Developing Informatics Technologies to Model Cancer Gene Regulation

Xiaole Shirley Liu

Dept of Biostatistics & Computational Biology
Center for Functional Cancer Epigenetics
Dana-Farber Cancer Institute
Harvard School of Public Health
Broad Institute of Harvard and MIT

DANA-FARBER CANCER INSTITUTE

CENTER for FUNCTIONAL CANCER EPIGENETICS

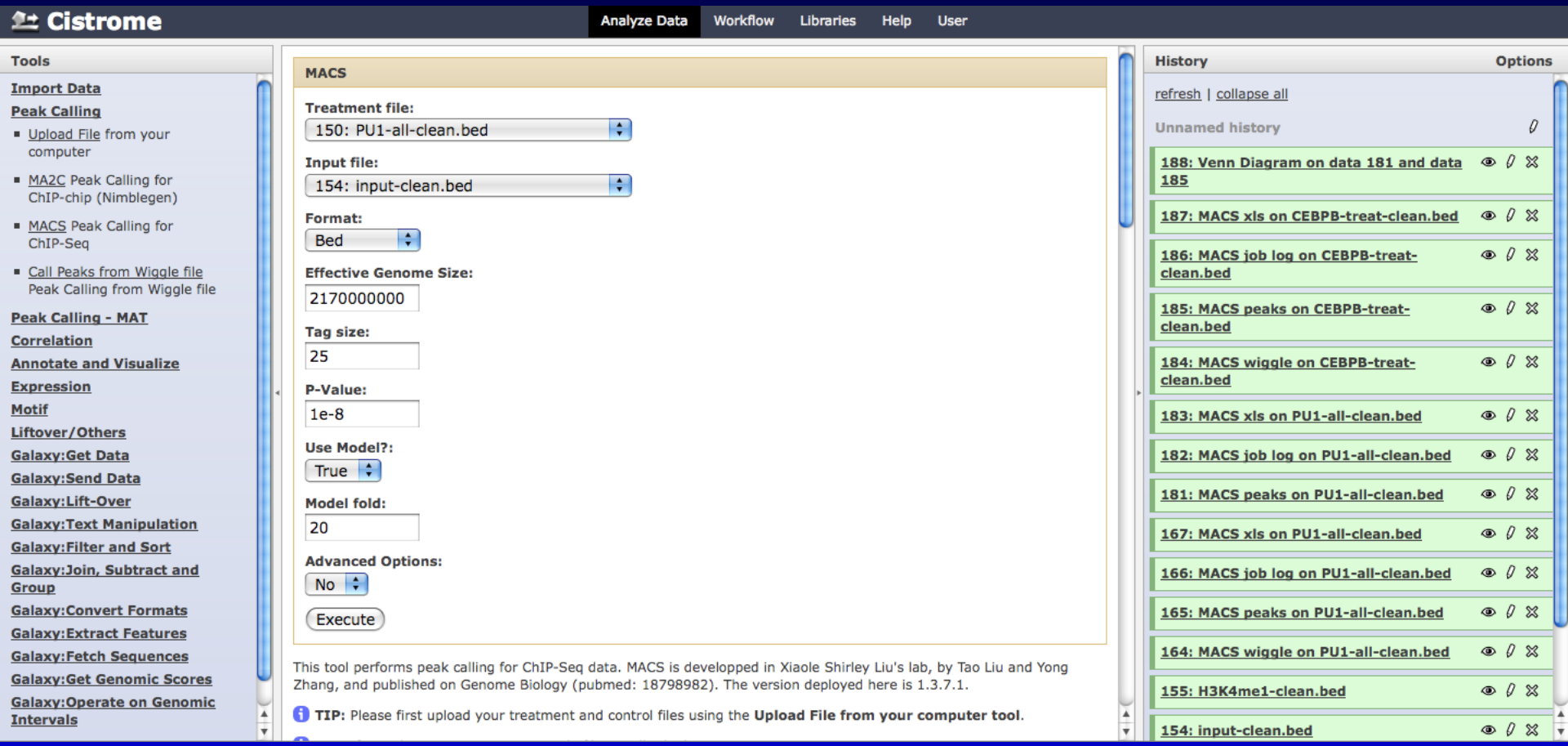
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

BROAD INSTITUTE

Public Data

- ChIP-seq and related data in human / mouse:
  - Factor ChIP-seq: > 11K, > 700 different TFs
  - Histone Mark ChIP-seq: > 10K
  - Chromatin accessibility: > 1300
- TCGA: gene expression profiles in ~ 12K tumors and 31 cancer types
- Model transcriptional and epigenetic gene regulation

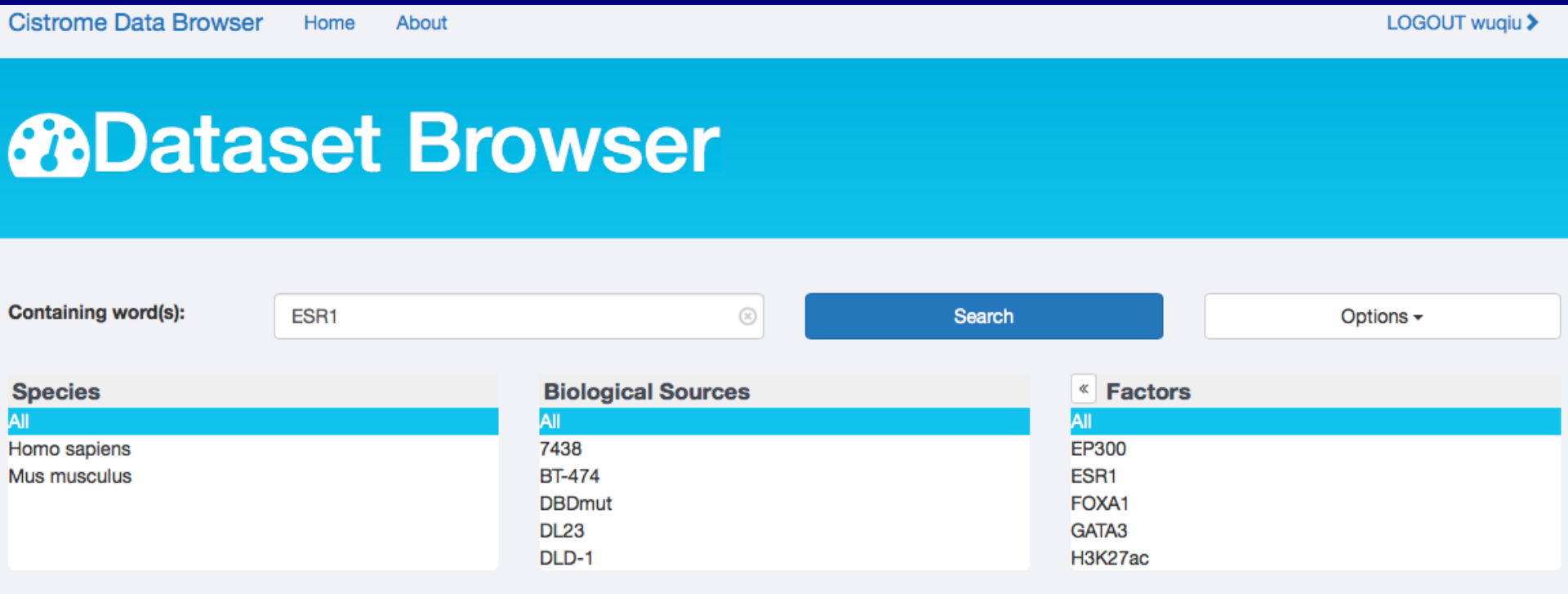
Cistrome: cis-elements bound by trans-factor
Cistrome Analysis Pipeline: http://cistrome.org/ap/



Liu et al, Genome Biol 2011; > 7K users and ~10T user data

Cistrome DB
http://cistrome.org/db/

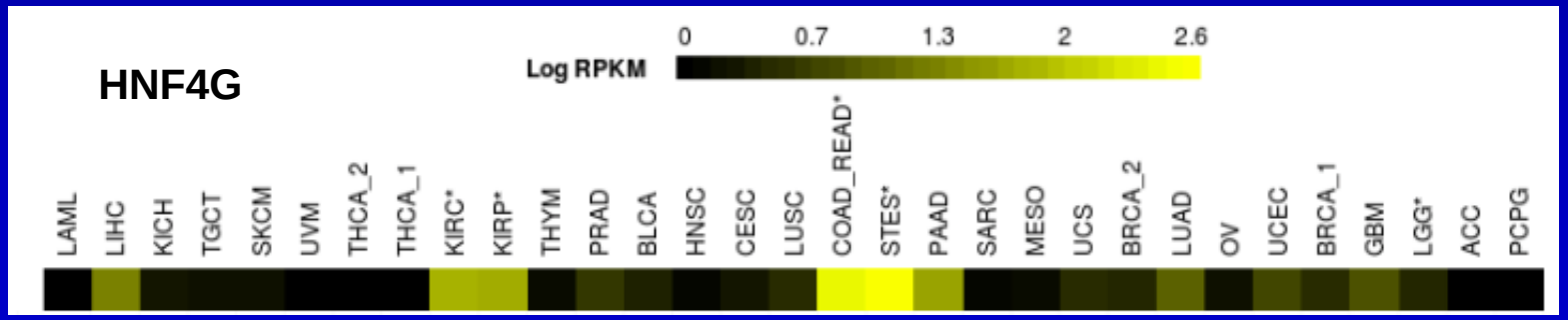
- All the ChIP/DNase-seq data human (hg38) / mouse (mm10) data as of Dec 2015 processed (peaks, signals, QC)



Factors Influencing Cancer-Specific Transcription Factor Targets

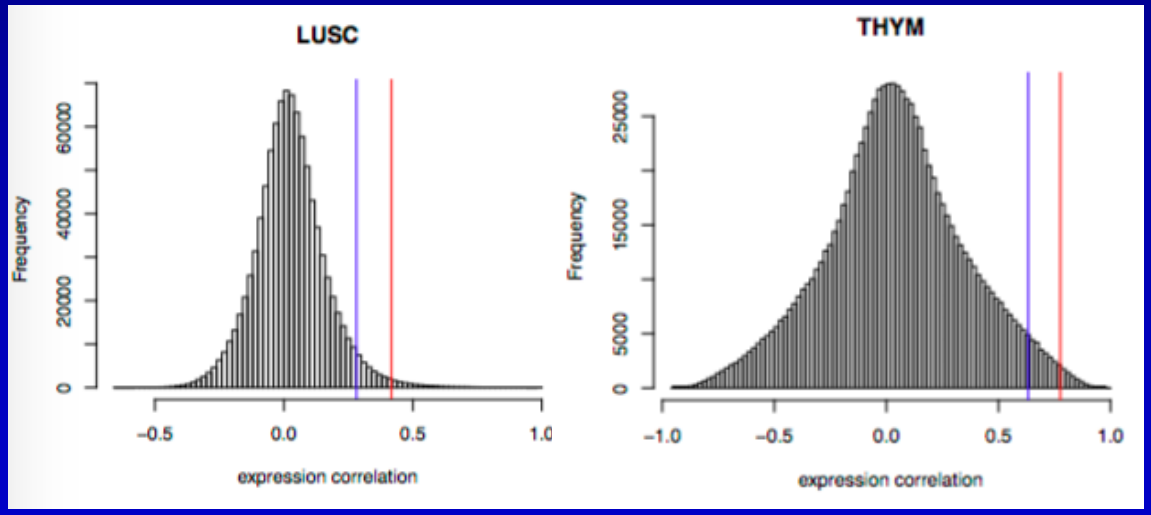
- Re-cluster all TCGA gene expression
  - Separate strong subtypes (luminal vs basal BRCA)
  - Combine similar cancer types (READ + COAD)

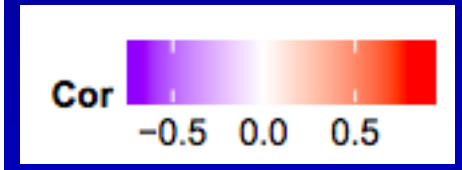
- TF expression level in each cancer type
  - Eliminate TF if median TF expression RPKM < 1



Factors Influencing Cancer-Specific Transcription Factor Targets

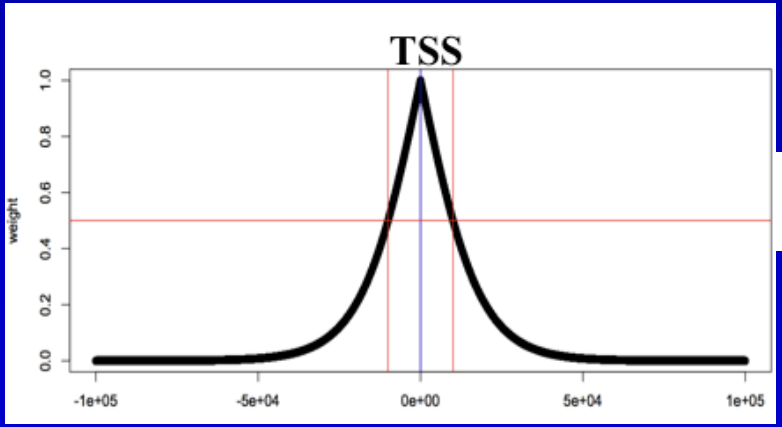
- Gene expression correlation between TF and putative targets
- Cutoff (5%) based on cancer-specific null distribution

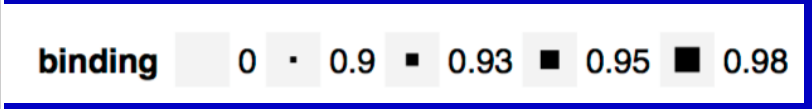


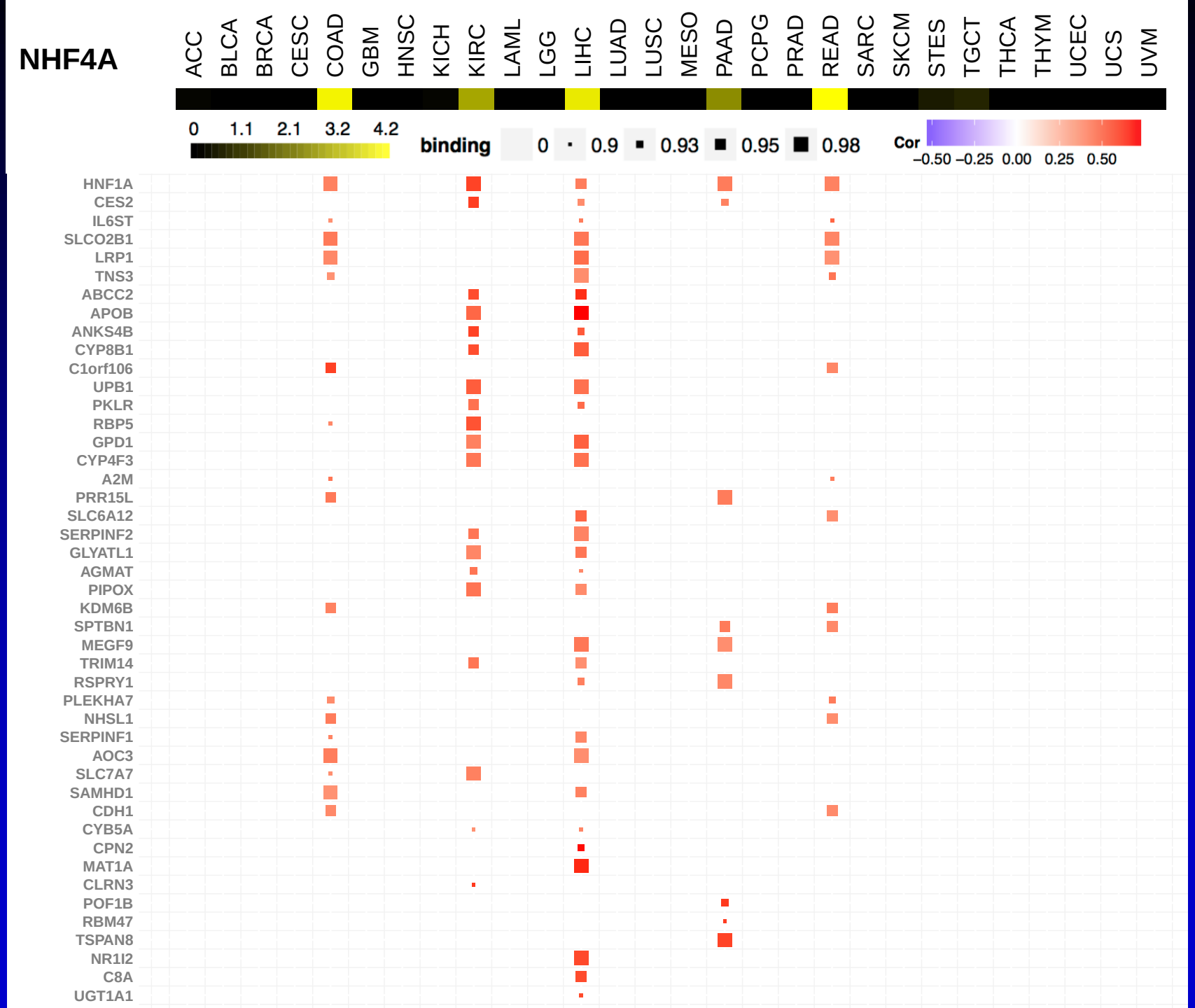


Factors Influencing Cancer-Specific Transcription Factor Targets

- For each TF ChIP-seq, calculate regulatory potential, the sum of binding sites weighted by distance to TSS with exponential decay
- For each cancer, use logistic regression to select TF ChIP-seq data that best explain the top correlated genes with TF

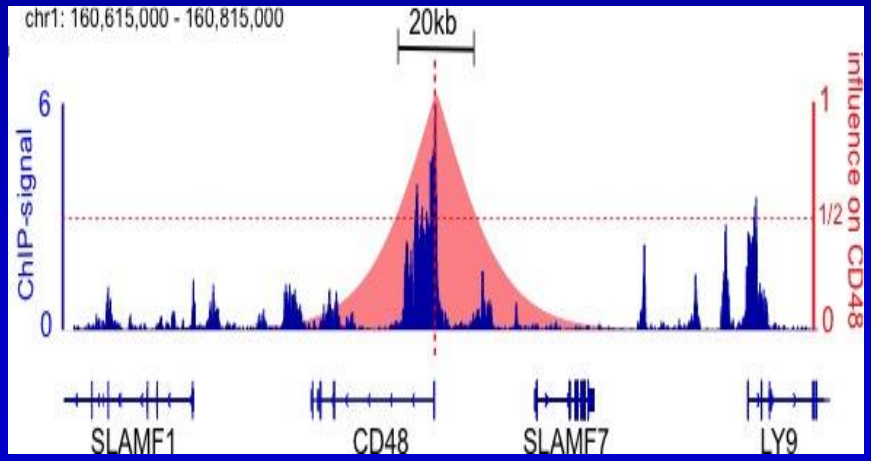





NHF4A
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
LAML
LGG
LIHC
LUAD
LUSC
MESO
PAAD
PCPG
PRAD
READ
SARC
SKCM
STES
TGCT
THCA
THYM
UCEC
UCS
UVM
0
1.1
2.1
3.2
4.2
binding
0
0.9
0.93
0.95
0.98
Cor
−0.50 −0.25 0.00 0.25 0.50
HNF1A
CES2
IL6ST
SLCO2B1
LRP1
TNS3
ABCC2
APOB
ANKS4B
CYP8B1
C1orf106
UPB1
PKLR
RBP5
GPD1
CYP4F3
A2M
PRR15L
SLC6A12
SERPINF2
GLYATL1
AGMAT
PIPOX
KDM6B
SPTBN1
MEGF9
TRIM14
RSPRY1
PLEKHA7
NHSL1
SERPINF1
AOC3
SLC7A7
SAMHD1
CDH1
CYB5A
CPN2
MAT1A
CLRN3
POF1B
RBM47
TSPAN8
NR1I2
C8A
UGT1A1

Reuse Public H3K27ac

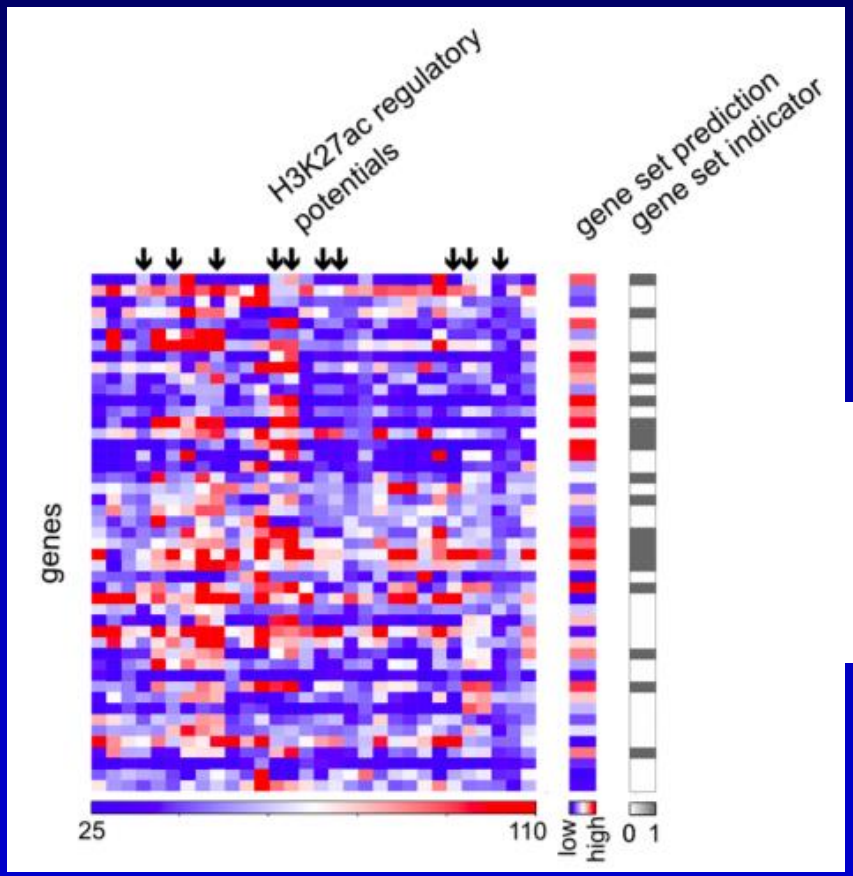
- Many laboratories have differential gene expression without matched H3K27ac ChIP-seq
- Regulatory Potential from H3K27ac
  - No peak calls, distance weight all H3K27ac reads by their distance to TSS with 10kb exponential decay



MARGE

Model-based Analysis of Regulation of Gene Expression

- Find relevant public H3K27ac ChIP-seq data that best represent the gene list



Logistic regression

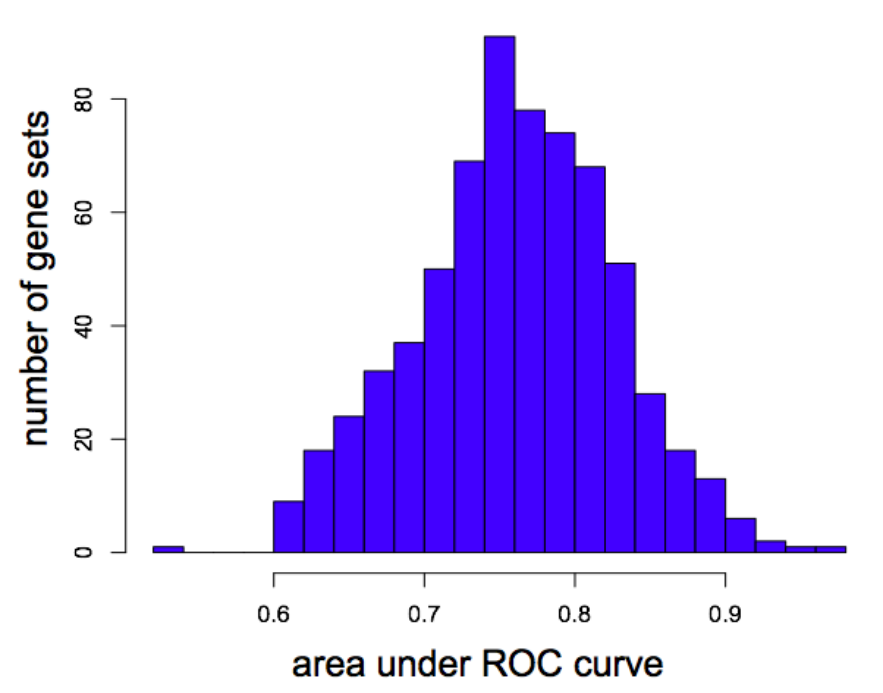
$$y_i \sim \alpha_0 + \sum_j a_j \sqrt{RP_{ij}}$$

The Power of Public H3K27ac Data

- Given a gene list (e.g. up-regulated upon perturbation), MARGE uses use regression to retrieve informative H3K27ac ChIP-seq profiles in the public to explain the genes

Train on genes in odd chromosomes and test on genes in even chromosomes

MSigDB
Molecular Signatures Database

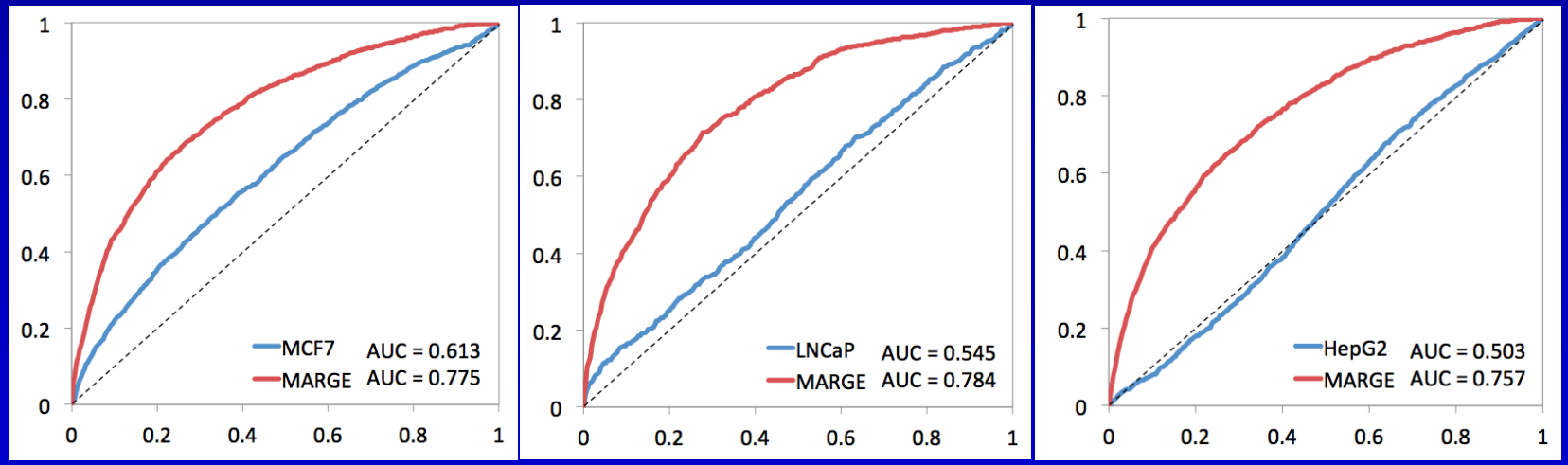


Wang, Zang, et al, revision under review

H3K27ac ChIP-seq for TCGA

- No histone mark ChIP-seq data for TCGA
- Given genes up-regulated in one cancer type, could we find relevant public H3K27ac ChIP-seq collections?
- Use DNase-seq peaks to improve H3K27ac resolution

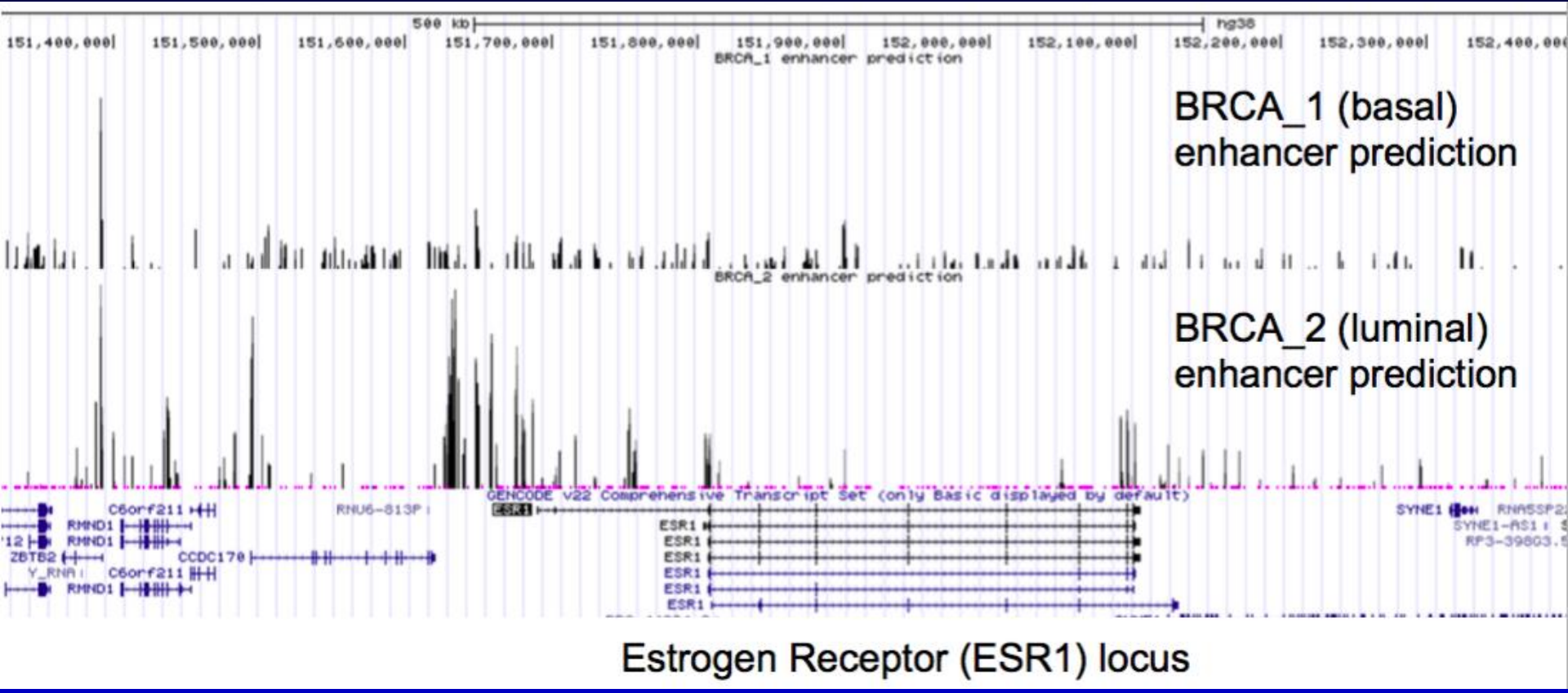
Breast cancer ***Prostate cancer*** ***Liver Cancer***



TCGA Cancer Enhancer Prediction (Beta Version)

No	ID	Description	# Cancer samples	# Normal samples	# Cancer genes	Enhancer Prediction		
1	BLCA	Bladder urothelial carcinoma	242	11	2724	download	Visualize WashU Browser	Visualize UCSC Browser
2	BRCA_1	Breast invasive carcinoma 1 (basal)	199	17	1860	download	Visualize WashU Browser	Visualize UCSC Browser
3	BRCA_2	Breast invasive carcinoma 2 (luminal)	844	94	1760	download	Visualize WashU Browser	Visualize UCSC Browser
4	COAD_READ*	Colon & colorectal adenocarcinoma	349	30	2031	download	Visualize WashU Browser	Visualize UCSC Browser
5	HNSC	Head and Neck squamous cell carcinoma	335	33	2165	download	Visualize WashU Browser	Visualize UCSC Browser
6	KICH	Kidney chromophobe	64	24	1504	download	Visualize WashU Browser	Visualize UCSC Browser
7	KIRC*	Kidney renal clear cell carcinoma	501	69	2727	download	Visualize WashU Browser	Visualize UCSC Browser
8	KIRP*	Kidney renal papillary cell carcinoma	287	32	1691	download	Visualize WashU Browser	Visualize UCSC Browser
9	LIHC	Liver hepatocellular carcinoma	354	47	4220	download	Visualize WashU Browser	Visualize UCSC Browser
10	LUAD	Lung adenocarcinoma	467	51	2161	download	Visualize WashU Browser	Visualize UCSC Browser
11	LUSC	Lung squamous cell carcinoma	356	34	3744	download	Visualize WashU Browser	Visualize UCSC Browser
12	PRAD	Prostate adenocarcinoma	490	52	967	download	Visualize WashU Browser	Visualize UCSC Browser
13	STES*	Stomach & Esophageal carcinoma	362	33	4162	download	Visualize WashU Browser	Visualize UCSC Browser
14	THCA_1	Thyroid carcinoma 1	357	42	1723	download	Visualize WashU Browser	Visualize UCSC Browser

Cancer-Specific Enhancer Profile



Estrogen Receptor (ESR1) locus

Mei, Zang et al, ms under preparation

Summary

- Use regulatory potential (exponential decay with distance) to model binding site and H3K27ac effect on gene expression
- MARGE uses logistic regression to retrieve public H3K27ac ChIP-seq to explain differential expression gene set of interest
- Can predict enhancer profiles for up-regulated genes in each TCGA cancer type

- http://cistrome.org/CistromeCancer/

Acknowledgement

Shirley Liu Laboratory

- Shenglin Mei
- Chongzhi Zang
- Cliff Meyer
- Su Wang
- Qian Qin
- Qiu Wu
- Jingyu Fan
- Hanfei Sun
- Rongbin Zheng
- Jian Ma
- Len Taing
- Muyuan Zhu
- Jiaxin Wu
- Anya Zhang

Collaborators

- Myles Brown
- Jun Liu
- Yong Zhang
- Tao Liu
- Fugen Li
- Henry Long
- ITCR and colleagues

YouTube Video

https://www.youtube.com/watch?v=pQ02wg8MI6o