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

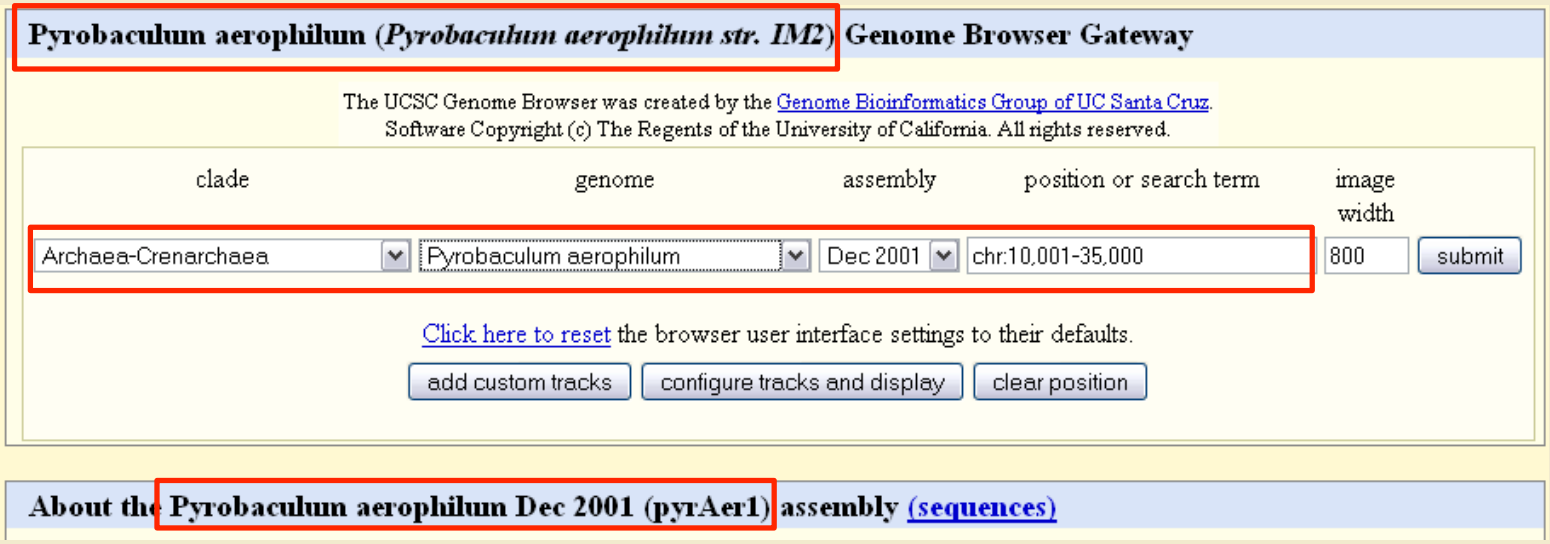
centraldb

- A database in MySQL containing all genome info

```
mysql> select * from centraldb.dbDb where name = 'pyrAer1';
+---------+-------------+-----------------+----------------------+----------------+---------+----------+-----------------------
| name    | description | nibPath         | organism             | defaultPos     | active | orderKey | genome
| scientificName | htmlPath                                | hgNearOk | hgPbOk | sourceName |
+---------+-------------+-----------------+----------------------+----------------+---------+----------+-----------------------
| pyrAer1 | Dec 2001    | /gbdb/pyrAer1 | Pyrobaculum aerophilum | chr:10001-35000 |      1 |      340 | Pyrobaculum aerophilum
| Pyrobaculum aerophilum str. IM2 | /gbdb/pyrAer1/html/description.html |        0 |      0 |            |
+---------+-------------+-----------------+----------------------+----------------+---------+----------+-----------------------
```

```
mysql> select * from centraldb.blatServers where db = 'pyrAer1';
+---------+-----------------------+-------+---------+---------+
| db      | host                  | port  | isTrans | canPcr |
+---------+-----------------------+-------+---------+---------+
| pyrAer1 | lowepub.cse.ucsc.edu | 23370 |       1 |      0 |
| pyrAer1 | lowepub.cse.ucsc.edu | 23369 |       0 |      1 |
+---------+-----------------------+-------+---------+---------+
```

hgGateway



- 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

```
track refSeq
shortLabel Genbank RefSeq
longLabel Genbank RefSeq Gene Annotations
group genes
priority 2.1
visibility pack
color 0,100,100
type genePred gpProtCodePep
nextItemButton on
track scaffolds
shortLabel Scaffolds
longLabel Assembly Scaffolds
group map
priority 1.8
visibility pack
color 120,0,0
type bed 6
```

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

trackDb table

```
mysql> select * from trackDb where tableName = 'refSeq';
+-----------+-----------------+----------+-----------------------------+-----------+----------+--------+------
| tableName | shortLabel      | type     | longLabel                   | visibility | priority | colorR | color
G | colorB | altColorR | altColorG | altColorB | useScore | private | restrictCount | restrictList | url | html | grp   | canPac
k | settings
                                                                                                                |
+-----------+-----------------+----------+-----------------------------+-----------+----------+--------+------
| refSeq    | Genbank RefSeq | genePred gpProtCodePep | Genbank RefSeq Gene Annotations |          3 |      2.1 |      0 |    10
0 |   100 |       127 |       177 |       177 |       0 |       0 |             0 |              |     |      | genes |
1 | color 0,100,100
group genes
longLabel Genbank RefSeq Gene Annotations
nextItemButton on
priority 2.1
shortLabel Genbank RefSeq
track refSeq
type genePred gpProtCodePep
visibility pack
+-----------+-----------------+----------+-----------------------------+-----------+----------+--------+------
------------------------------------------------------------+
```

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

```
faToTwoBit chr1.fa [chr2.fa ...] \
  /gbdb/newGenome/newGenome.2bit
```

- Make sure the input FASTA files have UNIX LF character

Setup Genome Database

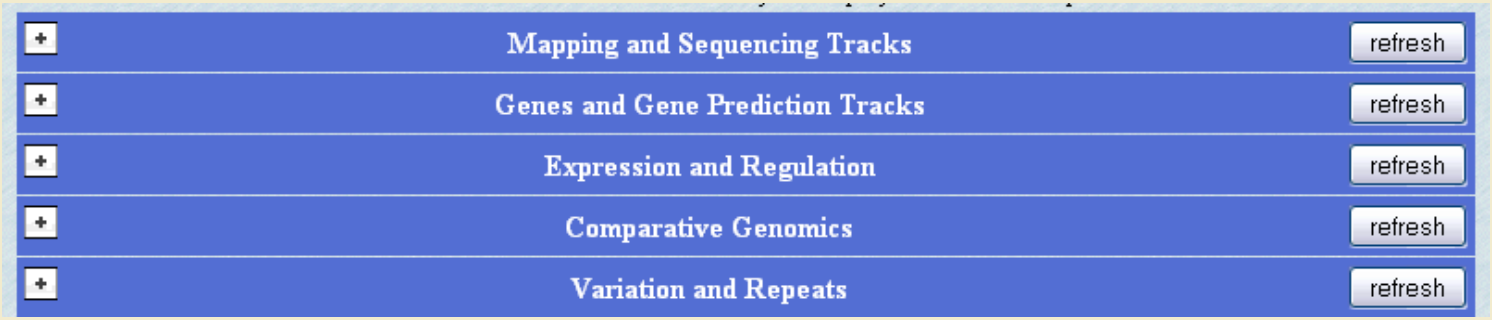
- Create a MySQL database for the genome assembly

```
hgsql "" -e "create database if not exists \
  newGenome"
```

- Create a group table for the new database

```
cd ~/kent/src/hg/lib
hgsql 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

```
hgsql '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

```
hgsql 'centraldb' -e 'INSERT INTO defaultDb
  (genome, name) \
  VALUES("Pyrobaculum aerophilum", "pyrAer1");'
```

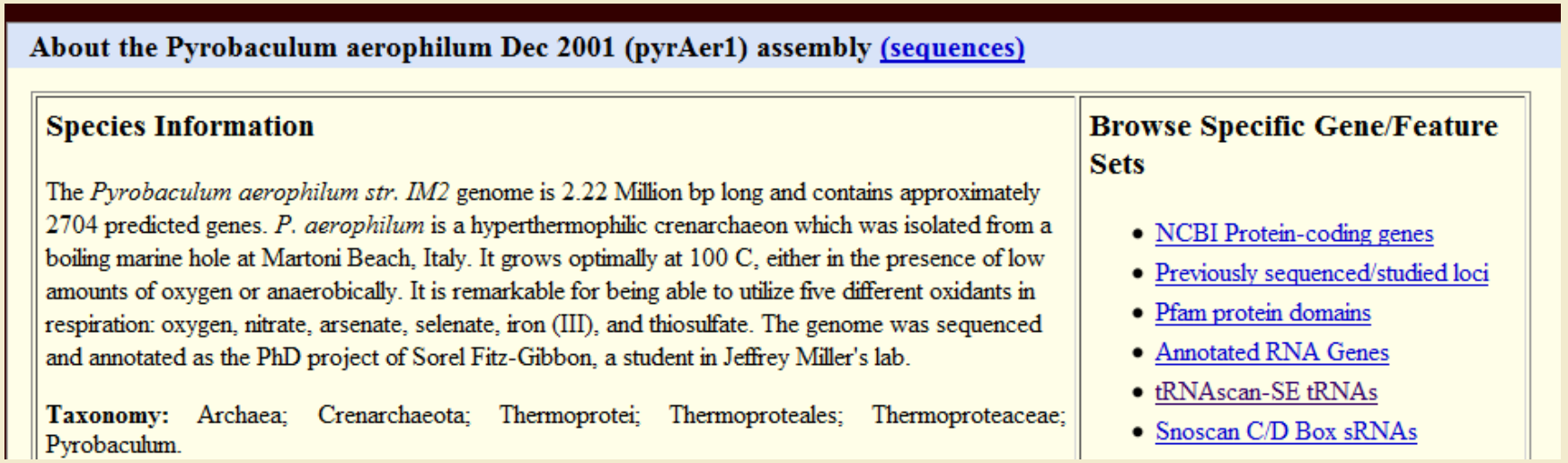
- Add an entry into `centraldb.genomeClade` table

```
hgsql '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

```
hgsql 'centraldb' -e 'INSERT INTO clade \
  (name, label, priority) \
  VALUES("archaea-crenarchaeota", \
  "Archaea-Crenarchaea", 1);'
```

Add Genome Description



About the Pyrobaculum aerophilum Dec 2001 (pyrAer1) assembly (sequences)

Species Information

The *Pyrobaculum aerophilum str. IM2* genome is 2.22 Million bp long and contains approximately 2704 predicted genes. *P. aerophilum* is a hyperthermophilic crenarchaeon which was isolated from a boiling marine hole at Martoni Beach, Italy. It grows optimally at 100 C, either in the presence of low amounts of oxygen or anaerobically. It is remarkable for being able to utilize five different oxidants in respiration: oxygen, nitrate, arsenate, selenate, iron (III), and thiosulfate. The genome was sequenced and annotated as the PhD project of Sorel Fitz-Gibbon, a student in Jeffrey Miller's lab.

Taxonomy: Archaea; Crenarchaeota; Thermoprotei; Thermoproteales; Thermoproteaceae; Pyrobaculum.

Browse Specific Gene/Feature Sets

- NCBI Protein-coding genes
- Previously sequenced/studied loci
- Pfam protein domains
- Annotated RNA Genes
- tRNAscan-SE tRNAs
- Snoscan C/D Box sRNAs

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,

```
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

```
hgsql 'centraldb' -e 'INSERT INTO blatServers
  (db, host, port, isTrans, canPcr) VALUES
  ("newGenome", “blat_host.cse.ucsc.edu",
  12345, 0, 1);'
hgsql '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”