Navigating sequenced genomes

Bing Zhang
Department of Biomedical Informatics
Vanderbilt University
bing.zhang@vanderbilt.edu
Learning objectives

- To recognize the need for genome browsers
- To appreciate the power of genome browsing
- To be able to browse data in the UCSC genome browser
Genome and genome project

- **Genome**: the complete genetic material of an organism.
  - *E. coli* genome: 4.7 million base pairs of DNA
  - Human genome: 3 billion base pairs of DNA (~3G)
  - Encodes heritable characteristics

- **Genome project**: scientific endeavors that ultimately aim to determine the complete genome sequence of an organism

![Graph showing completely sequenced genomes from 1995 to 2009](http://www.genomesonline.org)

Human genome project (1990-2003)
Genome assembly

- **Genome assembly**: the process of taking a large number of short DNA sequences, all of which were generated by a shotgun sequencing project, and putting them back together to create a representation of the original chromosomes from which the DNA originated.

- NCBI provides annotated assemblies of public genome sequence data, so that individual researchers do not have to piece together extended segments of a genome.

- **Build**: a run of the genome assembly.
  - The data are frozen at the start of the build process.
  - Mouse: NCBI37 (2007); NCBI36 (2006); NCBI35 (2005) ……
Genome annotation

- **Genome annotation**: the process of attaching different types of biological information to assembled sequences
  - Identifying elements on the genome, e.g. gene prediction; regulatory element prediction, etc.
  - Attaching biological information to these element, e.g. modification; expression; conservation, etc.
Same genome build, different annotations

- EBI: Ensembl project => Ensembl Gene
- Sanger: Human And Vertebrate Analysis aNd Annotation (HAVANA) project
- NCBI: Reference Sequence (RefSeq) project => Entrez Gene
- Sanger: Vertebrate Genome Annotation (VEGA) project: high quality manual annotation, mostly based on the HAVANA project
- NCBI: Consensus CDS (CCDS) project: core set of human and mouse protein coding regions that are consistently annotated and of high quality.
Why genome browsing?

- Facilitate genomic analysis by presenting alignment, experimental and annotation data in the context of genomic DNA sequences.
- Combines the power of data visualization and data aggregation:
  - Human mind can assimilate visual data in great detail.
  - Biological data are noisy and complementary.
  - Putting different types of data in the same context allows easy identification of real biological signals and common themes.
Genome browsers

- Widely used genome browsers
  - UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgGateway
  - Ensembl genome browser: http://www.ensembl.org/index.html
  - JGI Integrated Microbial Genomes (IMG): http://img.jgi.doe.gov
  - Gbrowse from the Generic Model Organism Database (GMOD): http://gmod.org/wiki/Gbrowse

- Specific genome browsers
  - VISTA Enhancer Browser for non-coding elements: http://enhancer.lbl.gov
  - 1000 Genome Browser for multiple individuals: http://browser.1000genomes.org
  - IGV (Integrative Genomics Viewer) with high scalability: http://www.broadinstitute.org/software/igv/home
Basic organization of a typical genome browser

- Annotations are graphically depicted along the genomic assembly
- Different types of annotations
  - Mapping and sequencing annotations: assembly, gap, GC percent details, etc.
  - Gene annotations: alignments of annotated genes from different projects as well as mRNAs and ESTs from GenBank
  - Comparative genomic information: pairwise genomic alignments, multiple sequence alignment
  - Others: expression, regulation, variation and phenotype information
- Each form of annotation is displayed as a track
- You can add your own data as a customized track
UCSC genome browser

- Accessing the genome browser and related tools at UCSC
- Making queries
- Controlling the Browser graphic
- Managing annotation tracks
- Exploring default tracks for TP53
- Exporting data
UCSC genome browser and related tools

- [http://genome.ucsc.edu/](http://genome.ucsc.edu/)
- Microbial genomes
- BLAT: a fast sequence-alignment tool similar to BLAST
- Table Browser: convenient text-based access to the database underlying the Genome Browser, similar to BioMart
- Genome Graphs: a tool that allows uploading and displaying genome-wide data sets
- Gene Sorter: expression, homology, and other information on groups of genes that can be related in many ways
- Proteome Browser: protein property data and links to a wealth of related information
- Custom Tracks: uploading your own annotation data for display in the browser
Making queries

- Select clade, genome, and assembly
- Search by position or search term
  - A genome position can be specified by the accession number of a sequenced genomic clone (e.g. AC008101), an mRNA or EST or STS marker (e.g. AF083811, AA205474, or D16S3046), a chromosomal coordinate range (e.g. chr3:1-1000000), or keywords from the GenBank description of an mRNA (e.g. zinc finger).
- Search by gene symbol, e.g. TP53
Controlling the Browser graphic

- Navigating
- Search box
- Select region to zoom in
- Click for track description, drag to rearrange
- Zooming
- Chromosome Ideogram
- Reverse

Applied Bioinformatics, Spring 2011
Managing annotation tracks

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

### Mapping and Sequencing Tracks
- **Base Position**
- **Chromosome Band**
- **STS Markers**
- **FISH Clones**
- **Recomb Rate**
- **Map Contigs**
- **Assembly**
- **GRC Map Contigs**
- **Gap**
- **BAC End Pairs**
- **Fosmid End Pairs**
- **GC Percent**
- **GRC Patch Release**
- **Hg18 Diff**
- **Short Match**
- **Restr Enzymes**
- **Wiki Track**
- **Mapability**

### Phenotype and Disease Associations

### Genes and Gene Prediction Tracks
- **UCSC Genes**
- **Alt Events**
- **Gencode Genes**
- **CCDS**
- **RefSeq Genes**
- **Other RefSeq**
- **MGC Genes**
- **ORFeome Clones**
- **TransMap...**
- **Vega Genes**
- **Ensembl Genes**
- **N-SCAN**
- **SGP Genes**
- **Geneid Genes**
- **GenScan Genes**
- **Exoniphy**
- **tRNA Genes**
- **H-Inv 7.0**
- **EvoFold**
- **sno/miRNA**
- **IKMC Genes Mapped**

### mRNA and EST Tracks

### Expression

### Regulation

### Comparative Genomics

### Neandertal Assembly and Analysis

### Variation and Repeats

Click for description

Visibility control

- hide
- dense
- squish
- pack
- full
Default tracks for TP53

- Base position
- UCSC genes
- RefSeq genes
- Human mRNAs
- Human ESTs
- Conservation
- SNPs
- Repeat mask

- Exon, UTR
- Exon, coding
- Intron and direction
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

Display
- Colors:
  - Black: matches in the Protein Data Bank (PDB)
  - Dark blue: reviewed or validated by the UniProt Knowledgebase
  - Medium blue: other RefSeq transcripts
  - Light blue: non-RefSeq transcripts

Methods

Credits

Data Use Restrictions

References

**Description**

The UCSC Genes track shows gene predictions based on data from RefSeq, GenBank, CCDS and UniProt. This is a moderately conservative set of predictions, requiring the gene to include both protein-coding and potential non-coding transcript. Some of these non-coding transcripts may actually code for protein, but the evidence for the associated protein is about twice as many splice variants.

**Display Conventions and Configurations**

This track is general followed the display conventions for gene prediction tracks. The genes for putative non-coding genes and unstranded regions are represented by relatively thin lines.

- Black: gene has a corresponding entry in the Protein Data Bank (PDB)
- Dark blue: gene has been reviewed or validated by either the RefSeq, SwissProt or CCDS staff
- Medium blue: other RefSeq transcripts
- Light blue: non-RefSeq transcripts

This track contains an optional exon coloring feature that allows users to quickly validate and compare gene predictions. To display exon colors, select the gene exon description.

**Methods**

The UCSC Genes are built using a multi-step pipeline:

1. RefSeq and GenBank RNAs are aligned to the genome with BLAST, keeping only the best alignments for each RNA and discarding alignments of less than 50% identity.
2. ALIGNments are broken up at inserts, and each isoform is represented by a unique alignment.
3. Splicing graphs are created for each set of overlapping alignments. The graph includes edges for each exon or intron, and a vertex for each splice site, start, and end. Each exon if a non-coding RNA is also included in the graph. The graph is then used to generate all unique transcripts. The traversal is guided by the initial RNAs to avoid a combinatorial explosion in alternative splicing. All RefSeq transcripts require either two RNAs or two additional lines of evidence beyond the single RNA.
4. Protein predictions are generated. For non-RefSeq transcripts we use the RefSeq database to determine if the transcript is protein-coding and if so, the locations of the start and stop codons are predicted in other species. As many evidence if not available from a human transcript and the NCBI RefSeq database.
5. The corresponding UniProt protein is found, if any.
6. The transcript is assigned a human transcript accession.

**Credits**

The UCSC Genes track was produced at UCSC using a computational pipeline developed by Jim Kent, Chuck Smit and Mark Diekhans. It is based on data from NCBI RefSeq. It contributes to them.

**Data Use Restrictions**

Copyright information from the [Uniprot website](http://www.uniprot.org).

References

Exporting data
Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

<table>
<thead>
<tr>
<th>clade:</th>
<th>Mammal</th>
<th>genome:</th>
<th>Human</th>
<th>assembly:</th>
<th>Feb. 2009 (GRCh37/hg19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group:</td>
<td>Genes and Gene Prediction Tracks</td>
<td>track:</td>
<td>UCSC Genes</td>
<td>add custom tracks</td>
<td></td>
</tr>
<tr>
<td>table:</td>
<td>knownGene</td>
<td>describe table schema</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>region:</td>
<td>genome position</td>
<td>chr17:7575972-7590863</td>
<td>lookup</td>
<td>define regions</td>
<td></td>
</tr>
<tr>
<td>identifiers (names/accessions):</td>
<td>paste list</td>
<td>upload list</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>filter:</td>
<td>create</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intersection:</td>
<td>create</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correlation:</td>
<td>create</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output format:</td>
<td>all fields from selected table</td>
<td>Send output to Galaxy GREAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output file:</td>
<td>(leave blank to keep output in browser)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file type returned:</td>
<td>plain text gzip compressed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To reset all user cart settings (including custom tracks), click here.
Summary

- Genome browsers combine the power of data visualization and data aggregation. They facilitate genomic analysis by presenting alignment, experimental and annotation data in the context of genomic DNA sequences.

http://genome.ucsc.edu/training.html

Fujita et al. NAR 39:D876-882, 2011