Comparative Genomics
~interpreting sequence similarity across species

Andrea Angstadt

Bioinformatics Course 4.20.11

(http://genome.ucsc.edu/)
Overview

• Demonstrate online freely available programs for translation of genomic information obtained from model organism studies to the human counterpart

• Focus on the
  – UCSC genome browser (http://genome.ucsc.edu/)
  – AutoGRAPH program (http://autograph.genouest.org/)
Comparative Genomics Importance

• Ethical and time consuming issues produce need for genomic analysis of biomedical model organisms of human diseases

• Tools allow
  – study of evolutionary changes among organisms
  – Identify genes that are conserved across species
  – Aid in new strategies for conserving rare and endangered species
  – Pinpoint ‘disease’ associated genetic abnormalities
Comparative Genomic Analysis

- **Ensembl**
  (http://useast.ensembl.org/index.html)
- **NCBI**
  - Homologene
- **VISTA**
  (http://genome.lbl.gov/vista/index.shtml)
- **UCSC genome browser**
  (http://genome.ucsc.edu/)
  - Human, chimp, orangutan, rhesus, marmoset mouse, rat, guinea pigs, rabbit, car, panda, dog, horse, pig, cow, elephant, opossum, platypus, and many more..
- **AutoGRAPH**
  (http://autograph.genouest.org/)
  - Human, dog, cow, chimp, mouse, rat
Ensembl

~information on individual eukaryotic genomes
Ensembl Genomes

~created in 2009 for bacteria, protists, fungi, plants and invertebrate metazoa
NCBI
HomoloGene
~produces alignment information for specific genes across several species

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of Genes</th>
<th>HomoGene Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>19,943</td>
<td>18,981</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>25,096</td>
<td>16,850</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>15,766</td>
<td>16,769</td>
</tr>
<tr>
<td>Eos taurus</td>
<td>22,049</td>
<td>16,180</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>25,388</td>
<td>21,766</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>21,991</td>
<td>19,229</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>17,969</td>
<td>15,142</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>25,690</td>
<td>21,084</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>13,827</td>
<td>9,282</td>
</tr>
<tr>
<td>Anopheles gambiense</td>
<td>12,460</td>
<td>8,867</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>20,132</td>
<td>6,878</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe</td>
<td>5,043</td>
<td>3,225</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>5,880</td>
<td>4,861</td>
</tr>
<tr>
<td>Klyveromyces lactis</td>
<td>5,335</td>
<td>4,459</td>
</tr>
<tr>
<td>Eremothecium galloti</td>
<td>4,722</td>
<td>3,928</td>
</tr>
<tr>
<td>Magnaporthis grisea</td>
<td>12,832</td>
<td>7,330</td>
</tr>
<tr>
<td>Neurospora crassa</td>
<td>9,821</td>
<td>6,207</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>27,309</td>
<td>19,961</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>26,887</td>
<td>17,276</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>5,265</td>
<td>1,962</td>
</tr>
</tbody>
</table>

**Tip of the Day**

Use [protein gi] in your search query to restrict search results to that particular protein. e.g. 4550361
[protein gi] [More Tips]

**What's New**

HomoloGene release 65 includes:
- updated annotations for the following species: Homo sapiens (NCBI release 37.2), Danio rerio (NCBI release 4.1), Drosophila melanogaster (NCBI release 9.3), Caenorhabditis elegans (NCBI release 9.1), Arabidopsis thaliana (NCBI release 9.1).
NCBI HomoloGene

~produces alignment information for specific genes (TP53) across several species
VISTA

~genome wide multiple and pairwise alignments

Whole Genome Comparative Analysis of the Human Mar. 2006 Genome

From this page you can access the results of:

the multiple alignments

the pairwise alignments of the

Human Mar. 2006 genome produced by the International Human Genome Sequencing Consortium with the following genomes:
- The Fugu rubripes v 4.0 genome produced by the JGI
- The Chimpanzee Mar. 2006 genome produced by the Chimpanzee Sequencing and Analysis Consortium
- The Chicken May 2006 genome produced by the Washington University School of Medicine in St. Louis
VISTA
~genome wide multiple and pairwise alignments
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. VisiGene 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 (CBSE) 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

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

04 March 2011 - The Meaning of Red: Consensus on color scheme for CNVs

Consensus was reached on color standards to represent CNV loss (red) and gain (blue) in genomics databases by DGV, NCBI, DECIPHER, UCSC and ISCA during the 1st annual ISCA Scientific Conference.

At the recent ISCA Scientific Conference in Atlanta, GA, Nigel Carter, representing DECIPHER, made a plea for a standardized color scheme for representation of CNV loss (deletion) or gain (duplication). Discussions in person and by email with leaders of major CNV/genomic databases (including DGV, dbVar, ISCA, and the UCSC Genome Browser) resulted in the recommendation to utilize red to represent loss/deletion (consistent with the original convention for CGH on metaphase chromosomes) and blue to represent gain/duplication (avoiding green which is difficult for color-blind individuals to discriminate from red).

Announcement reprinted from the website for the International Standards for Cytogenomic Arrays (ISCA) Consortium. The UCSC Genome Browser is now conforming to this standard for the DECIPHER track and the DGV track on human assemblies (which was released today) and on any future tracks derived from dbVar.

Thanks to Nigel Carter, Steve Scherer, Deanna Church, Angie Hinichs, Fan Hsu, Pauline Fujita and Robert Kuhn.

01 March 2011 - Updated Chimpanzee Browser Released
UCSC genome browser

About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium.

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, chrUa_g000212, chr3:1-1000000, chr3:1000000+2000, RH18061:RH80175, D16S3046, AA205474, AC008101

Genome Browser Response:

chr7: Displays all of chromosome 7
chrUa_g000212: Displays all of the unplaced contig g000212
chr3:1-1000000: Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000: Displays a region of chr 3 that spans 2000 bases, starting with position 1000000
RH18061:RH80175: Displays region between genome landmarks, such as the STS markers RH18061 and RH80175. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046: Displays region around STS marker D16S3046 from the Genethon/Marshalfield 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
UCSC genome browser interface
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 OpenHello 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>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>clade:</td>
<td>Mammal</td>
</tr>
<tr>
<td>genome:</td>
<td>Human</td>
</tr>
<tr>
<td>assembly:</td>
<td>Feb 2009 (GRCh37/hg19)</td>
</tr>
<tr>
<td>group:</td>
<td>Genes and Gene Prediction Tracks</td>
</tr>
<tr>
<td>track:</td>
<td>RefSeq Genes</td>
</tr>
<tr>
<td>table:</td>
<td>refGene</td>
</tr>
<tr>
<td>region:</td>
<td>genome: C position:</td>
</tr>
<tr>
<td>identifiers:</td>
<td>names/accessions:</td>
</tr>
<tr>
<td>filter:</td>
<td>create</td>
</tr>
<tr>
<td>intersection:</td>
<td>create</td>
</tr>
<tr>
<td>output format:</td>
<td>selected fields from primary and related tables</td>
</tr>
<tr>
<td>output file:</td>
<td>HumanRefSeqGenes</td>
</tr>
<tr>
<td>file type returned:</td>
<td>G plain text, G zip compressed</td>
</tr>
</tbody>
</table>

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

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User’s Guide.

- clade: Specifies which clade the organism is in.
- genome: Specifies which organism data to use.
Table browser

• Obtain information from UCSC genome browser in a text file format
  – BAC end pairs reads, bp positions, Refseq and UCSC gene IDs, transmap information
  – **Example:** You have several regions of interest or significant regions from microarray data analyzing copy number aberrations in dog osteosarcoma cells.
    – Lack of gene annotation in dog genome causes difficulty when trying to identify genes located within these ‘significant regions’
Table browser

- Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

1. Clade: Mammal, genome: Dog, assembly: May 2005 (Broad/canFam2)
2. Group: Genes and Gene Prediction Class, Track: TransMap Ref Gene
3. Table: transMapAlnRefSeq
   - Clicking on describe table schema provide summary
### Table browser

#### Describe table schema

**Database:** canFam2  **Primary Table:** transMapAhRefSeq  **Row Count:** 94,876

**Format description:** Summary info about a patSpace alignment

<table>
<thead>
<tr>
<th>Field</th>
<th>Example</th>
<th>SQL type</th>
<th>Info</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bin</td>
<td>76</td>
<td>smallint() unsigned</td>
<td>range</td>
<td>Indexing field to speed chromosome range queries.</td>
</tr>
<tr>
<td>matches</td>
<td>4253</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of bases that match that aren't repeats</td>
</tr>
<tr>
<td>misMatches</td>
<td>1233</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of bases that don't match</td>
</tr>
<tr>
<td>repMatches</td>
<td>1039</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of bases that match but are part of repeats</td>
</tr>
<tr>
<td>nCount</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of 'N' bases</td>
</tr>
<tr>
<td>qNumInsert</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of inserts in query</td>
</tr>
<tr>
<td>qBaseInsert</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of bases inserted in query</td>
</tr>
<tr>
<td>tNumInsert</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of inserts in target</td>
</tr>
<tr>
<td>tBaseInsert</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of bases inserted in target</td>
</tr>
<tr>
<td>strand</td>
<td>.</td>
<td>char(2)</td>
<td>values</td>
<td>+ or - for strand. First character query, second target (optional)</td>
</tr>
<tr>
<td>qName</td>
<td>NM_006208.2.1.1</td>
<td>varchar (255)</td>
<td>values</td>
<td>Query sequence name</td>
</tr>
<tr>
<td>qSize</td>
<td>7442</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Query sequence size</td>
</tr>
<tr>
<td>qStart</td>
<td>0</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Alignment start position in query</td>
</tr>
<tr>
<td>qEnd</td>
<td>7440</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Alignment end position in query</td>
</tr>
<tr>
<td>tName</td>
<td>chr1</td>
<td>varchar (255)</td>
<td>values</td>
<td>Target sequence name</td>
</tr>
<tr>
<td>tSize</td>
<td>125616256</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Target sequence size</td>
</tr>
<tr>
<td>tStart</td>
<td>3247833</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Alignment start position in target</td>
</tr>
<tr>
<td>tEnd</td>
<td>3321670</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Alignment end position in target</td>
</tr>
<tr>
<td>blockCount</td>
<td>134</td>
<td>int(10) unsigned</td>
<td>range</td>
<td>Number of blocks in alignment</td>
</tr>
<tr>
<td>blockSizes</td>
<td>11,37,17,26,61,8,34,13,38,4...</td>
<td>longblob</td>
<td></td>
<td>Size of each block</td>
</tr>
<tr>
<td>qStarts</td>
<td>2,13,54,73,100,161,169,206,...</td>
<td>longblob</td>
<td></td>
<td>Start of each block in query.</td>
</tr>
<tr>
<td>tStarts</td>
<td>3247833,3247846,3247883,324...</td>
<td>longblob</td>
<td></td>
<td>Start of each block in target.</td>
</tr>
</tbody>
</table>
Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

4. Define regions: Paste list provided (DogPostions.txt)

5. Filter: Select hgFixed.transMapSrcRefSeq based filter and then db=hg19

6. Output format: selected fields from primary and related tables

7. Output file: I’ve named it HumanRefSeq_Mapped to Dog

8. Click on get output
Table browser

### Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the [Table Browser User’s Guide](#).

- **clade**: Specifies which clade the organism is in.
- **genome**: Specifies which organism data to use.

### 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>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>clade</td>
<td>Mammal</td>
</tr>
<tr>
<td>genome</td>
<td>Dog</td>
</tr>
<tr>
<td>assembly</td>
<td>May 2005 (Broad/canFam2)</td>
</tr>
<tr>
<td>group</td>
<td>Genes and Gene Prediction Tracks</td>
</tr>
<tr>
<td>track</td>
<td>TransMap RefGene</td>
</tr>
<tr>
<td>table</td>
<td>transMapRefSeq</td>
</tr>
<tr>
<td>region</td>
<td>chr1:11072309-11078926</td>
</tr>
<tr>
<td>identifiers</td>
<td>paste list, upload list</td>
</tr>
<tr>
<td>filter</td>
<td>edit, clear</td>
</tr>
<tr>
<td>intersection</td>
<td>create</td>
</tr>
<tr>
<td>correlation</td>
<td>create</td>
</tr>
<tr>
<td>output format</td>
<td>selected fields from primary and related tables</td>
</tr>
<tr>
<td>Send output</td>
<td>Galaxy, GREAT</td>
</tr>
<tr>
<td>output file</td>
<td>HumanRefSeq_MappedtoDog</td>
</tr>
<tr>
<td>file type</td>
<td>plain text, gzip compressed</td>
</tr>
<tr>
<td>get output</td>
<td></td>
</tr>
<tr>
<td>summary/stats</td>
<td></td>
</tr>
</tbody>
</table>

To reset all user cart settings (including custom tracks), [click here](#).
Table browser

• Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

9. Selected fields from canFam2.transmapAlnRefSeq
   • tName (chromosome name)
   • tSize
   • tStart
   • tEnd

10. Selected fields from hgFixed.transMapSrcRefSeq
    • Db
    • Id
    • Chrom
    • chromStart
    • chromEnd
Table browser

- Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

  - Save the file and open it in excel. This contains all of the human RefSeq genes that map to those orthologous regions in the Dog genome (hgFixed.transMapMapSrcRefSeq.id)

  - Next step: What are the gene names associated with those accession numbers
Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

1. Open the file that I have provided called: ‘HumanReqSeq_MappedtoDog_Example’
2. Go back to the table browser homepage.
3. Clade: Mammal, genome: Human, assembly: Feb 2009 (GRCh37/hg19)
4. Group: Genes and Gene Prediction Class, Track: RefSeq Genes
5. Table: refGene
Table browser

• Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

  4. Identifies: Paste Column from file provided (Human RefSeq ID), hit submit
     ▪ You may receive an error that some do not exist, as this database is updated constantly that’s ok

  5. Filter: No filters are necessary here

  6. Output format: selected fields from primary and related tables

  7. Output file: I’ve named it HumanRefSeqGenes

  8. Click on get output
Table browser

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade**: Specifies which clade the organism is in.
- **genome**: Specifies which organism data to use.
- **assembly**: Specifies which version of the organism's genome sequence to use.
Example: Identifying Human RefSeq genes which reside in orthologous regions of the dog genome

9. Selected fields from Hg19.refGene
   - name
   - chrom
   - txStart
   - txEnd
   - cdsStart
   - cdsEnd
   - name2

10. This saved file contains all the names of the human RefSeq genes that mapped to those regions in the dog genome
   - These genes can be analyzed for functionality by programs such as Database for Annotation, Visualization, and Integrated Discovery (DAVID)
     - Complete program listing on Gene Ontology website (http://www.geneontology.org/GO.tools.microarray.shtml)
UCSC genome browser

- Several ‘Help’ tools available
AutoGRAPH
Visualizing Chromosomal Synteny

- Purpose:

AutoGRAPH is an integrated web server for multi-species comparative genomics analysis. It is designed for constructing and visualizing synteny maps between two or three species, determination and display of macrosynteny and microsynteny relationships among species, and for highlighting evolutionary breakpoints.

The web server constructs synteny maps by pairwise comparison of marker/anchor orders between a reference chromosome and one or two tested genome(s). It permits users to visualize and characterize several features:
- Conserved segments (CS)
- Conserved Segments Ordered (CSO)
- breakpoints
- marker density in these regions, inferred locations for 1:0 orthologous relationships...

AutoGRAPH is a versatile tool and can be utilized for the integration and comparison of different maps (meiotic, RH...) with sequence resources within a single species.

- Launch AutoGRAPH:

User can run AutoGRAPH with:
- Pre inserted dataset: corresponds to data for 6 genomes of interest (Human, chimp, rat, mouse, dog and cow) downloaded from Ensembl web site (v.42) (BioMart v0.5)
- Personal dataset: Users insert their own datasets.

Example of a 3 ways comparison for dog chromosome 34 vs. Human (right) and Mouse (left) genomes.

- This work is referenced in:
AutoGRAPH

- Two options on home page, pre-inserted dataset or your personal dataset
  - Personal dataset needs to be in Autograph file format or as a GFF.
  - Clicking on ‘help’ will demonstrate these options
- For an example select pre-inserted dataset
  - Compare 2 datasets (1 reference vs. 1 test)
  - Reference Species: human, chromosome 6, dataset set 2: dog
  - Click on submit
AutoGRAPH

• On your screen will appear a synteny map and a list of breakpoints, conserved segments, and conserved segments ordered locations.

• The tutorial is also a good reference:
  http://autograph.genouest.org/Tutorial.php
Summary

• A multitude of programs exists for comparative genomic analysis with Ensembl, NCBI, and the UCSC genome browser as the 3 largest browsers.

• Each program offers unique methodologies and most often you’ll need to choose the one that best fits your needs.