MDB and CV

In the Shadow of
The Shadow of
Greatness

Brian and Tim
March 2011

by Popular Demand?

- What is the MDB and CV?
- Why would you want to know?
- How do you use them?
- Why, What and How

Why?

Metadata and the Forces of Control vs. Chaos

http://getsmart.wikia.com

- But tracks already have metadata in the trackDb.
- The trackDb is almost all display settings.
- The "Description Page" (trackName.html) is free form which is great. But sometimes a more controlled form is useful.
- The trackDb is only for tracks. But what about metadata tied to non-track items like downloadable files?
- The **MDB** stands for ***Meta Data** bro.*
- The **CV** stands for ***Controlled Vocabulary***, and defines common terms used in a precise way across multiple tracks.

Where can you see the MDB?

NHGRI Bi-Pro Track Settings

ENCODE NHGRI Elnitski Bidirectional Promoters (▲All Regulation tracks)

Display mode: hide Submit

Filter by Annotation (select multiple items - help)

All

View table: schema, downloads, metadata

ENCODE NHGRI Elnitski Bidirectional Promoters

shortLabel: NHGRI Bi-Pro

Principal Investigator on grant: Elnitski

Lab producing data: NHGRI

Experiment (Assay) type: BiP

Cell, tissue or DNA sample: Reference_Genome

species: hg18

Submission ID: 294

Date submitted to UCSC: 2009-04-27

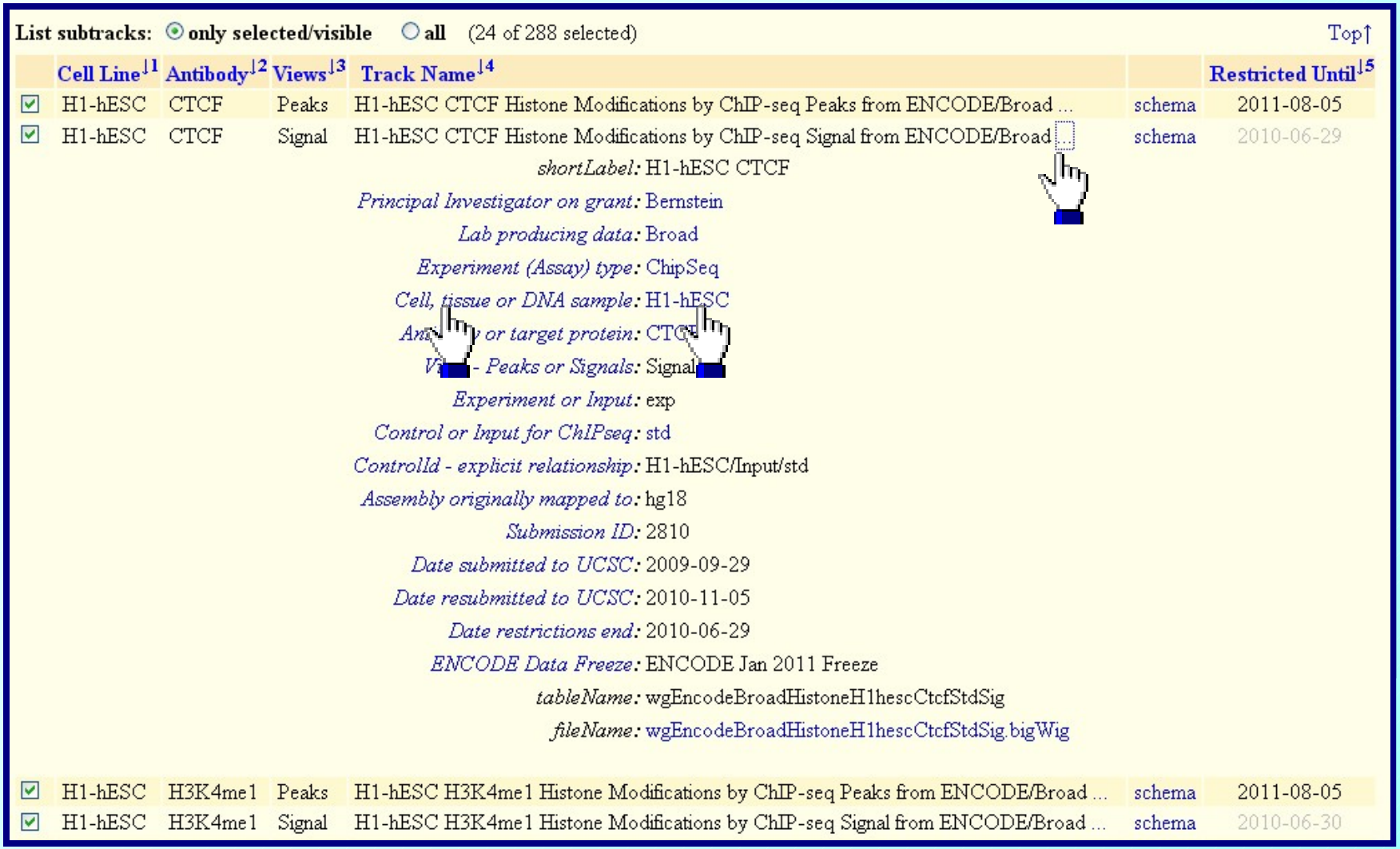
Date restrictions end: 2010-01-27

ENCODE Data Freeze: "ENCODE July 2009 Freeze"

fileName: wgEncodeNhgriBip.hg18.bed8.gz

Data last updated: 2009-04-27

“ ”
. . .


List subtracks: only selected/visible all (24 of 288 selected) Top↑
Cell Line↓1 Antibody↓2 Views↓3 Track Name↓4 Restricted Until↓5
H1-hESC CTCF Peaks H1-hESC CTCF Histone Modifications by ChIP-seq Peaks from ENCODE/Broad ... schema 2011-08-05
H1-hESC CTCF Signal H1-hESC CTCF Histone Modifications by ChIP-seq Signal from ENCODE/Broad ... schema 2010-06-29
shortLabel: H1-hESC CTCF
Principal Investigator on grant: Bernstein
Lab producing data: Broad
Experiment (Assay) type: ChipSeq
Cell, tissue or DNA sample: H1-hESC
Peaks or Signals: Signal
Experiment or Input: exp
Control or Input for ChIPseq: std
ControlId - explicit relationship: H1-hESC/Input/std
Assembly originally mapped to: hg18
Submission ID: 2810
Date submitted to UCSC: 2009-09-29
Date resubmitted to UCSC: 2010-11-05
Date restrictions end: 2010-06-29
ENCODE Data Freeze: ENCODE Jan 2011 Freeze
tableName: wgEncodeBroadHistoneH1hescCtcfStdSig
fileName: wgEncodeBroadHistoneH1hescCtcfStdSig.bigWig
H1-hESC H3K4me1 Peaks H1-hESC H3K4me1 Histone Modifications by ChIP-seq Peaks from ENCODE/Broad ... schema 2011-08-05
H1-hESC H3K4me1 Signal H1-hESC H3K4me1 Histone Modifications by ChIP-seq Signal from ENCODE/Broad ... schema 2010-06-30

Controlled Vocabulary revealed

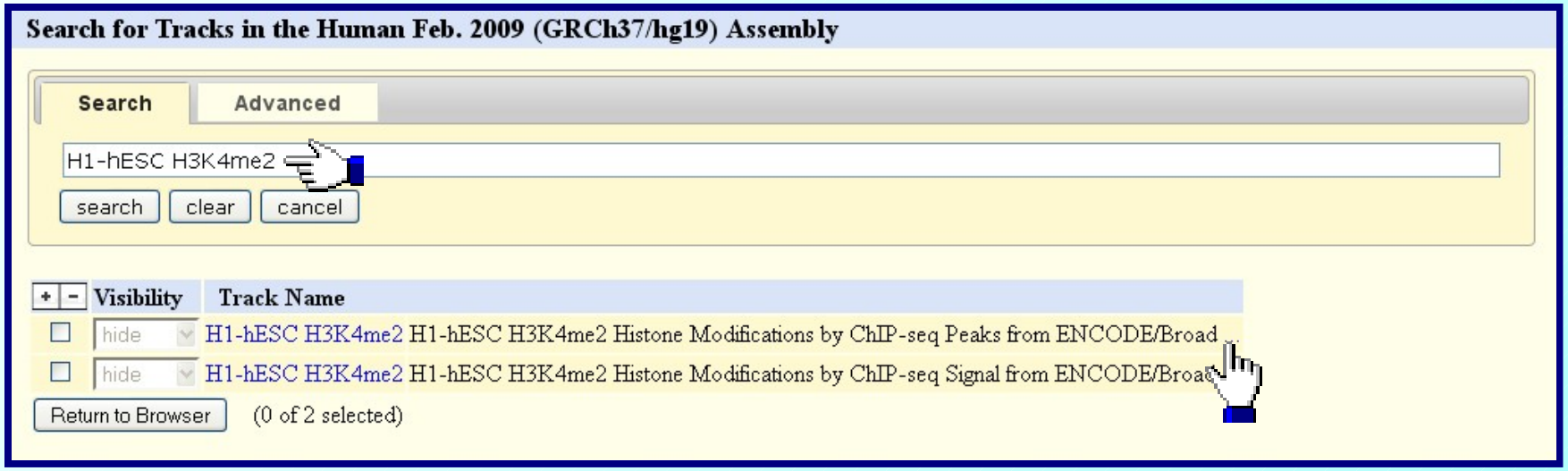
Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

Cell Line	Tier	Description	Lineage	Karyotype	Sex	Documents	Vendor ID	Term ID	*Label*
H1-hESC	1	Human Embryonic Stem Cells	embryonic stem cells	normal	M	protocol	Cellular Dynamics	CL:0000007	H1-hESC

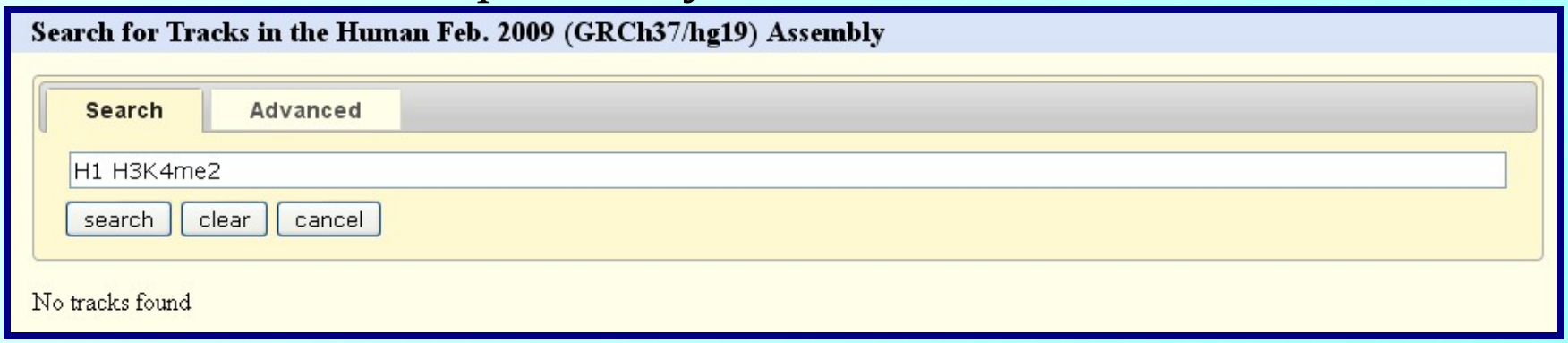
Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

Cell Line	Tier	Description	Lineage	Karyotype	Sex
8988T	3	human pancreas adenocarcinoma (PA-TU-8988T)	pancreatic adenocarcinoma	cancer	F
A549	3	epithelial cell line derived from a lung carcinoma tissue	"This line was initiated in 1972 by D.J. Giard, et al. through explant culture of lung carcinomatous tissue from a 58-year-old Caucasian male." - ATCC	cancer	M
AG04449	3	Fetal buttock/thigh fibroblast			M
AG04450	3	Fetal lung fibroblast			M
AG09309	3	Adult human toe fibroblast			F
AG09319	3	Adult human gum tissue fibroblasts			F
AG10803	3	Adult human abdominal skin fibroblasts			M
AoAF	3	Normal Human Aortic Adventitial Fibroblast Cells			F
AoSMC	3	aortic smooth muscle cells			U
Astrocy	3	Normal human astrocytes		normal	U

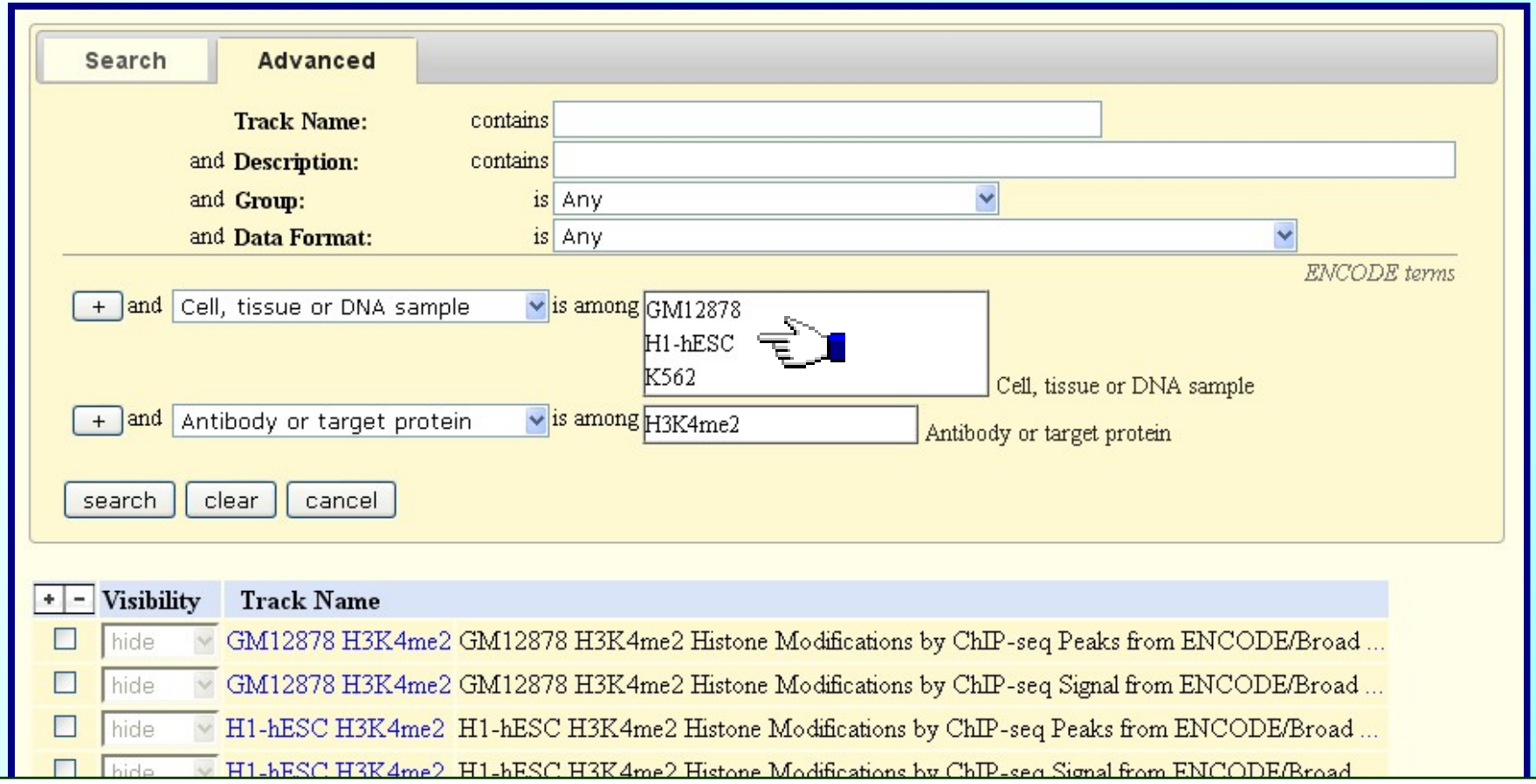
Track Search
“simple search”



- Searches for terms in the mdb
- Exact match is required so you have to know the term

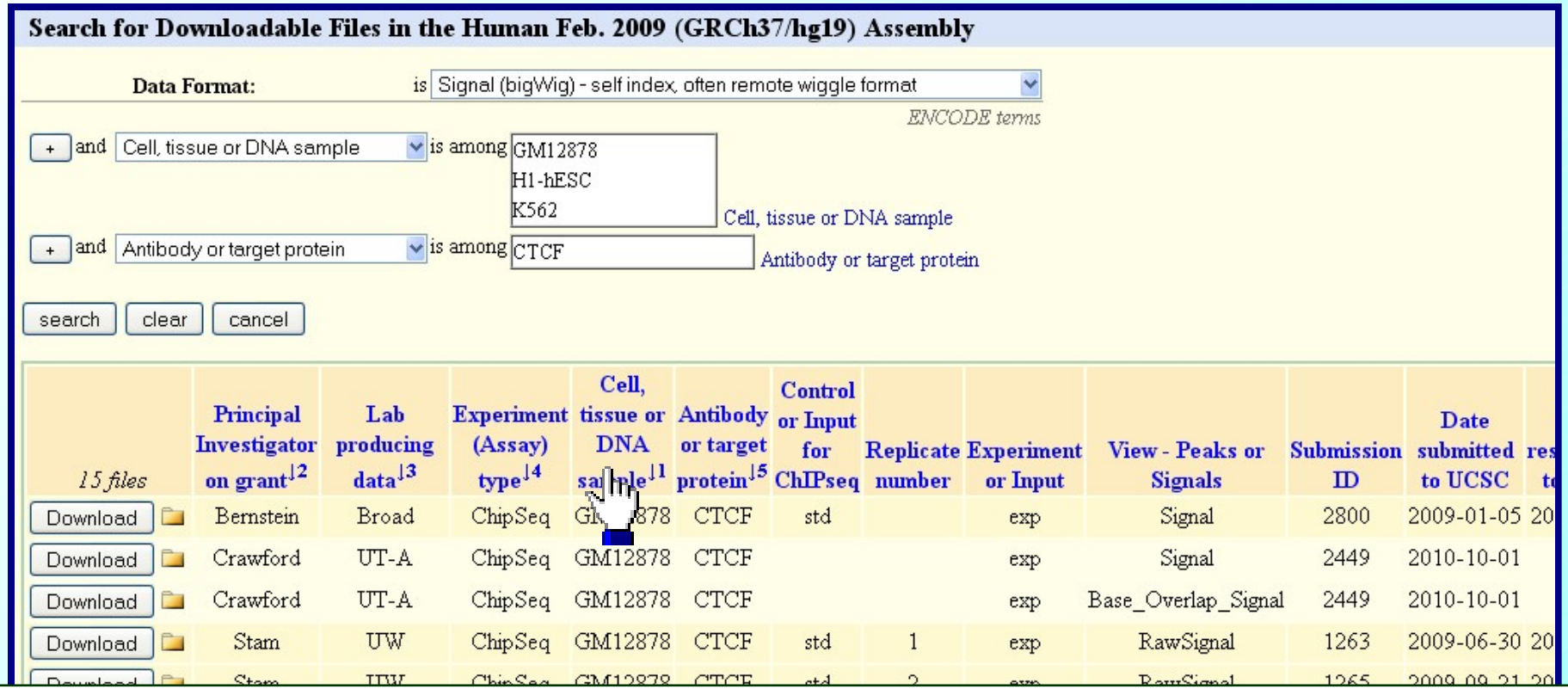


Track Search
"Advanced"



- Select exact terms. Select multiple terms.
- Titles and search methods are defined in cv.ra.

Downloadable Files Search



- Select multiple mdb terms to search.
- Result is list of downloadable files (currently only ENCODE composites).
- Table of results is sortable. Columns are discovered as common mdb terms. Titles are from cv.ra.

What?

http://www.subgenius.com/

Take it away Brian...

What is the metaDb Table?

MetaDb Attributes

- One active per assembly database
  - Sandbox versions (metaDb_braney)
  - Main version (metaDb)
- Attaches metadata to tables and files
  - Anything, usually experiment variables
    - Cell types, antibodies, replicate#, etc
- Consists of object with name/value pairs
- Currently used only for ENCODE

MetaDb Schema

```
mysql> desc metaDb;
+---------+----------------------+------+-----+---------+-------+
| Field   | Type                 | Null | Key | Default | Extra |
+---------+----------------------+------+-----+---------+-------+
| obj     | varchar(255)         | NO   | PRI | NULL    |       |
| var     | varchar(255)         | NO   | PRI | NULL    |       |
| varType | enum('txt','binary') | NO   |     | txt     |       |
| val     | longblob             | NO   |     | NULL    |       |
+---------+----------------------+------+-----+---------+-------+
4 rows in set (0.01 sec)
```

Sample Contents of metaDb

- 201474 rows in hg19

```
mysql> select * from metaDb limit 5;
+-------------------------------------------------+-------------+---------
| obj                                             | var         | varType | val
+-------------------------------------------------+-------------+---------
| wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal | cell        | txt | GM12878
| wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal | composite   | txt | wgEncodeAffyR
| wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal | dataType    | txt | RnaChip
| wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal | dataVersion | txt | ENCODE Feb..
| wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal | dateSubmit..| txt | 2009-03-10
```

Sample metaDb ra file

```
metaObject wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal
objType table
cell GM12878
composite wgEncodeAffyRnaChip
dataType RnaChip
dataVersion ENCODE Feb 2009 Freeze
dateSubmitted 2009-03-10
dateUnrestricted 2009-12-10
fileName wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal.broadPeak.gz
grant Gingeras
lab Affy
localization cell
origAssembly hg18
project wgEncode
rnaExtract total
subId 2121
tableName wgEncodeAffyRnaChipFiltTransfragsGm12878CellTotal
view FiltTransfrags

metaObject wgEncodeAffyRnaChipFiltTransfragsGm12878CytosolLongnonpolya
……
```

metaDb stats in hg19

- **201,474 rows**
- **10,018 distinct obj's**
- **Sample var's**
  - **10018 objType**
  - **10018 metaObject**
  - **10018 composite**
  - **10015 project**
  - **10015 lab**
  - **10015 grant**
  - **10015 fileName**
  - **10015 dataVersion**
  - **10015 dataType**
  - **10009 subId**
  - **10009 dateSubmitted**
  - **10008 view**
  - **10008 dateUnrestricted**
  - **10001 cell**
  - **8595 replicate**
  - **7267 origAssembly**
  - **6260 controlId**
  - **5942 antibody**

The Controlled Vocabulary

- Defines acceptable contents for fields in ENCODE submissions
- Used by track search
- Source exists in the a single file for all assemblies: cv.ra

Sample Contents of CV.ra

```
term K562
tag K562
type Cell Line
organism Human
descrip7on leukemia
7ssue Blood
vendorName ATCC
vendorId CCL-243
orderUrl
    h1p://www.atcc.org/ATCCAdvancedCatalogSearch/ProductDetails/tabid/452/Default.aspx?
ATCCNum=CCL243&Template=cellBiology
karyotype cancer
lineage "The continuous cell line K562 was established by Lozzio and Lozzio from the
    pleural effusion of a 53 year old
female with chronic myelogenous leukemia
in terminal blast crises." ATCC
termId BTO:0000664
termUrl h1p://www.ebi.ac.uk/ontology/lookup/browse.do?ontName=BTO&termId=BTO%3A
0000664
color 46,0,184
sex F
7er 1
protocol K562_protocol.pdf
```

cv.ra stats

- 1,305 terms
- Some terms
  - **234 Cell**
  - **199 Antibody**
  - **58 typeOfTerm**
  - **47 treatment**
  - **38 lab**
  - **31 dataType**
  - **19 grant**
  - **19 control**
  - **11 mapAlgorithm**
  - **10 seqPlatform**
  - **9 rnaExtract**
  - **9 localization**
  - **8 age**
  - **7 strain**
  - **7 species**
  - **7 readType**
  - **4 sex**

Meta-metadata: Type of Term

- If a term such as “H1-hESC” is defined as type “cell”, what do we know about all “cell” terms?
- Look in ***typeOfTerms*** in cv.ra.

```
term cellType
type typeOfTerm
label Cell, tissue or DNA sample
searchable multiSelect
cvDefined yes
validate cv or None # no cell is same as cell=None
priority 120
```

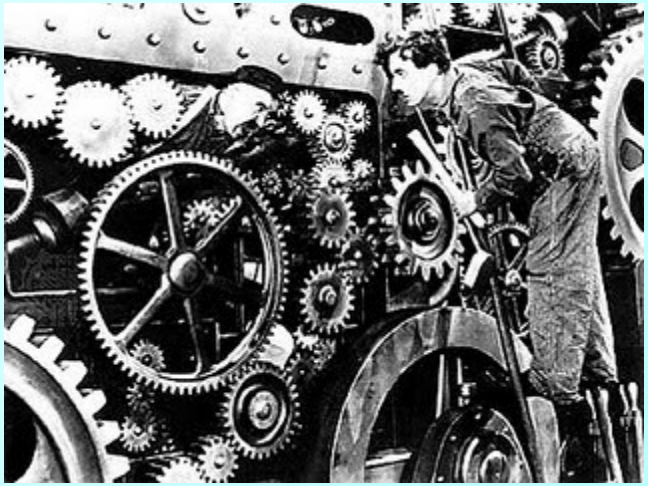
searchable (as seen in Track Search)

- ✓ select: choose one of several in drop-down box.
- ✓ multiSelect: choose multiple items in drop-down.
- ✓ freeText: typed in text will be bracketed with wildcards.
- ▪ date: (not yet implemented) by range.
- ▪ numeric: (not yet implemented) by range.

validate (as used by in mdbPrint)

- ✓ cv: defined in cv.ra.
- ✓ date: must be in YYYY-MM-DD format
- ✓ list: comma delimited list (yes,no,maybe)
- ✓ regex: regular expresion (e.g. **`^GS[M,E][0-9]*$`**)
- ✓ Also supported: none, exists, float, integer

How?



http://en.wikipedia.org/wiki/Modern_Times_(film)

Take it away Brian...

Structure of the MetaDb directories

- Source lives in the kent git repository
  - Parked in src/hg/makeDb/trackDb/org/assembly/metaDb/$release/*.ra
- Part of trackDb and friends
  - trackDb
  - metaDb
  - Trix files (index for track search built with cv.ra)
- metaDb RA files in Src/hg/makeDb/trackDb
  - Organism/assembly/metaDb/
    - Alpha, beta, public
      - » A bunch of RA files, now one per ENCODE composite

Current metaDb ra files in metaDb/alpha

makefile
wgEncodeAffyRnaChip.ra
wgEncodeBroadHistone.ra
wgEncodeBuOrchid.ra
wgEncodeCaltechRnaSeq.ra
wgEncodeCshlLongRnaSeq.ra
wgEncodeCshlShortRnaSeq.ra
wgEncodeDukeAffyExon.ra
wgEncodeGencode.ra
wgEncodeGisChiaPet.ra
wgEncodeGisDnaPet.ra
wgEncodeGisRnaPet.ra
wgEncodeGisRnaSeq.ra
wgEncodeHaibGenotype.ra
wgEncodeHaibMethyl27.ra
wgEncodeHaibMethylRrbs.ra
wgEncodeHaibMethylSeq.ra
wgEncodeHaibRnaSeq.ra
wgEncodeHaibTfbs.ra
wgEncodeMapability.ra
wgEncodeOpenChromChip.ra
wgEncodeOpenChromDnase.ra
wgEncodeOpenChromFaire.ra
wgEncodeOpenChromSynth.ra
wgEncodeRikenCage.ra
wgEncodeSunyAlbanyGeneSt.ra
wgEncodeSunyAlbanyTiling.ra
wgEncodeSunyRipSeq.ra
wgEncodeSunySwitchgear.ra
wgEncodeSydhHistone.ra
wgEncodeSydhNsome.ra
wgEncodeSydhTfbs.ra
wgEncodeUchicagoTfbs.ra
wgEncodeUmassDekker5C.ra
wgEncodeUncBsuProt.ra
wgEncodeUwAffyExonArray.ra
wgEncodeUwDgf.ra
wgEncodeUwDnase.ra
wgEncodeUwHistone.ra
wgEncodeUwTfbs.ra
wgEncodeYaleRnaSeq.ra

Three-state release mechanism

- Really four state!
- Sandbox version (pre-git push)
- Alpha version (on hgwdev)
  - built every morning by buildmeister/ on demand
  - On hgwdev
  - Developer managed
- Beta version (on hgwbeta)
  - QA staging and testing
  - QA managed
- Public version (on RR)
  - Public use!
  - QA managed

What happens when I do a make in trackDb?

- trackDb gets built
- metaDb is built
  - » Only if ra is newer
  - » Existing table dropped
  - » All ra files in associated release state directory read in
- If user (no argument)
  - » trackDb_user, metaDb_user, cv.ra copied to cgi-bin-user
- If alpha/beta/public
  - » trackDb, metaDb
  - » TRIX files are built
  - » Copies to cgi-bin

Tools

- mdbPrint

  -obj=, -composite=, -vars=, wildcards

  -count, -validate, -encodeExp

  -table

```
mdbPrint hg19 –vars=“cell=H1% antibody=H3K4me2”
```

- mdbUpdate

  -obj=, -composite=, -vars=, wildcards

  -var= -val=, -setVars=, -delete

  {fileName}

  -table, -recreate

  -test

```
mdbUpdate hg19 –obj=knownGene –setVars
  =“objType=table grant=Kent lab=UCSC
  strain=tooMuch sex=noneOfYourBusiness
  coolnessFactor=1000” -test
```

Tools (continued)

- hgEncodeVocab CGI

http://hgwdev.cse.ucsc.edu/cgi-bin/hgEncodeVocab?type=%22typeOfTerm%22

Term types defined: Types of terms used frequently in controlled vocabulary or metadata should be defined here.

TypeOfTerm	Description
accession	A generic GEO accession number provided by the producing lab.
age	The age of the organism used to produce tissue or cell line.
annotation	GENCODE specifies if an annotation is done manually or automatically.
Antibody	The antibody to a specific protein. Used in immuno-precipitation to target certain fractions of biological interest.
bioRep	Cross Transcriptome sample ID number.
cellType	Cell line or tissue used as the source of experimental material.
cellType *(for mouse)*	Cell line or tissue used as the source of experimental material. *(for mouse)*
composite	Related tracks in the UCSC Genome Browser are often grouped into a named composite track.
control	The type of control (or 'input') used in ChIP-seq experiments to remove background noise before peak calling.
controlId	This ID is used to explicitly tie a ChIP-seq experiment with the control/input that was used in peak calling. The specification may be a labExpId or a UCSC docId c
dataType	The types of experiments such as ChIP-seq, DNAse-seq and RNA-seq.
dataVersion	The ENCODE project declares specific data freezes for data to be used in papers or analysis.
dateResubmitted	Submitted data that was remapped to a new assembly, found to have errors or otherwise needed to be updated will have a date of resubmission.
dateSubmitted	Date that a particular file was originally submitted to the UCSC Genome Browser.
dateUnrestricted	ENCODE data is made publicly available but with restrictions on use for the first nine months since date submitted. After this date, the data is unrestricted.
dccInternalNotes	Notes about tracks that are internal to the DCC.
fileIndex	The name of the index file (.bai) that is associated with a particular bam file.
fileName	The name of a downloadable file associated with a particular track in the browser.
fragLength	DNA libraries built for ChIP-seq and similar experiments often involve fragmenting the DNA into lengths close to this size.
fragSize	length of GIS DNA PET fragments, which has different values than fragLength
freezeDate	Date when GENCODE froze data in order to submit to UCSC.
geoSample	GEO sample accession number applied to a single data set in a series of related data sets.

API

```
struct mdbObj
// The standard container of a single object's metadata.
    {
    struct mdbObj* next;     // Next in sl list of objects
    char *obj;               // Object name or ID
    struct mdbVar* vars;     // list of variables
    struct hash* varHash;    // may be NULL
    };

struct mdbVar
// The metadata var=val construct. Contained by mdbObj
    {
    struct mdbVar* next;     // Next in sl list of variables
    char *var;               // Metadata variable name.
    enum mdbVarType varType; // txt | binary
    char *val;               // Metadata value.
    };

struct mdbByVar
// When searching metadata by var=val pairs
```

API (continued)

```
struct mdbByVar *mdbByVarsLineParse(char *line);
// Parses a line of "var1=val1 var2=val2 into a mdbByVar object */

struct mdbObj *mdbObjsQueryByVars(*conn,char *table,*mdbByVars);
// Query the metadata table by one or more var=val pairs to find the
// distinct set of objs that satisfy ALL conditions.

struct mdbObj *mdbObjRepeatedSearch(*conn,slPair *varValPairs,...);
// Search the metaDb table for objs by var,val pairs.
// val may be comma delimited list for “is among”

void mdbObjsSortOnVars(struct mdbObj **pMdbObjs, char *vars);
// Sorts on comma delimited vars lists: fwd case-sensitive.

struct mdbObj *mdbObjsFilterByVars(struct mdbObj **pMdbObjs,char *vars,
                    boolean noneEqualsNotFound,boolean returnMatches);
// Filters mdb objects to only those that include/exclude var=val pairs
// Supports != ("var!=" means var not found).
```

API (continued)

```
struct mdbObj *mdbObjsCommonVars(struct mdbObj *mdbObjs);
// Returns a new mdbObj with all vars that are contained in every obj

int mdbObjsValidate(struct mdbObj *mdbObjs, boolean full);
// Validates vars and vals against cv.ra.

const struct mdbObj *metadataForTable(*db,*tdb,char *table);
// Returns the metadata for a table.  NEVER FREE THIS STRUCT!
// This is the main routine for CGIs to access metadata

struct hash *mdbCvTermHash(char *term);
// returns a hash of hashes of a term which should be defined in cv.ra

int mdbObjsSetToDb(*conn,*tableName,struct mdbObj *mdbObjs,
                                    boolean replace,boolean testOnly);
// Adds or updates metadata obj/var pairs into the named table.

void mdbObjsFree(struct mdbObj **mdbObjsPtr);
// Frees one or more metadata objects.
```

Thanks

- Kate Rosenbloom, Jim Kent, Cricket Sloan, Venkat Malladi, Melissa Cline, Katrina Learned, Venessa Kirkup, Brooke Rhead, Galt Barber, Larry Meyers, Krishna Roskin
- Special thanks to:
  - Melissa, the cv maven!
  - Cricket and Venkat for consistently pushing in the right direction.
  - Katrina, Venessa, Brooke and all of QA for expecting things to actually work.