Advanced UCSC Browser Functions

HSL Mar 24 2010

UCSC Genome Bioinformatics

Dr. Thomas Randall
tarandal@email.unc.edu
bioinformatics.unc.edu

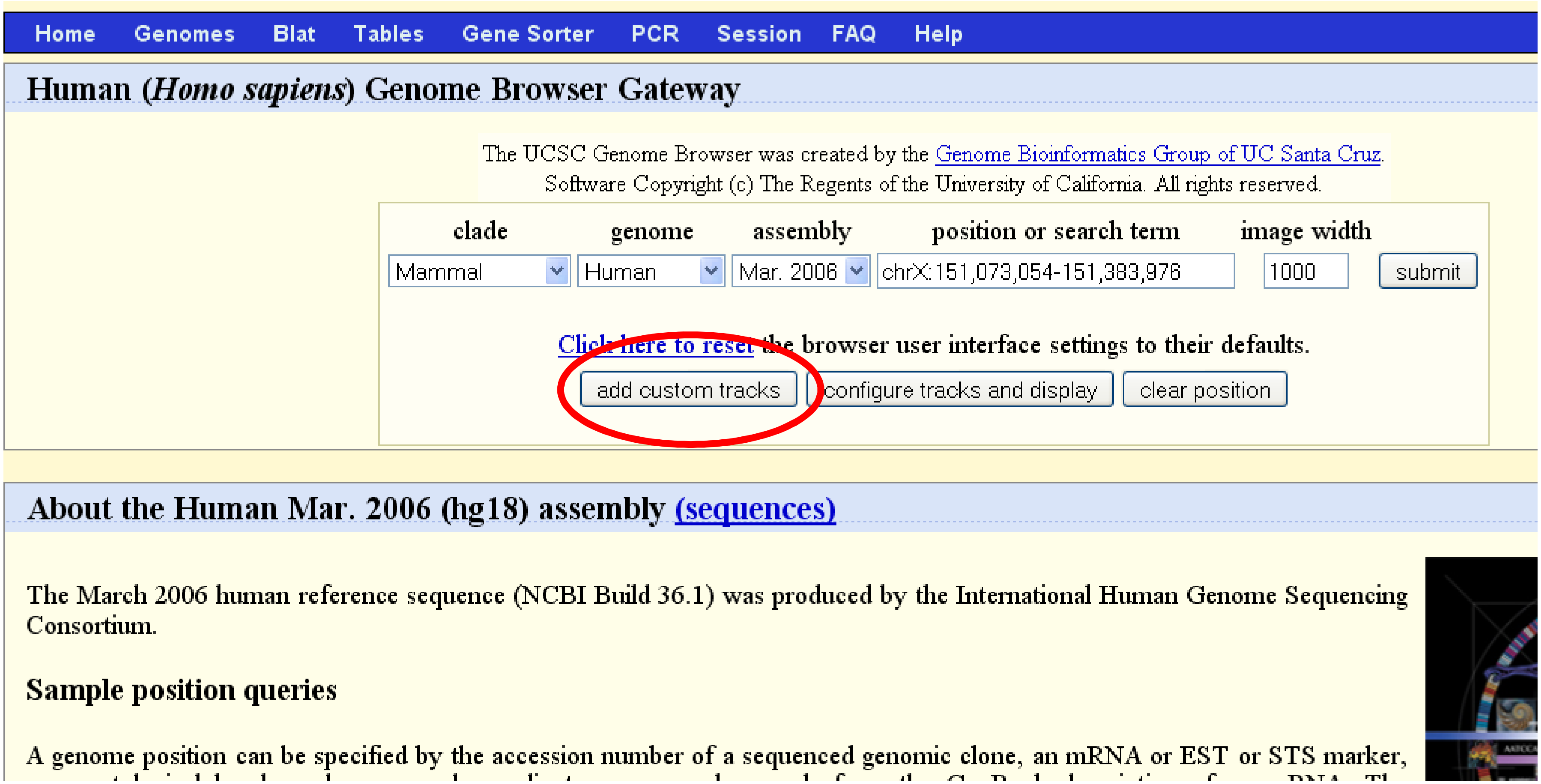
University of North Carolina at Chapel Hill
Center for Bioinformatics

UCSC Browser: http://genome.ucsc.edu

Overview

- **Custom Tracks – adding your own datasets**
- **Utilities – tools for manipulating files**
- **Saving session – saving alterations and custom tracks, sharing datasets**
- **Table Browser – large scale custom queries and downloads**
- **Galaxy – independent site with more custom tools, well integrated with UCSC**

Custom Tracks



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width	
Mammal	Human	Mar. 2006	chrX:151,073,054-151,383,976	1000	submit

Click here to reset the browser user interface settings to their defaults.

About the Human Mar. 2006 (hg18) assembly (sequences)

The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing 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,

If your own data is in the right format it can be displayed as a track

Custom Tracks

1) ctcfhg18.bed
www.unc.edu/~tarandal/HSL_Springfiles/UCSC_advanced
http://licr-renlab.ucsd.edu/download.html
The file ctcfhg18.bed is adapted from the file from Cell 128: 1231

2) Nature Genetics 38: 1289
http://research4.dfci.harvard.edu/brownlab//datasets/index.php?dir=ER_MCF7_whole_human_genome/

3) Also from above Cell paper
cftchg17.wig

4) primer data (primers.txt)
www.unc.edu/~tarandal/HSL_SpringFiles/UCSC_Advanced
we will make our own bed file

1) Map primers to genome with BLAT
2) Save results as a psl file
3) Load in UCSC Browser
4) Output in Table Browser as bed file

Minimal format for a bed file

Chr	start	stop
chr1	5319	6069
chr1	15612	16329
chr1	81077	82406
chr1	227508	228733
chr1	406299	406770
chr1	427582	428232
chr1	451635	451985
chr1	534463	536213
chr1	783006	783556
chr1	863362	863712
chr1	876627	877077
chr1	909263	909813
chr1	957868	958518

BED format – browser extensible data

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

The first three required BED fields are:

1. chrom *- The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).

2. chromStart * - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.

3. chromEnd *- The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. name - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.

5. score - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

6. strand - Defines the strand - either '+' or '-'.

7. thickStart - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).

8. thickEnd - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).

9. itemRgb - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.

10. blockCount - The number of blocks (exons) in the BED line.

11. blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.

12. blockStarts - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED

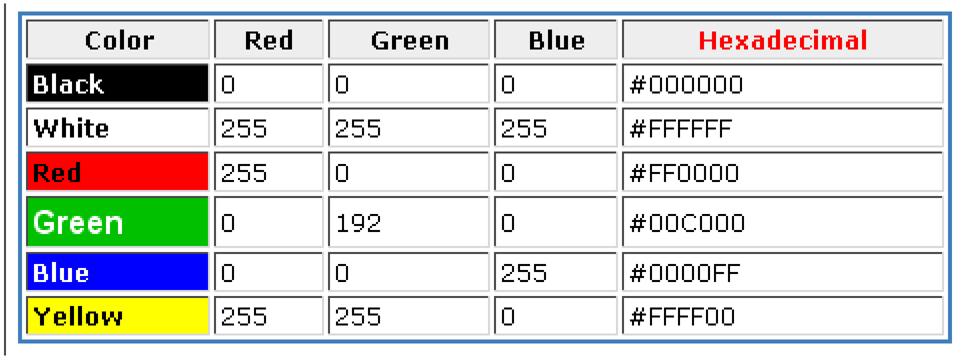
File formats for Custom Tracks

- Bed* (already covered)
- BedGraph - The bedGraph format allows display of continuous-valued data in track format. This display type is useful for probability scores and transcriptome data.
- GFF* (General Feature Format) lines are based on the GFF standard file format. GFF lines have nine required fields that *must* be tab-separated.
- GTF (Gene Transfer Format) is a refinement to GFF that tightens the specification.
- WIG* - The wiggle format is for display of dense, continuous data such as GC percent, probability scores, and transcriptome data.
- MAF - The multiple alignment format stores a series of multiple alignments in a format that is easy to parse and relatively easy to read. This format stores multiple alignments at the DNA level between entire genomes.
- PSL lines represent alignments, and are typically taken from files generated by BLAT or psLayout.

*****most commonly seen in papers, bed for discrete data, GFF for genome annotation, WIG for chip-chIP, chip-seq studies, more continuous and quantitative data

RGB color convention

**– all colors are specified by
some combination of values of red, green and blue
from 0 to 255**



Color	Red	Green	Blue	Hexadecimal
Black	0	0	0	#000000
White	255	255	255	#FFFFFF
Red	255	0	0	#FF0000
Green	0	192	0	#00C000
Blue	0	0	255	#0000FF
Yellow	255	255	0	#FFFF00

In “Manage Custom Tracks” > Name of Track
You can edit characteristics of track

Add color=255,0,0 to set color as **RED**

The Other RGB Color Chart

http://www.tayloredmktg.com/rgb/

bed, gff, wig files can get

BIG

bed files in Cell 129: 823 over 80 Mb

Example wig files (ctcfhg17.wig) on
www.unc.edu/~tarandal/HSL_Springfiles/UCSC_Advanced
Also GWAS datasets

Hard to map to the UCSC Browser as sending this through the internet takes time, the link may fail

Solution: local copy of the UCSC Browser, faster connection

Also, if you have a big file try the Duke mirror of the UCSC Browser
http://genome-mirror.duhs.duke.edu/

Making your own files

- Always create text files as your input
  - good general rule for all bioinformatics tools
- Save as plain text or tab delimited text – most tools recognize empty space as a tab
- Create and open in Wordpad, don’t use Notepad as the line breaks and delimiters are not recognized. For Macs, use Text Editor.
- Other text editors at http://bioinformatics.unc.edu/software/opensource/index.htm

Building our bed file

1. **Download or copy primers.txt**

2. **Open the UCSC Browser BLAT tool and input the above file. Run BLAT using default conditions.**

3. **Choose PSL as output type of BLAT, copy all and Paste into WordPad.**

4. **Save file as text as "primers.psl" to your desktop.**

5. **Load into custom tracks. Give a name.**

6. **Change color of track "color=0,0,255" (blue) "color=0,255", 0 (green) "color=255,0,0" (red)**

7. **Go to Table Browser and output this as a bed file.**

8. **Load into Ensembl by choosing Human and using the Manage your data function.**

PSL lines represent alignments, and are typically taken from files generated by BLAT or psLayout. See the BLAT documentation for more details. All of the following fields are required on each data line within a PSL file:

1.matches - Number of bases that match that aren't repeats

2.misMatches - Number of bases that don't match

3.repMatches - Number of bases that match but are part of repeats

4.nCount - Number of 'N' bases

5.qNumInsert - Number of inserts in query

6.qBaseInsert - Number of bases inserted in query

7.tNumInsert - Number of inserts in target

8.tBaseInsert - Number of bases inserted in target

9.strand - '+' or '-' for query strand. In mouse, second '+'or '-' is for genomic strand

10.qName - Query sequence name

11.qSize - Query sequence size

12.qStart - Alignment start position in query

13.qEnd - Alignment end position in query

14.tName - Target sequence name

15.tSize - Target sequence size

16.tStart - Alignment start position in target

17.tEnd - Alignment end position in target

18.blockCount - Number of blocks in the alignment

19.blockSizes - Comma-separated list of sizes of each block

20.qStarts - Comma-separated list of starting positions of each block in query

21.tStarts - Comma-separated list of starting positions of each block in target

Utilities

liftover - converts genome coordinates and genome annotation files between assemblies. The current version supports both forward and reverse conversions, as well as conversions between selected species.

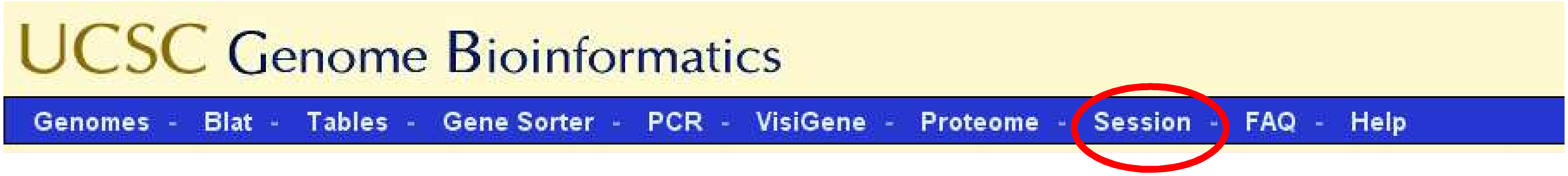
To go from hg16 to hg 18 one has to move one build at a time hg16 > hg17 > hg18

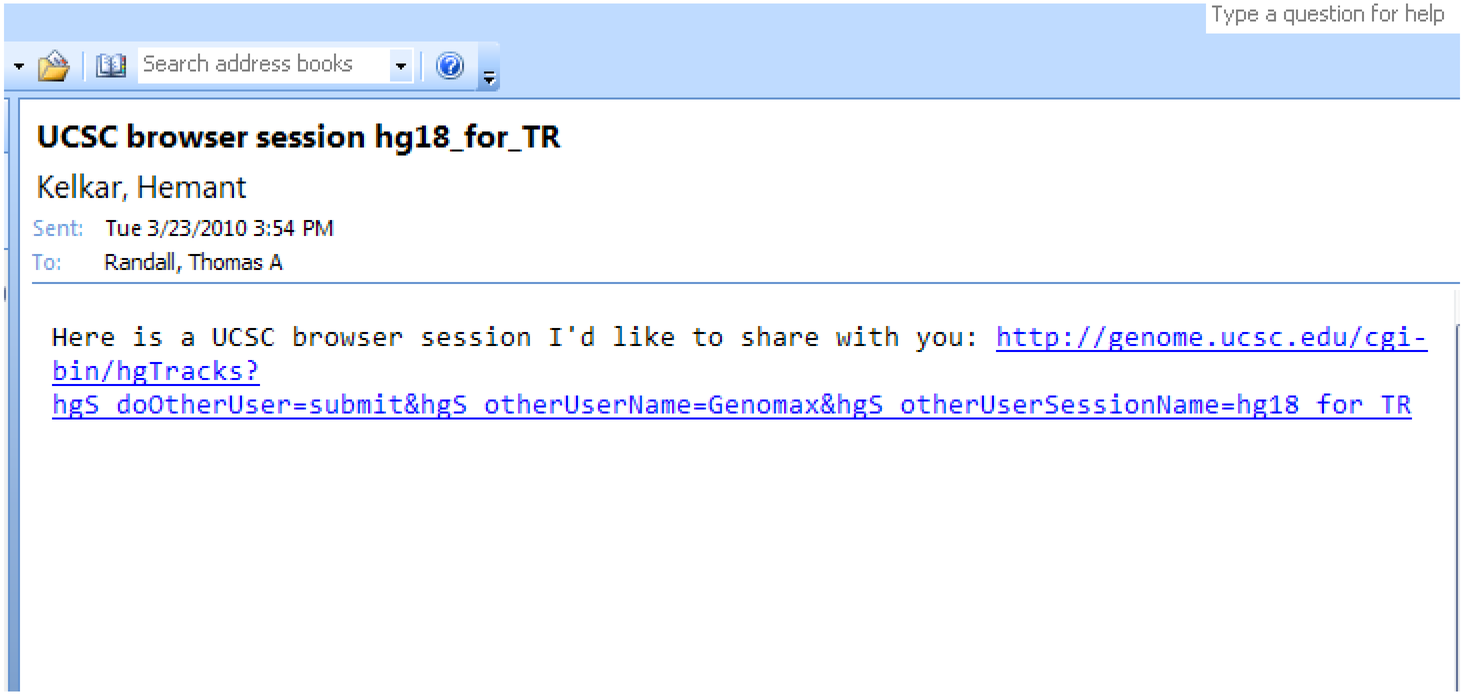
DNA and Protein duster: both removes formatting characters and other non-sequence-related characters from an input sequence.
Offers several configuration options for the output format.

Example: ctcf.bed (hg17) to hg18 Cell 128: 1231
(http://licr-renlab.ucsd.edu/download.html)

Saving sessions

- Custom tracks persist only ~48 hrs
- Use Save Session to keep custom tracks and to customize Genome Browser
- Create an account through Sessions
- Sessions persist for one year from last access time
- To save additions to a session, over-write old session of same name
- Can also save session to local file and reload, send to others – see HSLsession on website


UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help



UCSC browser session hg18_for_TR

Kelkar, Hemant

Sent: Tue 3/23/2010 3:54 PM

To: Randall, Thomas A

Here is a UCSC browser session I'd like to share with you: http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Genomax&hgS_otherUserSessionName=hg18_for_TR

Table Browser

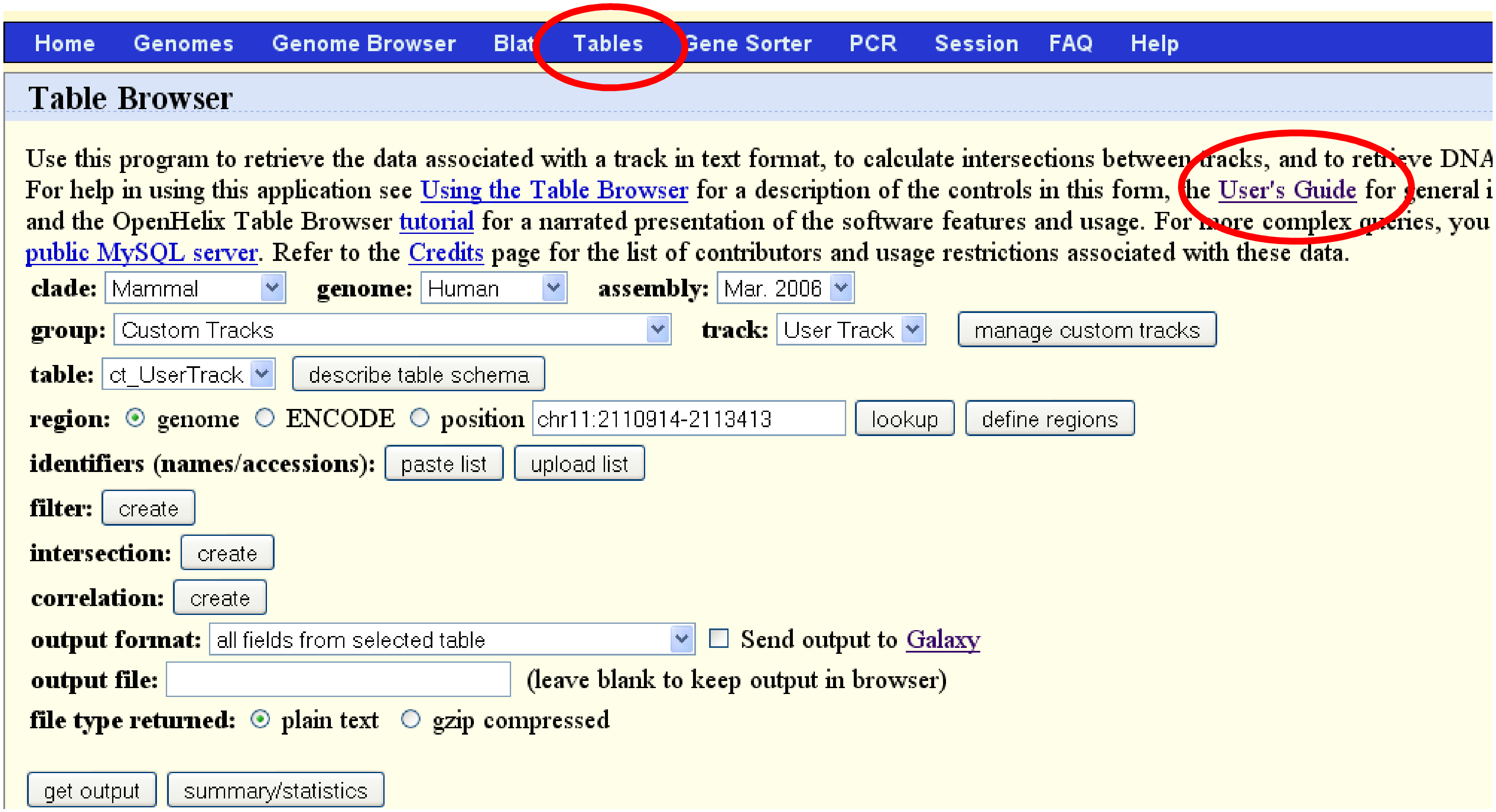


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA
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 i
and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you
public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

clade: Mammal **genome:** Human **assembly:** Mar. 2006

group: Custom Tracks **track:** User Track

table: ct_UserTrack

region: ● genome ○ ENCODE ○ position chr11:2110914-2113413

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: all fields from selected table □ Send output to Galaxy

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

All data (annotation tracks) from the Genome Browser is stored, and is available through the Table Browser via custom queries

Tips for Table Browser

- First two rows important to define data
- table – this is likely unimportant unless you are interested im MySQL databases
- Filter – used to restrict above selection to a more defined subset
- Intersection – used to get junctions of multiple annotation tracks
- Output format – important to specify type of output
- You can generate genome wide large datasets, gzip anything you think may be large
- Specify a name for the file or it will be loaded into the browser window
- Large datasets may time out – may need to go to Downloads section

Field descriptions for Table Browser

- **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.
- **group:** Selects the type of tracks to be displayed in the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the *group* list.
- **table:** Selects the SQL table data to use. This list shows all tables associated with the track specified in the *track* list.
- **describe table schema:** Displays schema information for the tables associated with the selected track.
- **region:** Restricts the query to a particular chromosome or region. Select *genome* to apply the query to the entire genome or *ENCODE* to examine only the ENCODE regions. To limit the query to a specific position, type a chromosome name, e.g. *chrX*, or a chromosome coordinate range, such as chrX:100000-200000, or a gene name or other id in the text box.
- **lookup:** Press this button after typing in a gene name or other id in the position text box to look up the chromosome position
- **identifiers** (selected tracks only)**:** Restricts the output to table data that match a list of identifiers, for instance RefSeq accessions for the RefSeq track. If no identifiers are entered, all table data within the specified region will be displayed.
- **filter:** Restricts the query to only those items that match certain criteria, e.g. genes with a single exon. Click the *Create* button to add a filter, the *Edit* button to modify an existing filter, or the *Clear* button to remove an existing filter.
- **intersection** (selected tracks only)**:** Combines the output of two queries into a single set of data based on specific join criteria. For example, this can be used to find all SNPs that intersect with RefSeq coding regions. The intersection can be configured to retain the existing alignment structure of the table with a specified amount of overlap, or discard the structure in favor of a simple list of position ranges using a base-pair intersection or union of the two data sets. The button functionalities are similar to those of the *filter* option.
- **output:** Specifies the output format (not all options are available for some tracks). Formats include:
  - ***all fields from selected table*** - data from the selected table displayed in a tab-separated format suitable for import into spreadsheets and relational databases. The ASCII format may be read in any web browser or text editor.
  - ***selected fields from primary and related tables*** - user-selected set of tab-separated fields from the selected table and (optionally) other related tables as well.
  - ***sequence*** - DNA (or protein sequence, in some cases) associated with the table.
  - ***BED*** - positions of data items in a standard UCSC Browser format.
  - ***GTF*** - positions of all data items in a standard gene prediction format. (Both BED and GTF formats can be used as the basis for custom tracks).
  - ***CDS FASTA alignment from multiple alignment*** - FASTA alignments of the CDS regions of a gene prediction track using any of the multiple alignment tracks for the current database. Output sequence can be in either nucleotide-space or translated to protein-space. Available only for genePred tracks.
  - ***custom track*** - customized Genome Browser annotation track based on the results of the query.
  - ***hyperlinks to Genome Browser*** - returns a page full of hyperlinks to the UCSC Genome Browser, one for each item in the table.
  - ***data points*** - the data points that make up a graph (aka wiggle) track.
  - ***MAF*** - multiple alignments in MAF format
- **Send output to Galaxy:** displays results of query in Galaxy, a framework for interactive genome analysis.
- **file type returned:** When a filename is entered in the "output file" text box, specifies the format of the output file:
  - ***plain text*** - data is in ASCII format
  - ***gzip compressed*** - data is compressed in gzip format
- **get output:** Submits a data query based on the specified parameters and returns the output.
- **summary/statistics:** Displays statistics about the data specified by the parameters.

Downloading individual genes with the Table Browser

- Go to the gene in the Genome Browser
- Select Tables to go to Table Browser
- Choose position (chromosomal position of genome browser is maintained)
- Choose output format as sequence
- Give file name
- Choose sequence type for download (genomic, protein or mRNA)
- Will output all sequences involved with particular track chosen – if there are three RefSeqs you will get three sequences
- If you want a specific identifier, paste it in.

Example 1
basic query

Dear Madam/Sir,

I am intending to obtain the sequence of promoter region, about -/+ 500bp around the transcription start site. Would you please tell me how to get those sequence in batch? Thanks a lot.

Best regards,
Rex

We will try 100 bp upstream so the download does not get too big

clade: Mammal; genome: Human; assembly: Mar. 2006; group: Genes and Gene Prediction Tracks; track: UCSC Genes; table: knownGene; region: "genome", output format: "sequence" output file “filename” and click "get output". On the next page you can select genomic sequence and then your promotor/upstream bases. Enter “100" into each box.

Example 2

intersection of two datasets

I want to know all the miRNAs that overlap with all known human genes

1) Select all genes
clade: Mammal; genome: Human; assembly: .Mar 2006;
group: Genes and Gene Prediction Tracks; track: Refseq Genes;
table: knownGene; region: "genome“

Create Intersection

2) Create intersection with all miRNAs
clade: Mammal; genome: Human; assembly: Mar 2006;
group: Genes and Gene Prediction Tracks;
track: sno/miRNAs; table: sno/miRNAs; use all default settings

output format: "sequence“ set file name and click "get output". On the next page select CDS exons. RESULT 65 human genes overlap with sno/miRNAs

Example 3

filtering a dataset

I want all genes on human chromosome 22 with more than 5 exons

1) Select all genes
clade: Mammal; genome: Human; assembly: Mar 2006;
group: Genes and Gene Prediction Tracks; track: UCSC Genes;
table: knownGene; region: "genome“

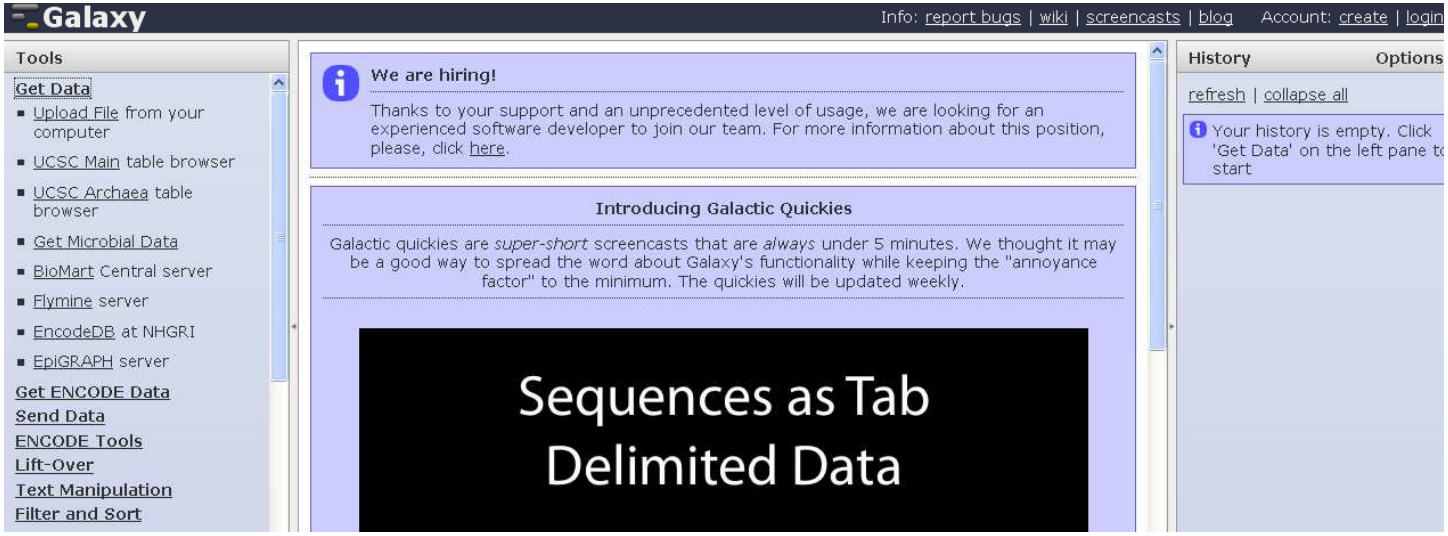
2) Create filter
chrom does match chr22
exonCount is > 5, submit

output format: "sequence“ set file name and click "get output". On the next page select CDS exons.   RESULT 851 chr22 genes overlap with >5 exons

No tool is an island

In genome queries, multiple tools required
Interaction between tools is often the limiting factor

e! Ensembl + NCBI + UCSC + Galaxy + ???



Galaxy Info: report bugs | wiki | screencasts | blog Account: create | login

Tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- Get Microbial Data
- BioMart Central server
- Flymine server
- EncodeDB at NHGRI
- EpiGRAPH server

Get ENCODE Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

We are hiring!

Thanks to your support and an unprecedented level of usage, we are looking for an experienced software developer to join our team. For more information about this position, please, click here.

Introducing Galactic Quickies

Galactic quickies are *super-short* screencasts that are *always* under 5 minutes. We thought it may be a good way to spread the word about Galaxy's functionality while keeping the "annoyance factor" to the minimum. The quickies will be updated weekly.

Sequences as Tab Delimited Data

History Options

refresh | collapse all

Your history is empty. Click 'Get Data' on the left pane to start

http://main.g2.bx.psu.edu/

see PLOS Comp Biol 4 e121 for discussion on the interaction of these tools

Galaxy examples

1) Use existing ctcf18.bed file.
This has coordinates for all binding sites of the CTCF protein in humans.
I want the sequence corresponding to those coordinates in order to determine if there are any conserved motifs.

Get Data – upload file, choose species and execute

Fetch sequences – extract genomic DNA

Save to desktop

RESULT – 13801 fasta formatted sequences for further motif analysis (separate tool)

Galaxy examples

Move output of Table Browser into Galaxy for further processing.

I want to know what the longest and shortest Refseq gene on chr22 are

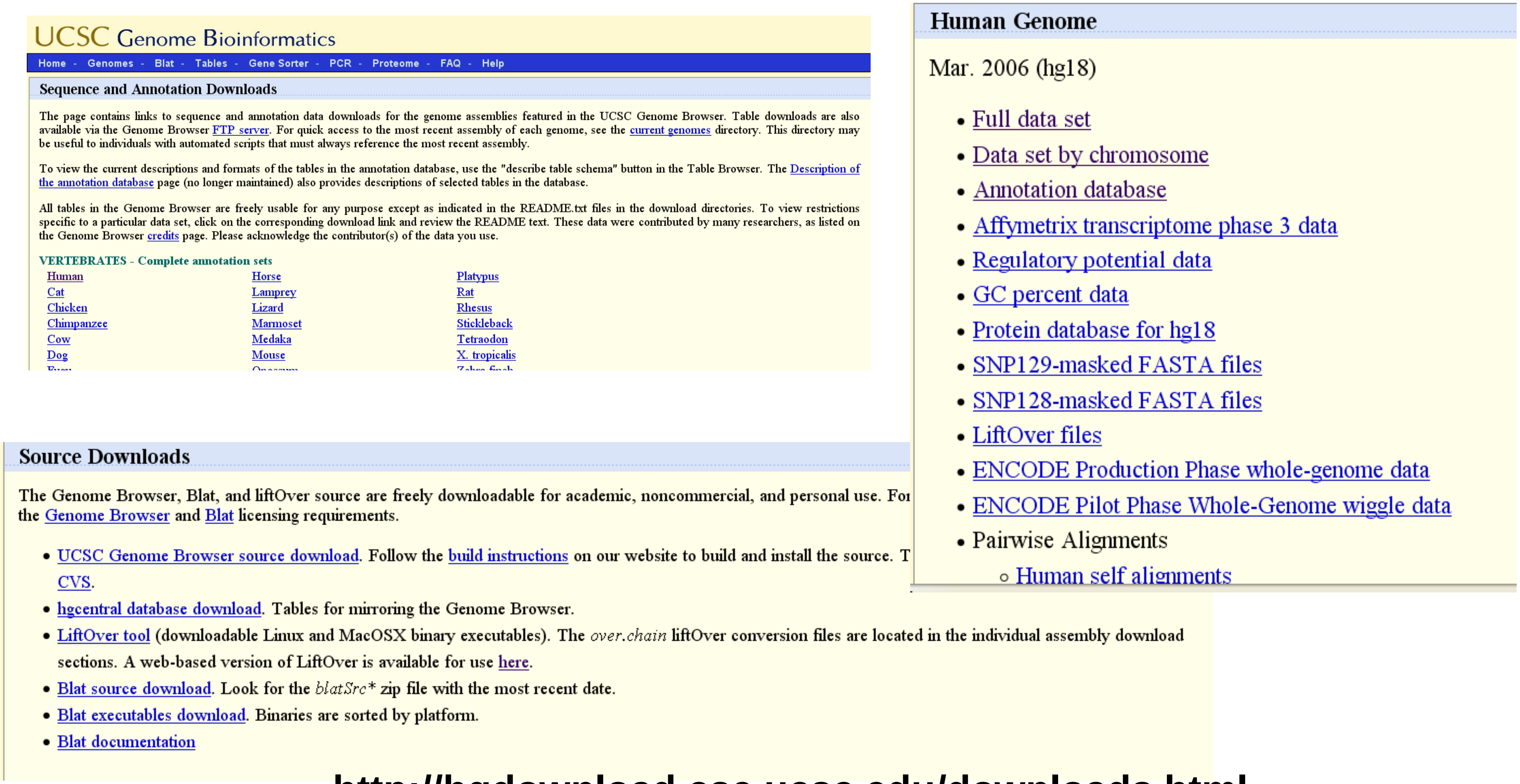
1) In Table Browser:
clade: Mammal; genome: Human; assembly: Mar 2006;
group: Genes and Gene Prediction Tracks; track: Refseq Genes;
table: knownGene; region: "genome“

Create filter

2) chrom does match chr22
Choose output to bed file and send query to Galaxy

3) In Galaxy
Fetch sequences – extract genomic DNA
FASTA manipulation – compute length
Filter and Sort – sort on c2, descending order

Downloads of all sequence data, tracks, all software


UCSC Genome Bioinformatics
Home - Genomes - Blat - Tables - Gene Sorter - PCR - Proteome - FAQ - Help

Sequence and Annotation Downloads

The page contains links to sequence and annotation data downloads for the genome assemblies featured in the UCSC Genome Browser. Table downloads are also available via the Genome Browser FTP server. For quick access to the most recent assembly of each genome, see the current genomes directory. This directory may be useful to individuals with automated scripts that must always reference the most recent assembly.

To view the current descriptions and formats of the tables in the annotation database, use the "describe table schema" button in the Table Browser. The Description of the annotation database page (no longer maintained) also provides descriptions of selected tables in the database.

All tables in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. To view restrictions specific to a particular data set, click on the corresponding download link and review the README text. These data were contributed by many researchers, as listed on the Genome Browser credits page. Please acknowledge the contributor(s) of the data you use.

VERTEBRATES - Complete annotation sets

Human
Cat
Chicken
Chimpanzee
Cow
Dog
Fugu
Horse
Lamprey
Lizard
Marmoset
Medaka
Mouse
Opossum
Platypus
Rat
Rhesus
Stickleback
Tetraodon
X. tropicalis
Zebra finch

Human Genome

Mar. 2006 (hg18)

- Full data set
- Data set by chromosome
- Annotation database
- Affymetrix transcriptome phase 3 data
- Regulatory potential data
- GC percent data
- Protein database for hg18
- SNP129-masked FASTA files
- SNP128-masked FASTA files
- LiftOver files
- ENCODE Production Phase whole-genome data
- ENCODE Pilot Phase Whole-Genome wiggle data
- Pairwise Alignments
  - Human self alignments

Source Downloads

The Genome Browser, Blat, and liftOver source are freely downloadable for academic, noncommercial, and personal use. For the Genome Browser and Blat licensing requirements.

- UCSC Genome Browser source download. Follow the build instructions on our website to build and install the source. T CVS.
- hgcentral database download. Tables for mirroring the Genome Browser.
- LiftOver tool (downloadable Linux and MacOSX binary executables). The *over.chain* liftOver conversion files are located in the individual assembly download sections. A web-based version of LiftOver is available for use here.
- Blat source download. Look for the *blatSrc** zip file with the most recent date.
- Blat executables download. Binaries are sorted by platform.
- Blat documentation

http://hgdownload.cse.ucsc.edu/downloads.html