Setting Up UCSC Genome Browser

Patricia Chan
Biomolecular Engineering
Overview

• UCSC genome browser runs on
  – 32-bit and 64-bit Linux/Unix-based system
  – CGI
  – MySQL database
  – Apache

• Programs are written in C and Javascript (JQuery)

• To install or mirror a genome browser on a new server
  – http://genomewiki.ucsc.edu/index.php/Browser_Installation
Where are the data?

- Data for each genome assembly are stored in 2 places
  - MySQL database
    - Each genome assembly has its own database
    - Examples: hg18, hg19, mm9
    - Most track data are stored in MySQL
  - /gbdb/<DB name>
    - Each genome assembly has its own local directory
    - Examples: /gbdb/hg18, /gbdb/hg19, /gbdb/mm9
    - Sequences, wiggle track data, and other large data source such as bam are stored in files
### centraldb

- A database in MySQL containing all genome info

```sql
mysql> select * from centraldb.dbDb where name = 'pyrAer1';

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
<th>nibPath</th>
<th>organism</th>
<th>hits.bed</th>
<th>defaultPos</th>
<th>active</th>
<th>orderKey</th>
<th>genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyrAer1</td>
<td>Dec 2001eTh</td>
<td>/gbdb/pyrAer1</td>
<td>Pyrobaculum aerophilum</td>
<td>chr:10001-35000</td>
<td>1</td>
<td>340</td>
<td>Pyrobaculum aerophilum</td>
<td></td>
</tr>
<tr>
<td>pyrAer1</td>
<td>Dec 2001eTh</td>
<td>/gbdb/pyrAer1/html/description.html</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

mysql> select * from centraldb.blotServers where db = 'pyrAer1';

<table>
<thead>
<tr>
<th>db</th>
<th>host</th>
<th>port</th>
<th>isTrans</th>
<th>canPcr</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyrAer1</td>
<td>lowepub.cse.ucsc.edu</td>
<td>23370</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pyrAer1</td>
<td>lowepub.cse.ucsc.edu</td>
<td>23369</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```
• centraldb may not be named as “centraldb”
• Get the database name from central.db entry in hg.conf or ask browser admin/developer
Track info

- Stored in a table called trackDb in each genome’s MySQL database
- Based on one or multiple trackDb.ra files as input source
- Global trackDb.ra
  - Contains track info that apply to all genomes
- Genome-specific trackDb.ra
  - Contains track info that are specific to a genome
**trackDb.ra Entries**

- Similar to track line for custom tracks
- [http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#TRACK](http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#TRACK)

```plaintext
track refSeq rDGCC7710 | tenax.cse
shortLabel Genbank RefSeq
longLabel Genbank RefSeq Gene Annotations
group genes
priority 2.1:scribble blatServers;
visibility pack
color 0,100,100 Type  |  Null  |  Key  |  Default
| genePred gpProtCodePep
nextItemButton omar(32)  |  NO  
| host  |  char(128)  |  NO  
| track scaffolds int(11)  |  NO  |  0
| shortLabel Scaffolds int(4)  |  NO  |  0
| longLabel Assembly Scaffolds NO  |  0  |  0
| group map
| priority 1.8 set (0.00 sec)
| visibility pack
color 120,0,0 rect * from centraldb.blatServers
type bed 6
```
### trackDb table

```sql
mysql> select * from trackDb where tableName = 'refSeq';
+--------------------------------------------------+
| shortLabel | longLabel | ggbBackColor | ggbBorderColor | ggbHighlightColor | ggbOutlineColor | ggbTextColor | ggbTheme | ggbType | ggbVisibility | ggbPriority | ggbColorR | ggbColorG | ggbColorB | ggbUseScore | ggbPrivate | ggbRestrictCount | ggbRestrictList | ggbUrl | ggbHtml | ggbGrp | ggbCanPac |
+--------------------------------------------------+
| refSeqRows | Genbank RefSeq | genePred gpProtCodePep | Genbank RefSeq Gene Annotations | 3 | 2.1 | 0 | 10 |
| 0 | www.100 | ur-doc | 127 | ls/stre177 | GCC7710 | 177 | ls-list.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | genes |
| 1 | color 0,100,100 | solococcus thermophilus | GCC7710 (Db: stretherDGCC7710, Abbr: Stre_ther DGCC7710) |
| group genes | Genbank RefSeq Gene Annotations |
+--------------------------------------------------+
```

nextItemButton on c.ucsc.edu closed.
priority 2.11332b95w: tulipa% eero
shortLabel Genbank RefSeq word:
track refSeqMay 22 16:24:39 2011 from c-24-130-132-29.hsd1.ca.comcast.net
type genePred gpProtCodePep
visibility pack:
support requests can be submitted via the web at
https://itrequest.ucsc.edu
+--------------------------------------------------+
| shortLabel | longLabel | ggbBackColor | ggbBorderColor | ggbHighlightColor | ggbOutlineColor | ggbTextColor | ggbTheme | ggbType | ggbVisibility | ggbPriority | ggbColorR | ggbColorG | ggbColorB | ggbUseScore | ggbPrivate | ggbRestrictCount | ggbRestrictList | ggbUrl | ggbHtml | ggbGrp | ggbCanPac |
+--------------------------------------------------+
```
Kent Code Base

• Need to get Kent source to set up browser
• Latest source code for all programs are in GIT
  
  http://genomewiki.ucsc.edu/index.php/Getting_Started_With_Git

  Put kent source tree in your home directory (or create symbolic link)
• Run make utils in ~/kent/src
• Binaries will be installed in ~/bin/$ {MACHTYPE}
Browser Configuration File

- `.hg.conf` – a hidden file
- Contains
  - MySQL user accounts and passwords
  - centraldb info
  - trackDb info
- Required by Kent applications to connect to MySQL
- Obtain this file from browser admin/developer
- Store it in your home directory
- Set `rw-------` permission
Prepare Genome Sequences

- Create `/gbdb/newGenome` directory for a new genome assembly
- Convert genome sequences from FASTA to 2bit format
  
  ```bash
  faToTwoBit chr1.fa [chr2.fa ...] \ 
  /gbdb/newGenome/newGenome.2bit
  ```
- Make sure the input FASTA files have UNIX LF character
Setup Genome Database

- Create a MySQL database for the genome assembly
  
  hgsqql "" -e "create database if not exists newGenome"

- Create a group table for the new database
  
  cd ~/kent/src/hg/lib
  hgsqql newGenome < grp.sql

  This is the table for creating these groups on the browser
Setup Genome Database (cont’d)

- Create a chromInfo table
  This makes the browser know the genome sequences
  
  `faSize -detailed chr1.fa [chr2.fa ...] > \chrominfo.tab`
  
  `hgsql newGenome < \~/kent/src/hg/lib/chromInfo.sql`
  
  `hgsql newGenome -e 'load data local infile "chrominfo.tab" into table chromInfo;'

  `hgsql newGenome -e 'update chromInfo set fileName = "/gbdb/newGenome/newGenome.2bit"'`
Make New Genome Available

• Add an entry into `centraldb.dbDb table`

  ```
  hgsq<sup>l</sup> 'centraldb' -e 'INSERT INTO dbDb 
  (name, description, nibPath, organism, 
  defaultPos, active, orderKey, genome, 
  scientificName, 
  htmlPath, hgNearOk, hgPbOk, sourceName) 
VALUES("pyrAer1", "Dec 2001", "/<gbdb/pyrAer1>", 
"Pyrobaculum aerophilum", 
"chr:10001-35000", 1, 310, 
"Pyrobaculum aerophilum", 
"Pyrobaculum aerophilum str. IM2", 
"/gbdb/pyrAer1/html/description.html", 0, 0, 
"NCBI");'
  ```
Make New Genome Available (cont’d)

• Add an entry into `centraldb.defaultDb` table
  ```
  hgsq1 'centraldb' -e 'INSERT INTO defaultDb (genome, name) VALUES("Pyrobaculum aerophilum", "pyrAer1");'
  ```
• Add an entry into `centraldb.genomeClade` table
  ```
  hgsq1 'centraldb' -e 'INSERT INTO genomeClade (genome, clade, priority) VALUES("Pyrobaculum aerophilum", "archaea-crenarchaeota", 85);'
  ```
• If the genome belongs to a clade that is not in the browser, add an entry into `centraldb.clade` table
  ```
  hgsq1 'centraldb' -e 'INSERT INTO clade (name, label, priority) VALUES("archaea-crenarchaeota", "Archaea-Crenarchaeae", 1);'
  ```
Add Genome Description

To add a description page for a genome, create a HTML file as /gbdb/newGenome/html/description.html
Track Configuration

- Each genome database needs to have a trackDb table
- The global trackDb.ra for UCSC Genome Browser is in `~/kent/src/hg/makeDb/trackDb`
- Genome-specific trackDb.ra is stored in `~/kent/src/hg/makeDb/trackDb/<DB name>`
- Can be stored at alternate location
Search Configuration

• A `hgFindSpec` table is required for specifying search criteria
• Search criteria for each track are also loaded from `trackDb.ra`
Track and Search Configuration

- To create `trackDb` and `hgFindSpec` table,
  
  ```bash
  mkdir ~/kent/src/hg/makeDb/trackDb/newGenome
  cd ~/kent/src/hg/makeDb/trackDb
  hgTrackDb -strict newGenome trackDb "~/kent/src/hg/lib/trackDb.sql"
  hgFindSpec -strict newGenome hgFindSpec "~/kent/src/hg/lib/hgFindSpec.sql"
  ```
Start BLAT Server

- To run BLAT, gfServer for each genome has to be started
- Insert 2 records into `centraldb.blatServers` table
  
  ```
  hgsqql 'centraldb' -e 'INSERT INTO blatServers
  (db, host, port, isTrans, canPcr) VALUES
  ("newGenome", "blat_host.cse.ucsc.edu",
  12345, 0, 1);'
  
  hgsqql 'centraldb' -e 'INSERT INTO blatServers
  (db, host, port, isTrans, canPcr) VALUES
  ("newGenome", "blat_host.cse.ucsc.edu",
  12346, 1, 0);'
  ```

- Make sure the port numbers are unique
Start BLAT Server (cont’d)

- If BLAT server is not going to run locally,
  
  ```
  rsync -v /gbdb/newGenome/newGenome.2bit \
   blat_host:/gbdb/newGenome/
  ```

- At the host machine, start BLAT server in the background
  
  ```
  cd /gbdb/newGenome
  
gfServer -tileSize=7 -canStop start \
   blat_host.cse.ucsc.edu 12345 \
   -stepSize=5 newGenome.2bit &

  gfServer -canStop start \
   blat_host.cse.ucsc.edu 12346 -trans \
   newGenome.2bit &
  ```
Stop BLAT Server

• Run the following to stop the BLAT server
  gfServer stop blat_host 12345
  gfServer stop blat_host 12346
Automation

• The steps discussed previously can be automated by writing some scripts
• For loading hundreds of microbial genomes, we developed a perl script called “make-browser”