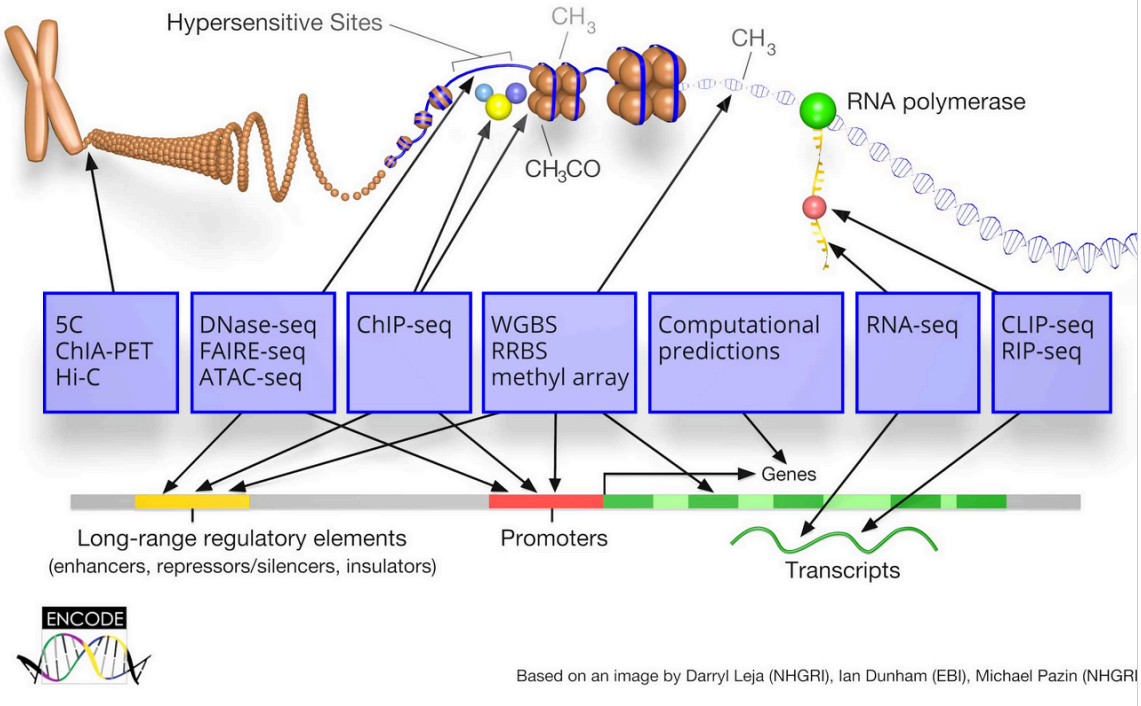
PennState

Metadata Madness: Lessons from ENCODE

Cheryl A. Keller
Penn State University

July 10, 2017

ENCODE: Encyclopedia of DNA Elements



HUMAN | MOUSE | WORM | FLY

BIOSAMPLE

ASSAY

131 results

Clear Filters

	ChIP-seq	total RNA-seq	polyA mRNA RNA-s...	ATAC-seq	RNA-seq	DNase-seq
immortalized cell line						
G1E-ER4	23	2	7	1		1
G1E	11	2	2	1		1
CH12.LX	8		1			
MEL cell line	4		2			
primary cell						
megakaryocyte	9	2	1	1		
erythroblast	8	2	1	1		
megakaryocyte-erythroid progenitor cell		2	1	1		
common myeloid progenitor		2		1		
erythroid progenitor cell		2		1		
+ See 9 more...						
stem cell						
E14TG2a.4	9					
hematopoietic stem cell		2		1		
leukemia stem cell			1			
tissue						
liver	2					
in vitro differentiated cells						
hematopoietic multipotent progenitor cell			1			
See all biosamples						

HAIB Production Group

Rick Myers (HudsonAlpha)

Ross Hardison (PSU)

Barbara Wold (Caltech)

Ali Mortazavi (UC Irvine)

Tim Reddy (Duke)

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

Metadata is not a new concept

Source: http://www.wisegeek.org/what-is-a-card-catalog.htm

Metadata provides information about other data

Descriptive metadata – Describes a resource for purposes of discovery and identification

Structural metadata – Describes how compound objects or data (i.e. databases) are organized and their relationships, versions, etc.

Administrative metadata – Describes how to manage a resource, how it was created and other technical details, etc.

Metadata considerations

Authentication of key biological resources

Accurate and complete lab records

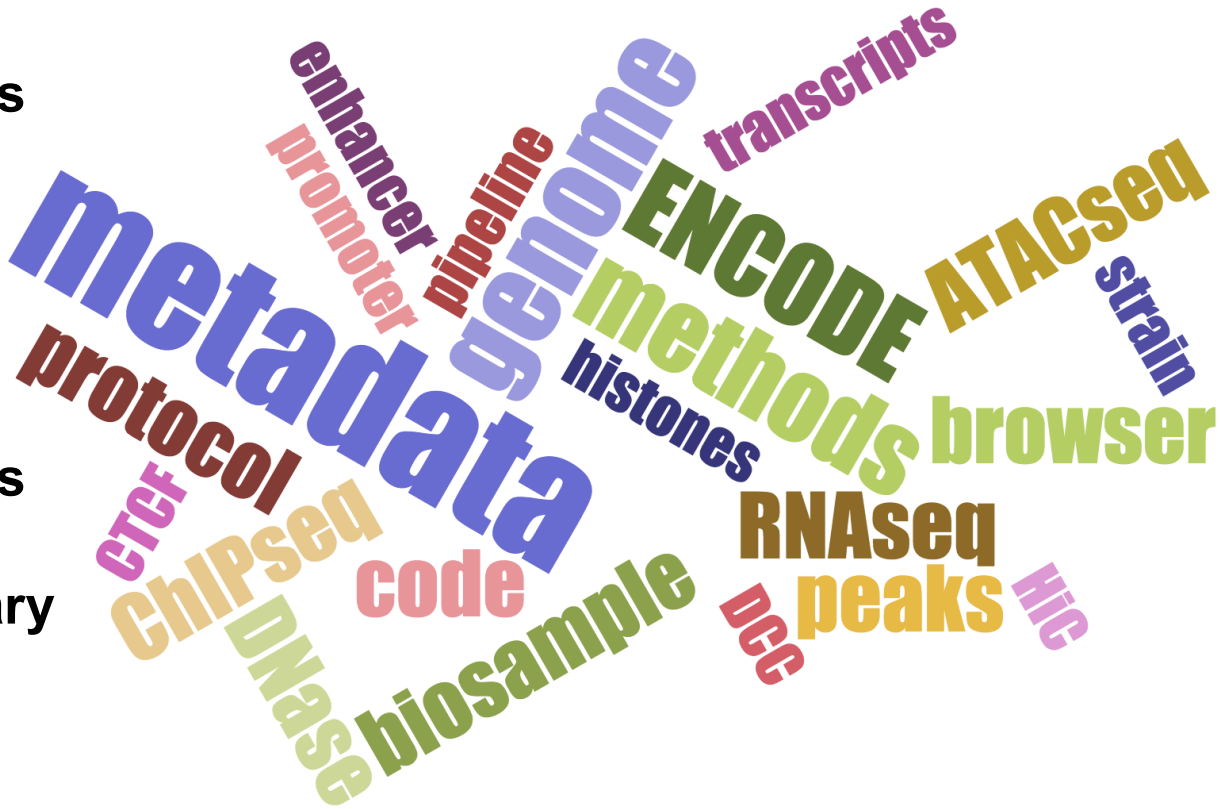
Consistent protocols

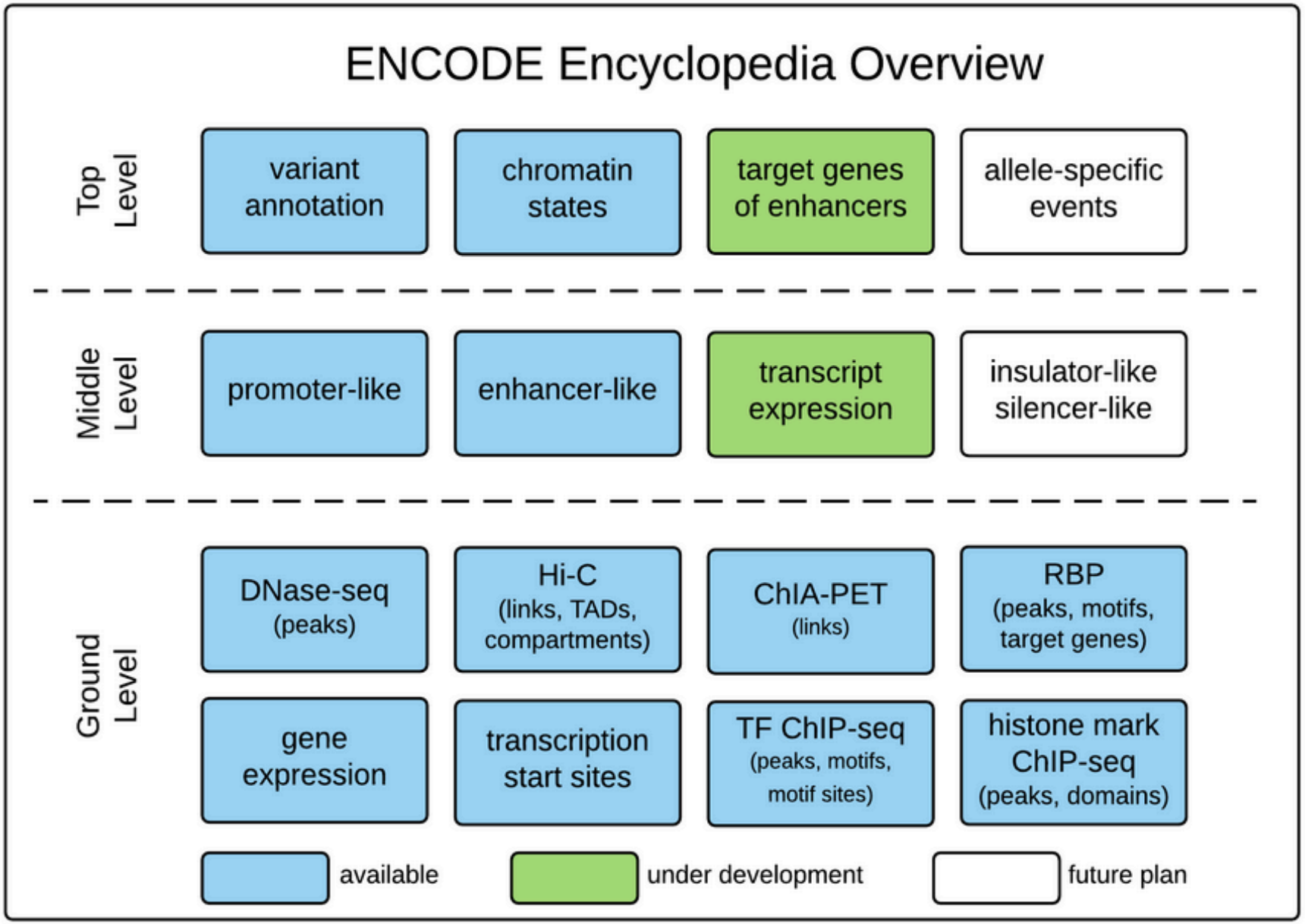
Sample tracking

Unique identifiers

Processing pipelines

Controlled vocabulary





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

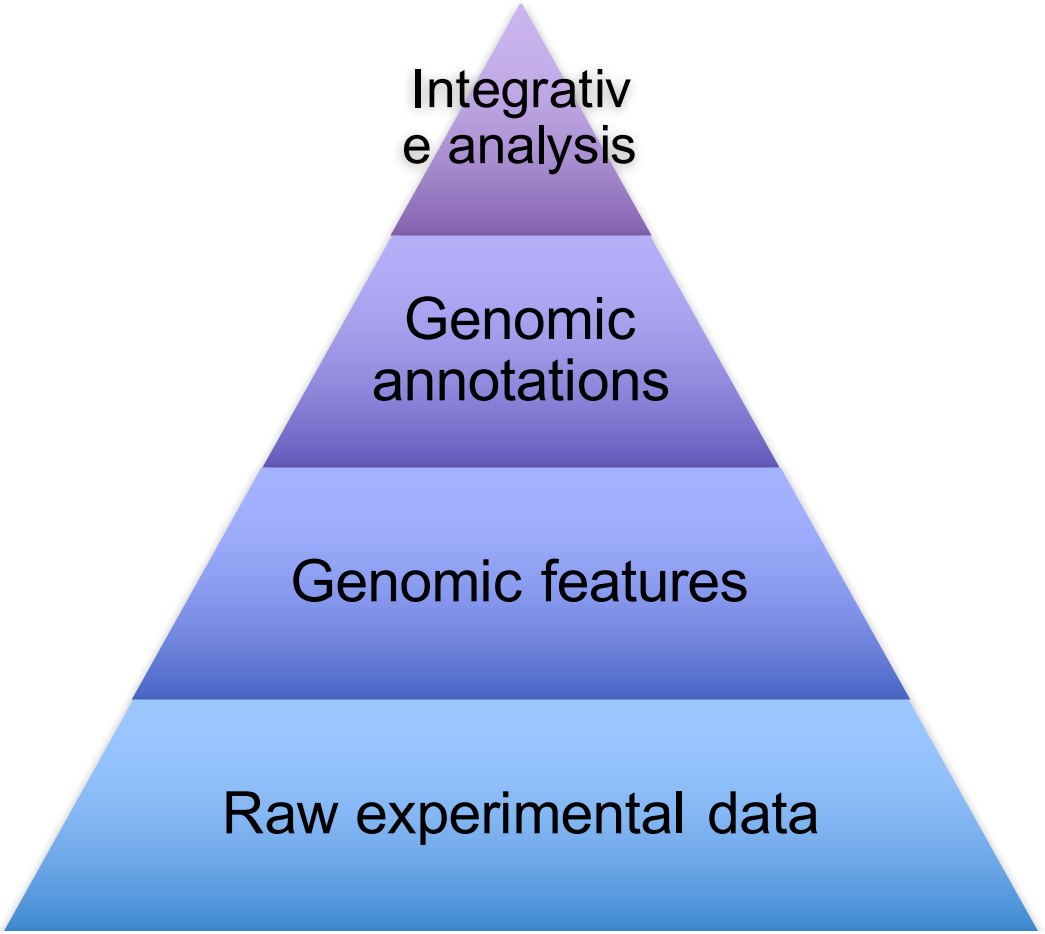
Each level is dependent on quality of lower levels

Top level

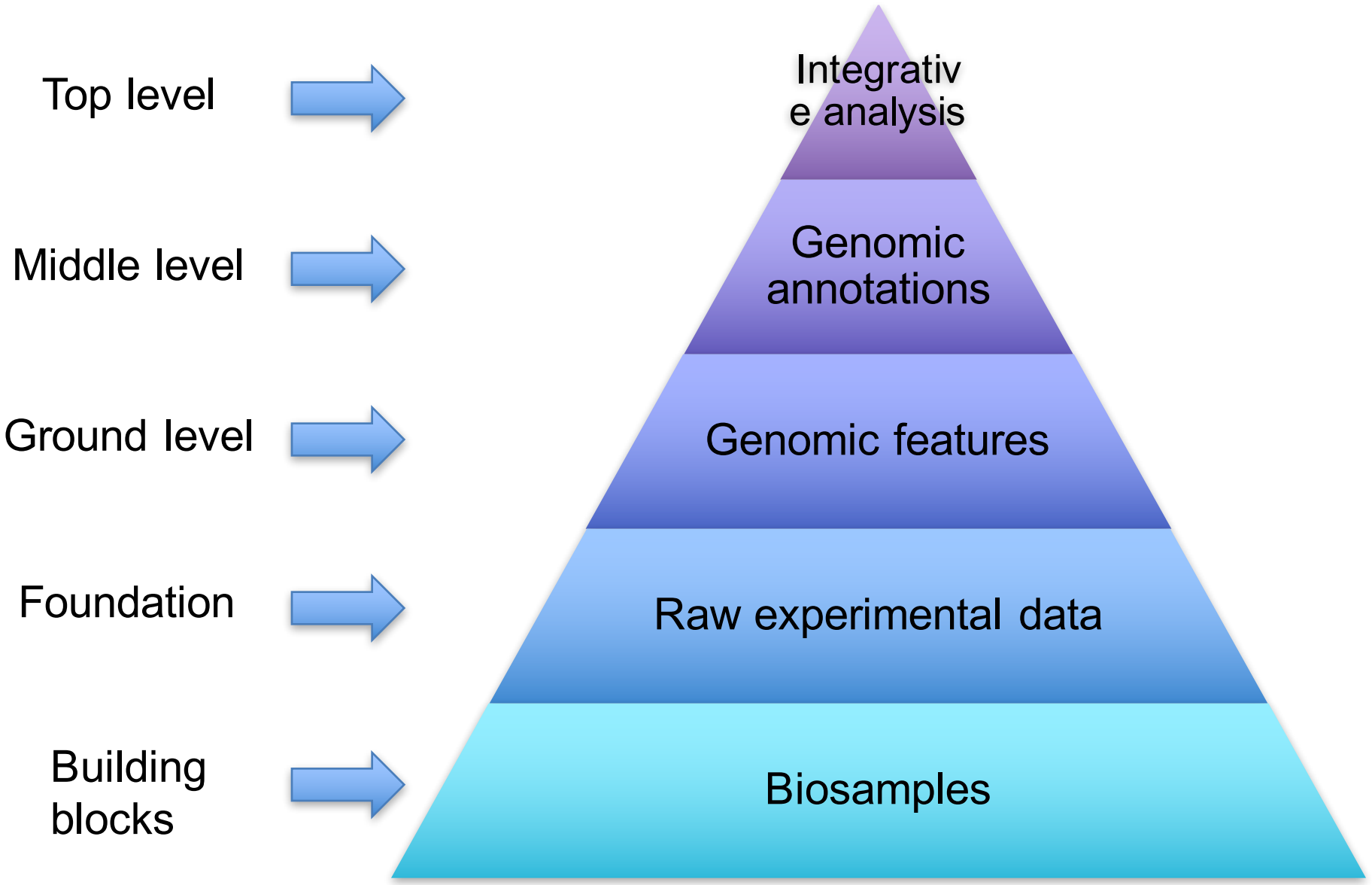
Middle level

Ground level

Foundation

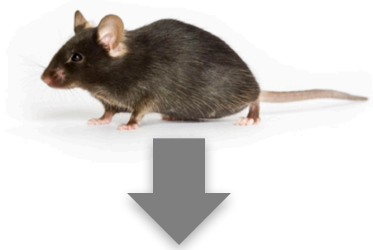


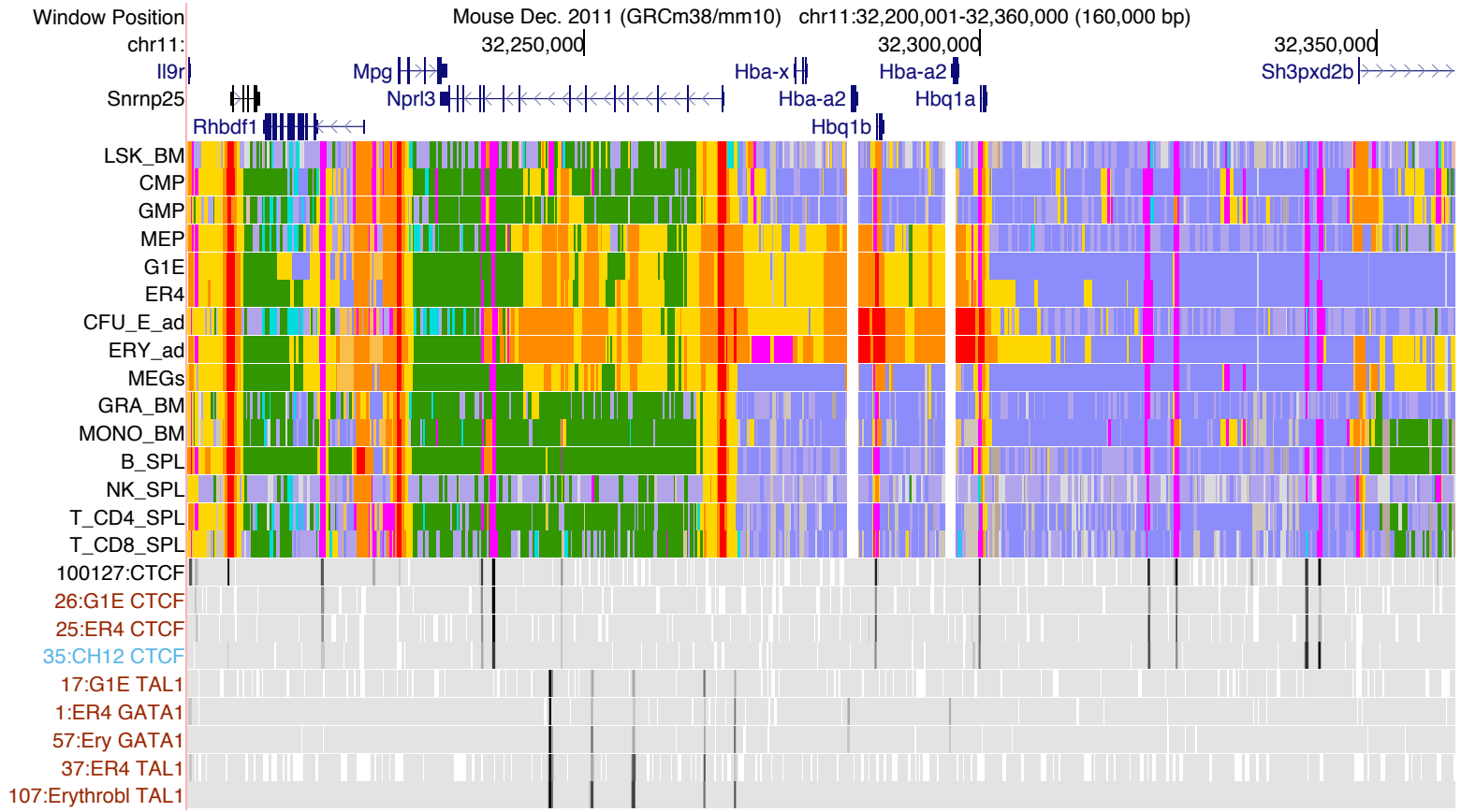
Metadata helps facilitate reproducibility



If your foundation is not solid….

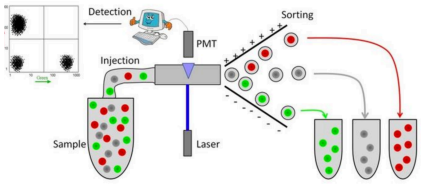
From mouse to machine (learning!)

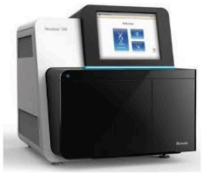


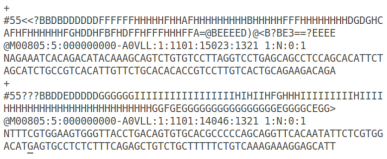


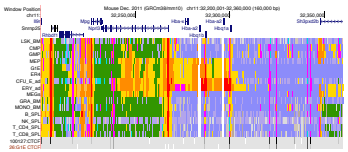
Yu Zhang

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

Belinda Giardine

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

Belinda Giardine

Hardison lab database

Hlab: Hardison lab database

Run detail page

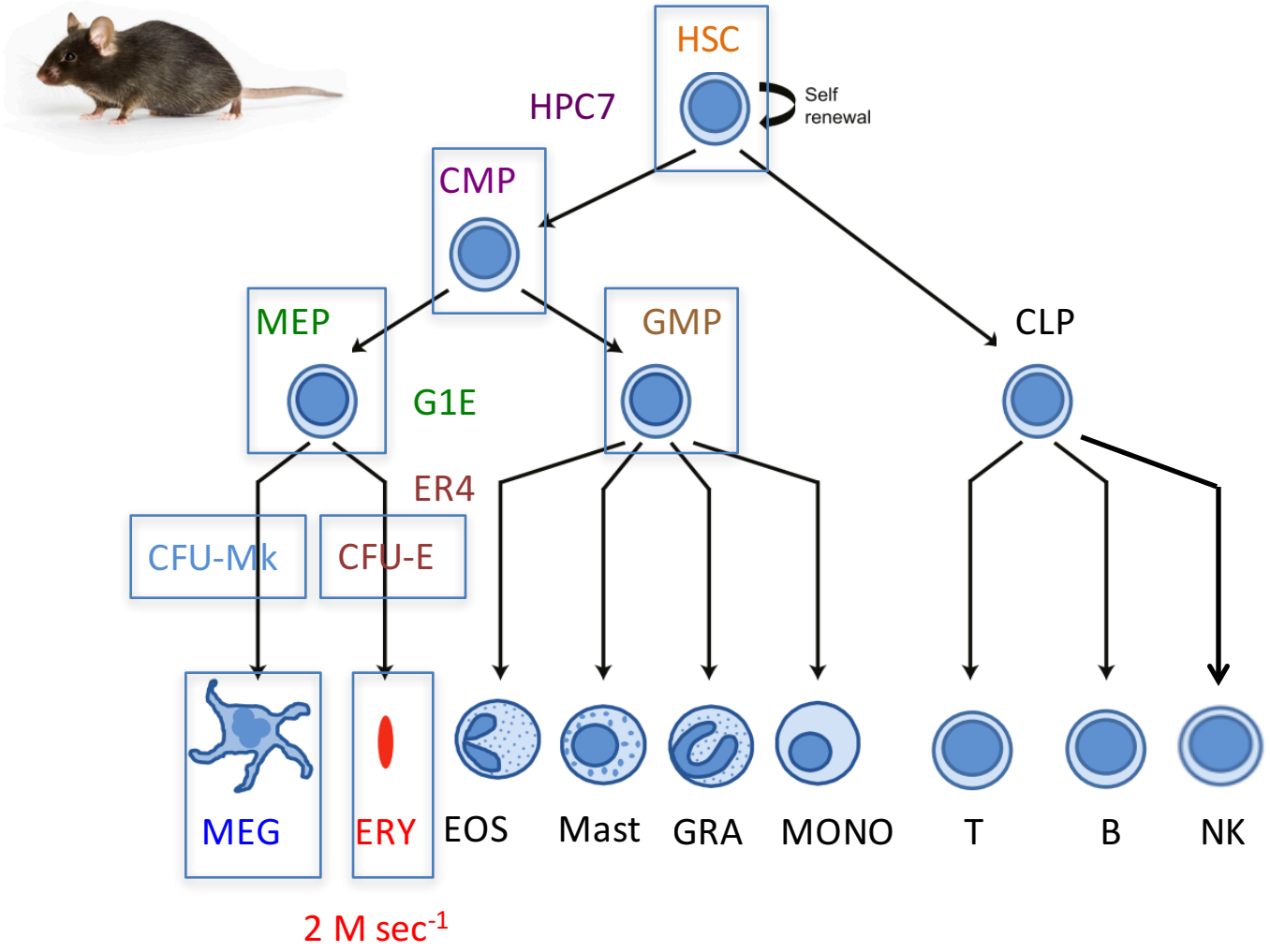
Run 49: 12/09/2014
Folder: 141209_SN407_0348BC5WJ9ACXX
Platform: HiSeq2000
Software: HSC1.5.15.1/RTA1.13.48
Recipe: 48|7
Read length: 48
Reads: single
Number of lanes: 8
Metrics: click here

Lanes:

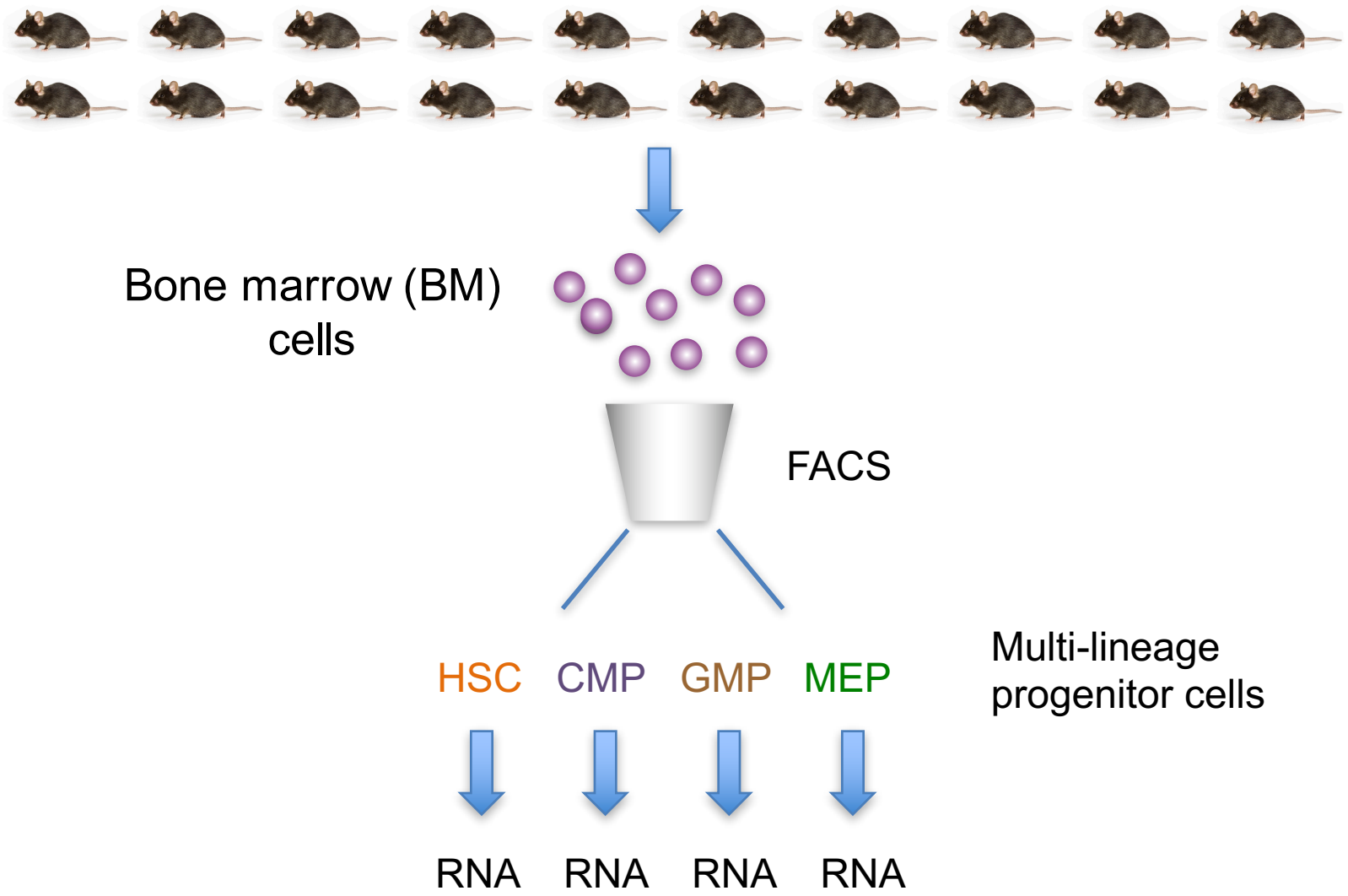
Lane	Libraries	#clusters	%PF	%Q30	Gb	Reads PF
1	708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725	934	83	93	10.8	211
2	708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725	930	83	93	10.9	213
3	708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725	935	83	93	11.0	214
4	708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725	936	83	93	10.9	213
5	708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725	929	84	93	11.0	215
6	670,671,672,673,674,692,693,694,695,696,702,703,704,705,706,707,732,733,734,735,736	830	90	95	10.8	230
7	670,671,672,673,674,692,693,694,695,696,702,703,704,705,706,707,732,733,734,735,736	842	90	95	11.0	233
8	670,671,672,673,674,692,693,694,695,696,702,703,704,705,706,707,732,733,734,735,736	835	90	95	10.8	231

Belinda Giardine

Simplified scheme of hematopoiesis



Biosamples for RNA-seq hematopioetic progenitors



Bodine lab

Biosamples for RNA-seq hematopioetic progenitors

= 20 biosamples

Bone marrow (BM) cells

species
strain
gender
age
source

FACS

HSC CMP GMP MEP

Multi-lineage progenitor cells

RNA RNA RNA RNA

Bodine lab

Biosamples for RNA-seq hematopioetic progenitors

Bone marrow (BM) cells

= 1 biosample (pool)

FACS

species
strain
gender
age
Source
mice
tissue

HSC CMP GMP MEP

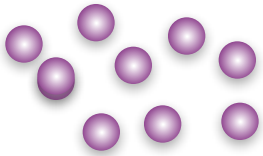
RNA RNA RNA RNA

Multi-lineage progenitor cells

Bodine lab

Biosamples for RNA-seq hematopioetic progenitors

Bone marrow (BM) cells



FACS

cell isolation protocol
sort markers
Ab sources

= 4 biosamples

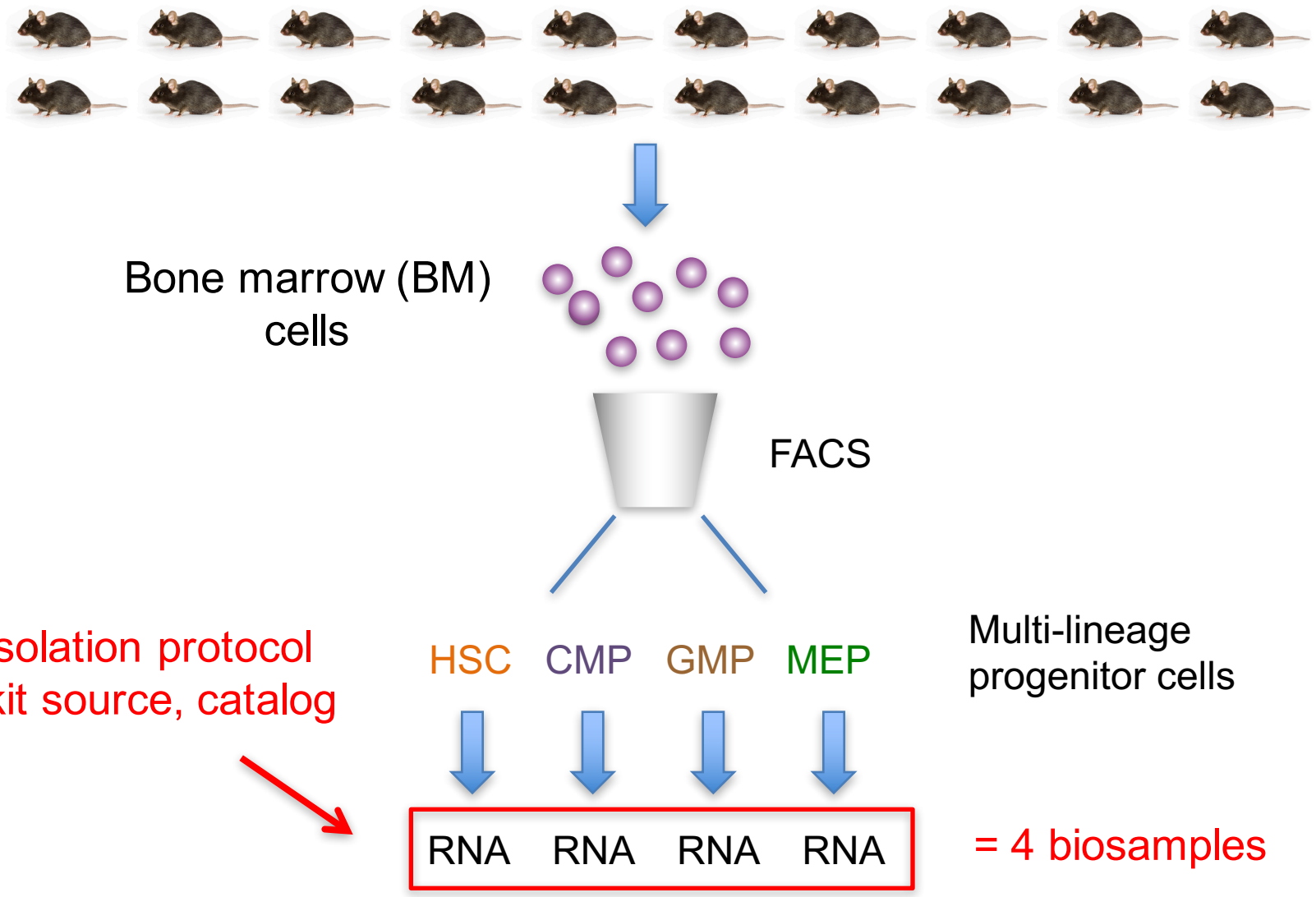
HSC CMP GMP MEP

Multi-lineage progenitor cells

Not everyone uses same sort markers!

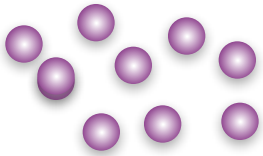
RNA RNA RNA RNA

Biosamples for RNA-seq hematopioetic progenitors

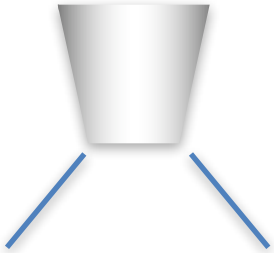


Biosamples for RNA-seq hematopioetic progenitors

Bone marrow (BM) cells



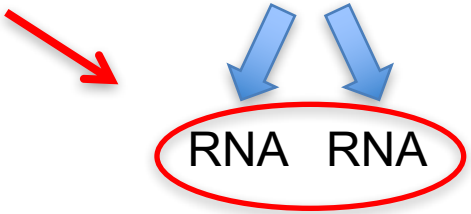
FACS



What if HSCs are split into two groups for RNA?

HSC CMP GMP MEP

Multi-lineage progenitor cells



Biological replicate
or
Technical replicate?

Biosample: Hematopioetic stem cell (HSC)

https://www.encodeproject.org/biosamples/ENCBS383DPY/

ENCBS383DPY / stem cell

Status: **released**

Summary		Attribution	
Term name:	hematopoietic stem cell	**Lab:**	Ross Hardison, PennState
Term ID:	CL:0000037	**Award PI:**	Richard Myers, HAIB
Summary:	*Mus musculus* strain C57BL/6J hematopoietic stem cell male adult (5-6 weeks)	**Submitted by:**	Belinda Giardine
Description:	Hematopoietic stem cell (HSC)	**Source:**	David Bodine
Life stage:	Adult	**Project:**	ENCODE
Age:	5-6 week	**Date obtained:**	2014-10-07
Separated from biosample:	ENCBS793NQI		

Strain information

Accession: ENCDO072AAA

Aliases: encode:C57BL6J, alexander-hoffmann:donor_of_macrophage, encode:Generic-C57BL6

Species: *Mus musculus*

Sex: Male

Strain reference: jaxmice.jax.org

Strain background: C57BL/6

Strain name: C57BL/6J

External resources: GEO:SAMN04284198 MGI.D:C57BL

Experiments using this biosample

Accession	Assay	Biosample term name	Target	Description	Lab
ENCSR085AJX	RNA-seq	hematopoietic stem cell		PSU mouse HSC 100ng rRNA-depleted RNA-seq via ScriptSeq	Ross Hardison, PennState

Displaying 1 of 1

Documents

Cell Isolation Protocol

Description: EH-Progenitor Isolation from Bone Marrow

EH-Progenitor Isolation from Bone Marr

General Protocol

Description: Illumina ScriptSeq library prep protocol- Hardison lab

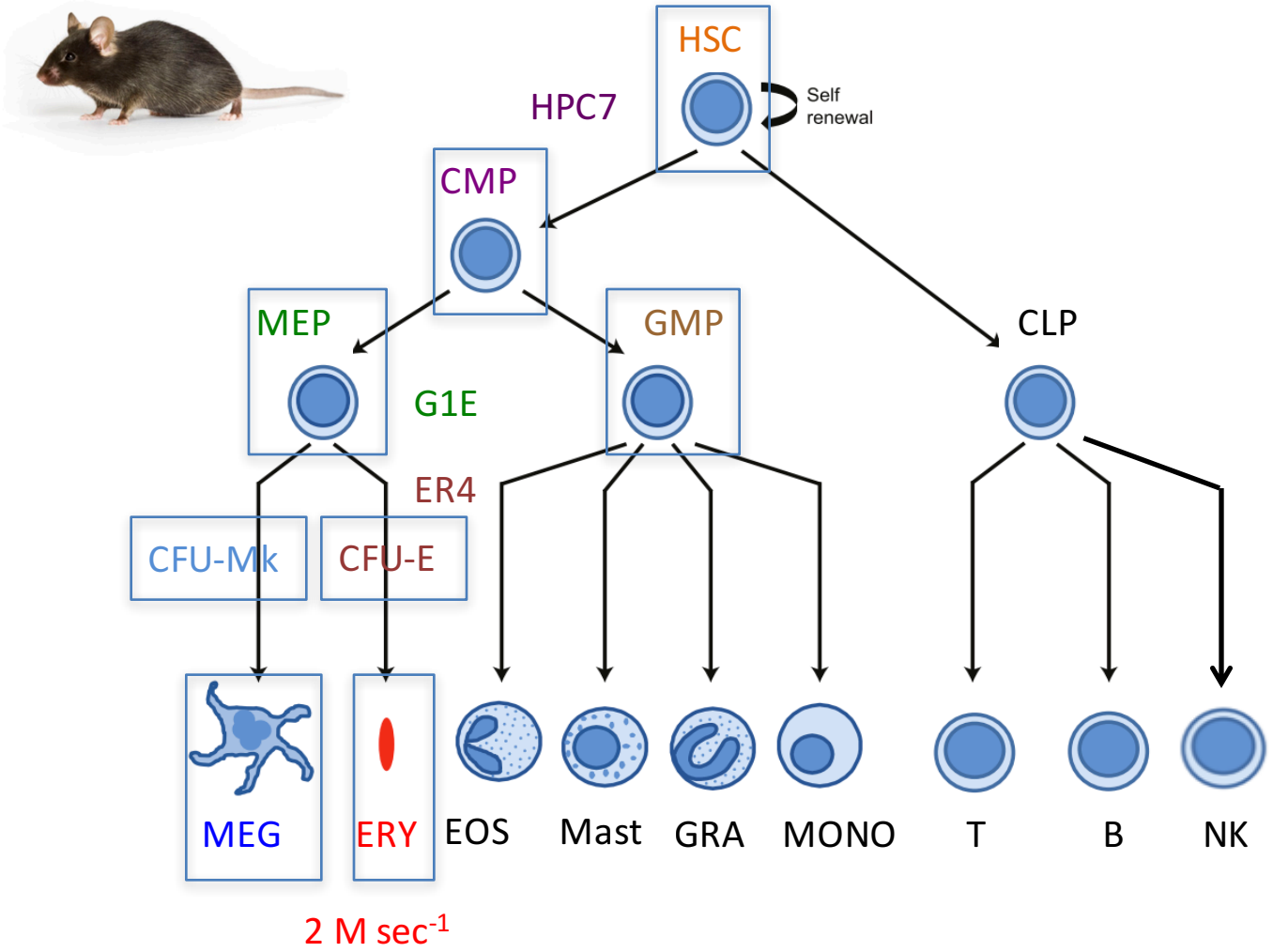
scriptseq-complete-kit-human-mouse-r

Extraction Protocol

Description: Total RNA purification for PureLink® RNA Mini Kit (Ambion)- Hardison lab

purelink_rna_mini_kit_man.pdf

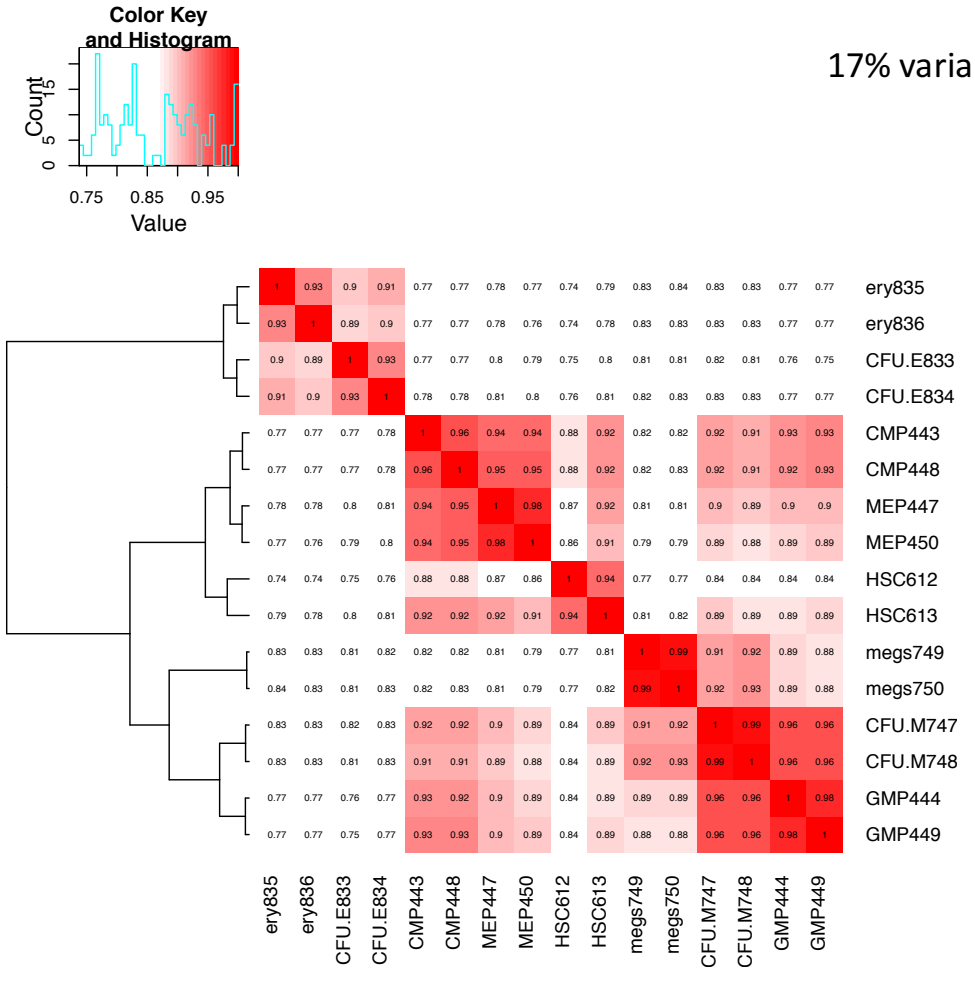
RNA-seq of hematopioetic progenitors

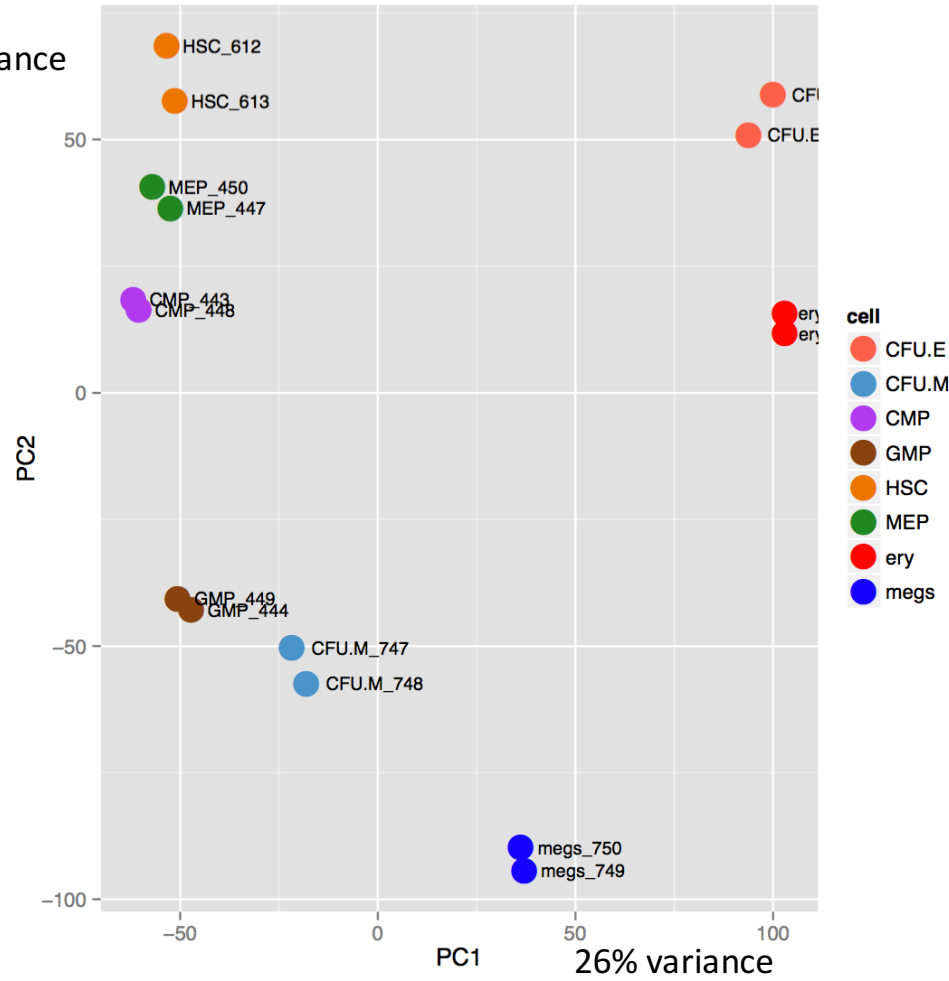


Totalscript RNA-seq data

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

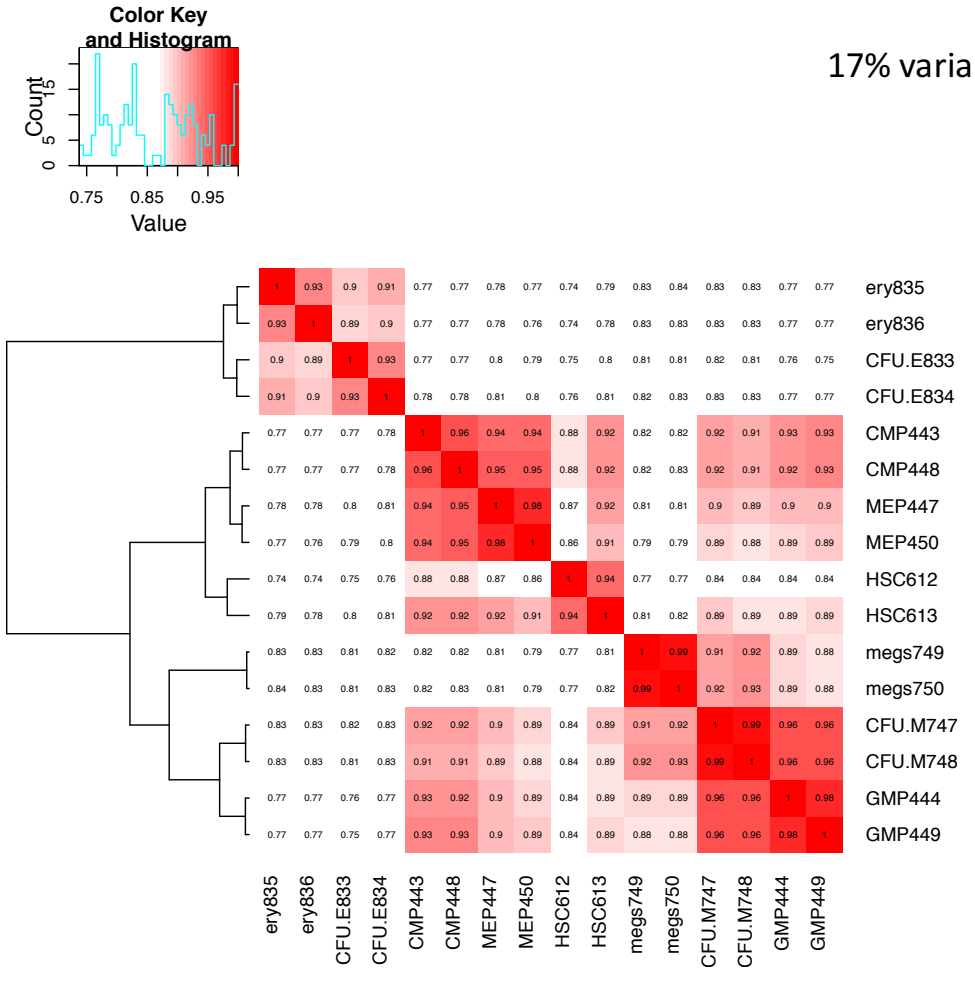
Erythroid cells separate from others in Totalscript

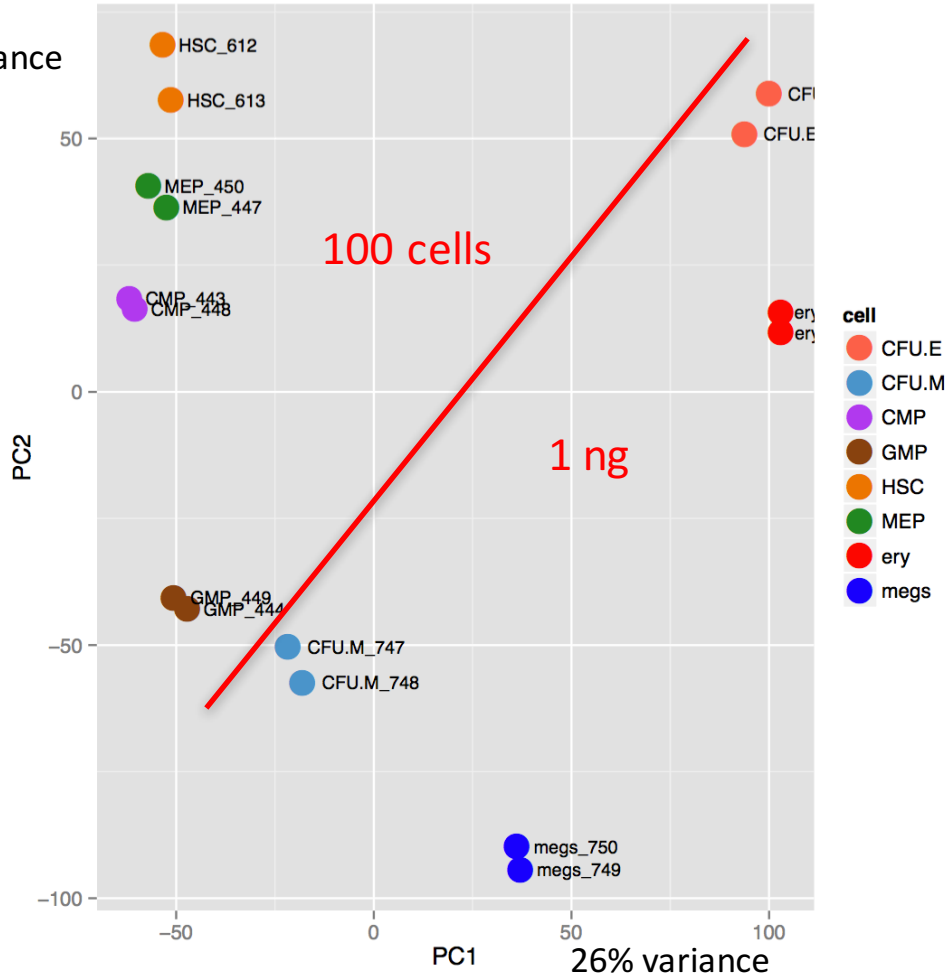




July 15, 2015

Erythroid cells separate from others in Totalscript



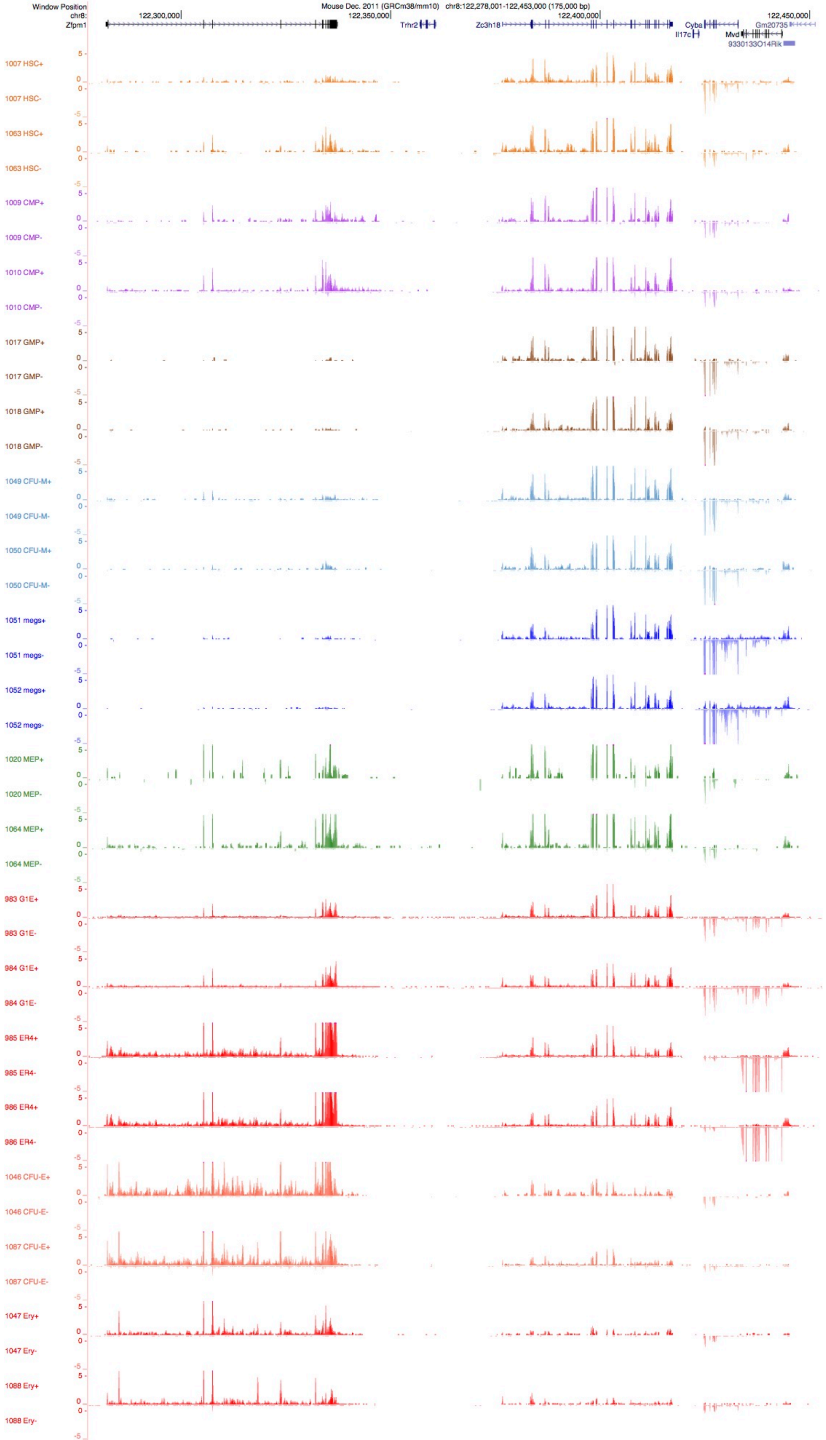


July 15, 2015

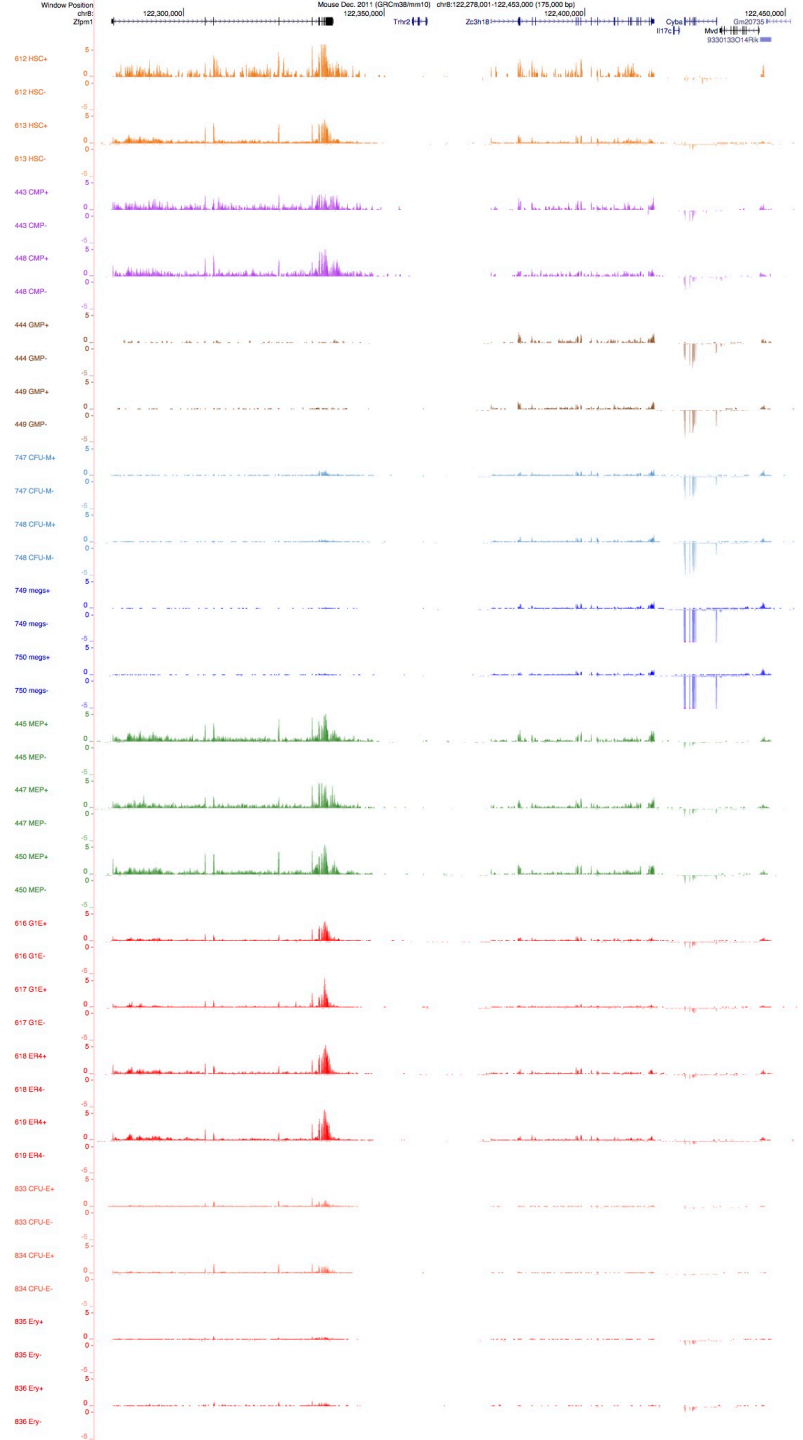
Scriptseq RNA-seq data

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

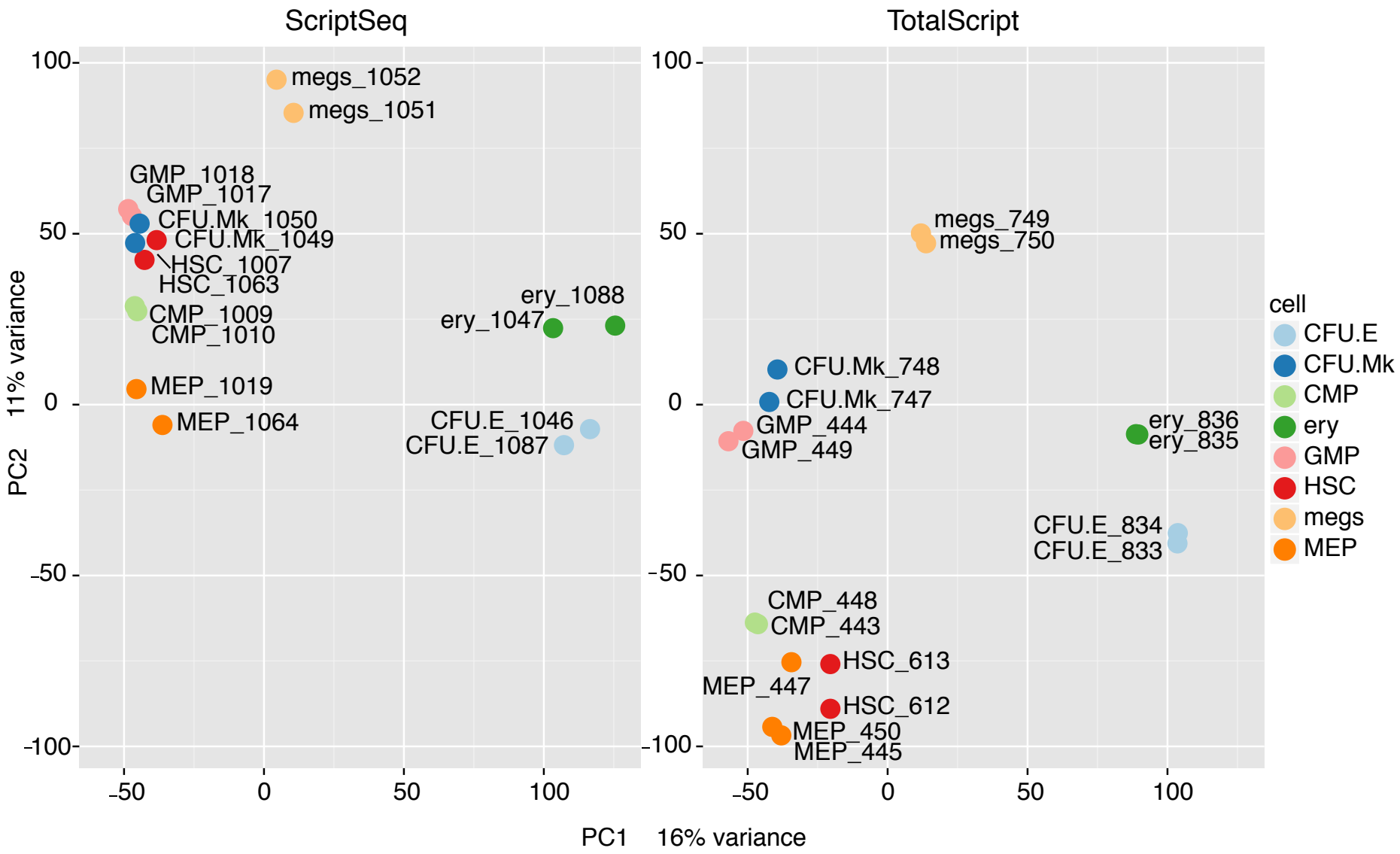
Scriptseq



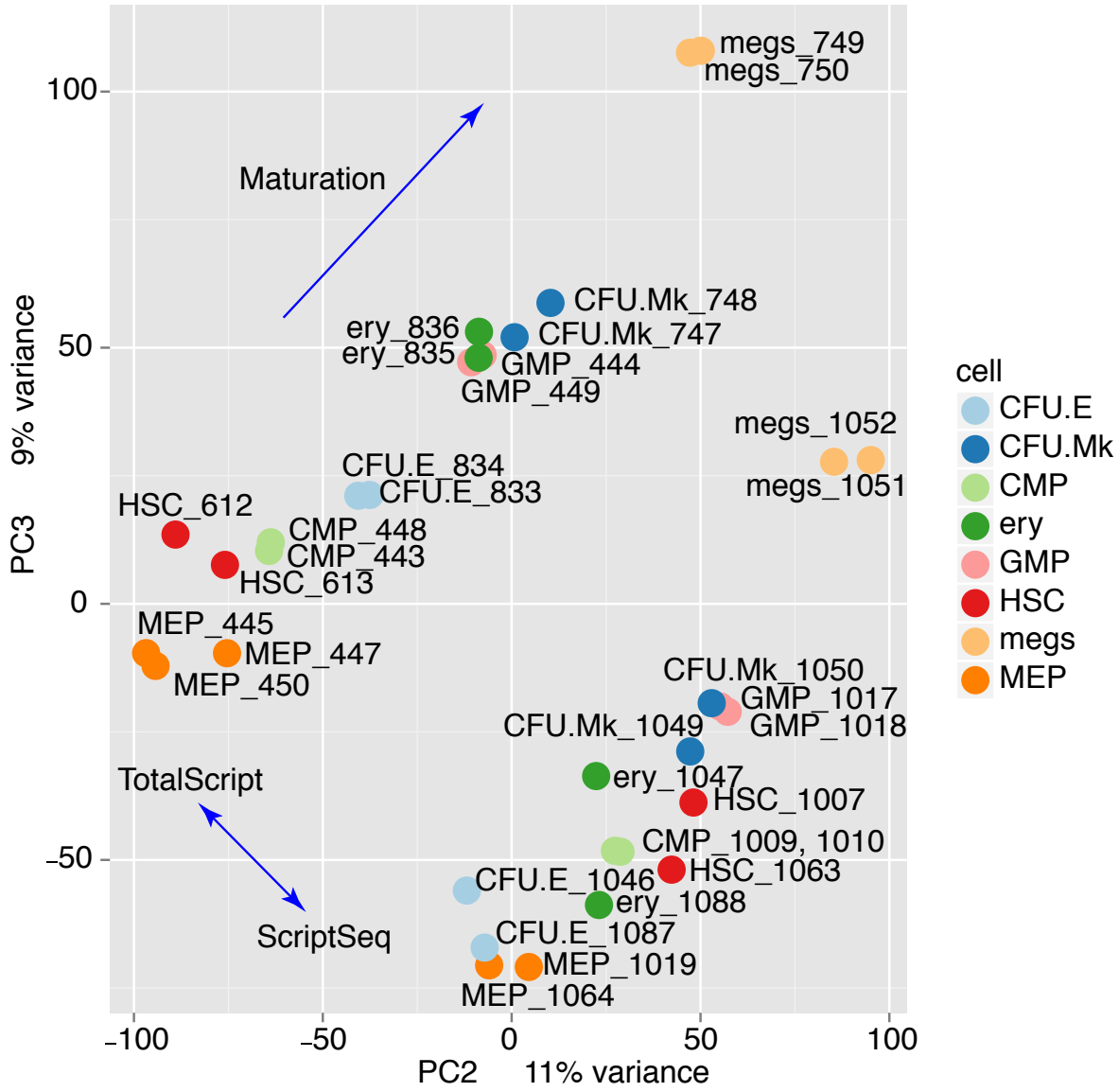
Totalscript



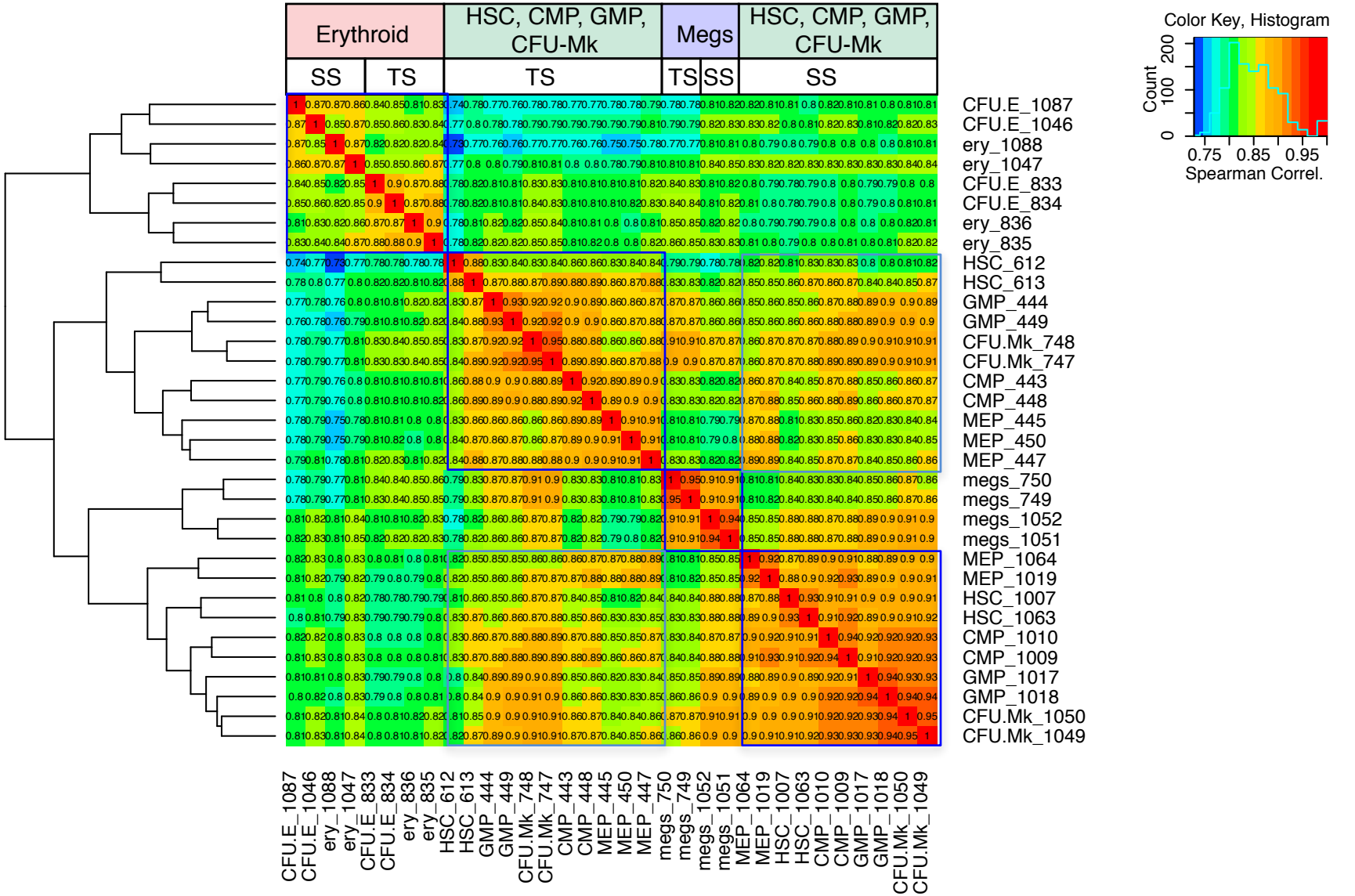
Scriptseq vs. Totalscript:
Similar relationships between cell types for PC1 vs PC2



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



Hierarchical clustering shows similar trends



ENCODE audits alert user to metadata concerns

https://www.encodeproject.org/data-standards/audits/

System of audits or flags provide additional information to research community about quality of the data

Flags may indicate:

- error in experimental metadata,
- data itself does not meet some aspect of the consortium standards

Color of the flags indicate severity of the problem

Flags, flags, everywhere!

https://www.encodeproject.org/matrix/?type=Experiment

Audit category:

extremely low spot score	244
extremely low read depth	108
control extremely low read depth	69
extremely low read length	36
inconsistent replicate	3

+ See more...

Audit category:

low read length	3000
mild to moderate bottlenecking	2701
low read depth	1974
moderate library complexity	1233
inconsistent platforms	1217

+ See more...

Audit category:

insufficient read depth	1555
control insufficient read depth	877
missing controlled_by	726
insufficient read length	482
partially characterized antibody	416

+ See more...

Audit category:

missing derived_from	3702
experiment not submitted to GEO	3330
biological replicates with identical biosample	1643
mismatched file status	1310
NTR assay	1283

+ See more...

Acknowledgements

Hardison lab (PSU)
Ross Hardison
Belinda Giardine
Amber Miller

Former lab members
Maria Long
Nergiz Dogan
Weisheng Wu
Christapher Morrisey
Tejaswini Mishra
Yong Cheng
Marta Bryska-Bishop
Swathi Kumar

Bodine lab (NHGRI)
David Bodine
Elisabeth Heuston
Stacie Anderson

Zhang lab (PSU)
Yu Zhang
Lin An

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