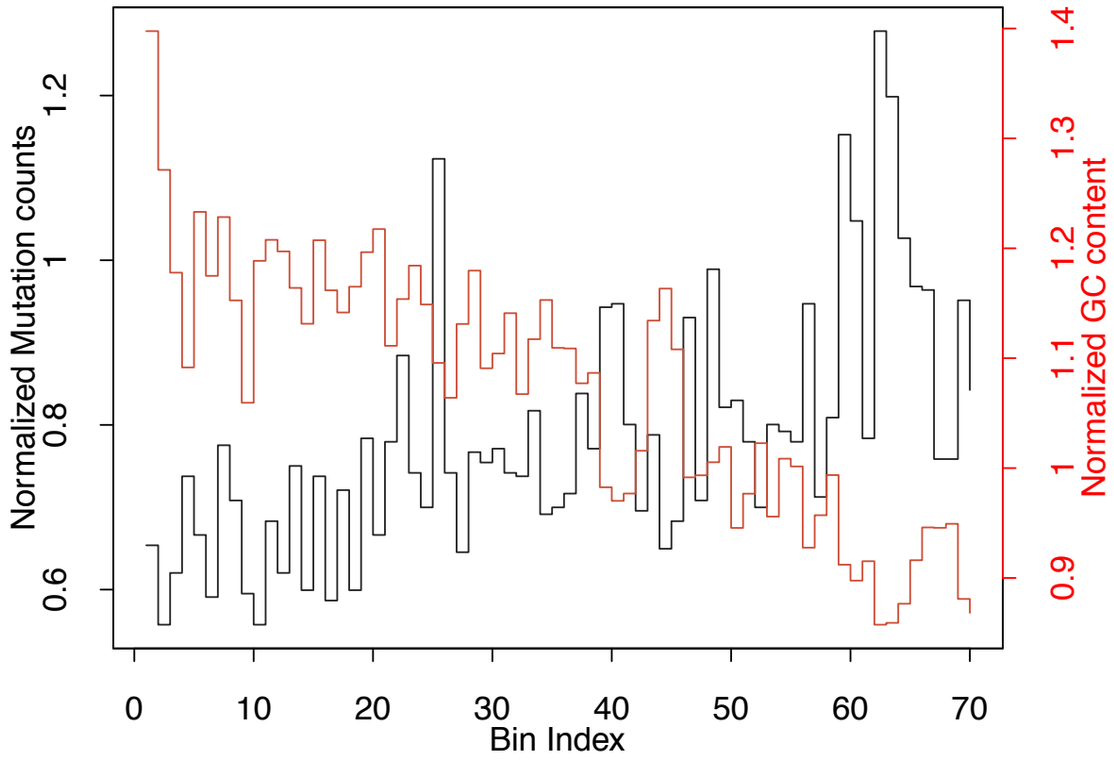
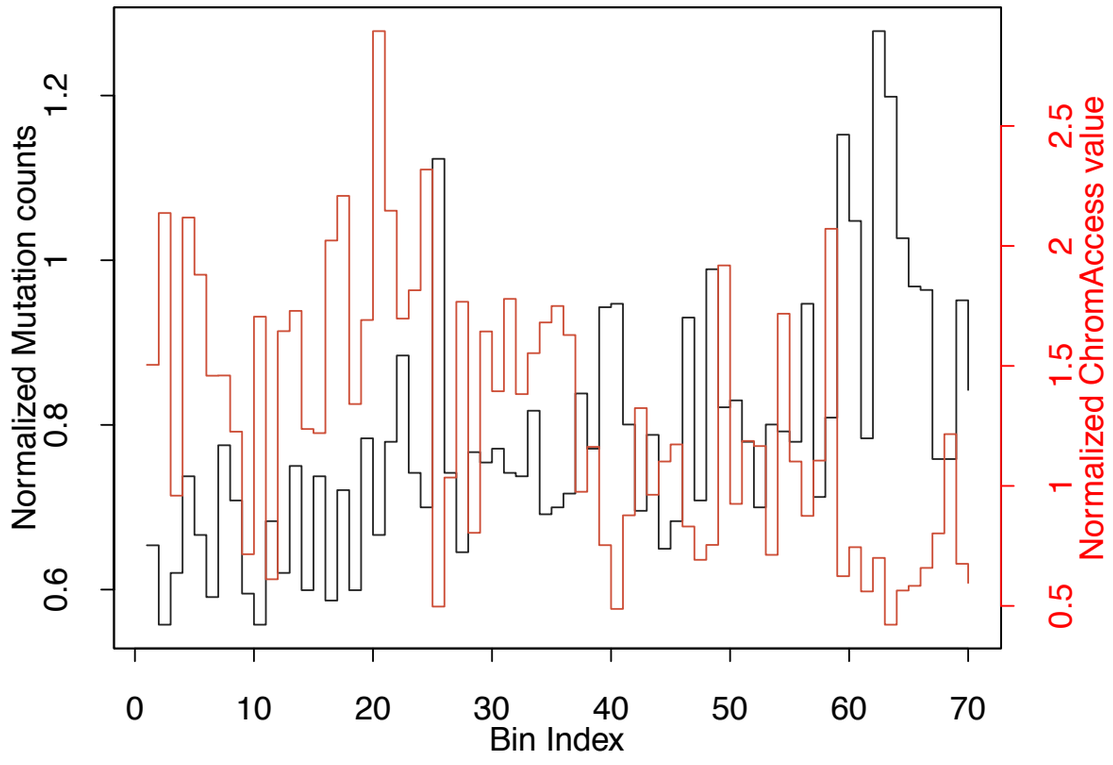
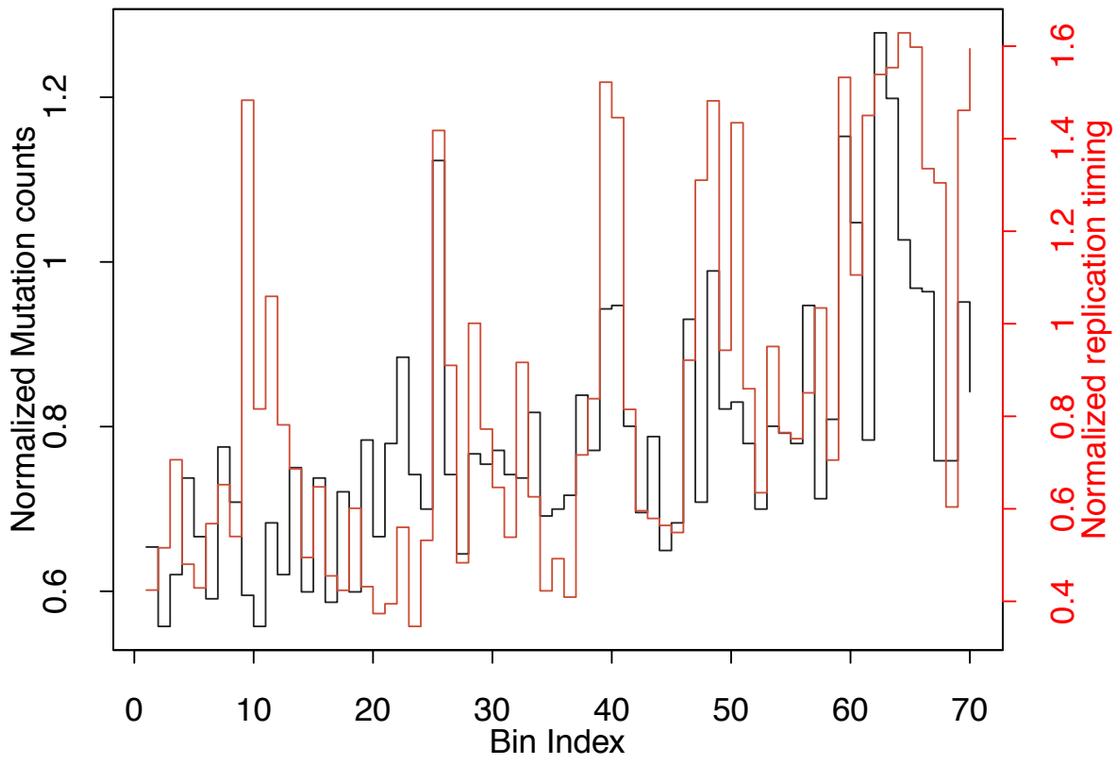
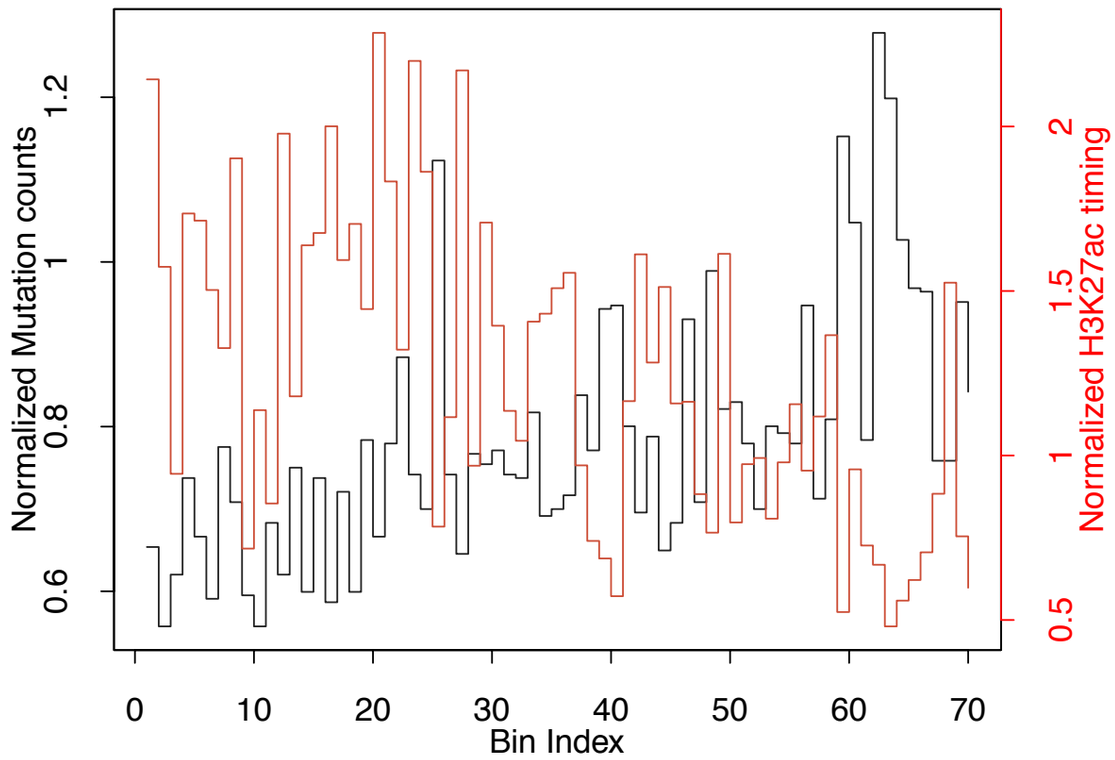
Encode Cancer Signal Matrix

Gerstein Lab
Donghoon Lee









Mutation rate has shown to correlate with various genomic features

Summary of Encode Cancer Data

Experiments

	ChIP-seq (Histone/TFs)	DNase-seq (open chromatin)	FAIRE-seq (open chromatin)	Repli-seq (replication timing)	RNA-seq	RRBS (methylation)
Encode 2 Released	802	61	19	30	107	34
Encode 3 Released	262	32	0	0	59	0
Revoked	15	0	0	0	2	0
Total	1,079	93	19	30	168	34

Last Updated: 2/16/2016

- Total of **73 human cancer cell lines** were identified in Encode (Curated w/ Shirley and Robert)

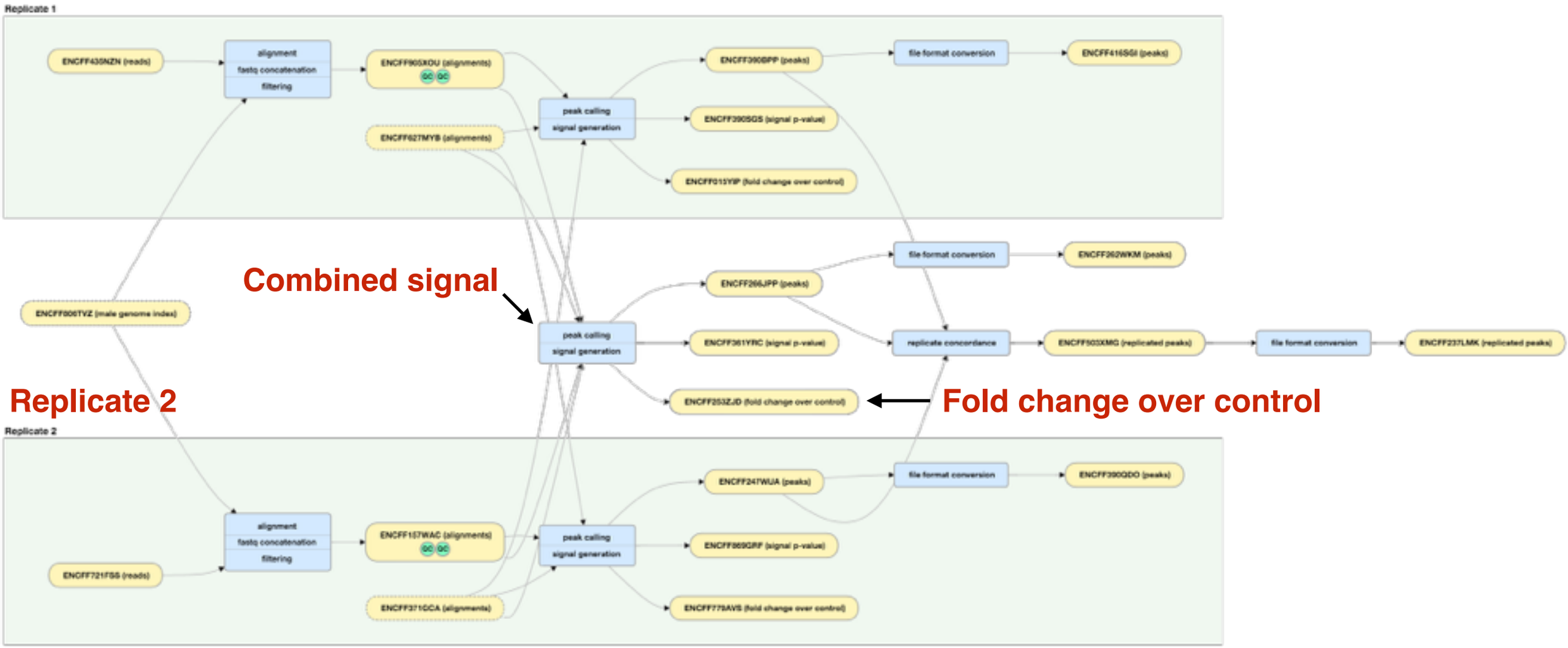
- We can build a covariate matrix by combining signals from various experiments, which can be used for mutation rate correction.

- However, it would be challenging to combine replicates and normalize signals produced from different labs/protocols.

- We need uniformly processed signals, which combine possible replicates and normalize across cell lines for different experiments.

Encode 3 Data Use

Replicate 1



http://www.encodeproject.org/experiments/ENCSR069XHI/

- For Encode 3, ChIP-seq experiments are uniformly processed using the new data processing pipeline.
- Eventually, all of Encode 3, and retroactively, Encode 2 data will be uniformly processed under the same ChIP-seq processing pipeline.
- However, currently, very **limited number of Encode ChIP-seq data has been processed** using this pipeline (from user's perspective) and this uniform processing pipeline **only applies to ChIP-seq data only**.

Encode 2 Uniform Signal

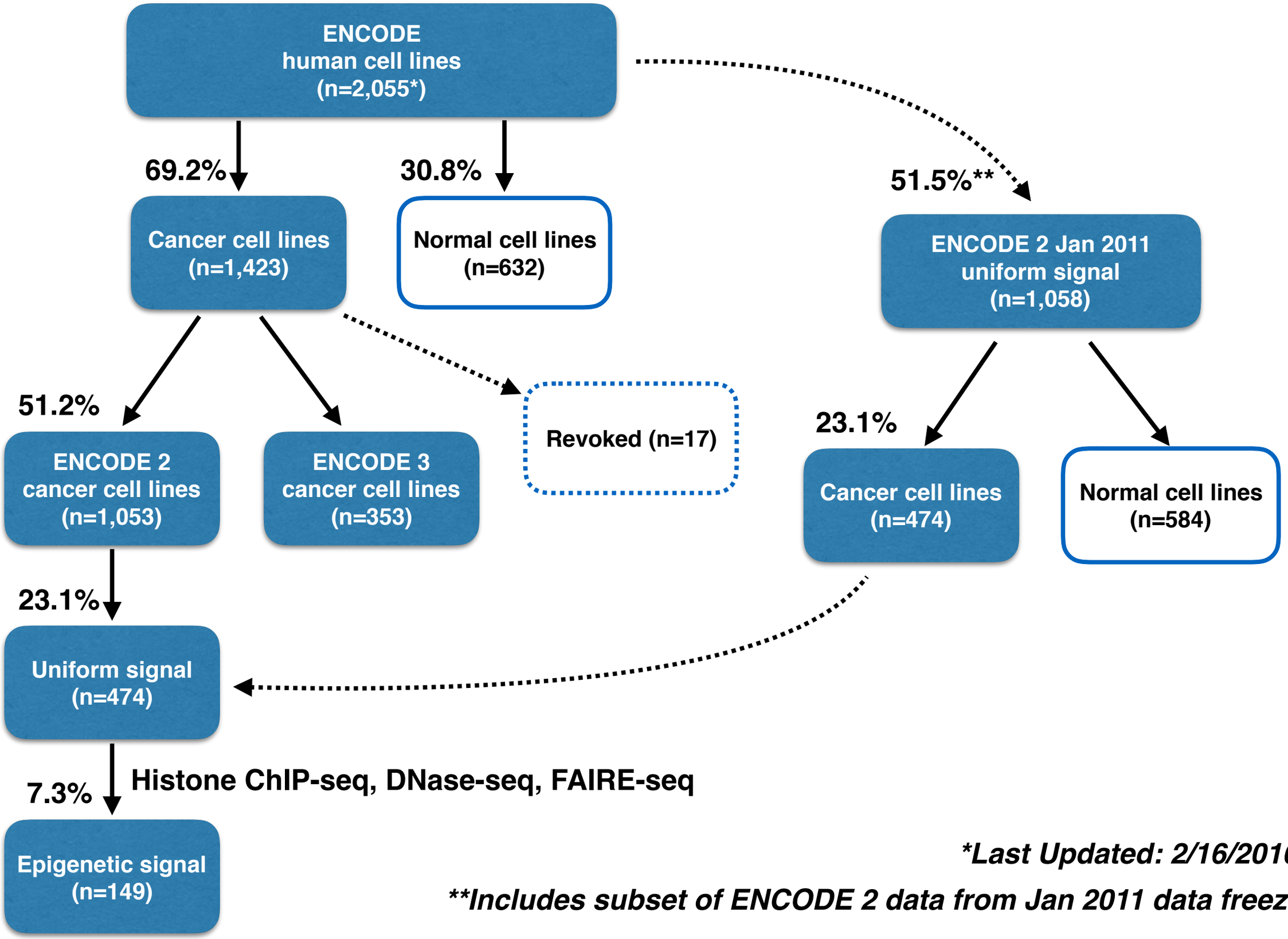
- For the time being, **Encode 2 uniform signals** (Jan 2011 freeze from EBI, consisting of DNase-seq, FAIRE, Histone, and TFBS) were used in the analysis.
  - *http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/signal/jan2011/bigwig/*
- Human genome (hg19) was divided into 1Mb bin (rows) and the signal was averaged over each bin

Encode 2 Uniform Signal

- **145** features (columns) across **25 cancer cell lines + 1 cancer primary cell**
  - 84 ChIP-seq signals were normalized into fold change over control
  - 32 DNase-seq, 11 FAIRE-seq, 1 MNase-seq (K562)
  - 17 features including GC content and dinucleotides such as CpG percentages
- Replication timing (Repli-seq) was omitted because the data was not available on Jan 2011 for uniform processing

Cell line	Term	Description	TCGA abbreviations	# Features	Y/N
8988T	8988T	pancreatic adenocarcinoma	PAAD	1	Y
A549	A549	lung cancer	LUSC or LUAD	3	
BE2_C	BE2C	neuroblastoma		1	
Caco-2	Caco-2	colorectal adenocarcinoma	COAD+READ	4	
CLL	B cell*	Chronic lymphocytic leukemia cell, B-lymphocyte		1	
CMK	CMK	acute megakaryocytic leukemia	AML	1	
Gliobla	H54	glioblastoma	GBM	4	Y
HCT-116	HCT116	colorectal carcinoma	COAD+READ	2	
HeLa-S3	HeLa-S3	cervical adenocarcinoma	CESC	23	
HepG2	HepG2	liver cancer/hepatoma/hepatocellular carcinoma	LIHC	20	Y
HL-60	HL-60	acute promyelocytic leukemia		2	
Huh-7	HuH-7	hepatocellular adenocarcinoma	LIHC	1	Y
Huh-7.5	HuH-7.5	hepatocellular adenocarcinoma	LIHC	1	Y
Jurkat	Jurkat	acute T cell leukemia		2	
K562	K562	chronic myelogenous leukemia (CML)		30	
LNCaP	LNCaP clone FGC	prostate adenocarcinoma	PRAD	3	Y
MCF-7	MCF-7	breast cancer	BRCA	10	Y
Medullo	medulloblastoma	medulloblastoma		2	
NB4	NB4	acute promyelocytic leukemia		2	
NT2-D1	NT2/D1	malignant pluripotent embryonal carcinoma		6	
PANC-1	Panc1	pancreatic carcinoma	PAAD	1	
SK-N-MC	SK-N-MC	Ewing's sarcoma		1	
SK-N-SH_RA	SK-N-SH	neuroblastoma		4	
T-47D	T47D	mammary ductal carcinoma	BRCA	1	Y
U2OS	U2OS	osteosarcoma		1	
WERI-Rb-1	WERI-Rb-1	retinoblastoma		1	

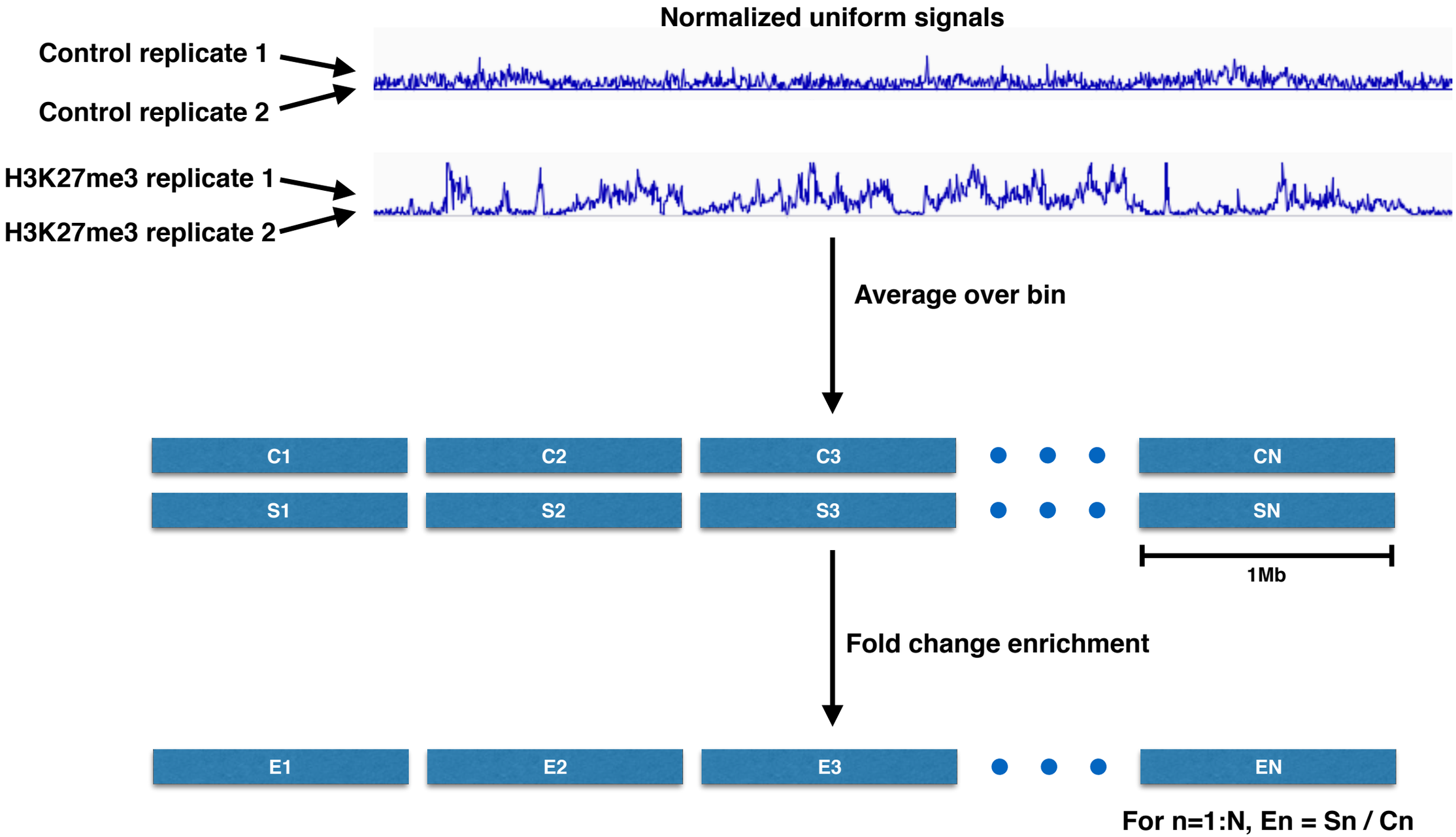
***Primary cell**



Last Updated: 2/16/2016

***Includes subset of ENCODE 2 data from Jan 2011 data freeze*

Example: Hep G2 H3K28me3 ChIP-seq



**For DNase-seq, FAIRE-seq, and MNase-seq, we assume the background is uniform*

- In addition to histone ChIP-seq data, **TFBS experiments of CTCF, Pol2, c-Myc** were included as features in the matrix

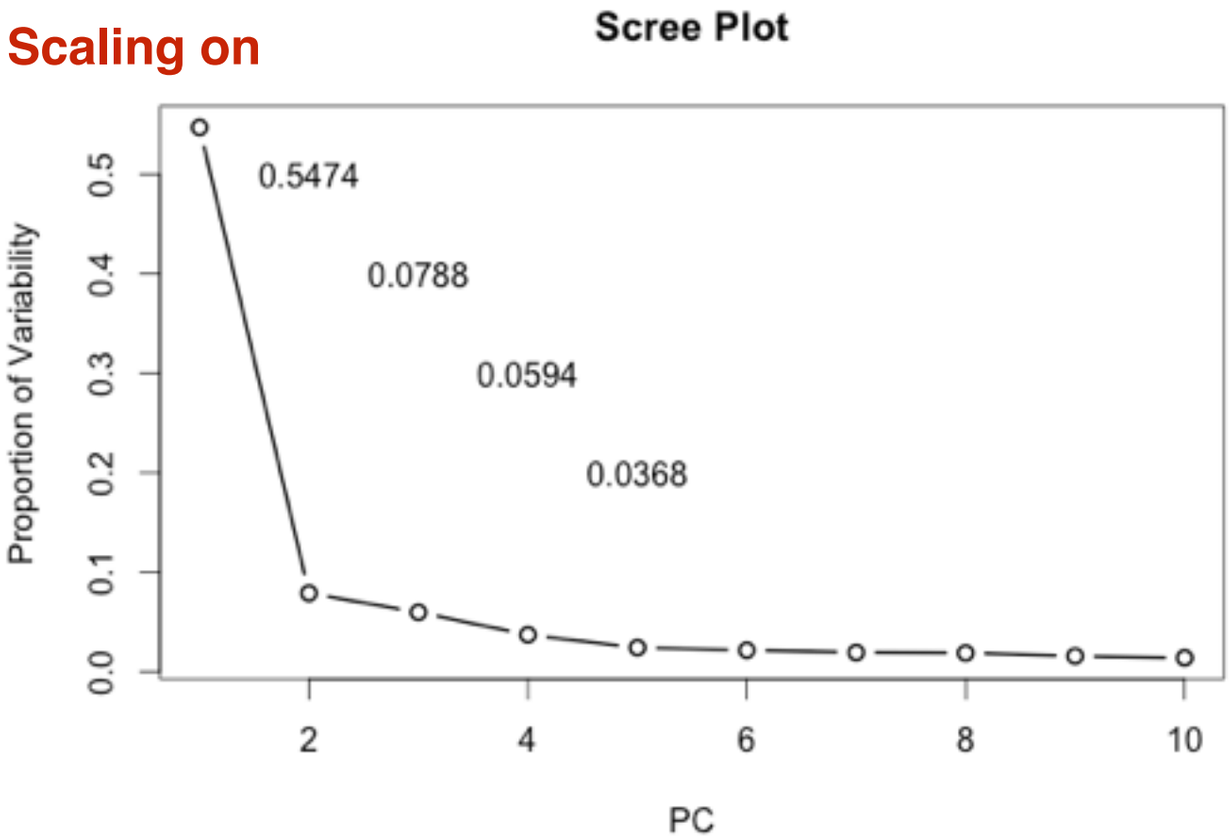
- These TFBSs have implications for chromatin regulation and these could indirectly contribute to mutation rate correction.

  - CTCF provides an anchor point for positioning nucleosomes, and CTCF is implicated in chromatin remodeling and interactions (Fu, Y et al, 2008, Phillips, J. E., & Corces, V. G., 2009).

  - Myc has been shown to regulate acetylation of histones H3 and H4 at several chromosomal loci (Bouchard et al. 2001, Frank et al. 2001, Nikiforov et al. 2002).

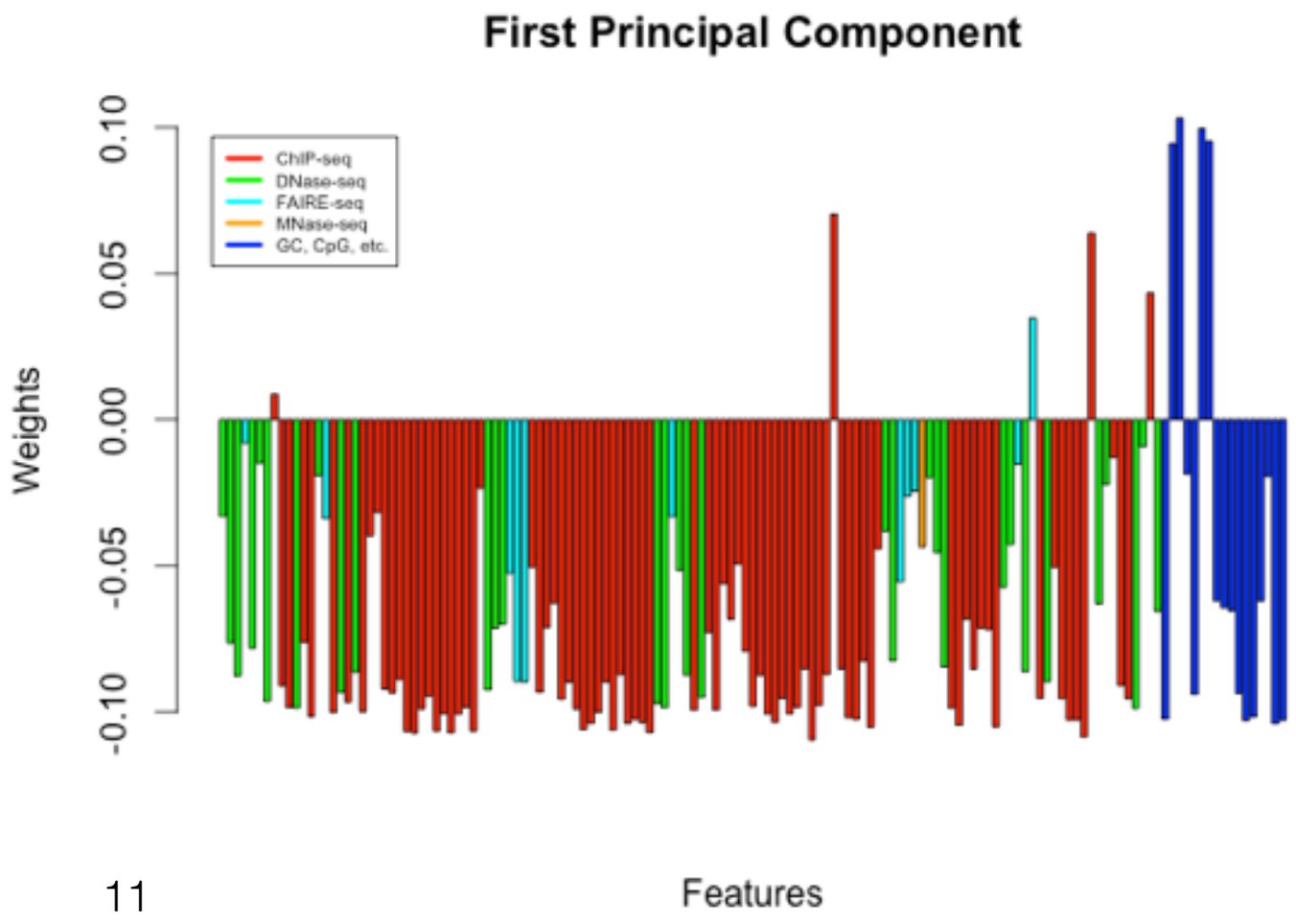
  - RNA polymerase II ChIP-seq associates with open chromatin (DNase hypersensitivity), histone acetyltransferases (HATs, P300-CBP), active histone marks, etc (Orphanides, G. et al., 2000).

Scaling on



PCA revealed the first 3 P.C. accounts for approx. 68% of total variability in the data

Coefficients (weights of rotation matrix) of the first P.C. shows no single feature is dominantly contributing to the total variability in the data



References

- Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature, 488(7412), 504–507. http://doi.org/10.1038/nature11273

- Fu, Y., Sinha, M., Peterson, C. L., & Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genetics, 4(7). http://doi.org/10.1371/journal.pgen.1000138

- Phillips, J. E., & Corces, V. G. (2009). CTCF: Master Weaver of the Genome. Cell. http://doi.org/10.1016/j.cell.2009.06.001

- Bouchard, C., Dittrich, O., Kiermaier, A., Dohmann, K., Menkel, A., Eilers, M., & Lüscher, B. (2001). Regulation of cyclin D2 gene expression by the Myc/Max/Mad network: Myc-dependent TRRAP recruitment and histone acetylation at the cyclin D2 promoter. Genes and Development, 15(16), 2042–2047. http://doi.org/10.1101/gad.907901

- Frank, S. R., Schroeder, M., Fernandez, P., Taubert, S., & Amati, B. (2001). Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. Genes and Development, 15(16), 2069–2082. http://doi.org/10.1101/gad.906601

- Nikiforov, M. A., Chandriani, S., Park, J., Kotenko, I., Matheos, D., Johnsson, A., … Cole, M. D. (2002). TRRAP-dependent and TRRAP-independent transcriptional activation by Myc family oncoproteins. Molecular and Cellular Biology, 22(14), 5054–63. http://doi.org/10.1128/MCB.22.14.5054-5063.2002

- Orphanides, G., & Reinberg, D. (2000). RNA polymerase II elongation through chromatin. Nature, 407(6803), 471–5. http://doi.org/10.1038/35035000

Acknowledgement

Mark Gerstein

Jing Zhang

Joel Rozowsky