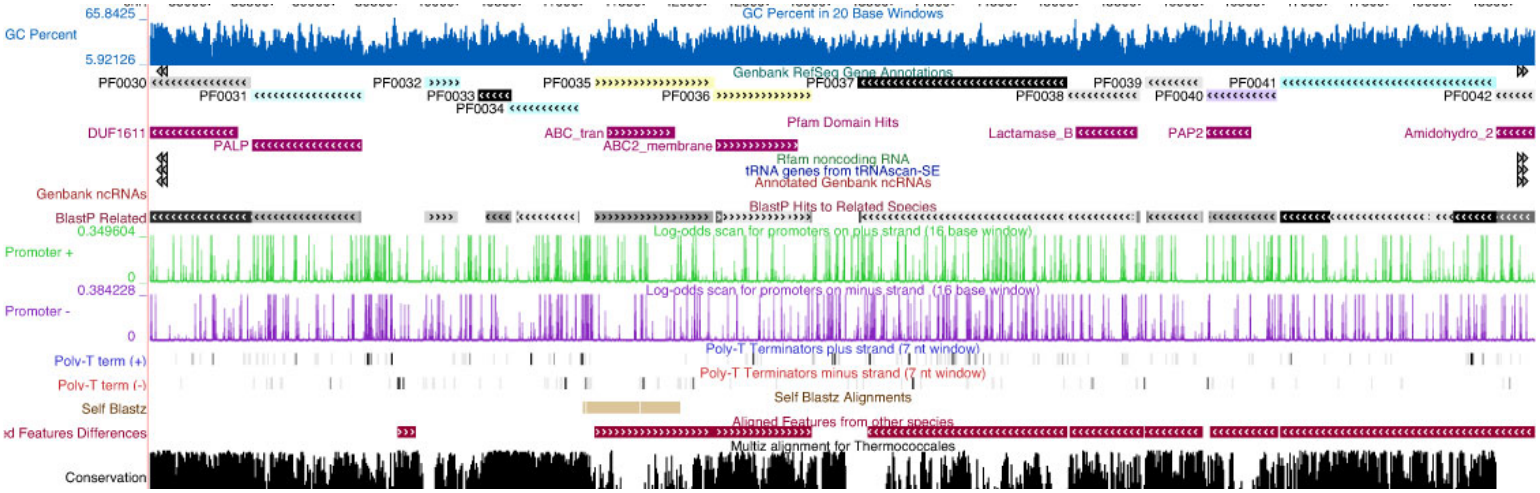
How to set up a genome browser



Patricia Chan

Overview

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

- Programs are mostly written in C with some Javascript

Kent Code Base

- Need to get Kent source to set up browser
- Latest source code for all programs are in CVS (will move to Git in June)

  `/projects/compbio/cvsroot/kent`
- 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

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 are stored in files

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

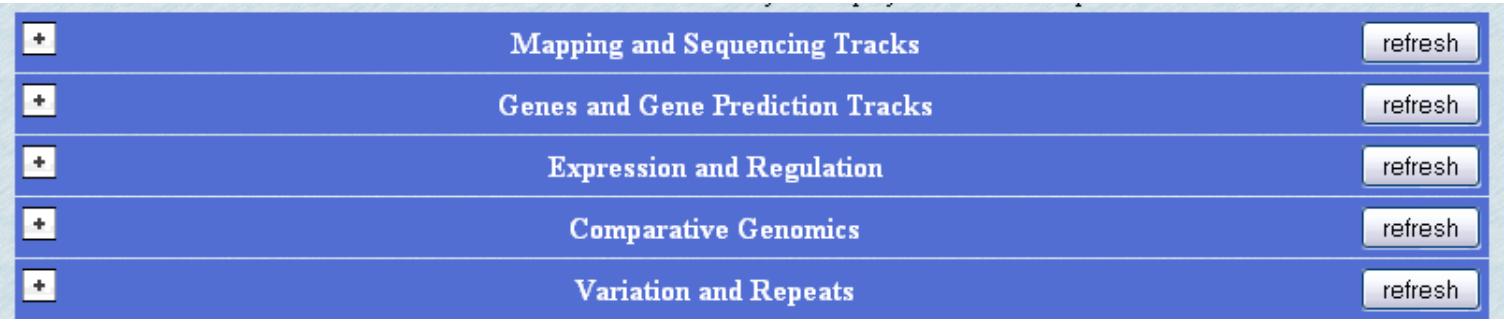
- Create a MySQL database for the genome assembly

```
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)

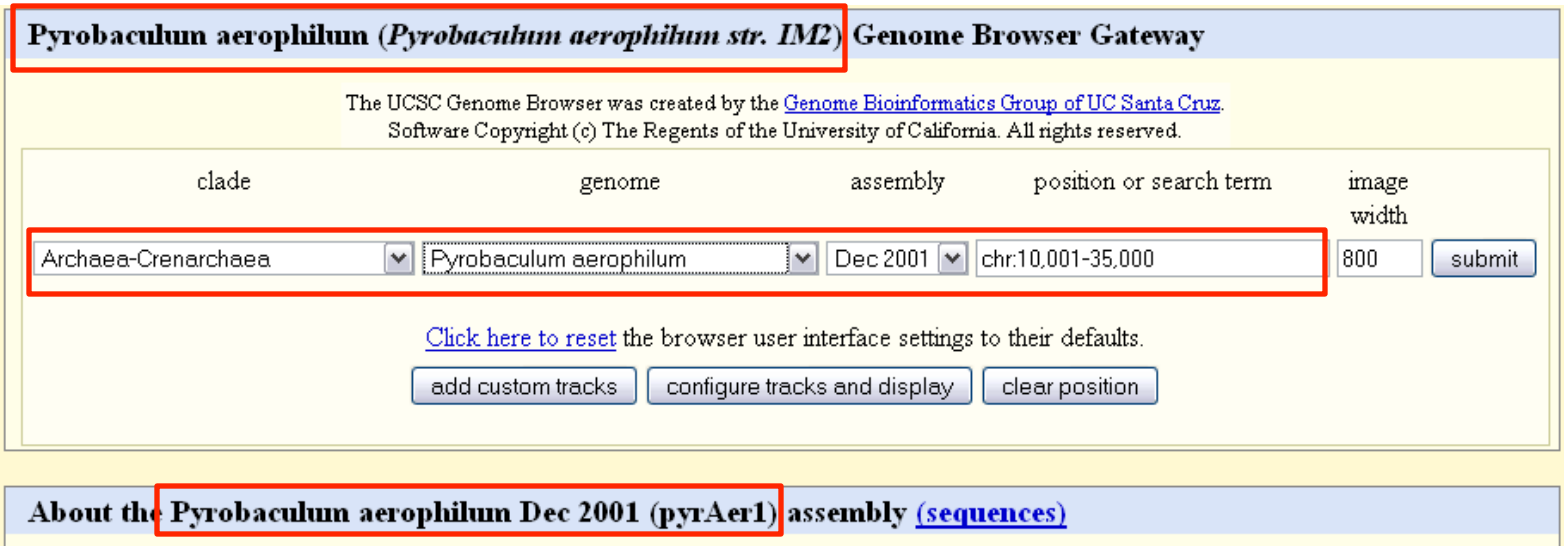
- Create a chromInfo table

  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

- centraldb database in MySQL contains information of all genomes in the browser



- centraldb may not be named as “centraldb”
- Get the database name from `central.db` entry in your `.hg.conf` or ask browser admin/developer

Make New Genome Available (cont'd)

- 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");'
```

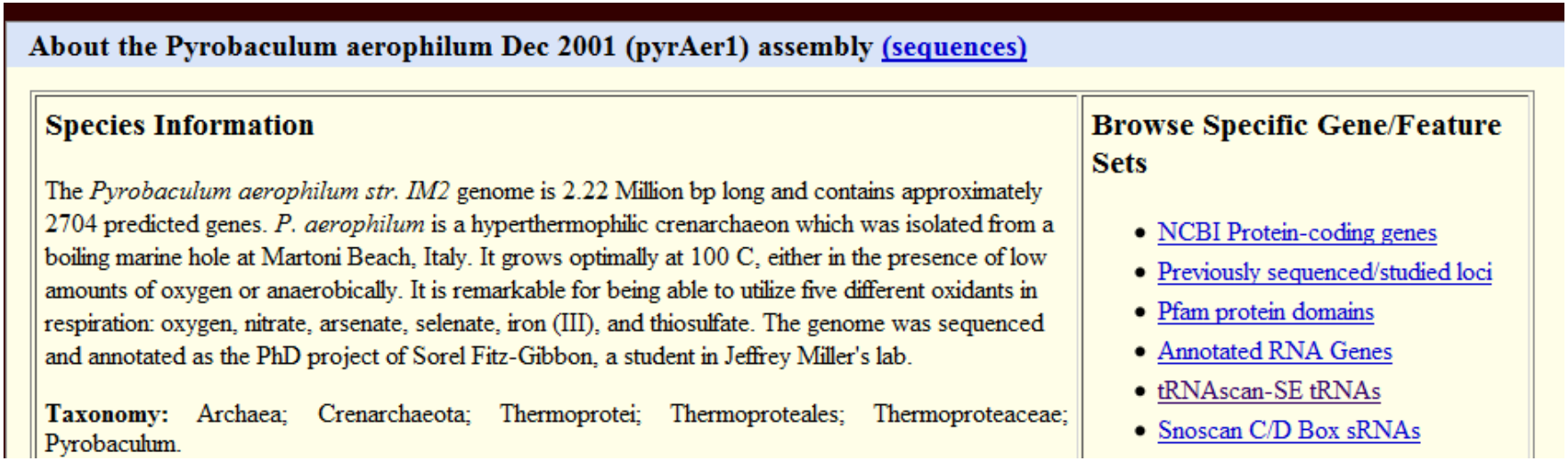
- Add an entry into `centraldb.genomeClade` table

```
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
- Track information are loaded from a file called `trackDb.ra`
- 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>`

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”

Demo

Adding Tracks

- UCSC Genome Browser supports different types of tracks, including bed, wiggle, psl, conservation, microarray, bam, etc
- Some involve more data formatting and setup
- Check out details at http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks

Adding Bed Track

- Create a bed file following the custom track description
- But, do not include browser line and track line

```
browser position chr7:127471196-127495720
browser hide all
track name="ColorByStrandDemo" description="Color by strand demonstration"
visibility=2 colorByStrand="255,0,0 0,0,255"
chr7    127471196  127472363  Pos1  0  +
chr7    127472363  127473530  Pos2  0  +
chr7    127473530  127474697  Pos3  0  +
chr7    127474697  127475864  Pos4  0  +
chr7    127475864  127477031  Neg1  0  -
chr7    127477031  127478198  Neg2  0  -
chr7    127478198  127479365  Neg3  0  -
chr7    127479365  127480532  Pos5  0  +
chr7    127480532  127481699  Neg4  0  -
```

- Load bed track into database

```
hgLoadBed -trimSqlTable newGenome
  newTrack newTrack.bed
```

Adding Track Info

- Add an entry of the track into `trackDb.ra`

```
track newTrack
shortLabel Gene Annotations
longLabel Gene Annotations from NCBI
group genes
priority 2.56
visibility dense
color 50,50,233
type bed 6
```

- Run

```
cd ~/kent/src/hg/makeDb/trackDb
hgTrackDb -strict newGenome trackDb \ ~/
  kent/src/hg/lib/trackDb.sql .
```

- Get details for creating trackDb entry at `~/kent/src/hg/makeDb/trackDb/README`

Adding Search Spec

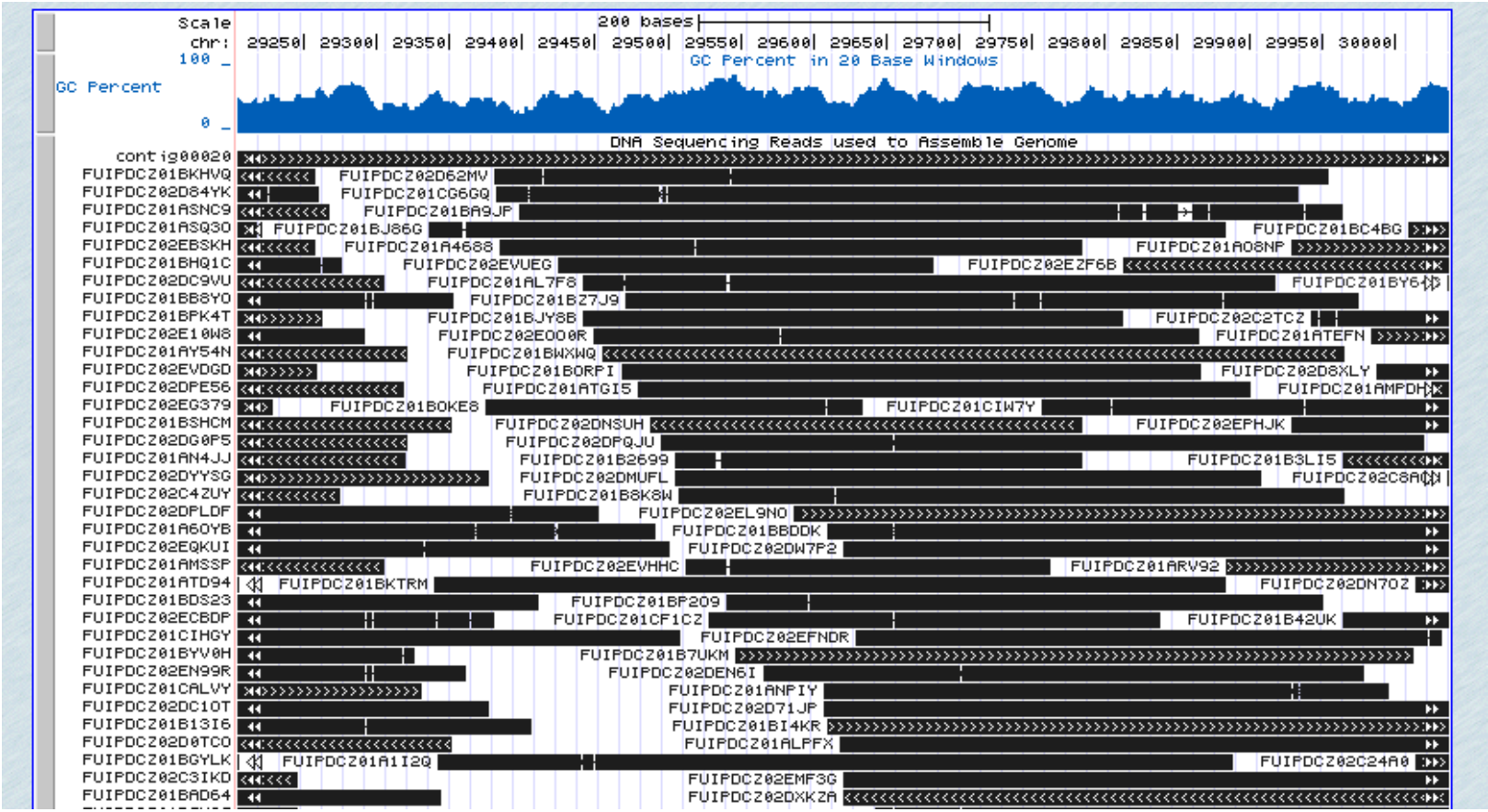
- To make the track data searchable, add a search entry into `trackDb.ra`

```
searchName newTrackSearch
searchTable newTrack
searchType bed
searchMethod exact
query select chrom,chromStart,chromEnd,name from %s where
  name like '%s'
searchDescription newTrack Gene Name
searchPriority 22
```

- Run

```
cd ~/kent/src/hg/makeDb/trackDb
hgFindSpec -strict newGenome hgFindSpec \
  ~/kent/src/hg/lib/hgFindSpec.sql .
```

Adding PSL Track



Sequencing reads align against assembled *Pyrobaculum oguniense*

Adding PSL Track

- Use PSL track to see BLAT search alignments
- Run

```
BLAT newGenome.fa query.fa pslTrack.psl
```

- Load PSL track into database

```
hgLoadPsl newGenome pslTrack.psl
```

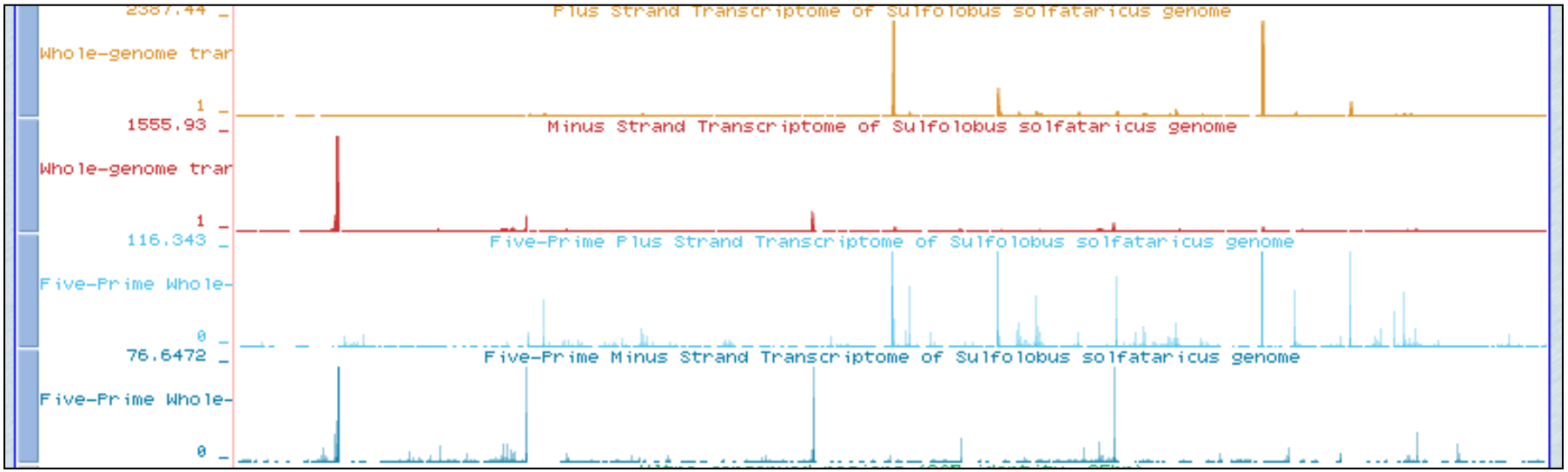
- Add trackDb.ra entry and load into database

Adding PSL Track (cont'd)

- To see pairwise alignments when clicking on track, load query sequence information into database

```
cp query.fa \ /gbdb/newGenome/
  query.fa
cd /gbdb/newGenome
hgLoadSeq newGenome query.fa
```

Adding Wiggle Track



- Wiggle track is a histogram
- Examples above show the RNA sequencing read coverage in *Sulfolobus solfataricus*

Adding Wiggle Track

- Create a WIG file following the custom track description
- But, do not include browser line and track line

```
fixedStep chrom=chr19 start=59307401 step=300 span=200
1000
 900
 800
 700
 600
 500
 400
 300
 200
 100
```

Adding Wiggle Track

- Convert ASCII WIG file to binary

```
wigEncode wigTrack.wig.ascii \
  wigTrack.wig wigTrack.wib
```

- Load wiggle track info into database

```
hgLoadWiggle newGenome wigTrack \
  wigTrack.wig
```

- Make binary WIB file accessable

```
cp wigTrack.wib \ /gbdb/newGenome/
  wib/
```

Demo