Introductory bioinformatics for human genomics

Introduction to the UCSC genome browser

Dominik Beck
NHMRC Peter Doherty Fellow

Stem Cell Group, Adult Cancer Program, Lowy Cancer Research Centre

Prince of Wales Clinical School, Faculty of Medicine, UNIVERSITY OF NEW SOUTH WALES, SYDNEY NSW 2052
CONTENT

- Background
- Genome Assemblies
- Annotation Tracks: data set that can be linked to the genome given some coordinates
- Associated Tools
- Practical Exercise
Genome Browser
http://genome.ucsc.edu/

24 October 2013 - Job Opening: UCSC Genome Browser Trainer

The Center for Biomolecular Science and Engineering (CBSE) at University of California Santa Cruz seeks an articulate, self-motivated educational in-person training in UCSC Genome Browser at universities, hospitals, institutes, and professional meetings in the United States and international browser experience ranging from novice users to bioinformatics specialists. Presentations include formal talks, problem-solving sessions, and interactive workshops.

This position requires a Master's degree in a biological science, depth in molecular biology, experience in a research environment, working knowledge of UCSC genome browser, and previous experience in teaching or training in a scientific environment. Preferred qualifications include a PhD in a relevant field, experience with video production, and knowledge of cell biology.

For more information and to apply for this position, see Job #1304619 on the UCSC Staff Employment website.

23 October 2013 - dbSNP Build 138 Available for hg19

We are pleased to announce the release of four tracks derived from NCBI dbSNP Build 138 data, available on the human assembly (GRCh37) corresponding coloring and filtering options in the Genome Browser.

As was the case for the annotations based on the previous dbSNP build 137, there are four tracks in this release. One is a track containing all subsets of this track and show interesting and easily defined subsets of dbSNP.

- Common SNPs (138): uniquely mapped variants that appear in at least 1% of the population or are 100% non-reference
- Flagged SNPs (138): uniquely mapped variants, excluding Common SNPs, that have been flagged by dbSNP as "clinically associated"
- Mult. SNPs (138): variants that have been mapped to more than one genomic location

By default, only the Common SNPs (138) are visible, other tracks must be made visible using the track controls.

You will find the four SNPs (138) tracks on the Human Feb. 2009 (GRCh37/hg19) browser in the "Variation and Repeats" group.

The tracks were produced at UCSC by Angie Hinrichs and Lurina Guravadoo. We'd like to thank the dbSNP group at NCBI for providing access.
Background

**Visualization of genomic data**

- Graphical viewpoint on the very large amount of genomic sequence produced by the Human Genome Project.

  Human Genome: 3,156,105,057 bp

- Focus turned from accumulating and assembling sequences to identifying and mapping functional landmarks

  Genetic markers
  Genes
  SNPs
  Points of regulation

- Visualization of Next-generation-sequencing data
Background

Client-side

Integrative Genomics Viewer*

- Application (Java) on the user’s machine
- Often difficult to install
- Does not have the extensive third-party data of the other browsers
- Much faster than web-based browsers

http://www.broadinstitute.org/igv/
Background

- Intronerator was developed by J. Kent to map the exon–intron structure of C. elegans RNAs mapped against genomic coordinates

Jim Kent
Background

- Draft human genome sequence became available at the UCSC in 2000
- Intronerator was used as the graphics engine
CONTENT

- Background
- Genome Assemblies
- Annotation Tracks
- Associated Tools
- Practical Exercise
The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request: chr7
Genome Browser Response: Displays all of chromosome 7

chr7
Displays all of chromosome 7

chr7:1-1000000
Displays first million bases of chr 7, counting from p-arm telomere

chr7:1000000+2000
Displays a region of chr7 that spans 2000 bases, starting with position 1000000

RH18061:RH01756
Displays region between genome landmarks, such as the STS markers RH18061 and RH01756, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.

D16S3045
Displays region around STS marker D16S3045 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.

AA205474
Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17

AC008101
Displays region of clone with GenBank accession AC008101

AF083811
Displays region of mRNA with GenBank accession number AF083811

PRNP
Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP

NM_017414
Displays the region of genome with RefSeq identifier NM_017414

NP_059110
Displays the region of genome with protein accession number NP_059110

pseudogene mRNA
Lists transcribed pseudogenes, but not cDNAs

homobox
Lists mRNAs for homologous homebox genes

zinc finger
Lists many zinc finger mRNAs

kruppel zinc finger
Lists only kruppel-like zinc fingers

huntington
Lists candidate genes associated with Huntington's disease

zahler
Lists mRNAs deposited by scientist named Zahler

Evans J.E.
Lists mRNAs deposited by co-author J.E. Evans
Genome Assemblies

- Regular updates to genome assemblies to close gaps in genomic sequence, troubleshoot assembly problems and otherwise improve the genome assemblies

- Shifting coordinates for known sequences and a potential for confusion and error among researchers, particularly when reading literature based on older versions.

- Frequently used assemblies hg18/hg19

- New assemblies increase genomic coverage 6-fold and have been deposited in GenBank.

- 127 genome assemblies have been released on 58 organisms (April 2012)
Configure Image

- image width: 600 pixels
- label area width: 17 characters
- text size: 12

- Display chromosome ideogram above main graphic
- Show light blue vertical guidelines
- Display labels to the left of items in tracks
- Display description above each track
- Show track controls under main graphic
- Next/previous item navigation
- Next/previous exon navigation

Configure Tracks on UCSC Genome Browser: Human Feb. 2009 (GRCh37/hg19)

Tracks: track search  hide all  show all  default  Groups: collapse all  expand all

Control track and group visibility more selectively below.

- Mapping and Sequencing Tracks
- Phenotype and Disease Associations
- Genes and Gene Prediction Tracks
- Literature
- mRNA and EST Tracks
- Expression
- Regulation
Annotation tracks
Annotation tracks

- The database may contain any data that can be mapped to genomic coordinates and therefore can be displayed in the Genome Browser

- Overview of tracks: [http://genome.ucsc.edu/cgi-bin/hgTracks](http://genome.ucsc.edu/cgi-bin/hgTracks)

- Three different categories:
  - computed at UCSC
  - computed elsewhere and displayed at UCSC
  - computed and hosted entirely elsewhere
Annotation tracks computed at UCSC

- Comparative genomic annotations as well as Convert and liftOver capabilities

- mRNAs and ESTs in GenBank are aligned to the reference assembly in separate tracks (75 million GenBank RNAs and ESTs, ~3 billion bases of the human reference assembly $\rightarrow$ 2 CPU-years of computing time)

- The Conservation composite track displays the results of the multiz algorithm that aligns the results from up to 46 pairwise Blastz alignments to the reference assembly (e.g. hg19 human assembly consumed 10 CPU-years)
Annotation tracks computed elsewhere and displayed at UCSC

Annotations that are not post-processed by the UCSC

- Probe sets for commercially available microarrays, copy-number variation from the Database of Genomic Variants or expression data from the GNF Expression Atlas
- Data Coordination Center for the ENCODE project allowing access to a large number of functional annotations in relation to gene regulation

Annotations that are post-processed by the UCSC

- dbSNP (Common SNPs, Flagged SNPs, Mult. SNPs)
- OMIM (OMIM Allelic Variant SNPs, OMIM Genes, OMIM Phenotypes)
Annotation tracks computed and hosted elsewhere

- Data tracks are hosted remotely (no data are stored at UCSC) and publicly available, e.g. Epigenomics Roadmap project [http://epigenome.wustl.edu/](http://epigenome.wustl.edu/)

**Roadmap Epigenomics Visualization Hub (VizHub)**

Roadmap Epigenomics Project | NCBI Epigenomics Gateway | Roadmap Data Coordination Center

**VizHub**

Roadmap Epigenomics Visualization Hub at Wash U

**About the UCSC Genome Bioinformatics Site**

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VizGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSI) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

**News**

16 August 2012 – Announcing a Genome Browser for the Medium ground finch

We have released a browser for the Medium ground finch, *Geospiza fortis*, reknowned as one of naturalist Charles Darwin’s Galapagos finches. This species, which has been the subject of many evolutionary studies, is one of a group of birds that evolved over a few million years from a single ancestral species into multiple species whose beak sizes and shapes are specialized for using different food resources. The phenotypic diversity of these birds contributed to Darwin’s theory of evolution. The significance of this genome assembly is described in the August 16,
Tracks from the Epigenome project
Associated Tools

- Tools other than the main graphic image account for 42% of traffic on the UCSC server
Sessions

Save Settings

Save current settings as named session:
- name: hg19
- allow this session to be loaded by others

Save current settings to a local file:
- file: UCSC_Session.txt
- file type returned: plain text

(leave file blank to get output in browser window)

Restore Settings

Use settings from another user's saved session:
- user:
- session name:

Use settings from a local file:
- file: C:\Users\23265235\Choose...

Use settings from a URL (http://..., ftp://...):

Sharing Sessions

There are several ways to share saved sessions with others.

- Each previously saved named session appears with Browser and Email links.
- Email link invokes your email tool with a message containing the Genome Browser page and a link to that session loaded. The resulting Genome Browser page can be bookmarked.
- You can also send email with a link to a saved settings file. For example, genome.ucsc.edu/cgi-bin/hgSession
- If a saved settings file is available from a web server, you can send email to others with a link like http://server/hgSession?hgS_doLoadUrl=submit&hgS_loadUrlName=U where U is the URL of your saved settings file.
Custom track
# Table Browser

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

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, etc. For more complex queries, you may want to use Galaxy or our public MySQL server. To execute a query, refer to the **Credits** page for the list of contributors and usage restrictions associated with these data. All table columns are explained below.

### Example Table

<table>
<thead>
<tr>
<th>#</th>
<th>chrom</th>
<th>chromStart</th>
<th>chromEnd</th>
<th>name</th>
<th>type</th>
<th>clinSign</th>
<th>phenotype</th>
<th>origin</th>
<th>otherIds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chr10</td>
<td>123247548</td>
<td>123247549</td>
<td>FGF2r2.c.1942G</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0035</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>chr10</td>
<td>123247608</td>
<td>123247609</td>
<td>FGF2r2.c.1882G</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>chr10</td>
<td>123256214</td>
<td>123256215</td>
<td>FGF2r2.c.1694A</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>chr10</td>
<td>123258104</td>
<td>123258105</td>
<td>FGF2r2.c.1576A</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>chr10</td>
<td>123274745</td>
<td>123274746</td>
<td>FGF2r2.c.1172T</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>chr10</td>
<td>123274776</td>
<td>123274777</td>
<td>FGF2r2.c.1141T</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0044</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>chr10</td>
<td>123274793</td>
<td>123274794</td>
<td>FGF2r2.c.1124A</td>
<td>single nucleotid other, pathogenic</td>
<td>somatic, uncertain</td>
<td>OMIM Allelic Variant: 176943.0015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>chr10</td>
<td>123274802</td>
<td>123274803</td>
<td>FGF2r2.c.1115C</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>chr10</td>
<td>123276855</td>
<td>123276856</td>
<td>FGF2r2.c.1061C</td>
<td>single nucleotid pathogenic</td>
<td>germline</td>
<td>OMIM Allelic Variant: 176943.0005</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exercise

UCSC Genome Browser

Epigenome Browser

UNSW
Exercise

**UCSC Genome Browser**

- Search for gene of interest
- Clean up the UCSC Browser tracks,
  - show UCSC genes
  - show RefSeq genes
- Find the conservation tracks
  - use “Track Search” and Drop-down tables
  - Visualize Conservation tracks
  - Visualize PhastCons tracks
- Find the dbSNP tracks,
  - Visualize the Flagged SNPs track
  - Modify the “Flagged SNP track” and apply the “squish” visualization
- Find the OMIM tracks
  - Visualize the OMIM AV SNP track.
  - Modify the OMIM AV SNP track and apply the “full” visualization
- Modify text sizes, browser resolution and track colors using the configure button
- Save session as a txt file
Exercise

Epigenome Browser: [http://vizhub.wustl.edu/](http://vizhub.wustl.edu/)

- Apply genome coordinates from gene of interest
- Clean up the Epigenome browser
  - show UCSC genes
  - show RefSeq genes
  - conservation tracks
- Find the Epigenome Atlas Data Complete Collection Composite Tracks
  - Visualize Footprinting tracks and adjust visualization
- Find the CpG and MRE sites
  - Visualize the CpG Island track and adjust visualization
- Find the Phenotype and Disease Associations tracks
  - Visualize GWAS Catalog and adjust visualization
- Find the Epigenome Atlas Data Complete Collection Integrative Tracks
  - Visualize Methylation Summary and hide MeDIP0seq and RRBS Summery
- Modify text sizes, browser resolution and track colors using the configure button
- Save session as a txt file