FILE REORGANIZATION AND OUR DATABASE

BY BELINDA GIARDINE

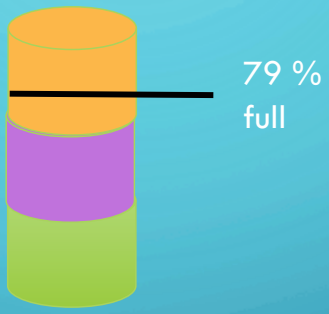
REASONS FOR FILE REORGANIZATION

- Organize the production (i.e. semi-final) data so that it can be found by someone other than the person who generated it.
- Divide the data into smaller chunks (volumes) to make backups and rearranging on physical disks easier.

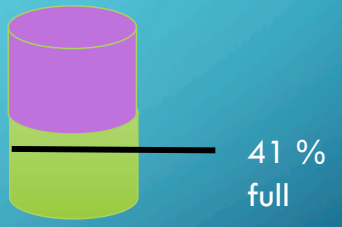
HARDWARE

Badger
Mal
Herbie
Desktops

fs8, fs9, fs10

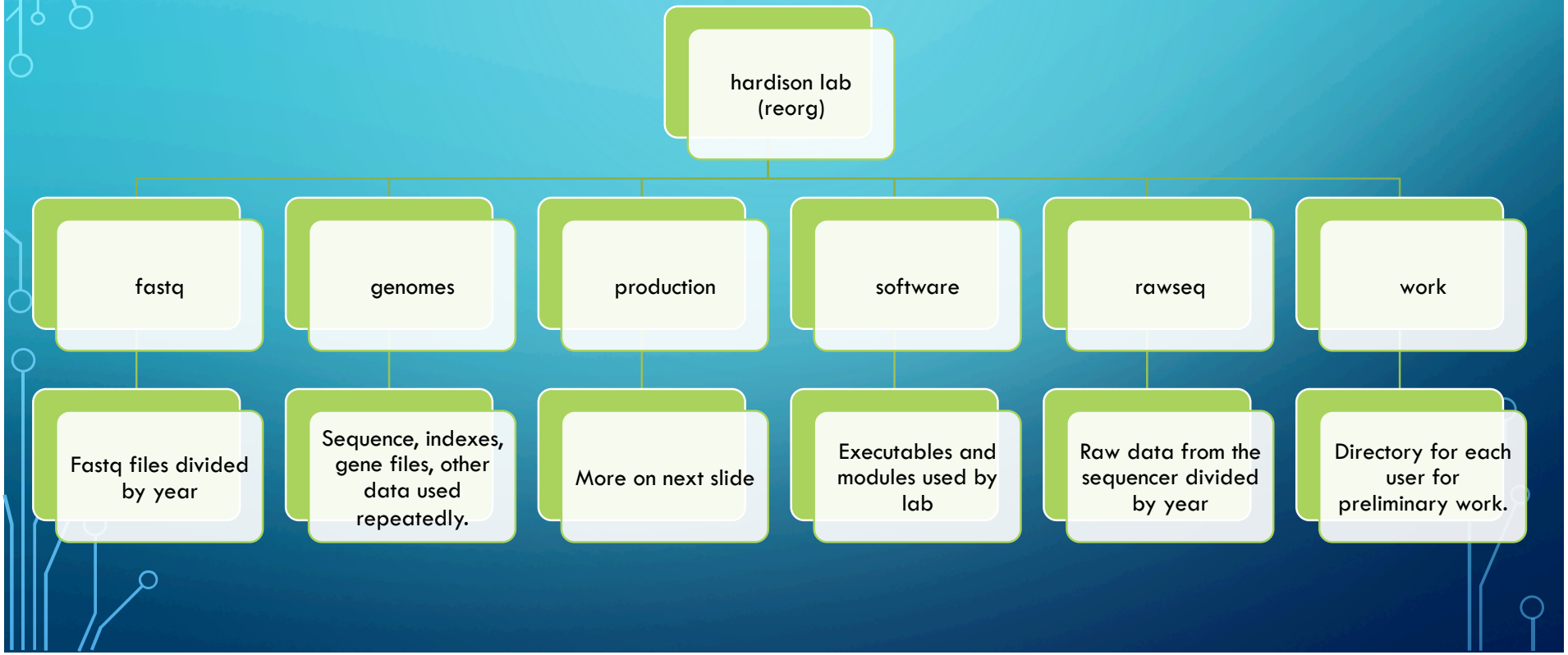


Tank
shared disk array
includes new
(~37T)

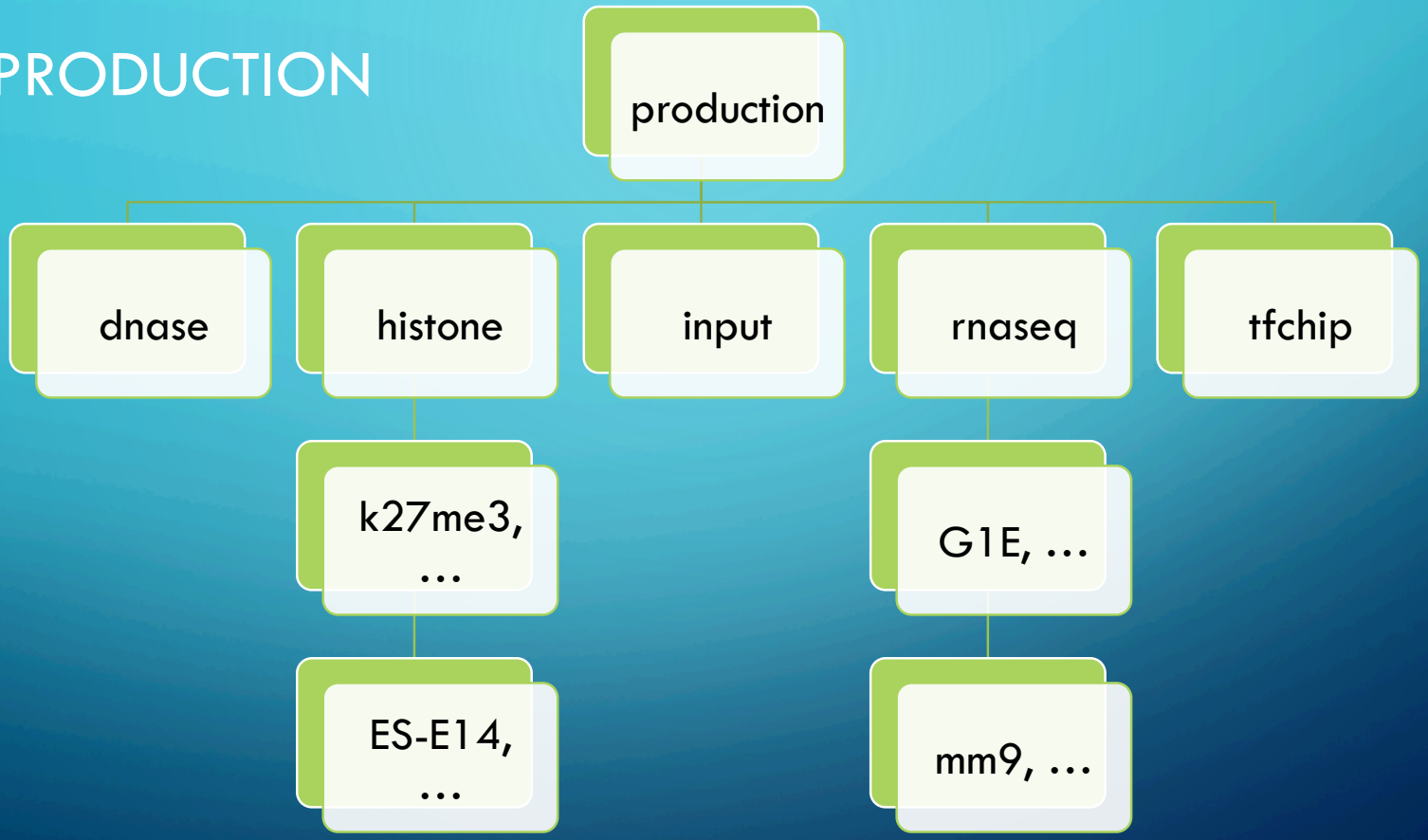


Jbod
Ross's disk array
includes newer
(~34T)

TOP LEVELS OF NEW DIRECTORY STRUCTURE



PRODUCTION


production
dnase
histone
input
rnaseq
tfchip
k27me3, …
ES-E14, …
G1E, …
mm9, …

QUESTIONS?

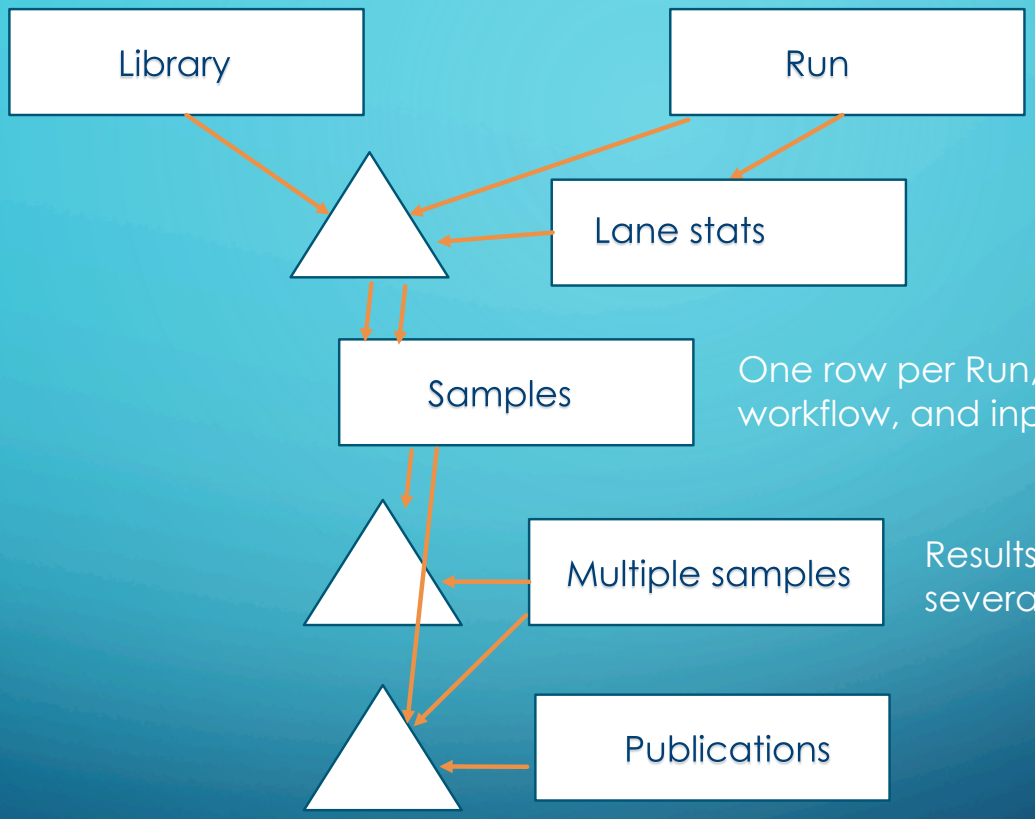
HLAB DATABASE

- Organize the metadata
  - Keep track of what data came from which run and library
  - What workflow was used and when
  - Where the results are in the file system
  - Where the results are displayed (browser)
- Query interface to find data

WHY WE NEED A DATABASE
QUESTIONS THAT HAVE COME UP RECENTLY

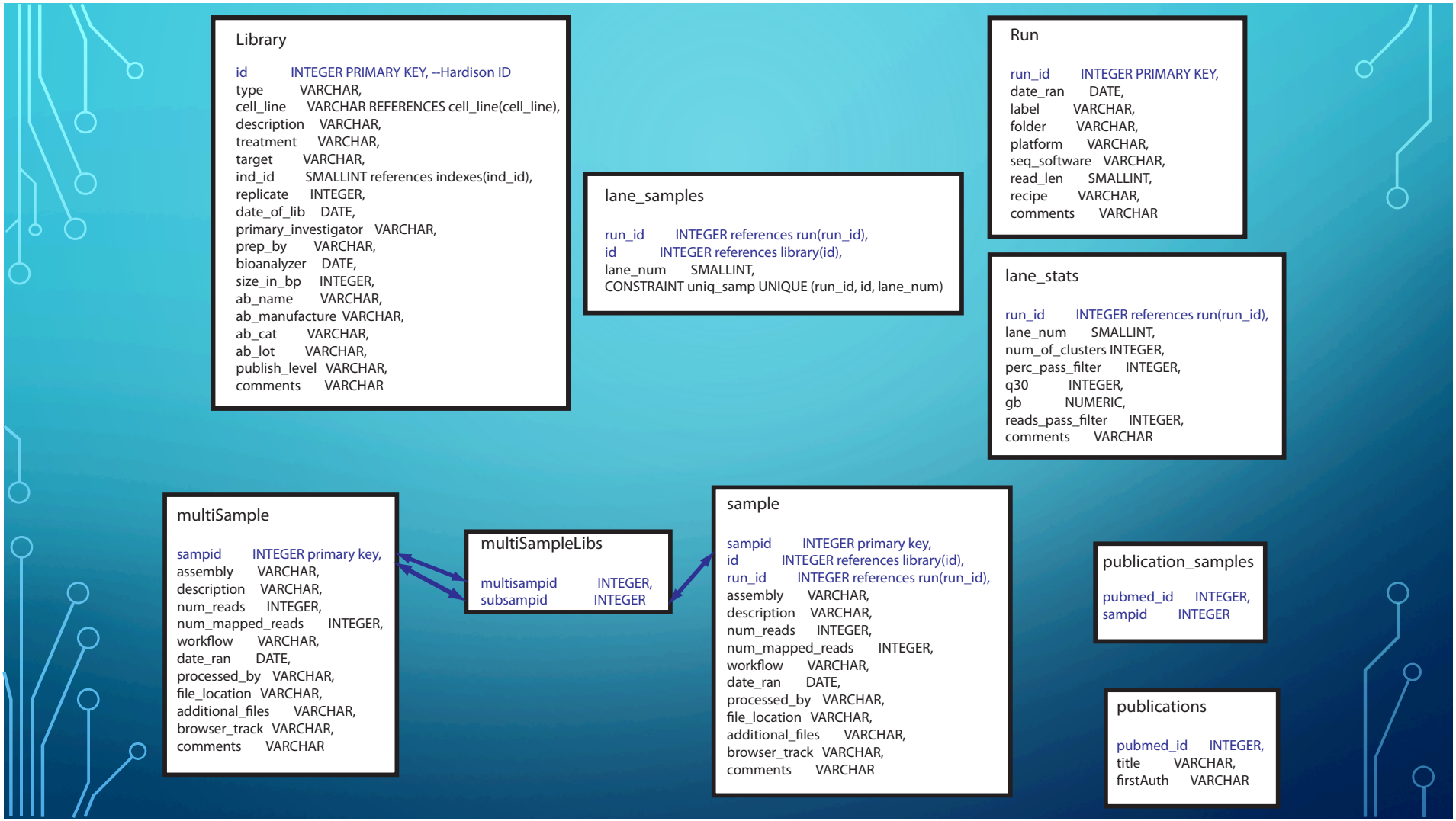
- Do you know which peaks were used in Swathi's paper? Pilon et. al, 2011
- Are there FPKMs for the Bodine RNA-seq?
- Where is the H3K4me3 bigwig in G1E?

TABLES



One row per Run, library, build, workflow, and inputs used.

Results from workflows that use several of the above samples


Library
id INTEGER PRIMARY KEY, --Hardison ID
type VARCHAR,
cell_line VARCHAR REFERENCES cell_line(cell_line),
description VARCHAR,
treatment VARCHAR,
target VARCHAR,
ind_id SMALLINT references indexes(ind_id),
replicate INTEGER,
date_of_lib DATE,
primary_investigator VARCHAR,
prep_by VARCHAR,
bioanalyzer DATE,
size_in_bp INTEGER,
ab_name VARCHAR,
ab_manufacture VARCHAR,
ab_cat VARCHAR,
ab_lot VARCHAR,
publish_level VARCHAR,
comments VARCHAR
lane_samples
run_id INTEGER references run(run_id),
id INTEGER references library(id),
lane_num SMALLINT,
CONSTRAINT uniq_samp UNIQUE (run_id, id, lane_num)
Run
run_id INTEGER PRIMARY KEY,
date_ran DATE,
label VARCHAR,
folder VARCHAR,
platform VARCHAR,
seq_software VARCHAR,
read_len SMALLINT,
recipe VARCHAR,
comments VARCHAR
lane_stats
run_id INTEGER references run(run_id),
lane_num SMALLINT,
num_of_clusters INTEGER,
perc_pass_filter INTEGER,
q30 INTEGER,
gb NUMERIC,
reads_pass_filter INTEGER,
comments VARCHAR
multiSample
sampid INTEGER primary key,
assembly VARCHAR,
description VARCHAR,
num_reads INTEGER,
num_mapped_reads INTEGER,
workflow VARCHAR,
date_ran DATE,
processed_by VARCHAR,
file_location VARCHAR,
additional_files VARCHAR,
browser_track VARCHAR,
comments VARCHAR
multiSampleLibs
multisampid INTEGER,
subsampid INTEGER
sample
sampid INTEGER primary key,
id INTEGER references library(id),
run_id INTEGER references run(run_id),
assembly VARCHAR,
description VARCHAR,
num_reads INTEGER,
num_mapped_reads INTEGER,
workflow VARCHAR,
date_ran DATE,
processed_by VARCHAR,
file_location VARCHAR,
additional_files VARCHAR,
browser_track VARCHAR,
comments VARCHAR
publication_samples
pubmed_id INTEGER,
sampid INTEGER
publications
pubmed_id INTEGER,
title VARCHAR,
firstAuth VARCHAR

QUERY PAGE

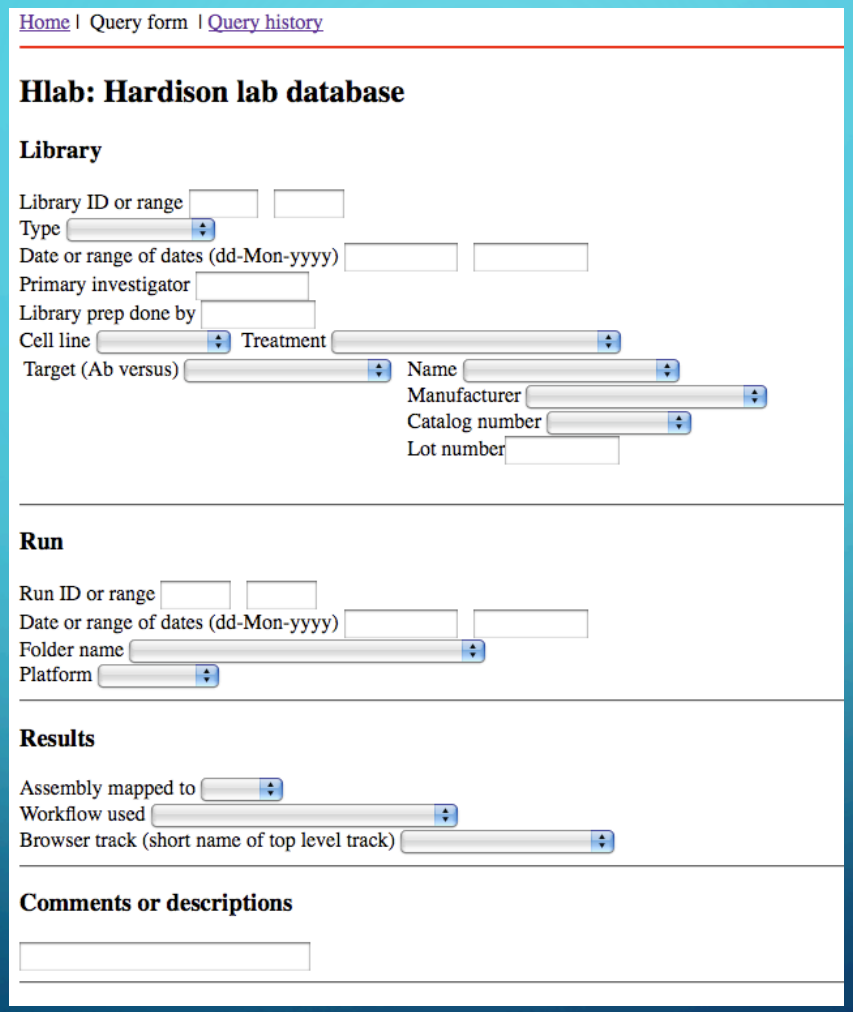
- Still under construction
  - globin.bx.psu.edu/hlab

QUESTIONS?
ANY SUGGESTIONS?

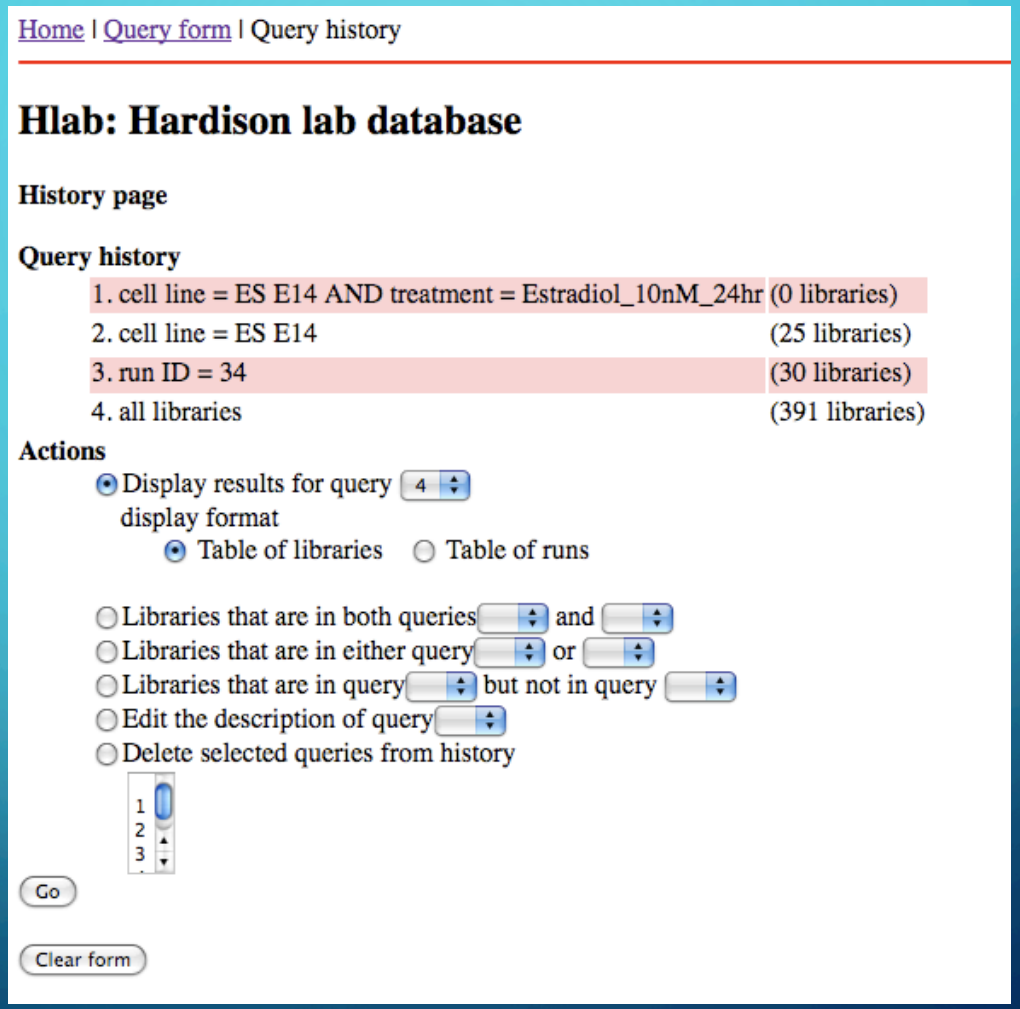
This presentation and other documentation can be found on the wiki. wiki.bx.psu.edu/HLab:Main

Thanks to Cheryl and Maria for the help with data entry, Cathy with help with the files, and the rest of the lab for help in finding the data.

QUERY FORM


Home | Query form | Query history
Hlab: Hardison lab database
Library
Library ID or range
Type
Date or range of dates (dd-Mon-yyyy)
Primary investigator
Library prep done by
Cell line
Treatment
Target (Ab versus)
Name
Manufacturer
Catalog number
Lot number
Run
Run ID or range
Date or range of dates (dd-Mon-yyyy)
Folder name
Platform
Results
Assembly mapped to
Workflow used
Browser track (short name of top level track)
Comments or descriptions

HISTORY PAGE



Home | Query form | Query history

Hlab: Hardison lab database

Library table

0	input	CH12	not used	none				07/15/2013	Hardison	Belinda	07/15/2013			
1	ChIP	G1E-ER4+E2	GATA1 G1E-ER4+E2	diffProtD_24hr	GATA1			05/12/2009	Hardison	Yong		206	GATA1	Santa Cru
2	input	G1E-ER4+E2		none	N/A			06/08/2009	Hardison	Yong		231		
3	ChIP	G1E-ER4+E2		Estradiol_10nM	H3K27me3			06/08/2009	Hardison	Weisheng		400	H3K27me3	Millipore
4	ChIP	G1E-ER4+E2		Estradiol_10nM	H3K27me3			06/08/2009	Hardison	Weisheng		219	H3K27me3	Millipore

Hlab: Hardison lab database

Library detail page

Library 289 H3K4me1 rep2
Type: ChIP
Cell: ES E14 Treatment: none
Target: H3K4me1
Replicate: 2
Index: AR012 CTTGTA(A)
Primary investigator: Hardison, library prep: Cheryl
Date: 10/29/2012
Bioanalyzer date: 12/03/2012
Size (bps): 358
Antibody Name: H3K4me1, Manufacturer: Abcam, Catalog#: ab8895, Lot#: 741326
Publish level: UCSC

Run
33 lane 6, 7

Processed data

Sample ID	Run	Assembly	Number of Reads	Mapped Reads	Workflow	Date	Processed by	Files	Additional files	Track	Comments
48:	33	mm9	53,670,532	52,039,566	histone workflow on biostar	02/01/2013	Belinda	/afs/bx.psu.edu/depot /data/hardison_lab/reorg /production/histone /H3K4me1/ESE14/mm9/		PSUreps Histone	macs for wiggles, sicer for peaks Experimented with different gap sizes in sicer (200 vs 600)
62: H3K4me1 ES E14 pool		mm9		98,040,643	histone workflow for pools on biostar	02/25/2013	Belinda	/afs/bx.psu.edu/depot /data/hardison_lab/new/work /histones/H3K4me1 /ESE14/mm9/		PSUreps Histone	samples used: 49, 48 gap size 200
71: ES-E14 chromHMM		mm9			chromHMM	03/12/2013	Belinda	/afs/bx.psu.edu/depot /data/hardison_lab/new/work /histones/chromHMM /ESE14/			samples used: 60, 59, 53, 52, 49, 48, 47, 46, 45, 44, 36, 35 Ran for 9 to 15 states