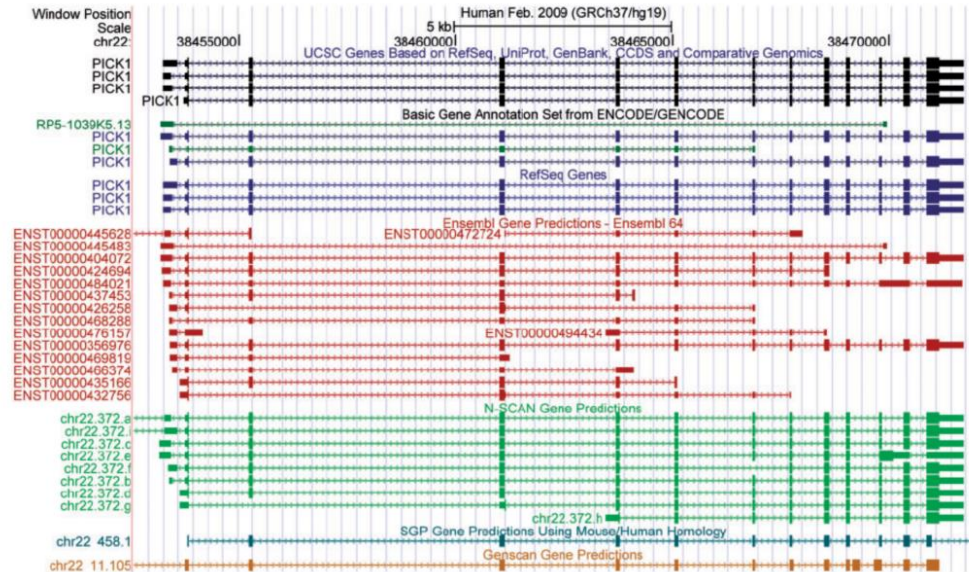


Figure 1: Screenshot of UCSC Genome Browser displaying human PICK1 gene region on chr22 in hg19 assembly. Different gene prediction algorithms predict different annotations in the region. By presenting multiple data sets of similar type, the user is able to more easily evaluate hypotheses. The different tracks often predict different 3'- and 5'-untranslated regions (half-height boxes on ends of annotations), coding regions (full-height boxes), introns (thin line with transcription-direction arrows) or start and end coordinates. The differences may be used to establish a level of confidence in an annotation not obtained from any single method.

ENCODE data in UCSC

Published online 30 October 2010

Nucleic Acids Research, 2011, Vol. 39, Database issue D871–D875
doi:10.1093/nar/gkq1017

ENCODE whole-genome data in the UCSC genome browser (2011 update)

Brian J. Raney^{1,*}, Melissa S. Cline¹, Kate R. Rosenbloom¹, Timothy R. Dreszer¹, Katrina Learned¹, Galt P. Barber¹, Laurence R. Meyer¹, Cricket A. Sloan¹, Venkat S. Malladi¹, Krishna M. Roskin¹, Bernard B. Suh¹, Angie S. Hinrichs¹, Hiram Clawson¹, Ann S. Zweig¹, Vanessa Kirkup¹, Pauline A. Fujita¹, Brooke Rhead¹, Kayla E. Smith¹, Andy Pohl¹, Robert M. Kuhn¹, Donna Karolchik¹, David Haussler^{1,2} and W. James Kent¹

¹Center for Biomolecular Science and Engineering, School of Engineering and ²Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

Received September 15, 2010; Accepted October 9, 2010

ABSTRACT

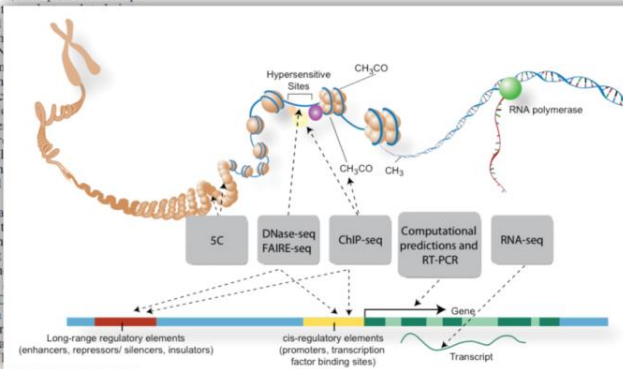
The ENCODE project is an international consortium with a goal of cataloguing all the functional elements in the human genome. The ENCODE Data Coordination Center (DCC) at the University of California, Santa Cruz serves as the central repository for ENCODE data. In this role, the DCC offers a collection of high-throughput, genome-wide data generated with technologies such as ChIP-Seq, RNA-Seq, DNA digestion and others. This data helps illuminate transcription factor-binding sites, histone marks, chromatin accessibility, DNA methylation, RNA expression, RNA binding and other cell-state indicators. It includes sequences with quality scores, alignments, signals calculated from the alignments, and in most cases, element or peak calls calculated from the signal data. Each data set is available for visualization and download via the UCSC Genome Browser (<http://genome.ucsc.edu/>). ENCODE data can also be retrieved using a metadata system that captures the experimental parameters of each assay. The ENCODE web portal at UCSC (<http://encodeproject.org/>) provides information about the ENCODE data and links for access.

into RNA, which might be spliced, transported to an appropriate cellular compartment

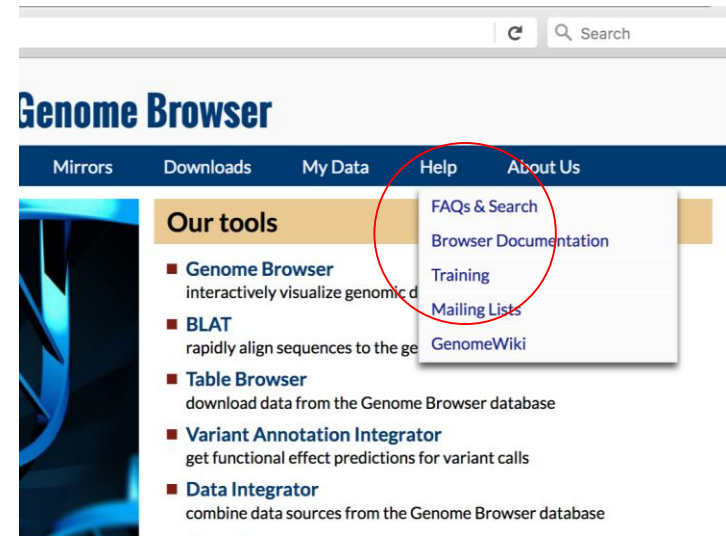
proteins. This process is regulated by DNA methylation, chromatin accessibility, transcription factors to the DNA and RNA. The traits are determined as differences in gene expression.

The goal of the ENCODE project is to catalog all the functional elements in the human genome. The DCC is to organize and display the data in the consortium, and to ensure specific quality standards when it comes to data. Before a lab submits any data, the DCC performs a series of quality checks. The DCC Quality Assurance team performs a series of

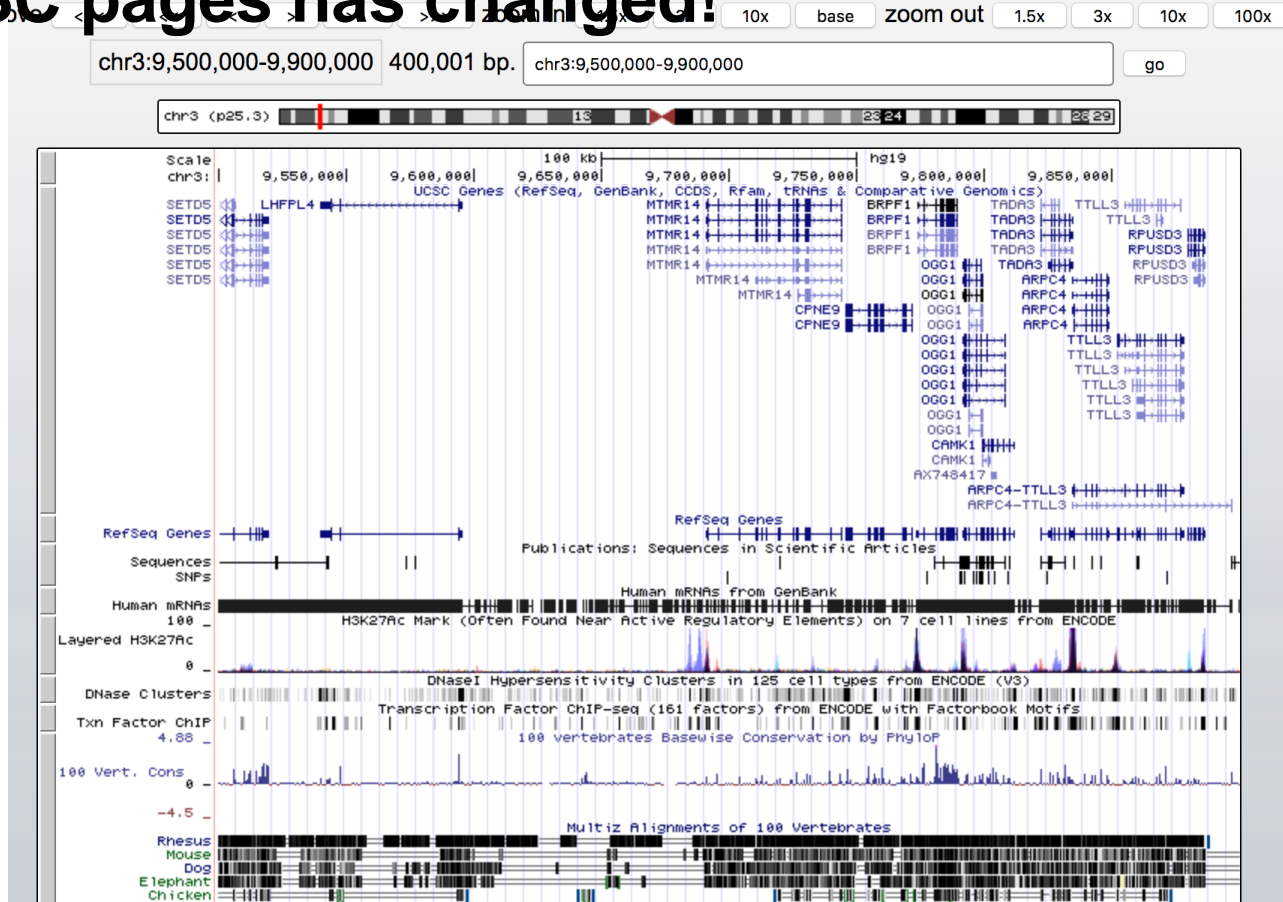
<http://genome.ucsc.edu/ENCODE/aboutScaleup.html>



UCSC pages has changed!



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



UCSC pages has changed!

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Help](#) [About Us](#)

Human Gene CPNE9 (uc003bsd.3) Description and Page Index

Description: Homo sapiens copine family member IX (CPNE9), mRNA.
Transcript (Including UTRs)
Position: hg19 chr3:9,745,510-9,771,592 **Size:** 26,083 **Total Exon Count:** 20 **Strand:** +
Coding Region
Position: hg19 chr3:9,745,681-9,771,376 **Size:** 25,696 **Coding Exon Count:** 20

Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD	Gene A
RNA-Seq Expression	Microarray Expression	RNA Structure	Protein Structure	Other Species	GO An
mRNA Descriptions	Other Names	Model Information	Methods		

Data last updated: 2013-06-14

☐ **Sequence and Links to Tools and Databases**

Genomic Sequence (chr3:9,745,510-9,771,592)	mRNA (may differ from genome)		Protein (553 aa)		
Gene Sorter	Genome Browser	Protein FASTA	Table Schema	BioGPS	CGAP
Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork	Gepis Tissue
HGNC	HPRD	Lynx	MGI	MOPED	neXtProt
PubMed	Stanford SOURCE	Treefam	UniProtKB		