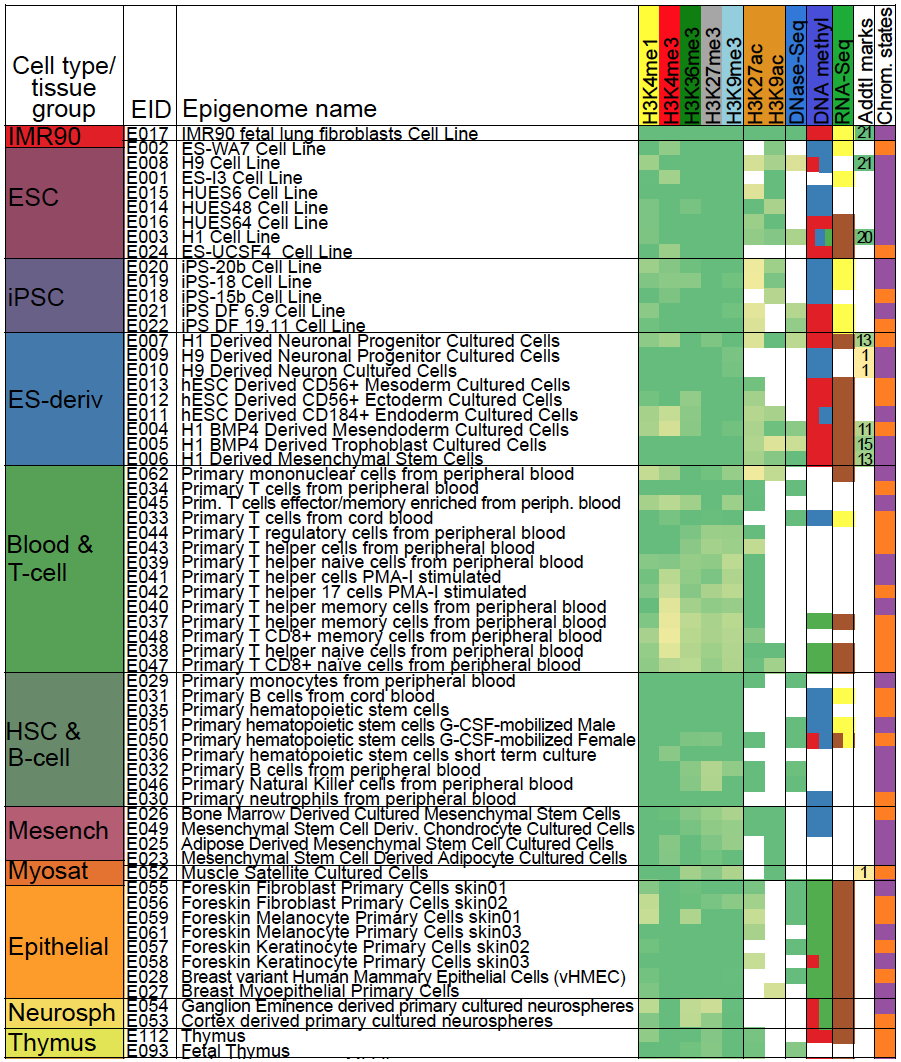
ENCODE3 chromatin state calls based on Roadmap Epigenomics ChromHMM models

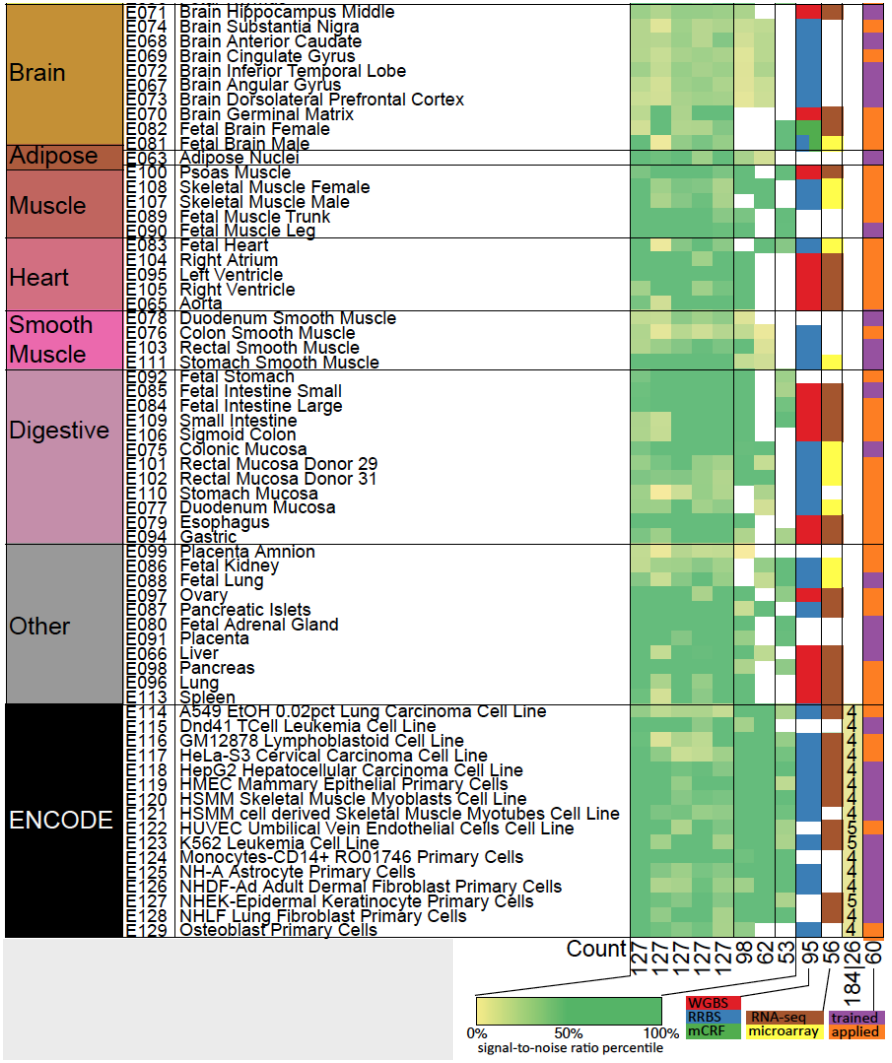
WM20151112

Goal: chromatin state calls for ENCODE3 data

- The latest ENCODE3 ChIP-seq data freeze contained a limited amount of new data. Instead of building new chromatin state models on these data, we may be able to use existing models.
- The Roadmap Epigenomics Project has released chromatin state models across 127 epigenomes, spanning 100+ cell types.
- These models can be applied to incoming ENCODE3 data, provided that these data are processed identically to what was used for Roadmap.
- Benefits: validate ENCODE3 cell types against large body of reference cell types, fairly rapid general QC, identify novel regions.

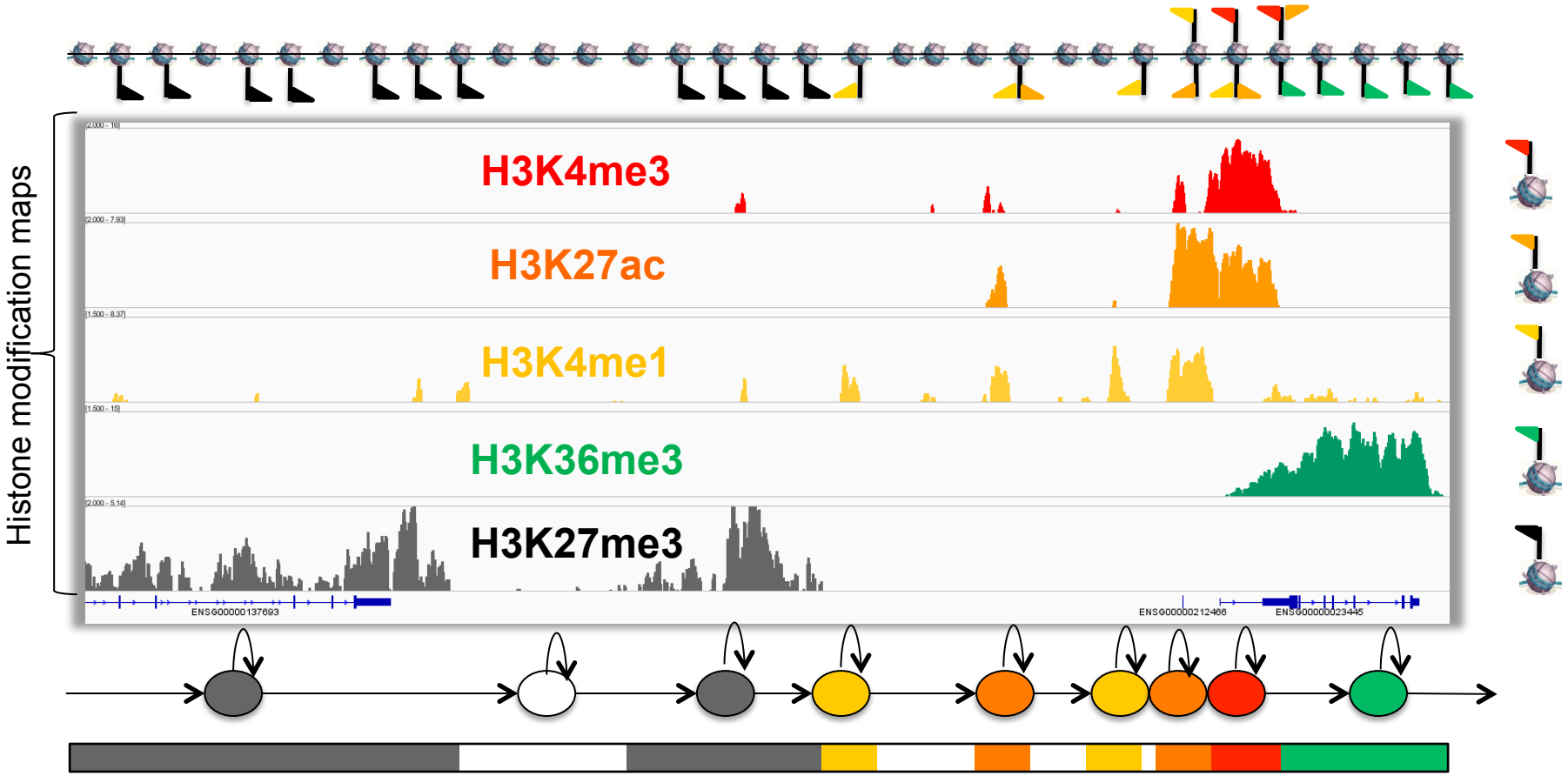
The Roadmap Epigenomics Project

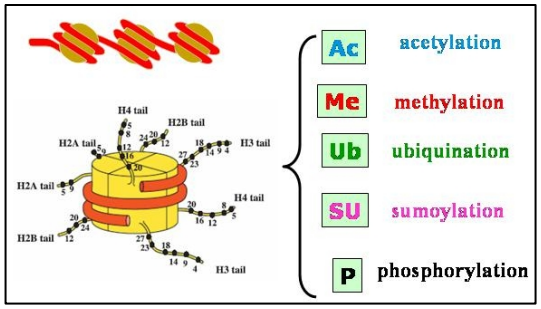




Performed a wide variety of epigenomics assays to increase our understanding of the role of epigenetics in gene regulation

Learning genome-wide chromatin state maps





www.integratedhealthcare.eu

Chromatin States: Hidden Markov model

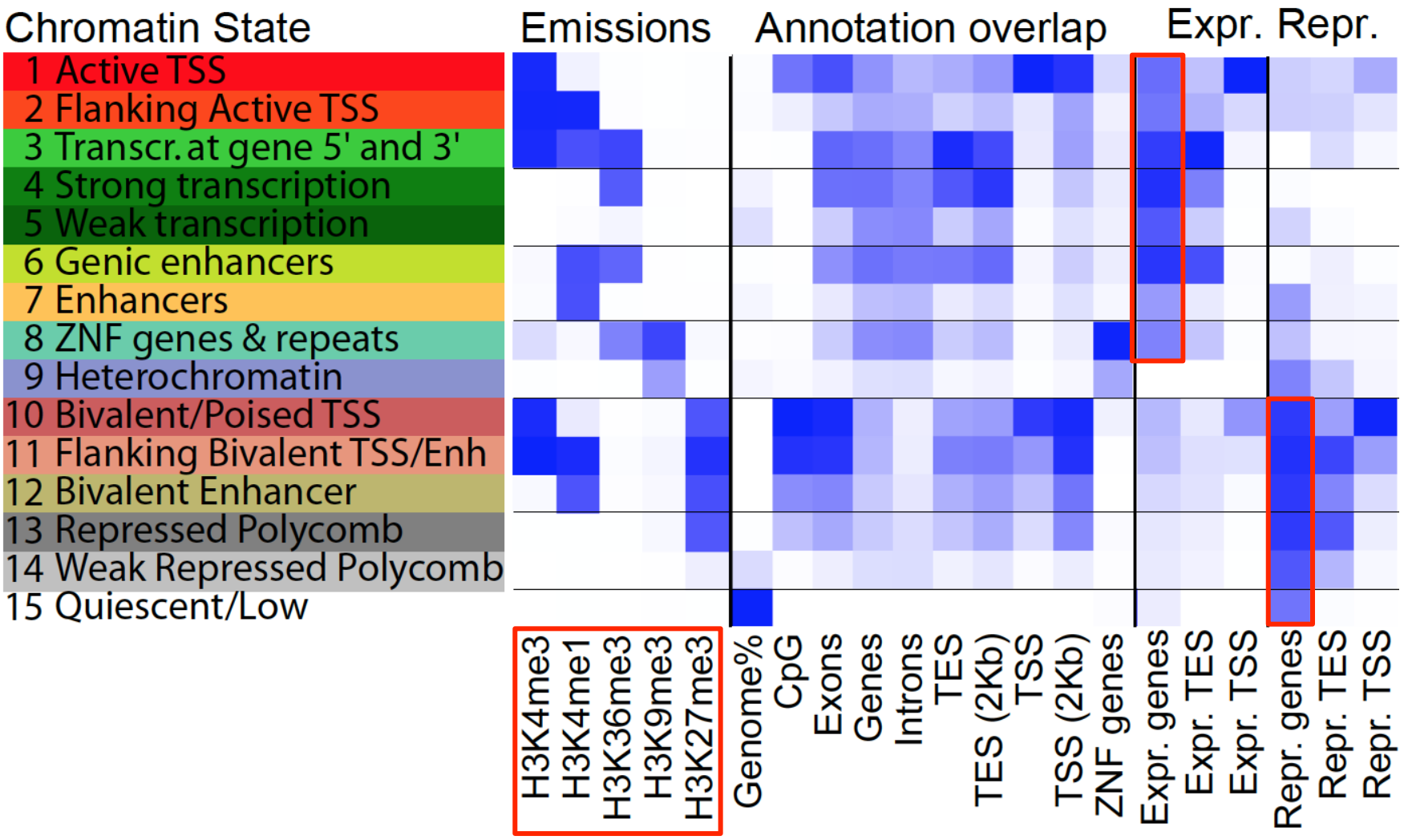
ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

Nature Methods **9**, 215–216 (2012) | doi:10.1038/nmeth.1906

Slide: Anshul Kundaje

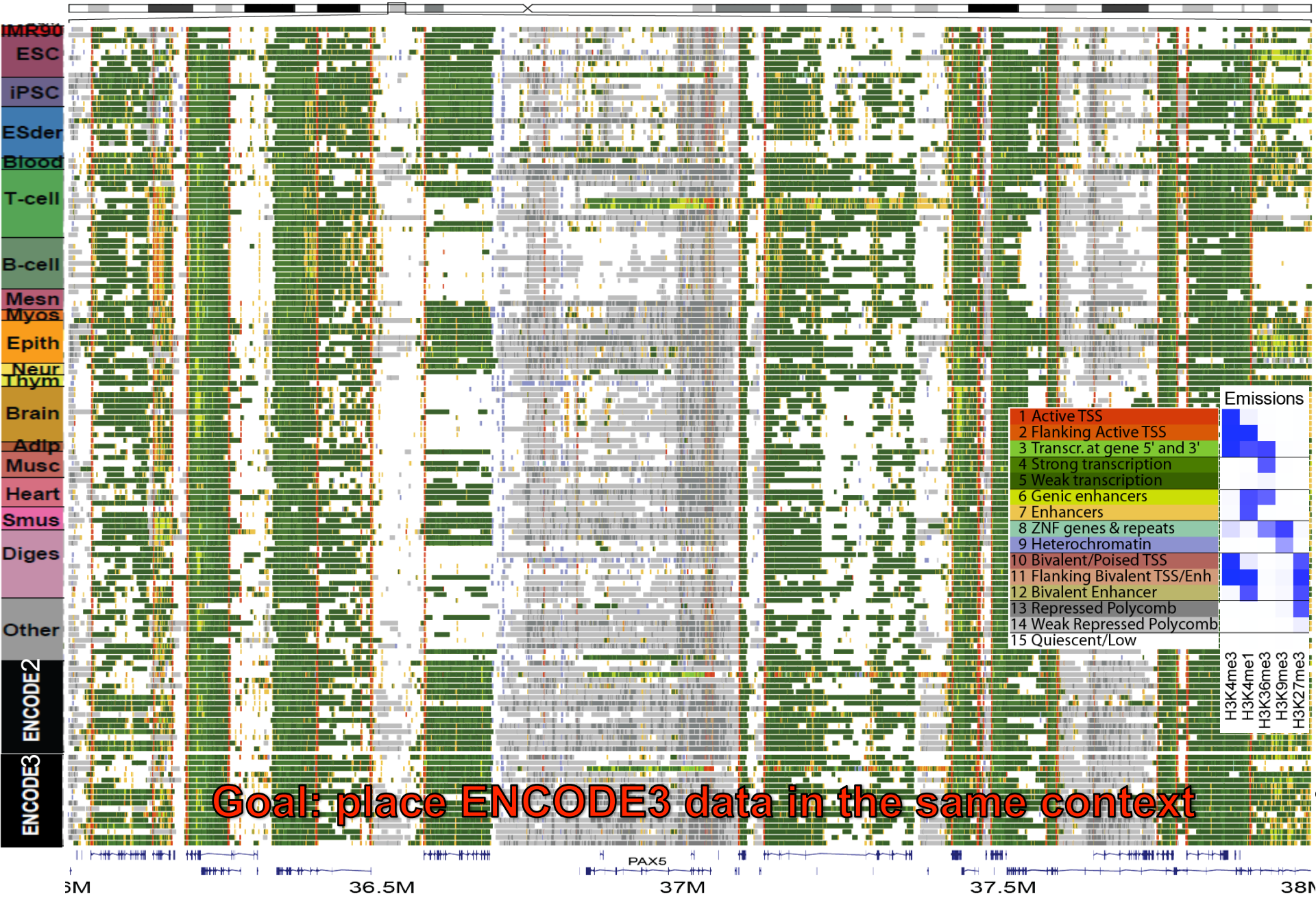
The 5 core histone modification marks have been used to build a ChromHMM model with 15 chromatin states



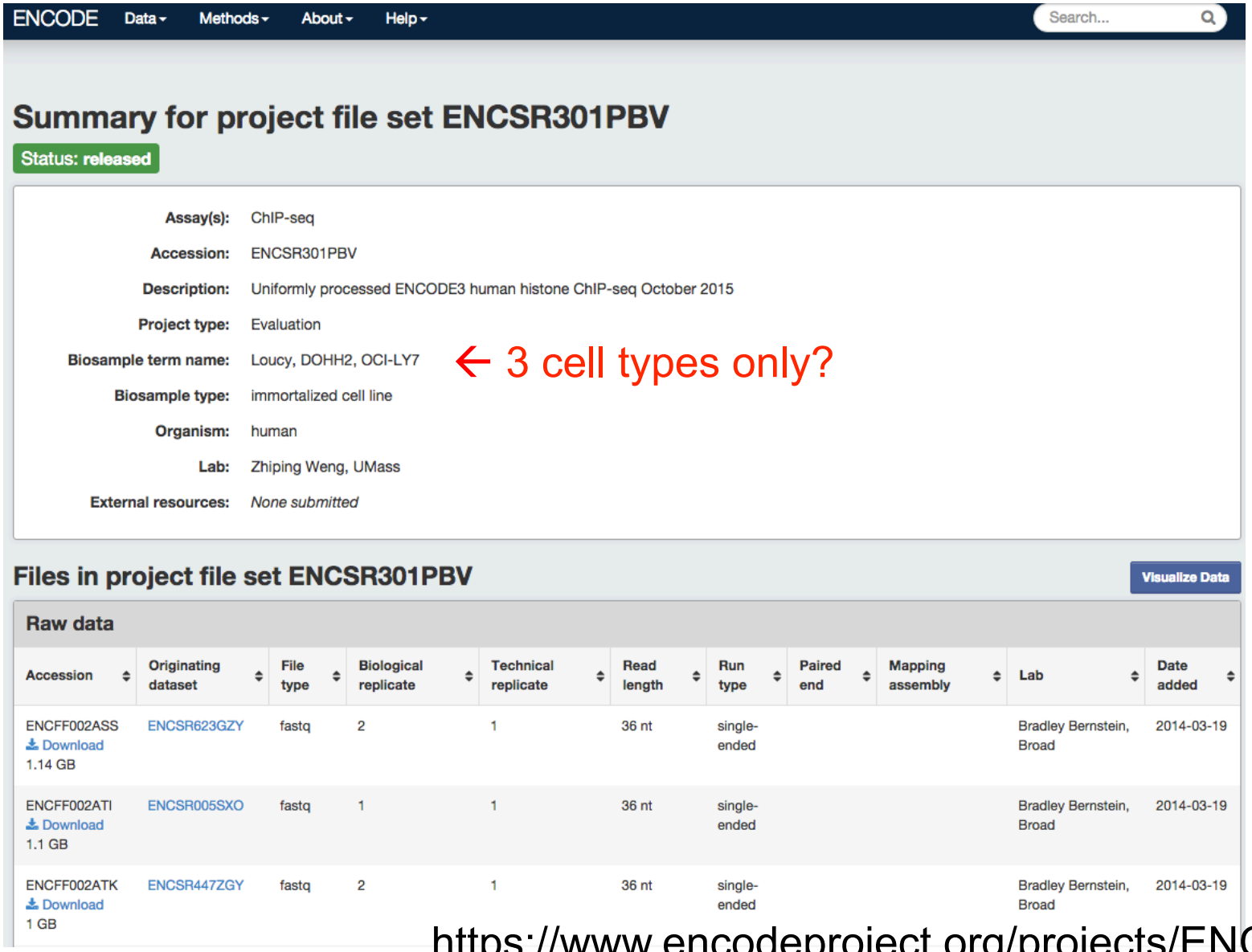
States make a clear distinction based on transcriptional activity

Anshul Kundaje, Jason Ernst

Chromatin state annotations across 127 epigenomes



ENCODE3 human ChIP-seq data on the DCC portal



etcetera

https://www.encodeproject.org/projects/ENCSR301PBV/

Navigated to via:
https://www.encodeproject.org/search/?type=Project&lab.title=Zhiping+Weng%2C+UMass&status=released

Instead, I'm working with 19 cell types

Epiblast	H3K27ac	H9	/targets/H3K27ac-human/	/files/ENCFF149OGB/,/files/ENCFF723ODQ/
PC3 prostate	H3K27me3	PC-3	/targets/H3K27me3-human/	/files/ENCFF281ZEA/,/files/ENCFF299BLB/
H1 derived neurons	H3K27ac	neural cell	/targets/H3K27ac-human/	/files/ENCFF002AZH/,/files/ENCFF002AZW/
Neuroectoderm	H3K4me3	neuroepithelial stem cell	/targets/H3K4me3-human/	/files/ENCFF149VFS/,/files/ENCFF287UOE/
Oci-Ly-1	H3K4me3	Oci-Ly-1	/targets/H3K4me3-human/	/files/ENCFF002AYK/,/files/ENCFF002BBL/
H1 derived neurons	H3K4me3	neural cell	/targets/H3K4me3-human/	/files/ENCFF002AUV/,/files/ENCFF002AYB/
Neuroectoderm	H3K36me3	neuroepithelial stem cell	/targets/H3K36me3-human/	/files/ENCFF860YOG/,/files/ENCFF755RTB/
Oci-Ly-1	H3K27ac	Oci-Ly-1	/targets/H3K27ac-human/	/files/ENCFF001ZZY/,/files/ENCFF002AUX/
A673	WCE	A673	/targets/Control-human/	/files/ENCFF061PAW/,/files/ENCFF885SBP/
Fibroblast of arm	H3K27me3	fibroblast of arm	/targets/H3K27me3-human/	/files/ENCFF728JQA/,/files/ENCFF013TOE/
Radial Glia - VZ	H3K4me3	radial glial cell	/targets/H3K4me3-human/	/files/ENCFF539YTE/,/files/ENCFF995NAN/
Fibroblast of arm	H3K27ac	fibroblast of arm	/targets/H3K27ac-human/	/files/ENCFF619XSH/,/files/ENCFF756MGK/
Oci-Ly-3	H3K27ac	Oci-Ly-3	/targets/H3K27ac-human/	/files/ENCFF002AUE/,/files/ENCFF002BCD/
PC9	H3K27me3	PC-9	/targets/H3K27me3-human/	/files/ENCFF246YHL/,/files/ENCFF869KWM/
Karpas-422	WCE	Karpas-422	/targets/Control-human/	/files/ENCFF002AXD/,/files/ENCFF002AZZ/
A673	H3K27ac	A673	/targets/H3K27ac-human/	/files/ENCFF452XVM/,/files/ENCFF811OHE/
Oci-Ly-1	H3K4me1	Oci-Ly-1	/targets/H3K4me1-human/	/files/ENCFF002ATO/,/files/ENCFF002AVR/

etcetera

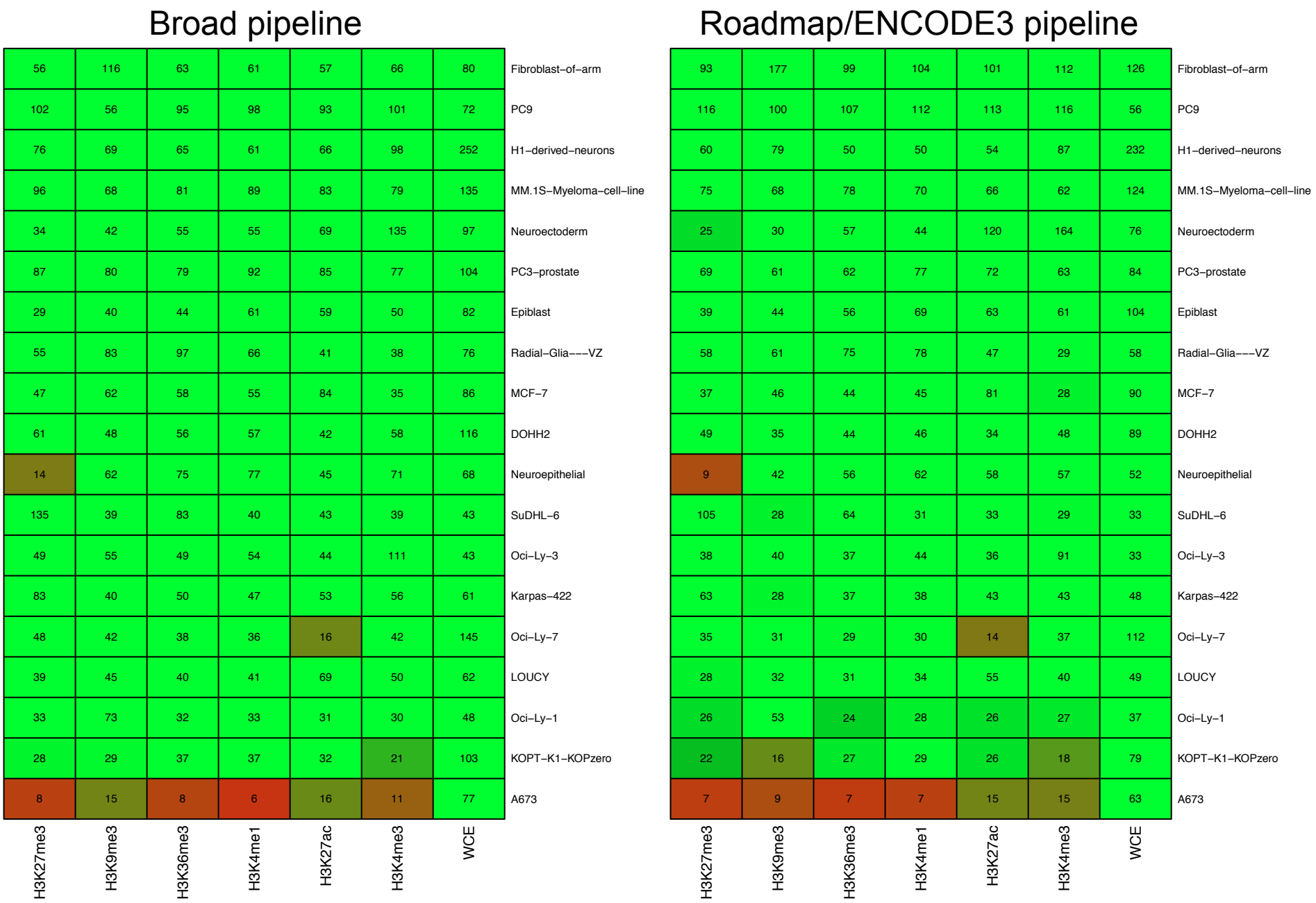
Pointers to data provided by Chuck Epstein & Noam Shoresh (Broad)

ENCODE3 chromatin state calls

- Current ENCODE3 data I'm working with:
  - 19 cell types (next slide)
  - 6 epitopes (K4me1/3, K27me3/ac, K9me3, K36me3) + WCE

- Strategy:
  - process all mapped sequencing reads of ENCODE3 data in a similar manner as done for Roadmap Epigenomics.
  - Remove duplicates & pool data across replicates.
  - Aim for at least 30M reads per cell type / epitope combination. If more, subsample to 30M.
  - Call chromatin states for processed ENCODE3 data, using Roadmap Epigenomics models.

QC: number of (non-duplicate) reads in millions



Processing pipeline

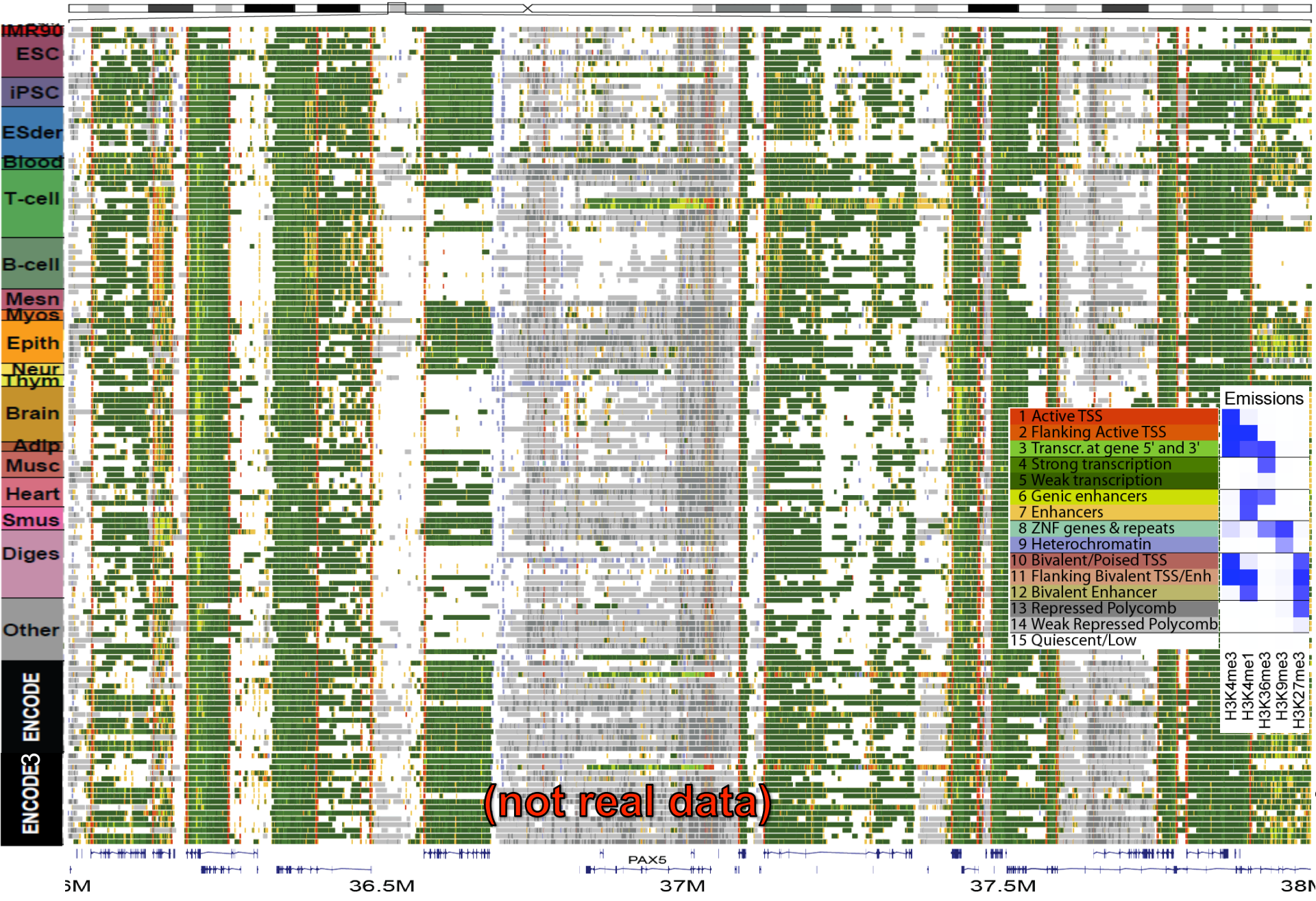
Largely based on code from Anshul and the (draft?) ENCODE3 processing pipeline*

Processing steps:

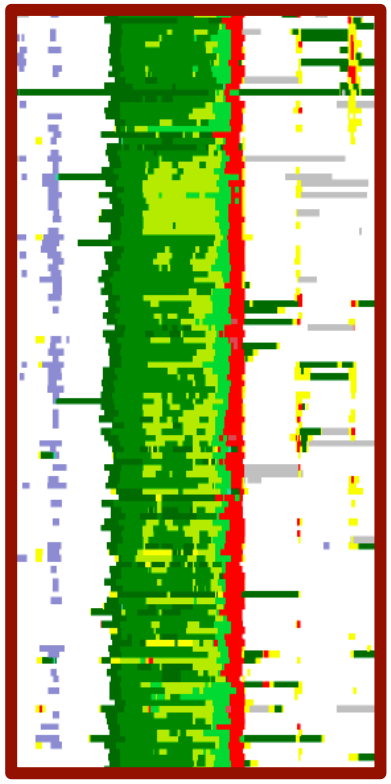
1. Remove un/mis-mapped, low quality reads, mark & remove duplicates, compute library complexity QC, filter for mappability
2. Pool reads across replicates and (if needed) sub-sample to 30M reads, use phantomPeakQualTools to obtain QC info
3. Binarize data using ChromHMM BinarizeBed, create segmentation based on existing Roadmap models and output states per 200bp bins

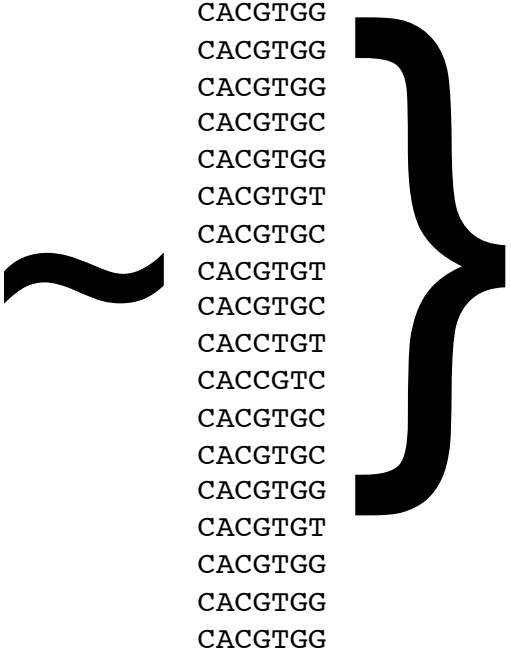
*: https://docs.google.com/document/d/1lG_Rd7fnYgRpSIqrIfuVlAz2dW1VaSQThzk836Db99c/edit#heading=h.9ecc41kilcvq

Bonus: **epilogos** for ENCODE3 data

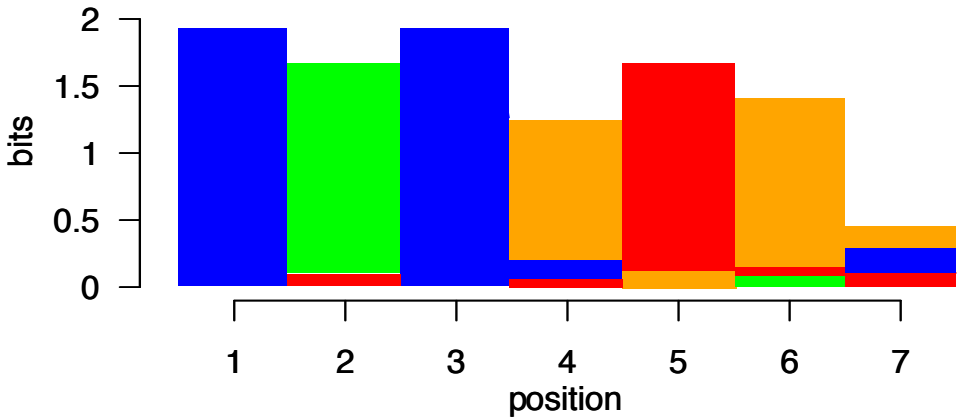


Chromatin state calls across many epigenomes can be viewed as an alignment of sequences with a finite alphabet





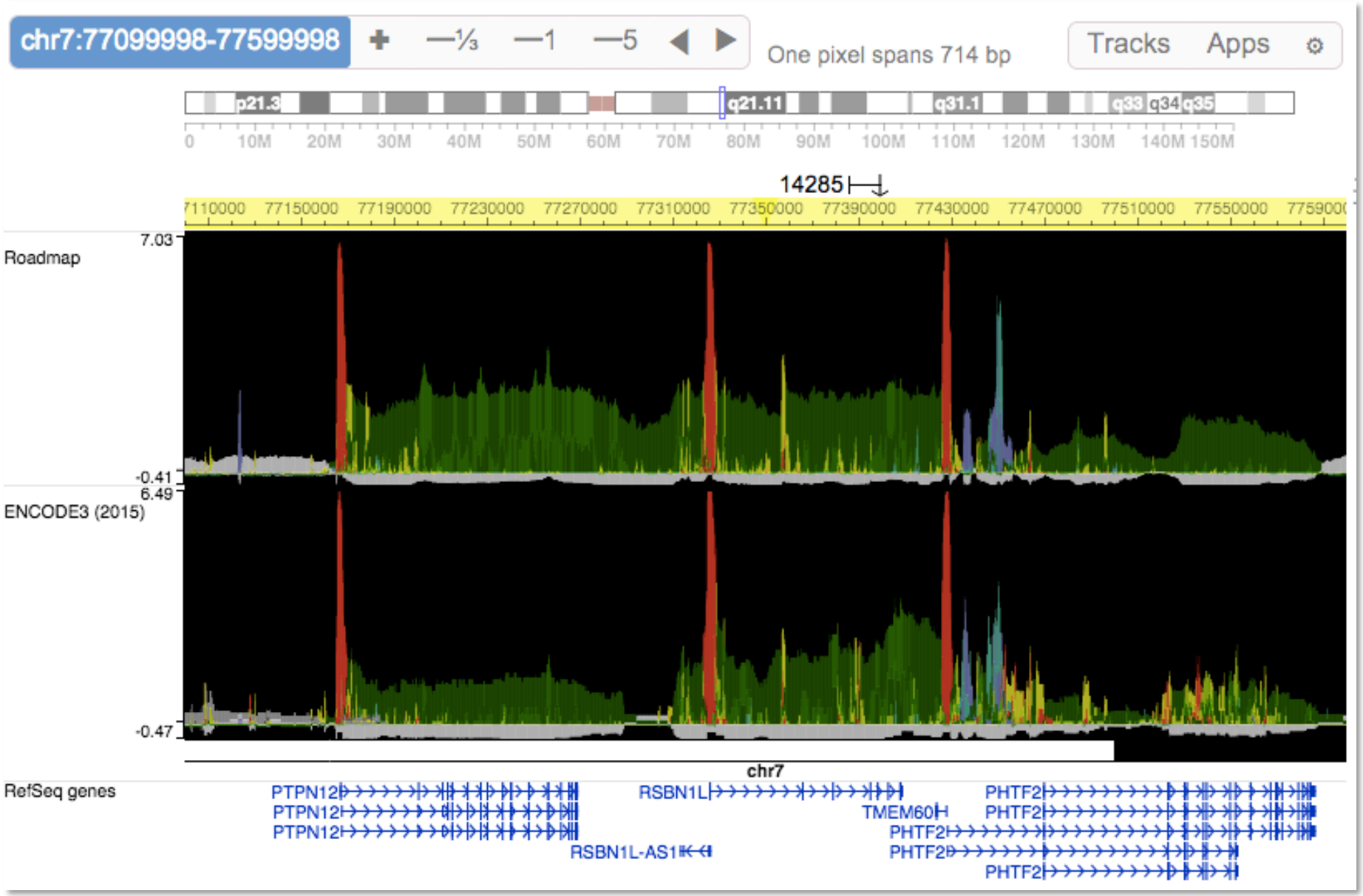
```
CTCTTAT
CACGTGC
CACGTGC
CACGTGG
CACGTGG
CACGTGG
CACGTGC
CACGTGG
CACGTGT
CACGTGC
CACGTGT
CACGTGC
CACCTGT
CACCGTC
CACGTGC
CACGTGC
CACGTGG
CACGTGT
CACGTGG
CACGTGG
CACGTGG
```



There are good ways of modeling such alignments: logos!

Information content of a region, considering background

ENCODE3 data **epilogos**



http://tinyurl.com/encode-vs-roadmap