PennState

Reproducibility Begins at the Bench

Cheryl A. Keller, PhD

Associate Research Professor

Dept of Biochemistry and Molecular Biology

@KellerCaponePhD

PSU Bootcamp on Reproducible Research

Penn State University
June 3, 2019

Reproducibility Begins at the Bench

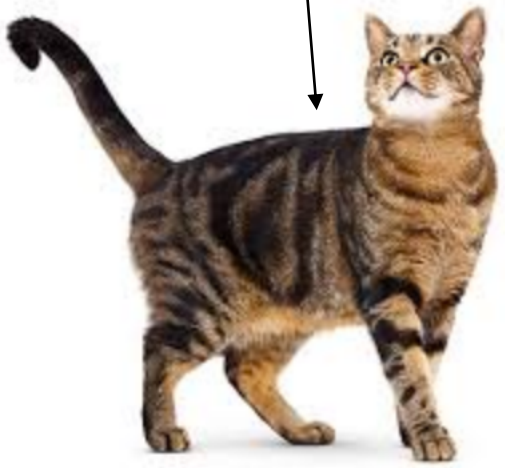
- Metadata
  - What is metadata? 🤔
  - Purposes of metadata
  - ENCODE data organization
  - Metadata collection
  - Hardison lab database
- Lessons from the bench
  - RNA-seq
  - ChIP-seq

What is metadata?

Metadata is constructed information, which means that it is of human invention and not found in nature.

Metadata is developed by people for a purpose or a function.

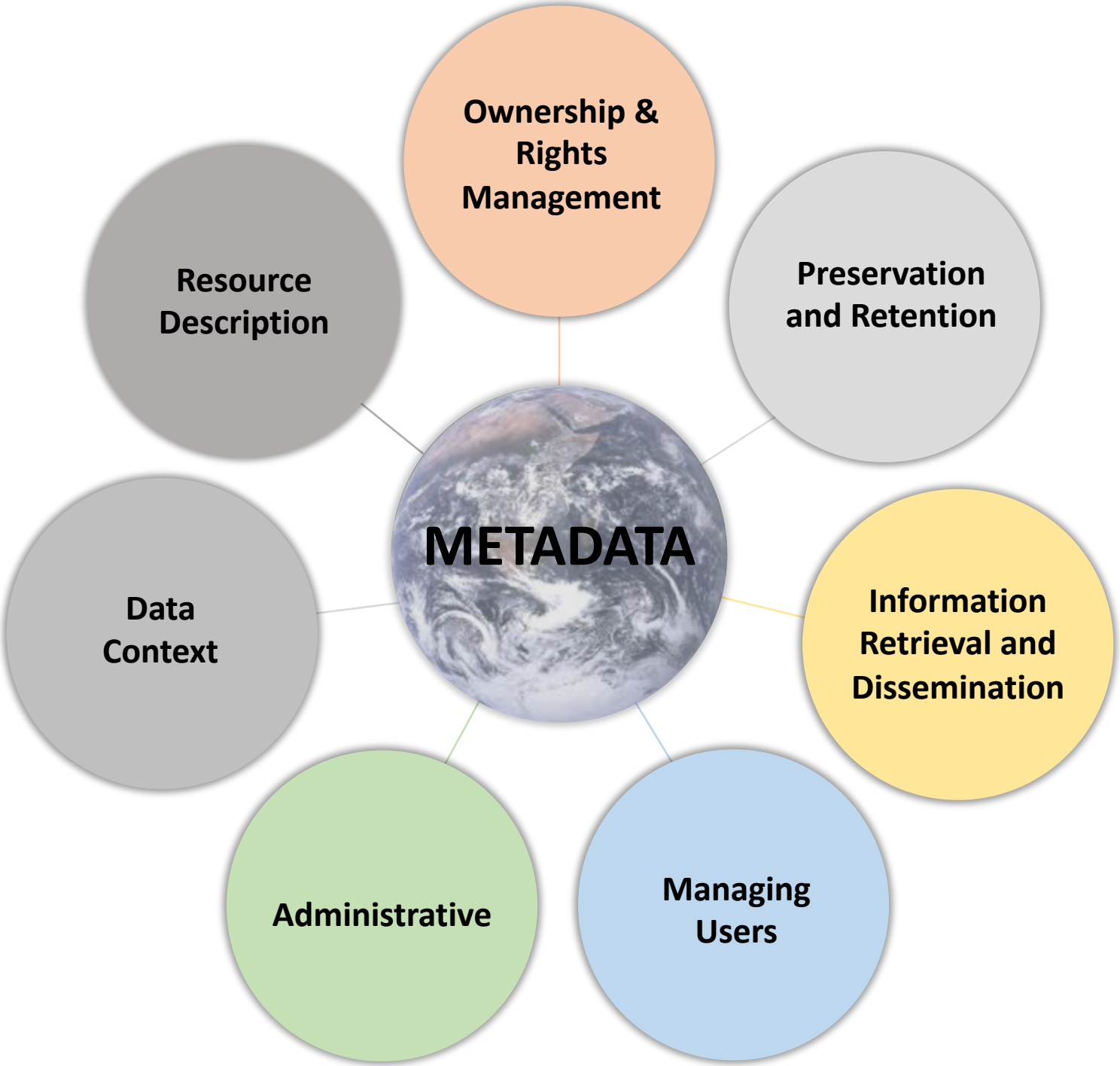
Data



Metadata

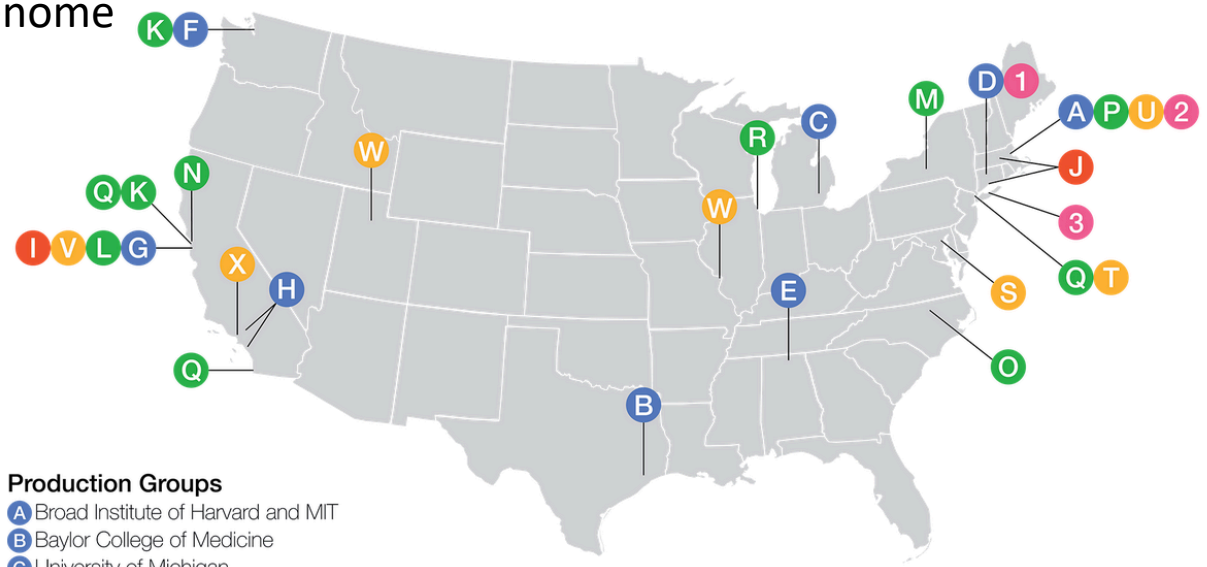
File name: Ali Cat
Genus: *Felis*
Species: *catus*
Date: 06 May 2015
File size: 3.6 kg
Owner: Crazy Cat Lady
Site: Trout Rd

Etc….


Ownership & Rights Management
Resource Description
Preservation and Retention
METADATA
Data Context
Information Retrieval and Dissemination
Administrative
Managing Users

ENCODE: Encyclopedia of DNA Elements

- Ongoing international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI
- The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome



Production Groups
- A Broad Institute of Harvard and MIT
- B Baylor College of Medicine
- C University of Michigan
- D The Jackson Laboratory
- E HudsonAlpha Institute for Biotechnology, University of Alabama in Huntsville
- F Altius Institute for Biomedical Sciences
- G Stanford University
- H California Institute of Technology, University of California, Irvine

Data Coordination Center
- I Stanford University

Data Analysis Center
- J University of Massachusetts Medical School; Yale University

Characterization Centers
- K University of California, San Francisco; University of Washington
- L Stanford University
- M Cornell University
- N Lawrence Berkeley National Laboratory
- O Duke University
- P Broad Institute of Harvard and MIT
- Q University of California, San Francisco; University of California, San Diego; Ludwig Institute for Cancer Research
- R University of Chicago

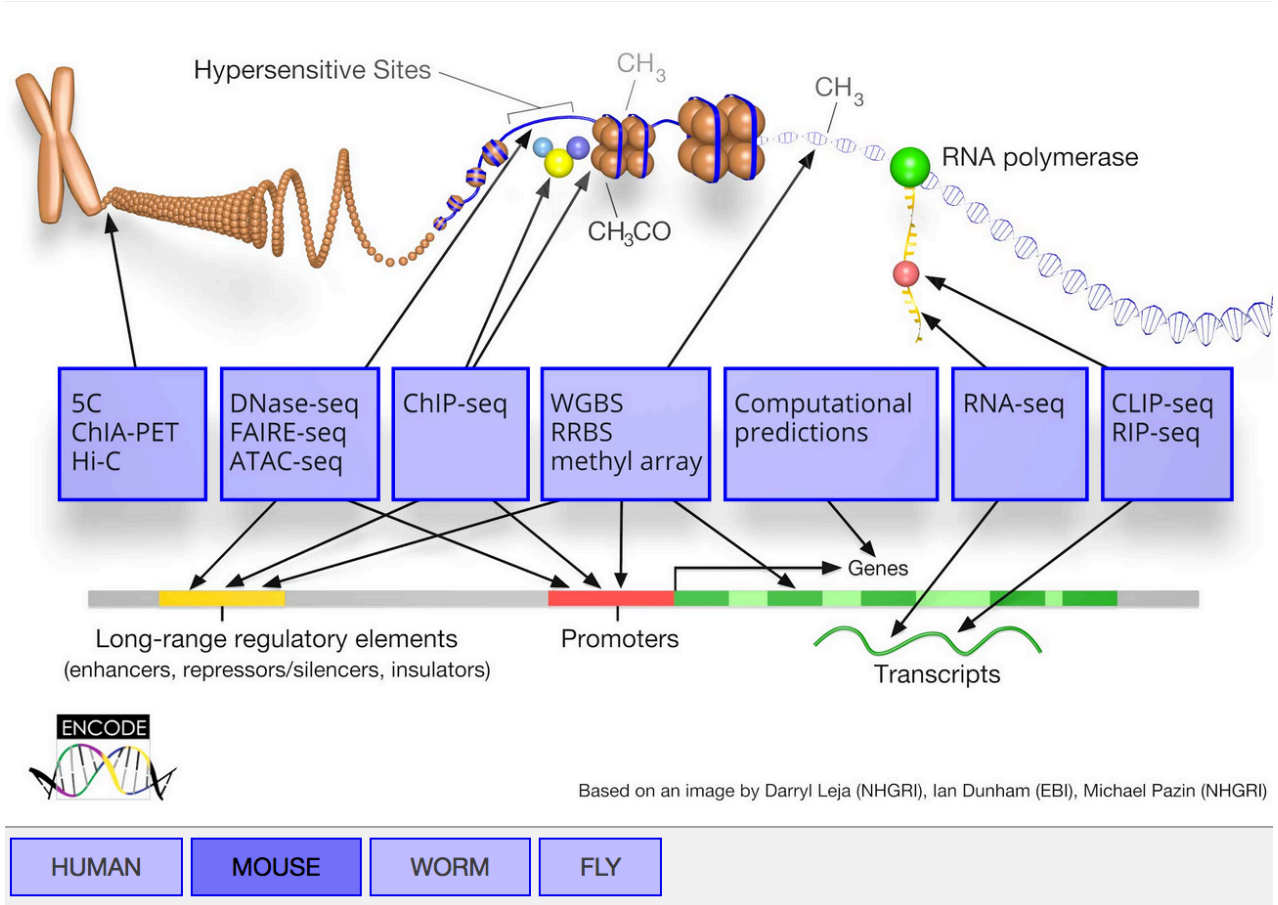
Computational Analysis Groups
- S Johns Hopkins University
- T Memorial Sloan Kettering Cancer Center
- U Harvard University; Brigham and Women's Hospital
- V Stanford University
- W Washington University; University of Utah
- X University of California, Los Angeles

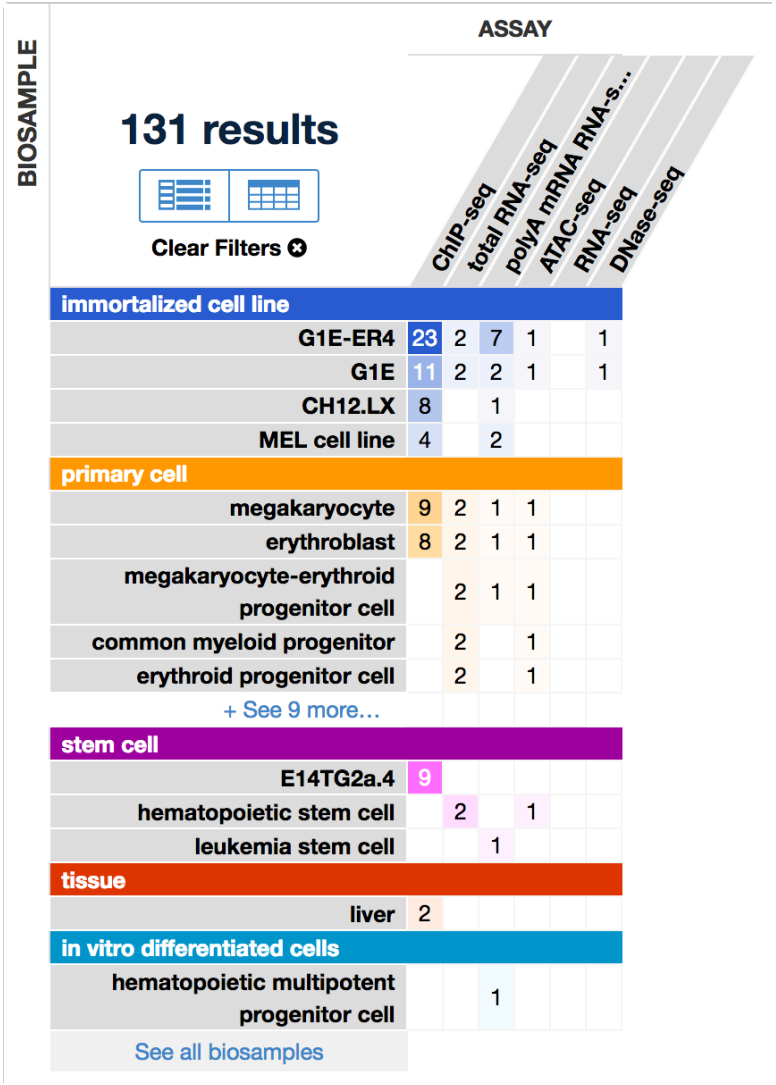
Affiliated Groups
- 1 University of Connecticut Health Center
- 2 Dana-Farber Cancer Institute
- 3 Cold Spring Harbor Laboratory

Map of participating investigators and affiliates for ENCODE 4.

Source: https://www.encodeproject.org/

ENCODE: Encyclopedia of DNA Elements

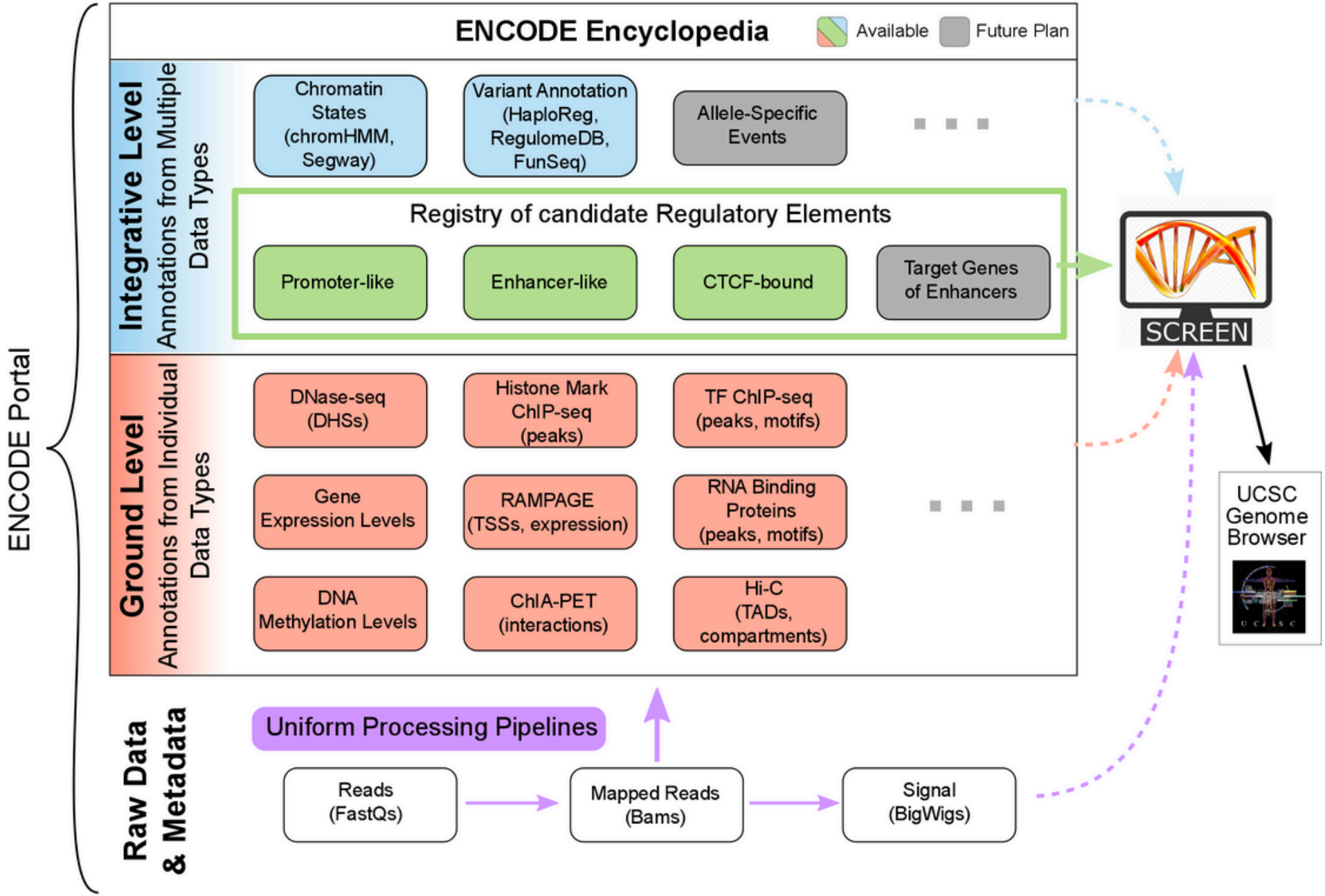




HAIB Production Group

Rick Myers (HudsonAlpha)
Ross Hardison (PSU)
Barbara Wold (Caltech)
Ali Mortazavi (UC Irvine)
Tim Reddy (Duke)

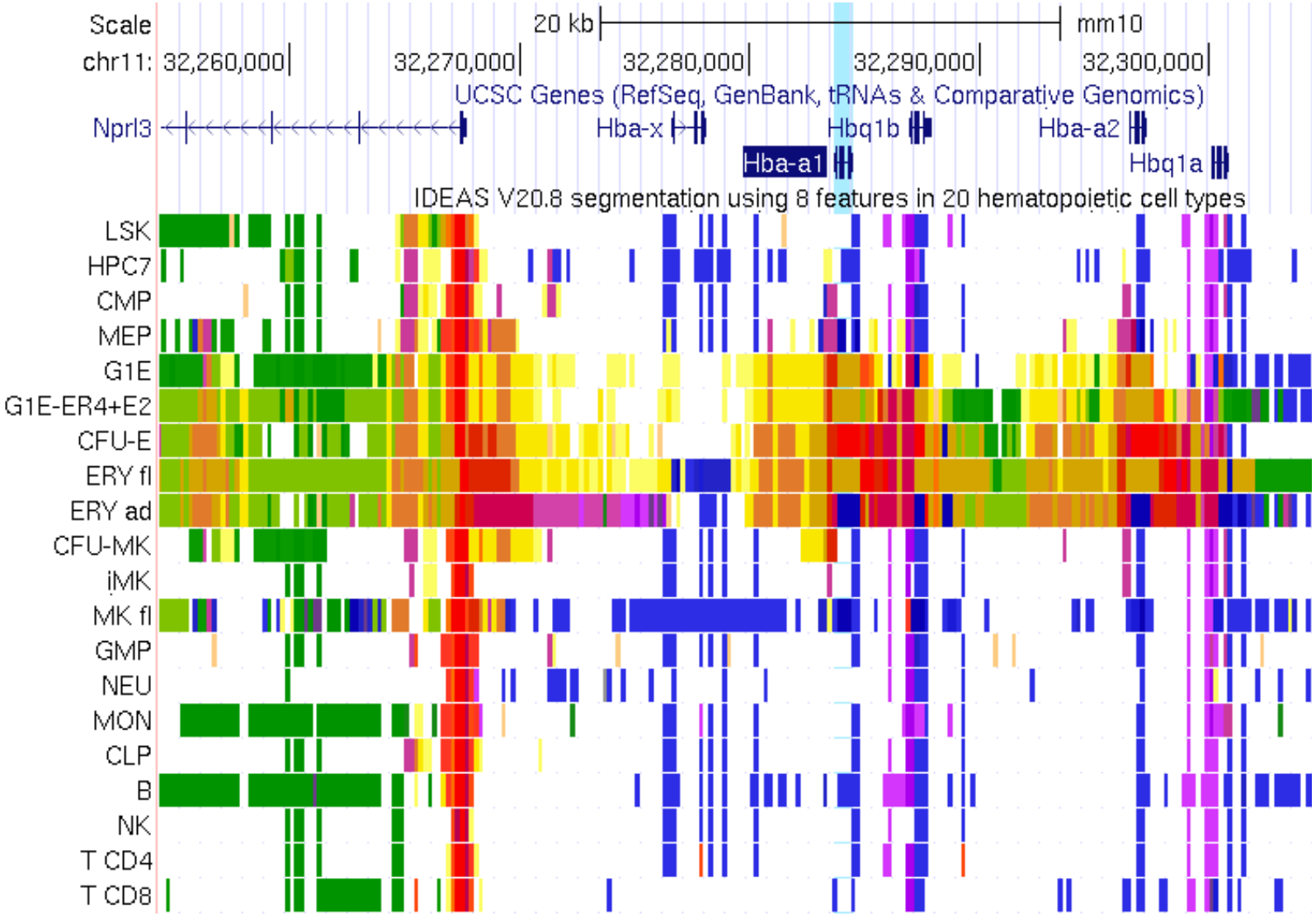
Source: https://www.encodeproject.org/



https://www.encodeproject.org/data/annotations/

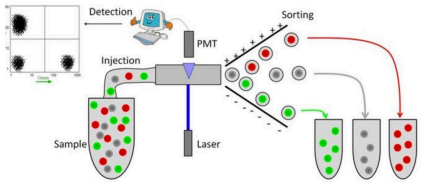
If your foundation is not solid….

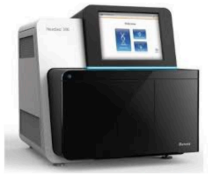
From mouse to machine (learning!)

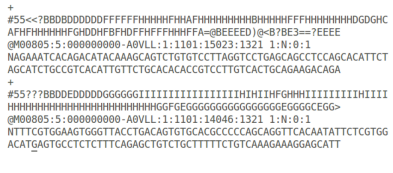


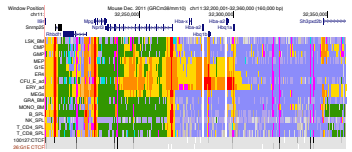
Yu Zhang, Lin An, Guanjue Xiang

Metadata must be collected at every step









Step		Example
Biosample	→	Strain, tissue
Sample processing	→	Antibodies
Library generation	→	Methods
Sequencing	→	Platform
Raw data	→	Q scores
Processed data	→	Pipeline

Hardison lab uses an internal database to store metadata

Hlab: Hardison lab database

Library detail page

unique identifier

Library 734 CTCF, G1E-ER4+E2, 20M,7 cycles
Type: ChIP
Cell: G1E-ER4+E2 Species: mouse, ID: 7079, Source: Mitch Weiss lab: weissmi@email.chop.edu
Starting amount: 20 M cells
Treatment: Estradiol_10nM_24hr
Target: CTCF
Index: AR019 GTGAAAC
Primary investigator: Hardison, library prep: Maria
Date: 11/24/2014
Number of cycles of PCR: 16
Bioanalyzer date: 11/24/2014
Fragmentation date: 11/11/2014
Size (bps): 336(205-502)
Antibody Name: CTCF, Manufacturer: Millipore, Catalog#: 07-729, Lot#: 1962117, ENCODE#:
Access status: none
Level of analysis: mapped

Hardison lab database

Run
49 09-Dec-2014 lane 6, 7, 8

Processed data

Product ID	Run	Assembly	Number of Reads	Mapped Reads	Filtered Reads	Workflow	Date	Processed by	Files	Additional files	Control, product ID	Track	ENCODE ID
807:	49	mm9	27,158,713	25,235,378		tfWorkflow on biostar	12/12/2014	Belinda	hardison_lab/reorg /production/tfchip /CTCF/ER4 /mm9/734/	er4_pooled_input.bam, blacklist.bed	418	Antibody tests	

Quality metrics (description)

Product ID	Percent GC	Total duplicate percentage	Percent of seqs remaining if deduplicated	Complexity	Percent mapped	NSC	RSC	FRIP | FRIT	Percent rRNA	Number of expressed genes	Number of reads mapping to spike-ins	Strand specificity	Spearman corr
807	41	25		0.95	93	1.06	2.07 QTag 2	0.024					

Reports
807 FastQC report
807 Cross-correlation

Publications
None found

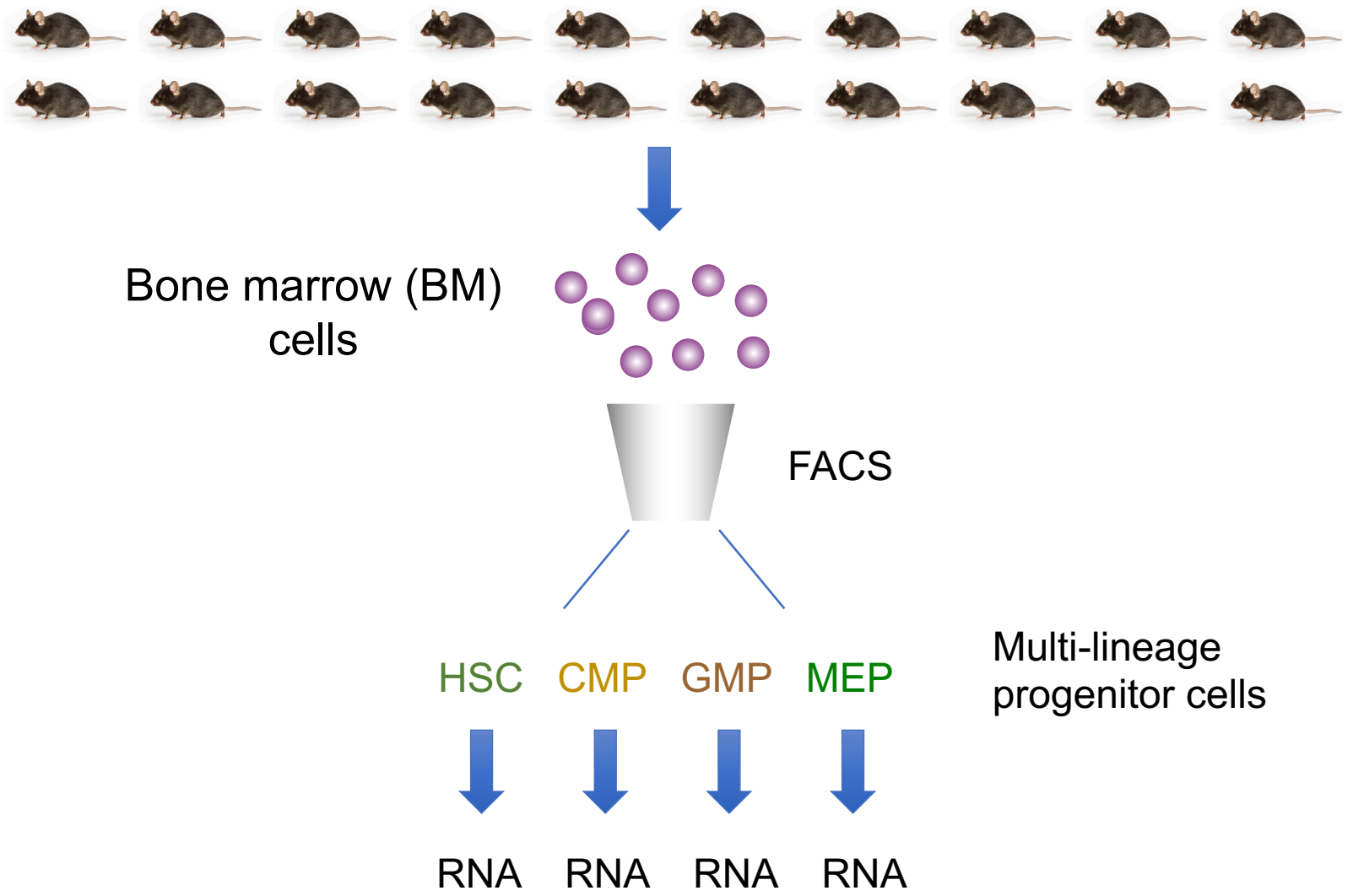
Reproducibility Begins at the Bench

- Metadata
  - What is metadata? 🤔
  - Purposes of metadata
  - ENCODE data organization
  - Hardison lab database

- Lessons from the bench
  - RNA-seq
  - ChIP-seq

Simplified scheme of hematopoiesis

Biosamples for RNA-seq hematopioetic progenitors



Biosamples for RNA-seq hematopioetic progenitors

= 20 biosamples

species
strain
gender
age
source

Bone marrow (BM) cells

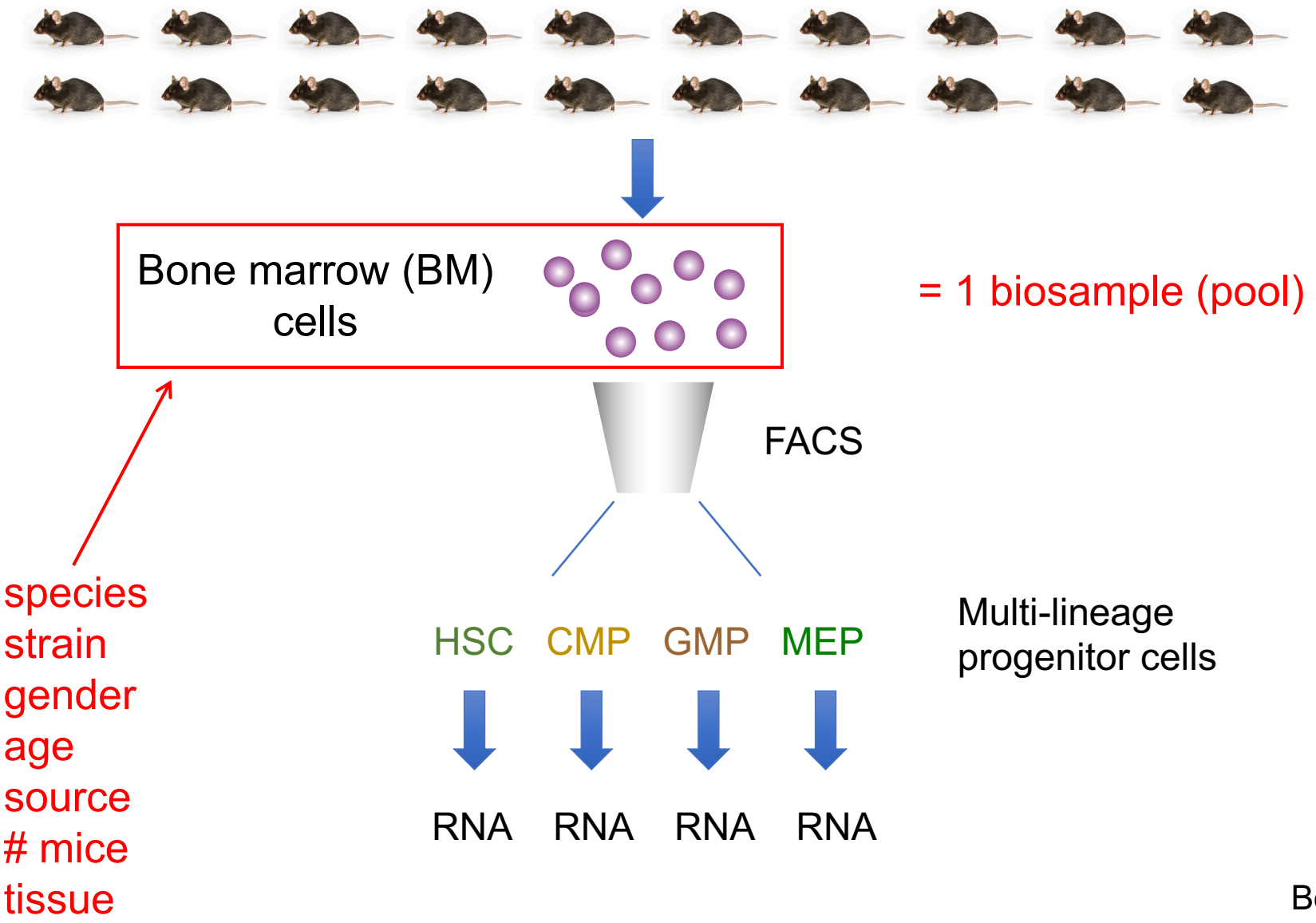
FACS

HSC CMP GMP MEP

Multi-lineage progenitor cells

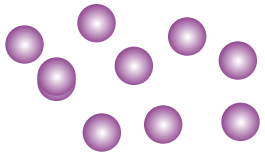
RNA RNA RNA RNA

Biosamples for RNA-seq hematopioetic progenitors



Biosamples for RNA-seq hematopioetic progenitors

Bone marrow (BM) cells



FACS

cell isolation protocol
sort markers
Ab sources

= 4 biosamples

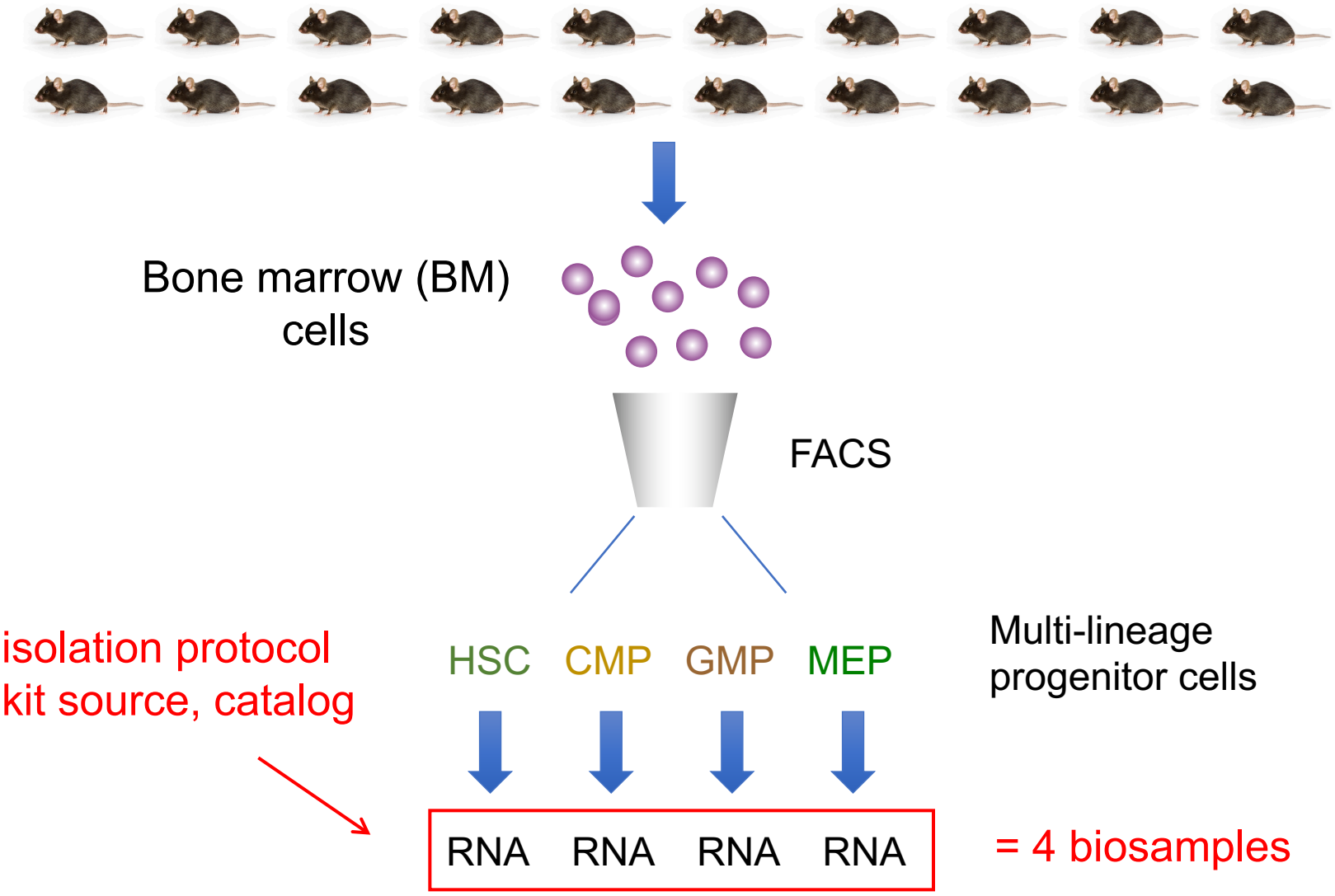
HSC CMP GMP MEP

Multi-lineage progenitor cells

Not everyone uses same sort markers!

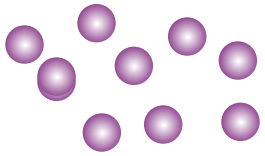
RNA RNA RNA RNA

Biosamples for RNA-seq hematopioetic progenitors



Biosamples for RNA-seq hematopioetic progenitors

Bone marrow (BM) cells

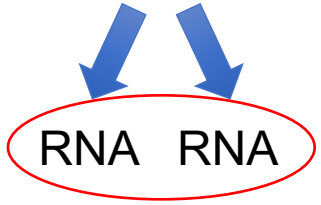


FACS

What if HSCs are split into two groups for RNA?

HSC CMP GMP MEP

Multi-lineage progenitor cells

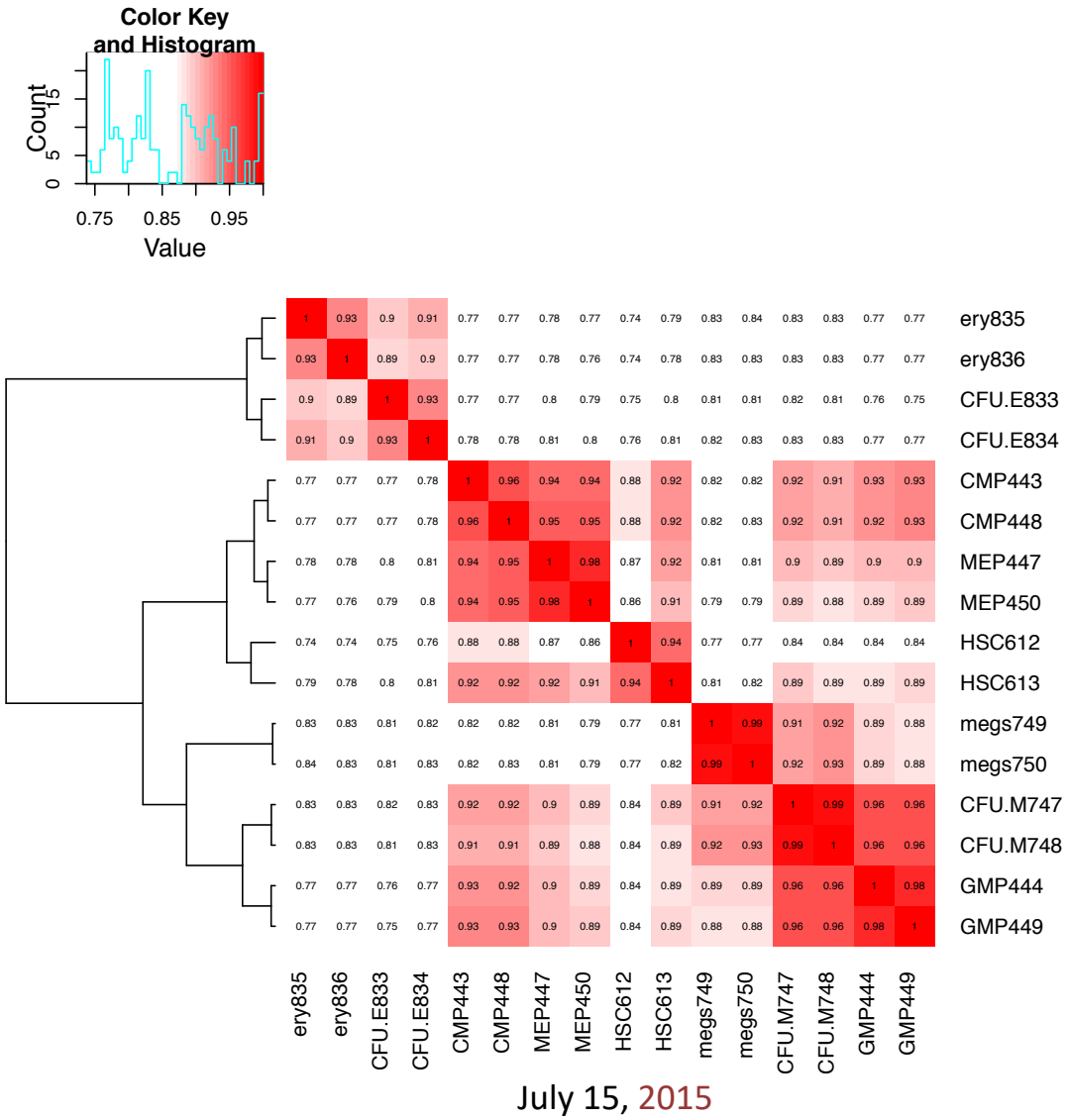


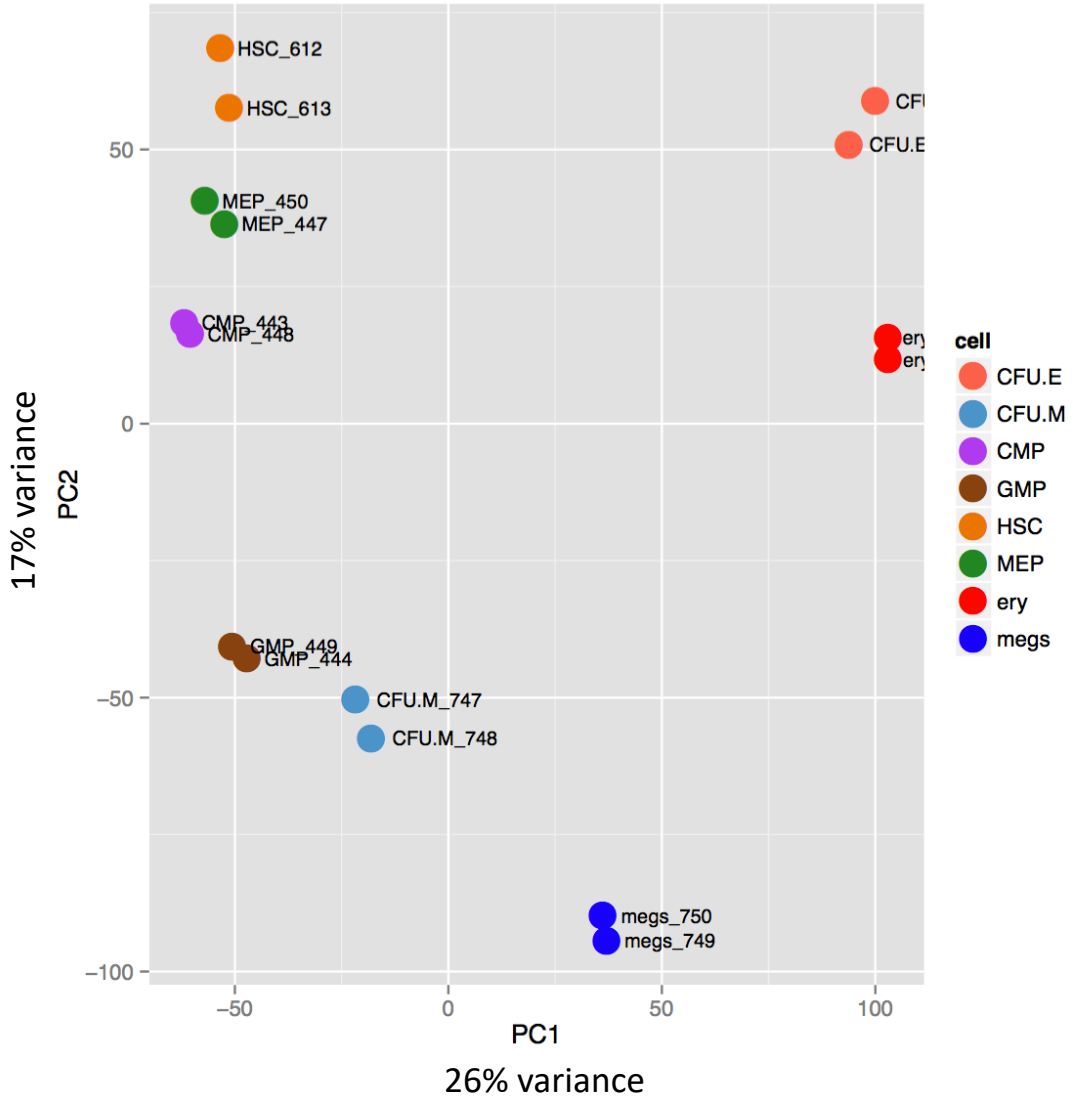
Biological replicate
or
Technical replicate?

Totalscript RNA-seq data

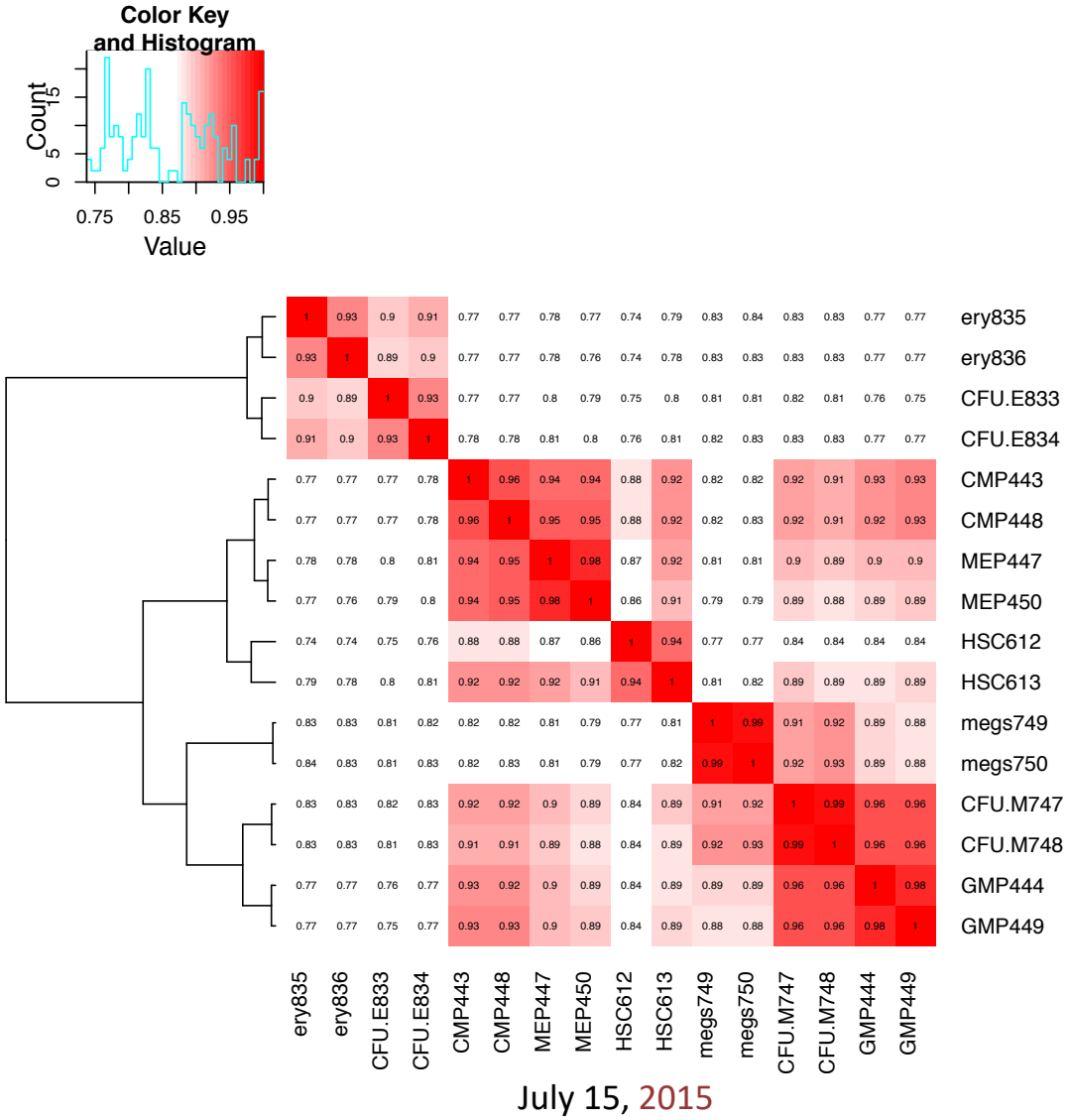
Cell	ID	raw reads	mapped	%mapped	#exp genes	#cells
HSC	612	166,587,096	125,034,591	75%	11135	100
HSC	613	97,306,522	78,586,012	81%	12192	100
CMP	443	111,962,046	74,489,046	67%	10,970	100
CMP	448	102,721,185	69,397,219	68%	11,275	100
GMP	444	122,112,996	97,336,358	80%	10,924	100
GMP	449	113,019,439	87,941,288	78%	11,057	100
MEP	445	101,659,452	81,295,641	80%	10,265	100
MEP	447	95,521,267	74,128,800	78%	10,888	100
MEP	450	165,368,955	133,544,605	81%	10,880	100
CFUE	833	254,709,319	187,485,391	74%	9835	1 ng
CFUE	834	223,556,580	157,862,009	71%	9874	1 ng
ERY	835	181,392,089	117,414,097	65%	9752	1 ng
ERY	836	204,160,404	164,319,341	80%	9533	1 ng
CFUM	747	238,358,635	177,046,832	74%	10968	1 ng
CFUM	748	240,411,833	175,613,074	73%	10847	1 ng
MEG	749	218,779,679	162,008,296	74%	10763	1 ng
MEG	750	231,552,233	161,186,992	70%	10865	1 ng

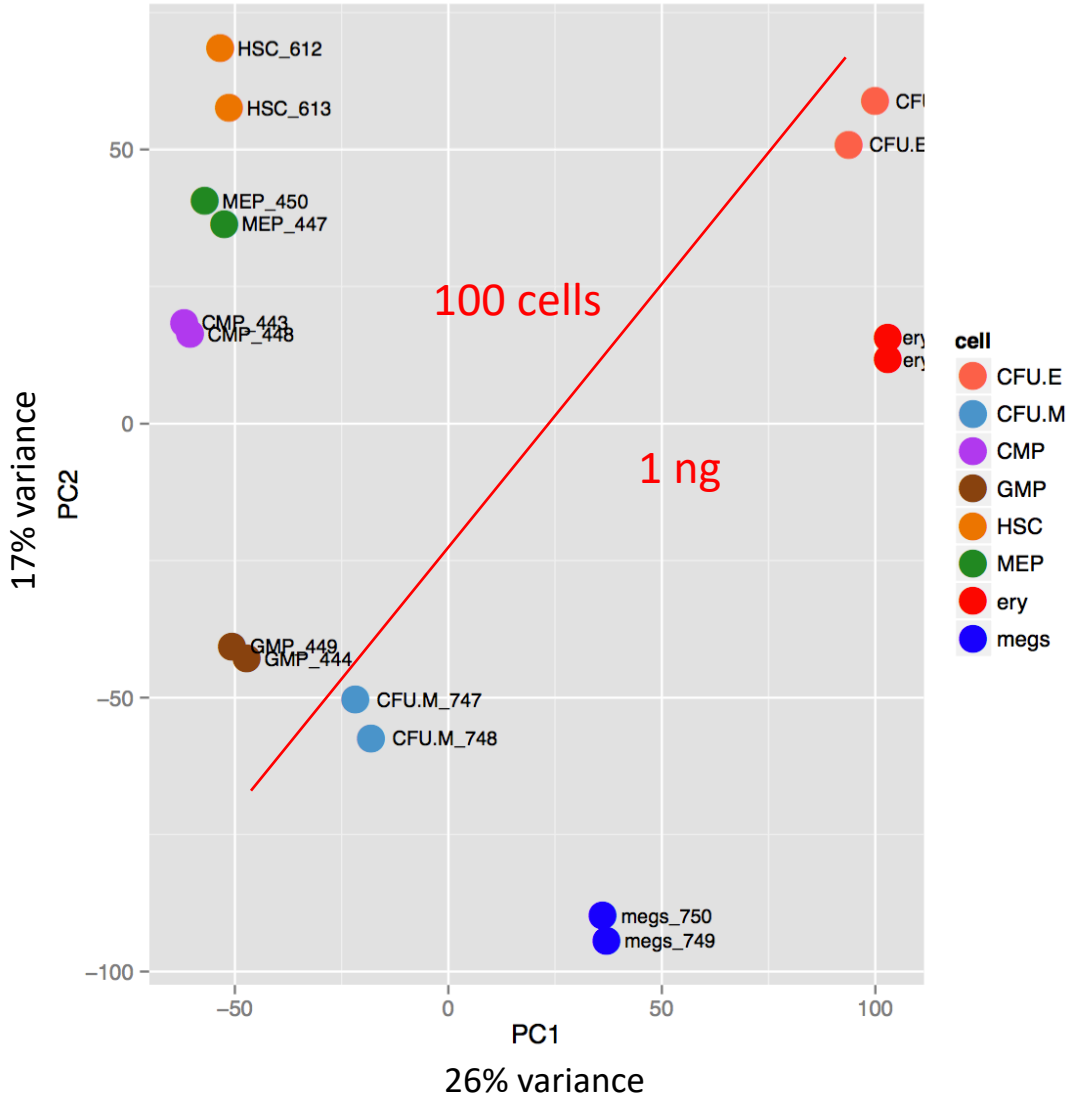
Erythroid cells separate from others in Totalscript





Erythroid cells separate from others in Totalscript

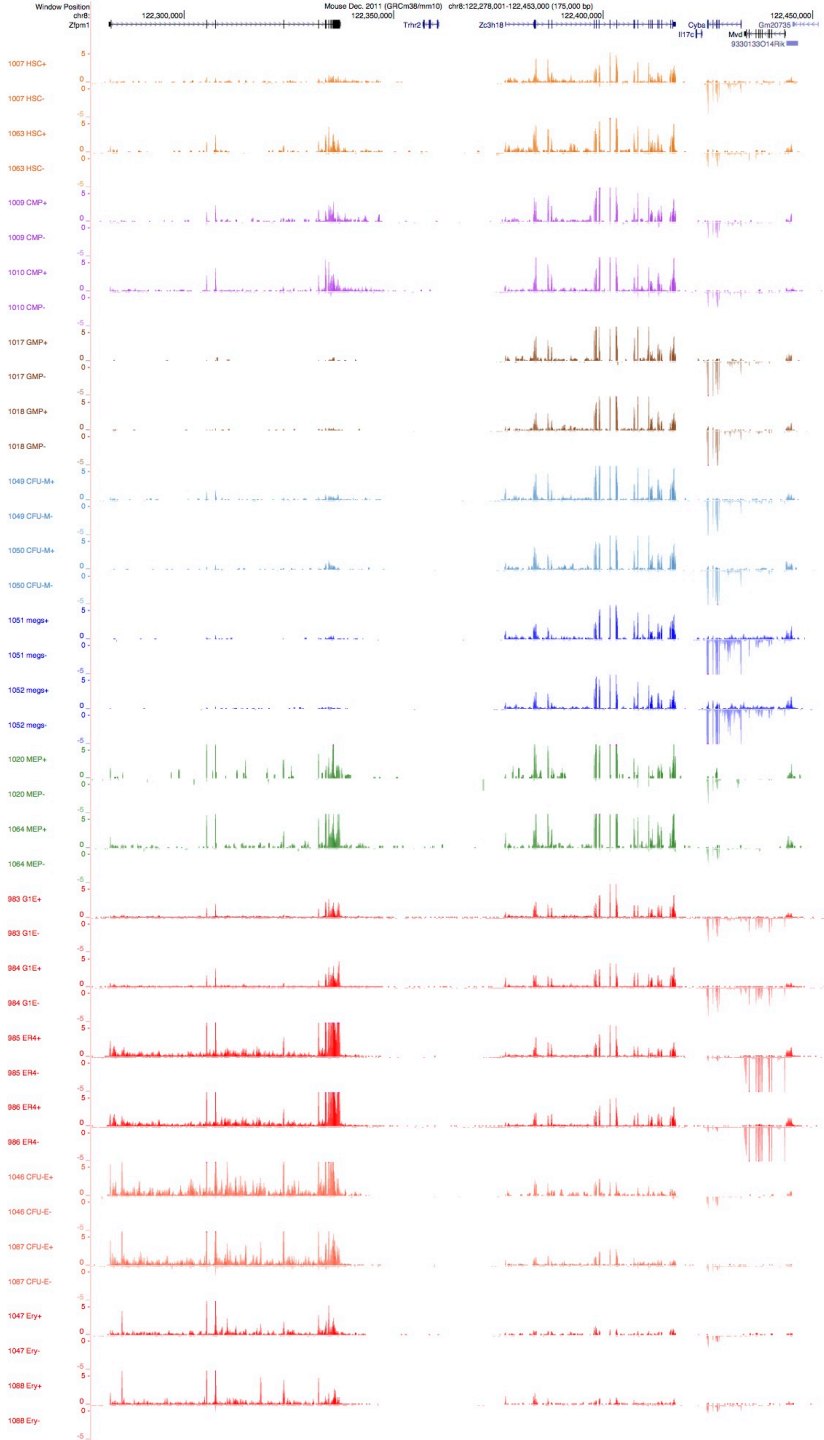




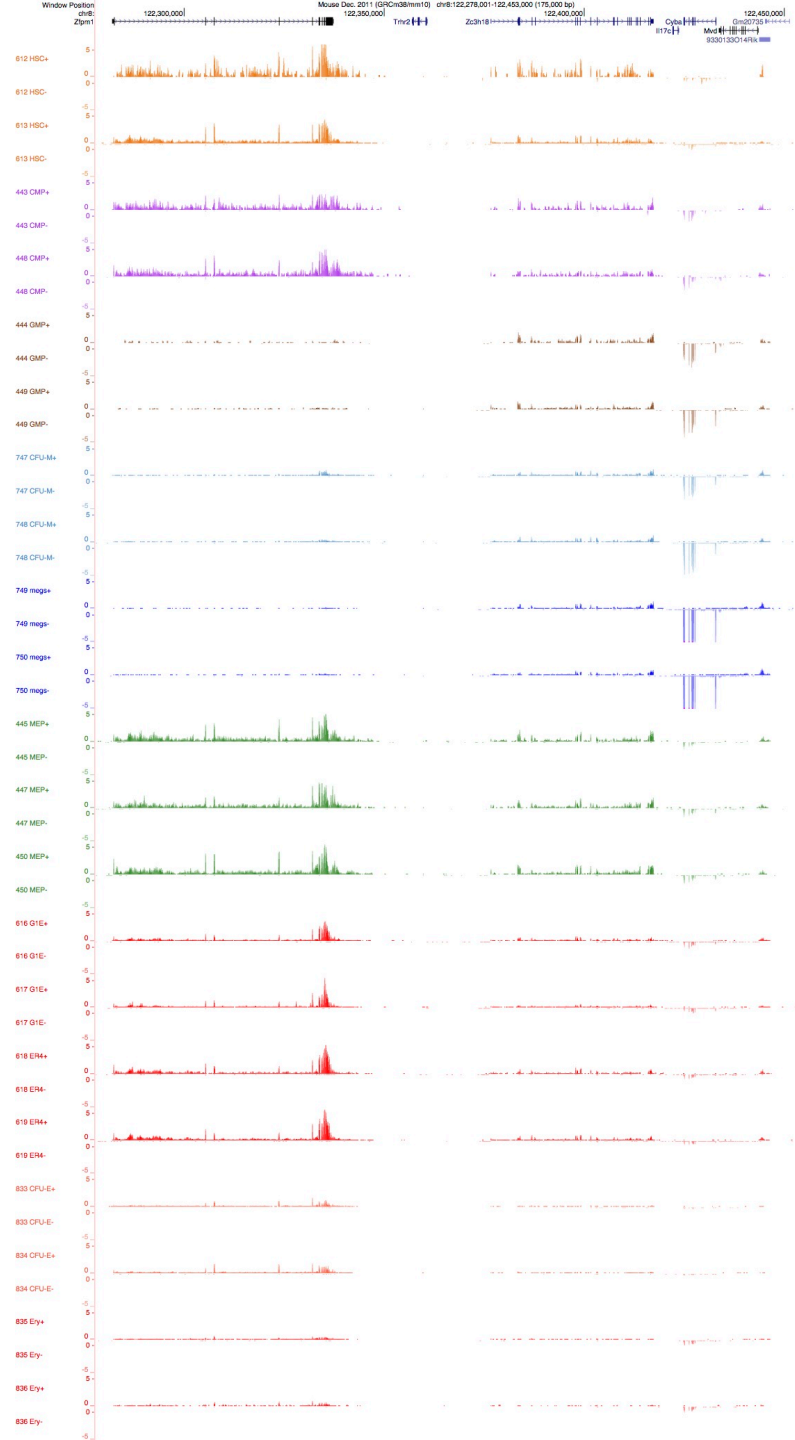
Scriptseq RNA-seq data

Cell	ID	raw reads	mapped reads	%mapped	#exp genes	RNA
HSC	1007	47,628,197	34,521,492	72%	10902	100 ng
HSC	1063	93,842,615	83,452,280	89%	10908	100 ng
CMP	1009	58,215,511	39,155,605	67%	10385	100 ng
CMP	1010	224,960,297	155,342,267	69%	10831	100 ng
GMP	1017	81,308,681	52,061,556	64%	9885	100 ng
GMP	1018	86,717,708	70,399,490	81%	10165	100 ng
MEP	1019	129,623,204	68,375,149	53%	10152	100 ng
MEP	1064	104,418,331	76,484,456	73%	9603	100 ng
CFUE	1046	141,468,597	123,367,182	87%	7062	100 ng
CFUE	1087	133,748,192	92,775,316	69%	6097	100 ng
ERY	1047	99,742,922	86,380,451	87%	6739	100 ng
ERY	1088	125,833,208	66,207,743	53%	5192	100 ng
CFUM	1049	72,198,257	64,512,086	89%	10722	100 ng
CFUM	1050	110,159,472	102,000,386	93%	10173	100 ng
MEG	1051	73,217,334	68,678,478	94%	10299	100 ng
MEG	1052	83,842,340	77,331,723	92%	10233	100 ng

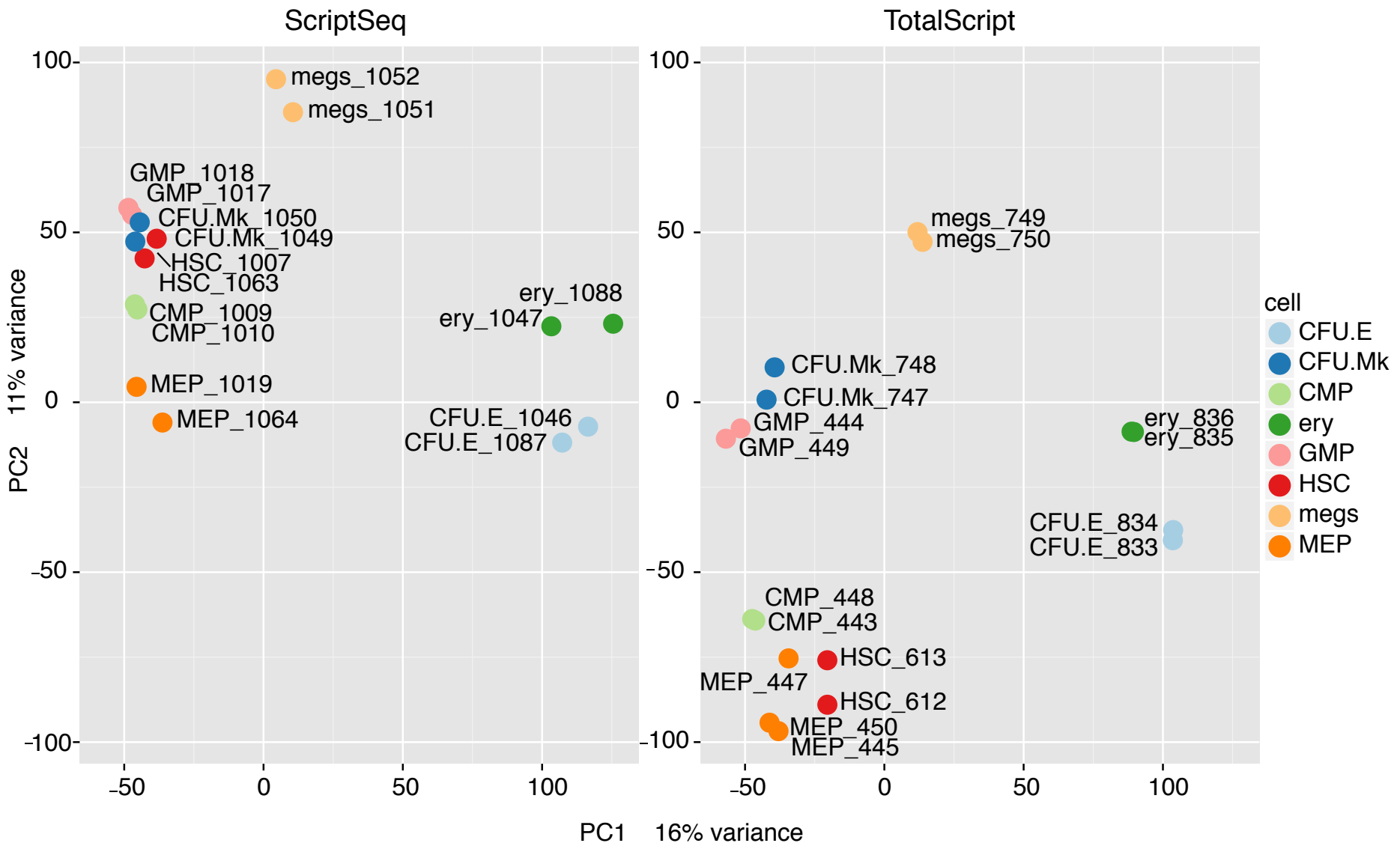
Scriptseq



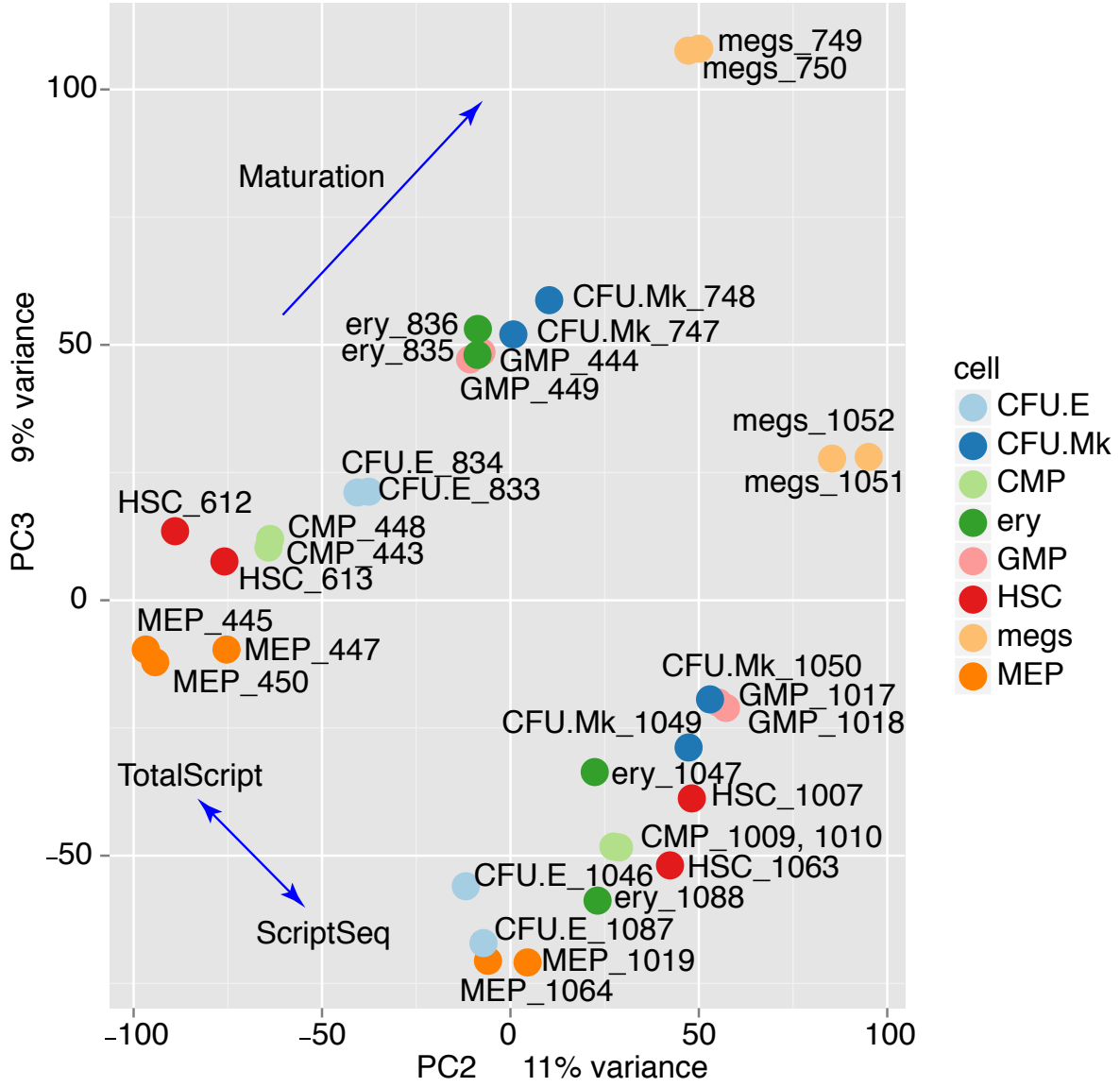
Totalscript



Scriptseq vs. Totalscript:
Similar relationships between cell types for PC1 vs PC2



Scriptseq vs. Totalscript:
Cells types separate by method PC2 vs PC3



Commercial technologies affect reproducibility

Totalscript (2013 through mid 2015)

- total RNAseq using an rRNA reduction method
- Tn based library prep
- *discontinued in mid-2015*

ScriptSeqv2 (mid 2015 through Oct 2018)

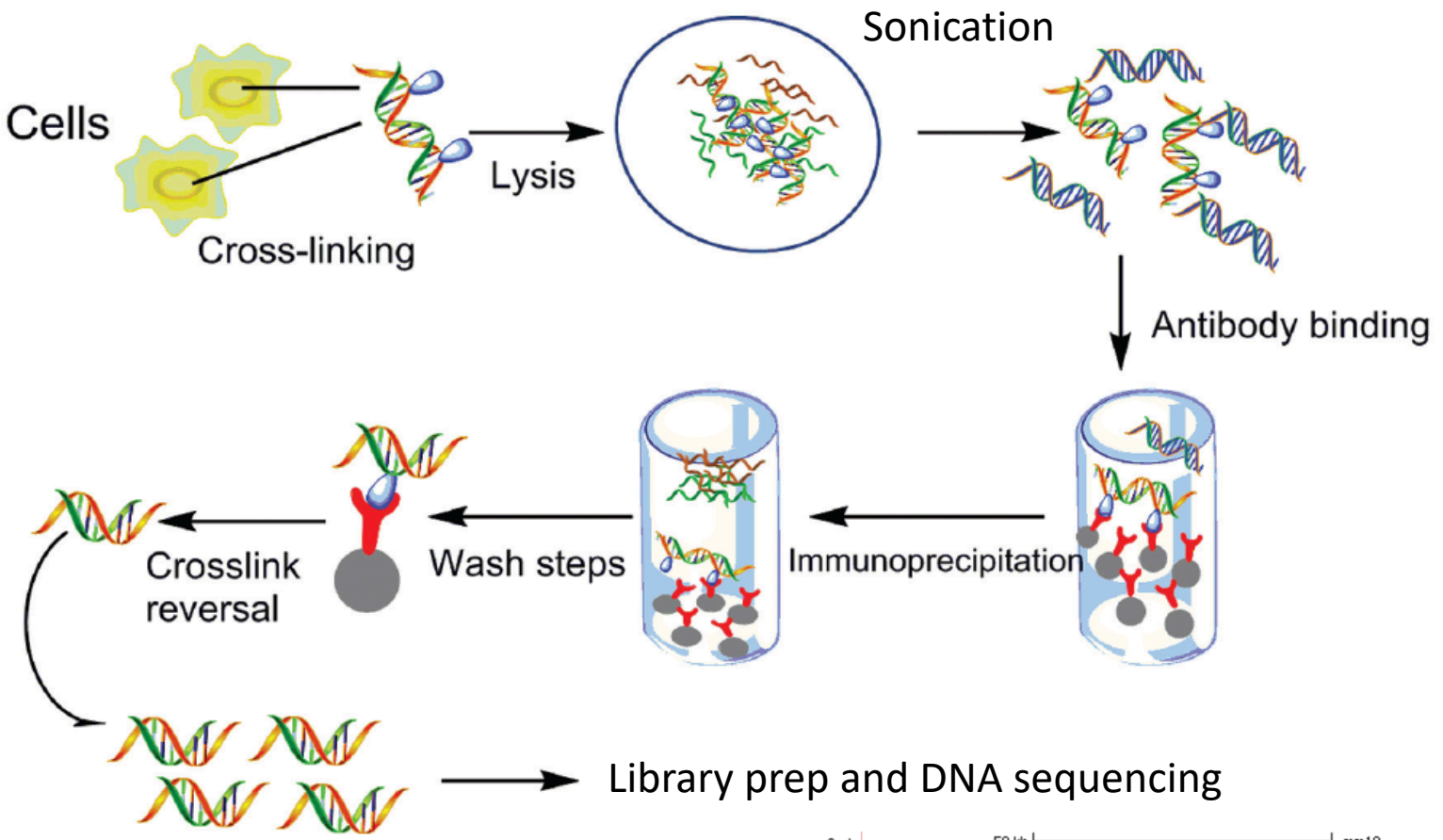
- total RNAseq using rRNA depletion (Ribo-Zero)
- standard fragmentation, cDNA synthesis, ligation, etc.
- *discontinued in Oct 2018*

TruSeq Total RNA-seq

- total RNAseq using rRNA depletion (Ribo-Zero, but different protocol)
- standard fragmentation, cDNA synthesis, ligation, etc.

Illumina just announced that they are discontinuing Ribo-Zero!!!

Overview of ChIP-seq



Adapted from Song et al., 2015. PMCID: PMC4588119

ChIP-seq dataset quality is highly variable

- Marinov et al. 2014* performed a uniform analysis of vertebrate transcription factor ChIP-seq datasets in the Gene Expression Omnibus (GEO) repository as of April 1, 2012.
- 917 ChIP-seq and 292 control libraries
- Datasets were non-ENCODE
- Many datasets only had 1 replicate
- Using ENCODE quality metrics, the study found:
  - The majority (55%) of datasets scored as being highly successful
  - Approximately 20% were of poor quality
  - Approximately 25% were of intermediate quality
- *Significant subset of control datasets displayed an enrichment structure similar to successful ChIP-seq data!*

*Marinov GK, Kundaje A, Park PJ, Wold BJ. 2014. Large-scale quality analysis of published ChIP-seq data. 2014. *G3 (Bethesda)*. Feb 19;4(2):209-23. doi: 10.1534/g3.113.008680.

High standards do not guarantee robustness

- One strength of ENCODE is its transparency and extensive data release policy
- Devailly et al. 2015* performed an systematic analysis of ChIP-seq data of transcription and epigenetic factors ENCODE.
- At the time of publication, 690 ChIP sequencing datasets TFs
- Focused on 57 experiments in which had multiple replicates, but the replicates were not merged by ENCODE. *Why?*
- They found that 18 of 57 showed a very low overlap between peak lists between replicates.
- High standards set for technical quality control achieved by the ENCODE consortium does not guarantee the robustness of the sample.

*Devailly G, Mantsoki A, Michoel T, Joshi A. 2015. Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. FEBS Lett Dec 21; 589(24): 3866–3870. doi: 10.1016/j.febslet.2015.11.027

Multiple variables influence ChIP-seq success

- Variables that *cannot* be controlled:
  - **Antisera** – low success rate
  - **Target factor** (eg. TF vs. histone mark) – expression levels, extent of interaction with DNA (i.e. direct binding vs. DNA-associated factor)
  - **Cell type** – abundance, isolation, growth conditions, extent of chromatin condensation

- Variables that *can* be controlled:
  - **Cell number** – except for cell types of limited quantity
  - **Fixation** – time of fixation and concentration of formaldehyde
  - **Sonication** – time and extent of fragmentation
  - **Library prep**
  - **Sample management**

How does chromatin sonication affect ChIP-seq success?

Factor	#Cells	#Cycles
CTCF	50M	3
CTCF	50M	5
CTCF	50M	7
CTCF	50M	10
CTCF	50M	15
CTCF	20M	3
CTCF	20M	5
CTCF	20M	7
CTCF	20M	10
CTCF	20M	15
TAL1	50M	3
TAL1	50M	5
TAL1	50M	7
TAL1	50M	10
TAL1	50M	15
TAL1	20M	3
TAL1	20M	5
TAL1	20M	7
TAL1	20M	10
TAL1	20M	15

We tested two TFs, CTCF and TAL1, both of which bind DNA, but exhibit different binding characteristics.

Four independent sets of ChIP-seq experiments

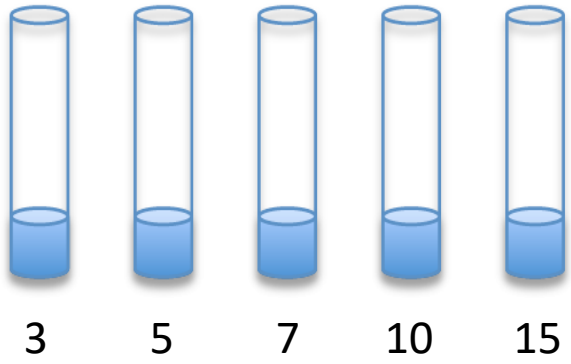
CTCF	50M cells
CTCF	20M cells
TAL1	50M cells
TAL1	20M cells

Five sonication conditions were used:

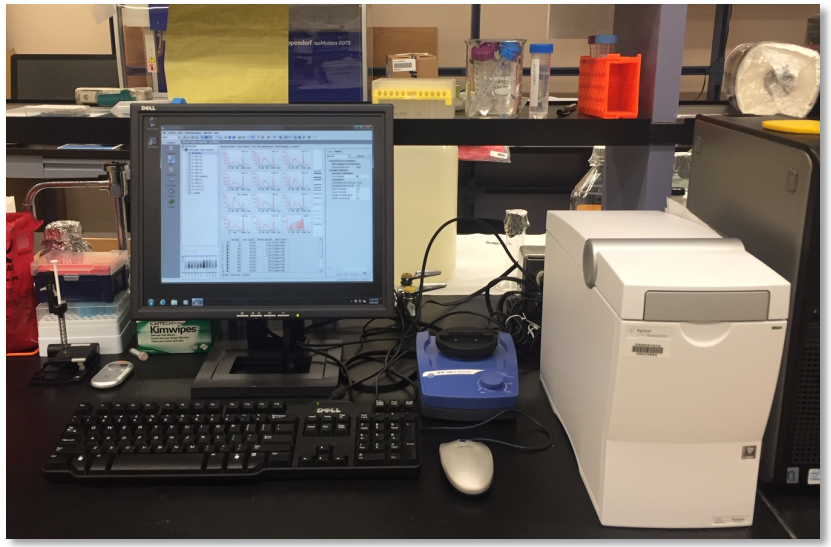
3, 5, 7, 10, 15 cycles of 30 sec on/off

All five conditions/experiment were sonicated at the same time, but in separate tubes

Cell concentration for each condition was the same (i.e. 20M cells/ml)



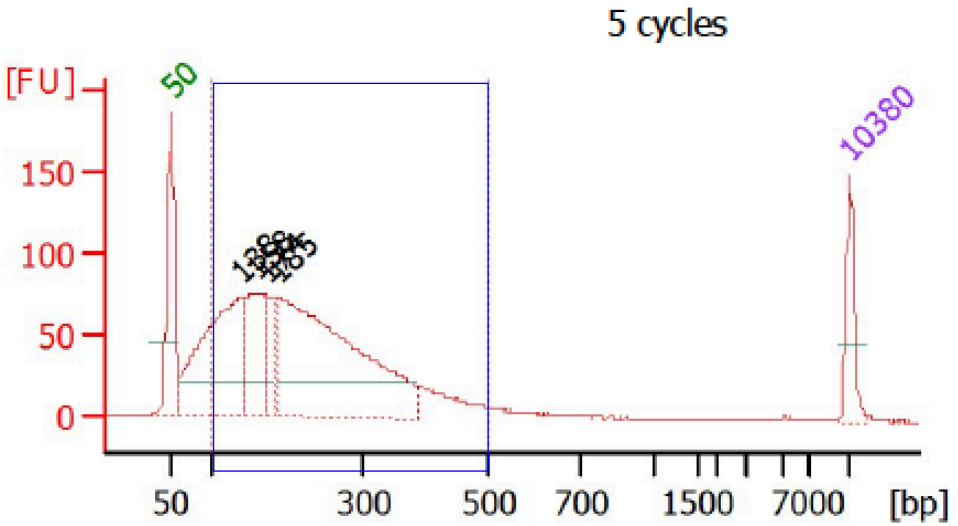
Detailed information about chromatin size can be acquired using the Agilent Bioanalyzer



Why use a window of 100-500 bp to estimate the average chromatin length?

- Only fragments in this range would possibly enter library

- Some cell types may have large MW heterochromatin that could skew average size when measured over total range



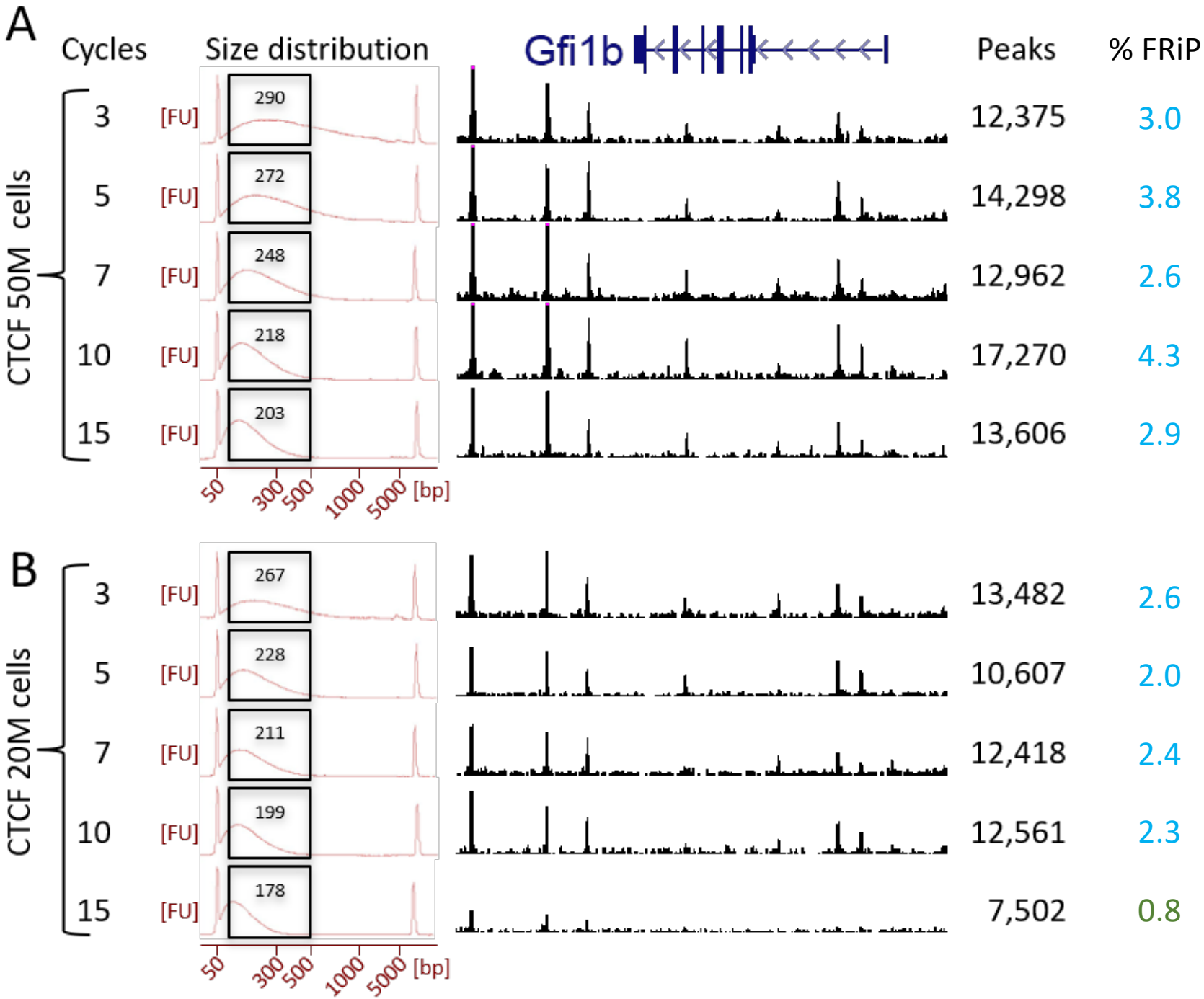
Overall Results for sample 2 : 5 cycles

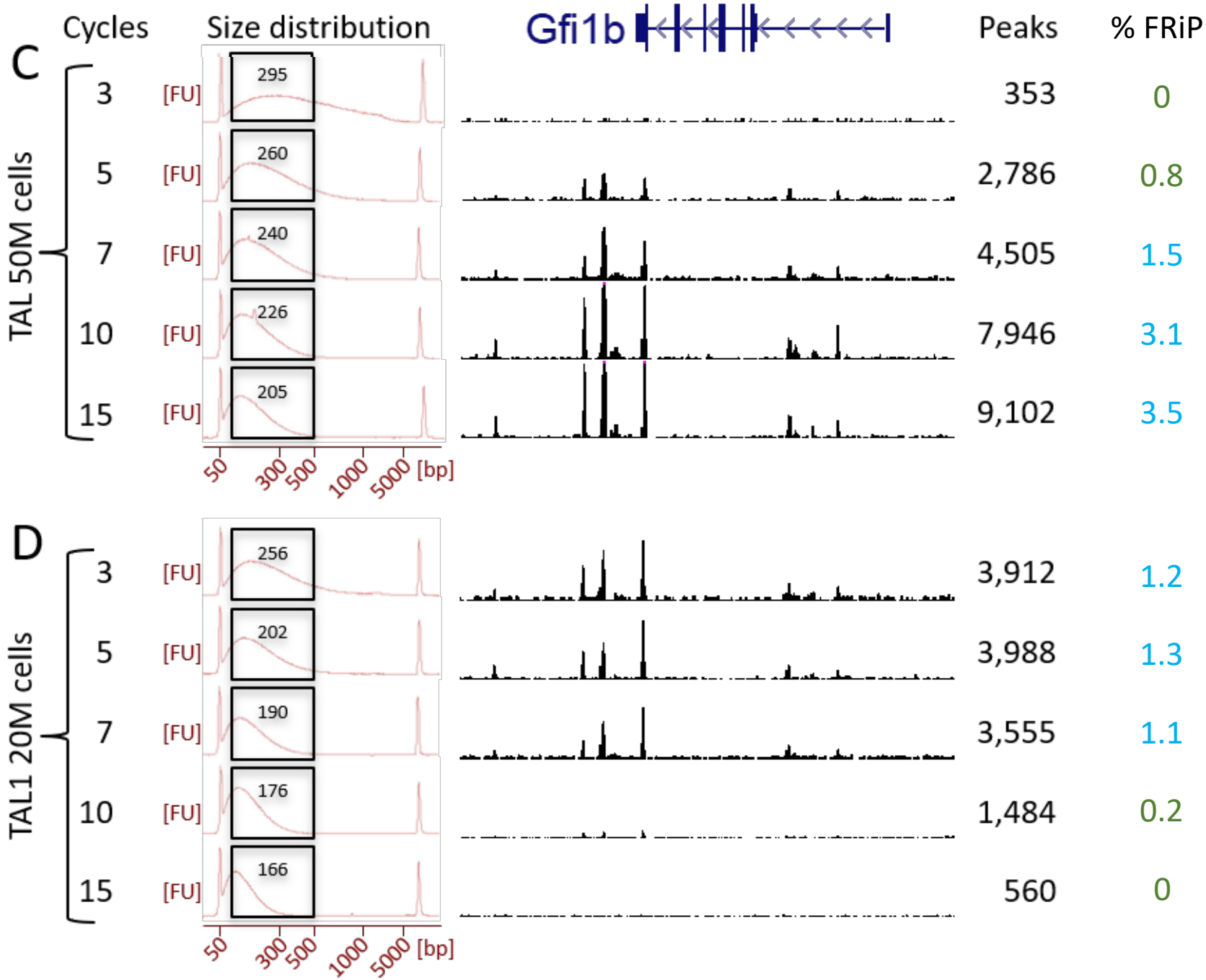
Number of peaks found: 4

Area 1: 921.4

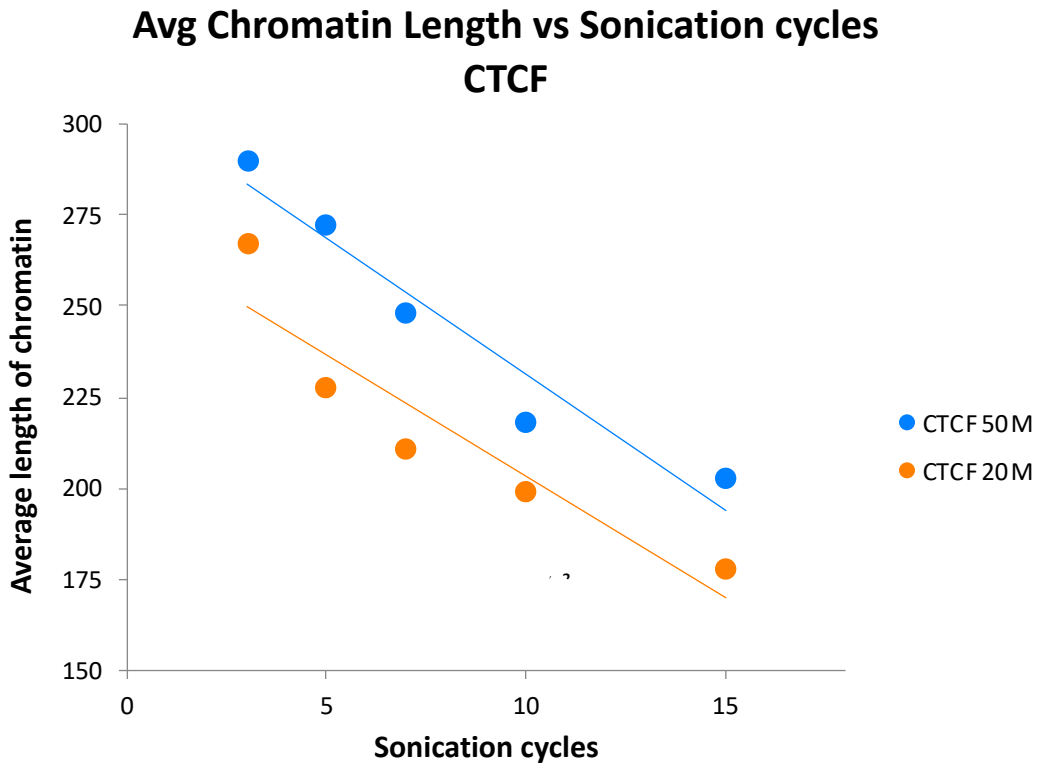
Region table for sample 2 : 5 cycles

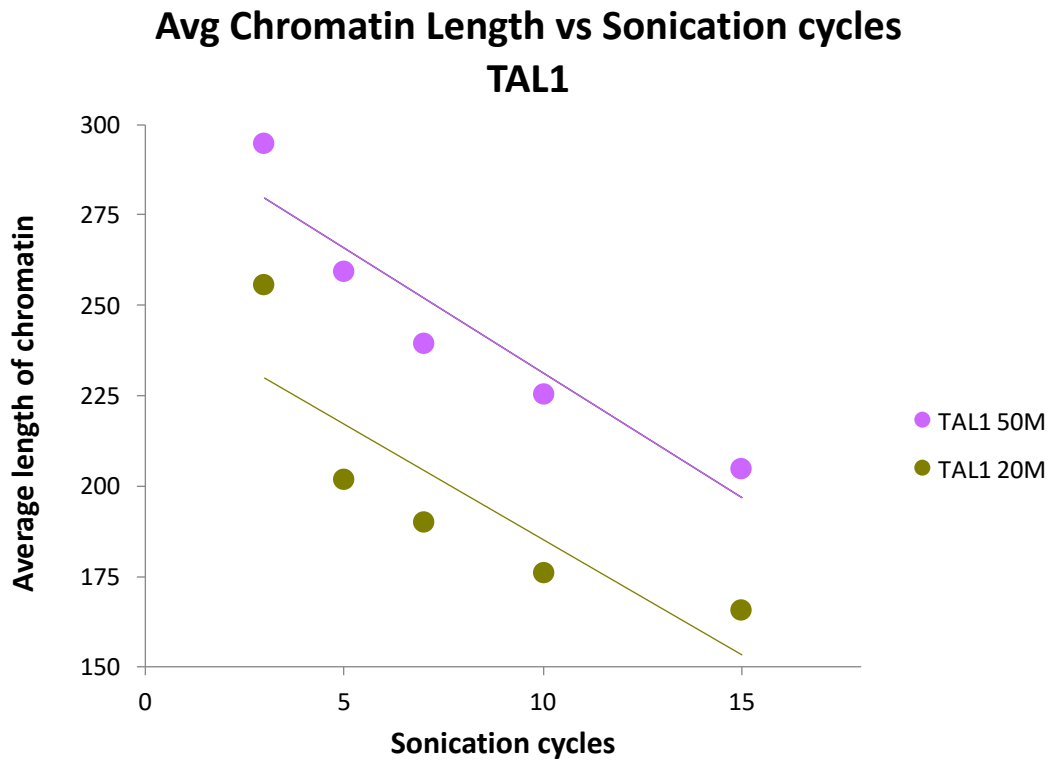
From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
100	500	921.4	84	228	39.8	64.74	


A
Cycles
Size distribution
Gfi1b
Peaks
% FRiP
CTCF 50M cells
3
5
7
10
15
[FU]
290
272
248
218
203
12,375
14,298
12,962
17,270
13,606
3.0
3.8
2.6
4.3
2.9
50
300
500
1000
5000
[bp]
B
CTCF 20M cells
3
5
7
10
15
[FU]
267
228
211
199
178
13,482
10,607
12,418
12,561
7,502
2.6
2.0
2.4
2.3
0.8
50
300
500
1000
5000
[bp]

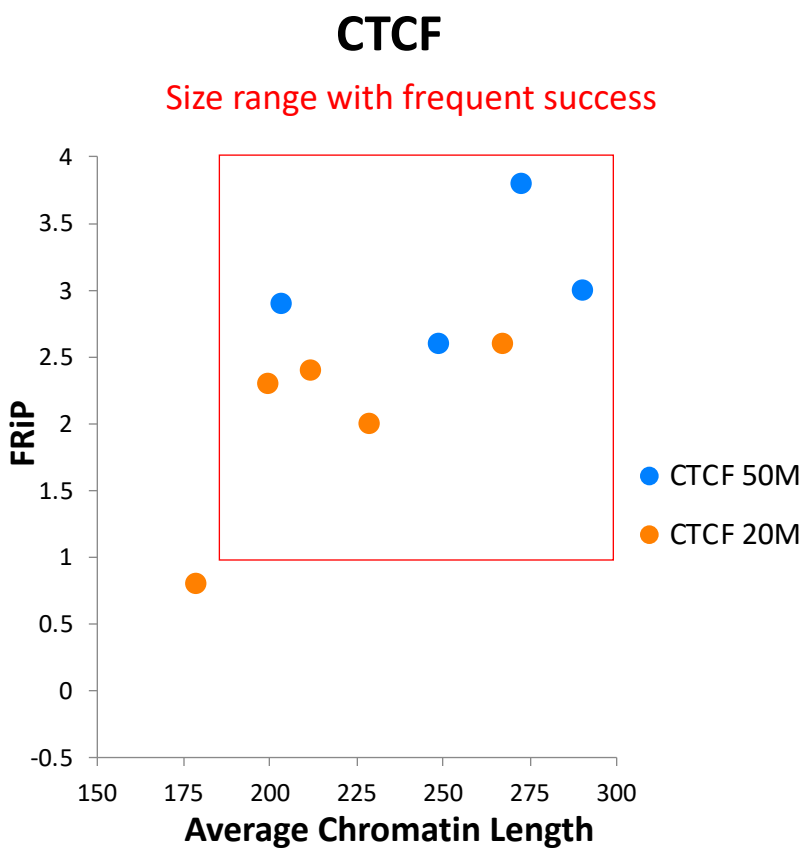

C
Cycles
Size distribution
Gfi1b
Peaks
% FRiP
TAL 50M cells
3
5
7
10
15
[FU]
295
260
240
226
205
353
2,786
4,505
7,946
9,102
0
0.8
1.5
3.1
3.5
50
300
500
1000
5000
[bp]
D
TAL1 20M cells
256
202
190
176
166
3,912
3,988
3,555
1,484
560
1.2
1.3
1.1
0.2
0

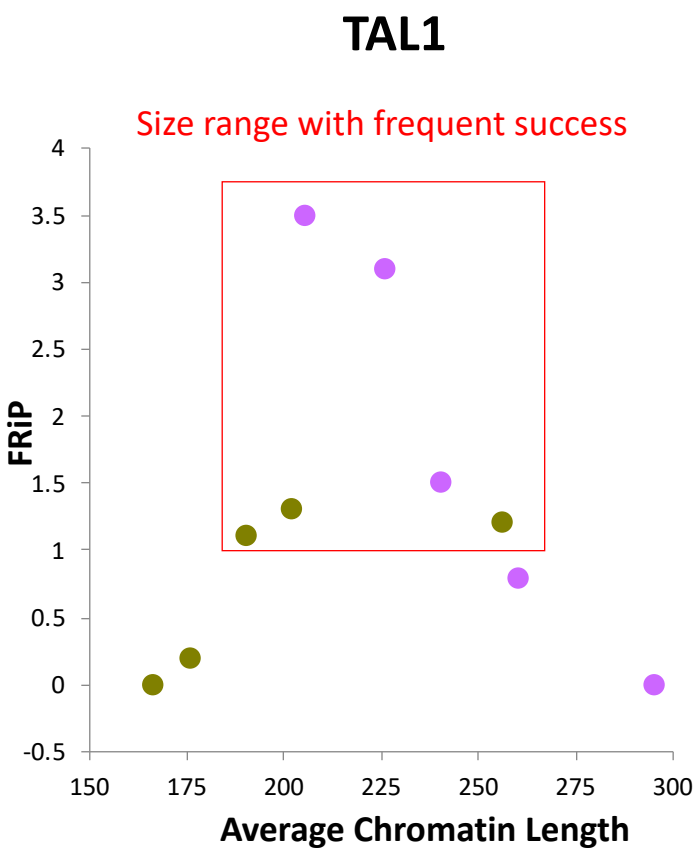
As expected, fragment length is inversely correlated with the number of sonication cycles





FRiP is influenced by chromatin length





Fragment sizes

Success rate

Dataset ID	Factor	Avg size	FRiP
706	TAL1	166	0
705	TAL1	176	0.2
736	CTCF	178	0.6
505	TAL1	183	0.3
704	TAL1	190	1.1
364	CTCF	191	0.1
485	CTCF	194	0.3
486	TAL1	194	0.5
366	CTCF	195	1.1
735	CTCF	199	2.3
703	TAL1	202	1.3
696	CTCF	203	2.9
674	TAL1	205	3.5
734	CTCF	211	2.4
365	CTCF	213	0.1
332	CTCF	214	3.7
333	CTCF	214	4.4
334	CTCF	214	0.7
335	CTCF	214	3.5
336	CTCF	214	2.4
337	CTCF	214	1.7
338	CTCF	214	1.9
339	CTCF	214	1.8
340	CTCF	214	1.9
695	CTCF	218	4.3
367	CTCF	226	0
368	CTCF	226	4.1
369	CTCF	226	4.7
673	TAL1	226	3.1
733	CTCF	228	2
707	TAL1	238	0.8
672	TAL1	240	1.5
694	CTCF	248	2.6
702	TAL1	256	1.2
671	TAL1	260	0.8
585	TAL1	263	0.4
732	CTCF	267	2.6
583	TAL1	269	1.1
693	CTCF	272	3.8
581	TAL1	273	0.1
584	TAL1	289	0.4
692	CTCF	290	3
670	TAL1	295	0
582	TAL1	297	0
580	TAL1	298	3.9

<200bp — 3/10 = 30%

200-250bp — 19/23 = 82.6%

>250bp — 6/12 = 50%

Analysis of sonication profiles can be used to predict success of ChIP-seq

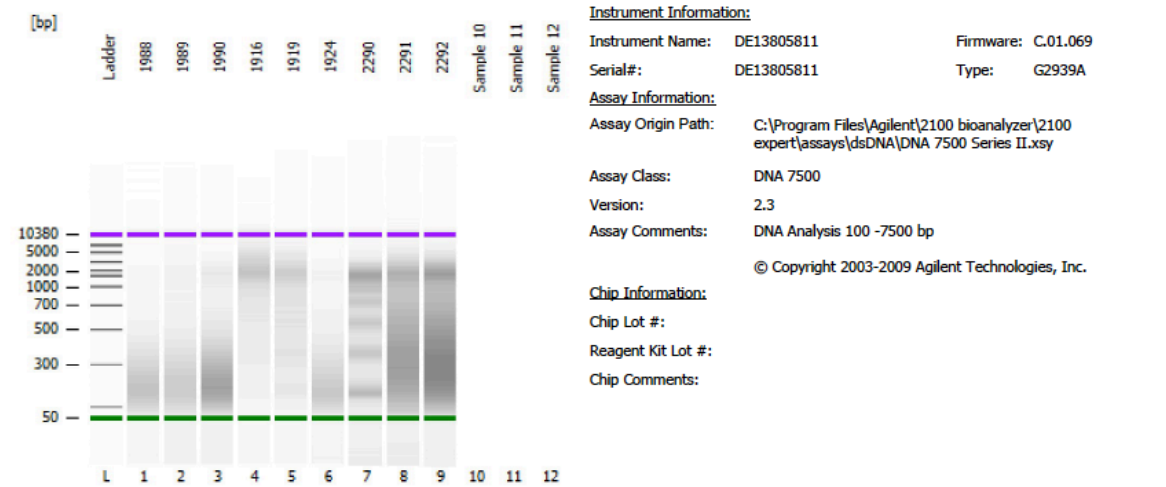
Retrospective analysis of ChIP-seq samples/datasets

Inclusion criteria:
G1E-ER4+E2
CTCF or TAL1
Corresponding input material

All datasets (n=45) were divided into three groups based on average chromatin length.

How many datasets in each size category had a FRiP ≥1.0%?

These samples were all sheared "the same way"



Instrument Information:

Instrument Name: DE13805811 Firmware: C.01.069

Serial#: DE13805811 Type: G2939A

Assay Information:

Assay Origin Path: C:\Program Files\Agilent\2100 bioanalyzer\2100 expert\assays\dsDNA\DNA 7500 Series II.xsy

Assay Class: DNA 7500

Version: 2.3

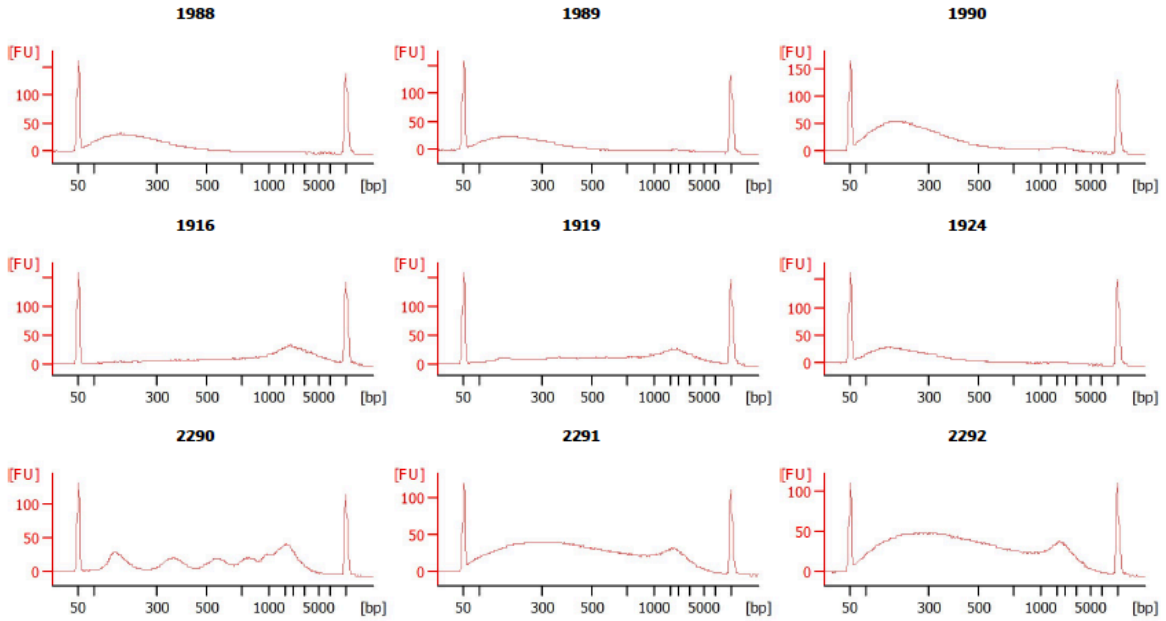
Assay Comments: DNA Analysis 100 -7500 bp

Chip Information:

Chip Lot #:

Reagent Kit Lot #:

Chip Comments:



Metadata considerations for data generation and analysis

Authentication of key biological resources

Accurate and complete lab records

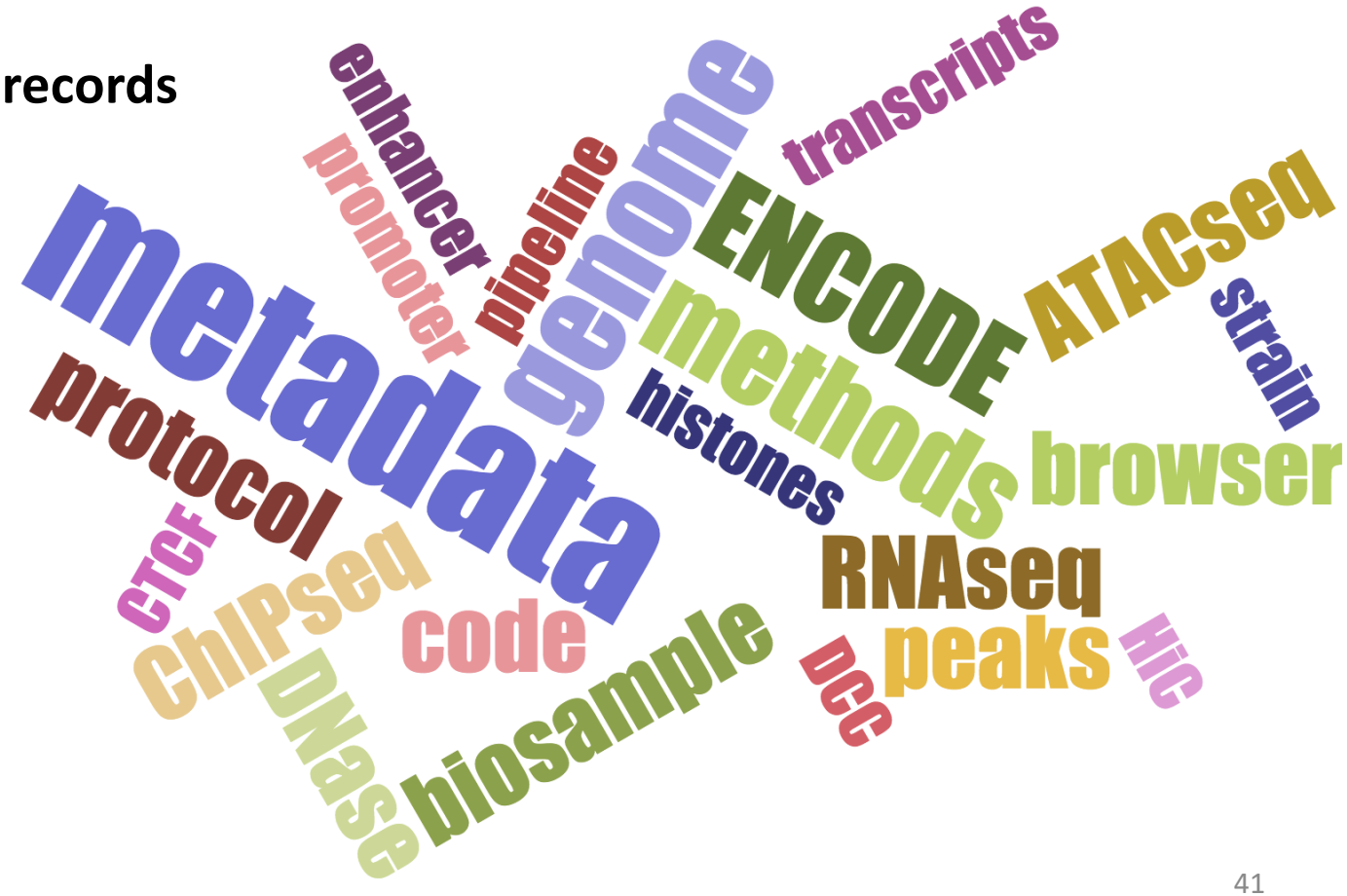
Consistent protocols

Sample tracking

Unique identifiers

Processing pipelines

Controlled vocabulary



Acknowledgements

Hardison lab (PSU)
Ross Hardison
Belinda Giardine
Alex Wixom
Guanjue Xiang
April Cockburn
Lin An (former)
Maria Long (former)

Bodine lab (NHGRI)
David Bodine
Elisabeth Heuston
Stacie Anderson

ENCODE3 production group
Rick Myers (HudsonAlpha)
Barbara Wold (Caltech)
Ali Mortazavi (UC Irvine)
Tim Reddy (Duke)

Yu Zhang (formerly at PSU)

ENCODE
NHGRI
NIH NIDDK