Beyond Genome Browser

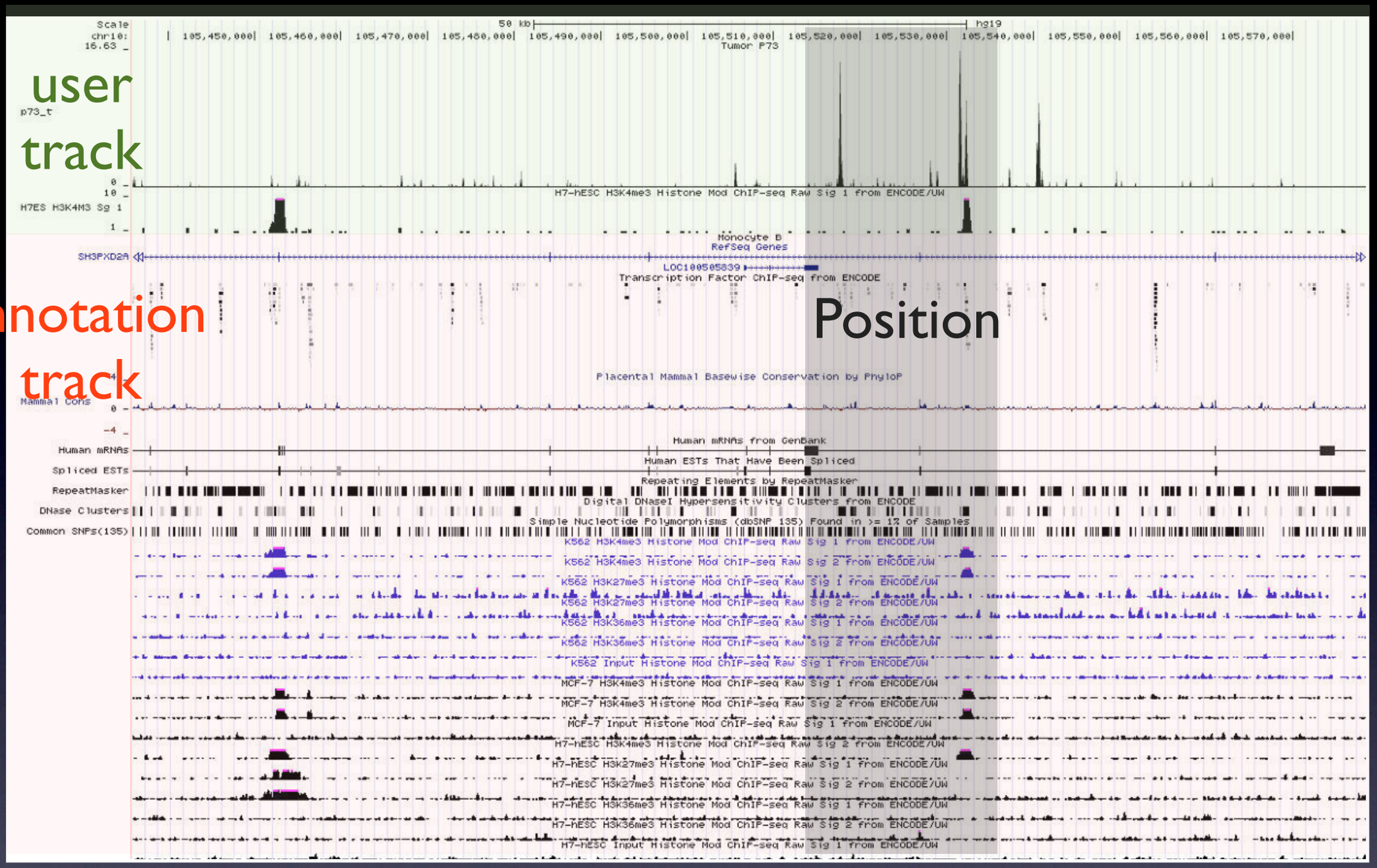
A more efficient system for browsing epigenetic data

Visiting student@DFCI
Hanfei Sun

3 parts

- Limits of Genome Browser
- Atlas, Recommendation and Comparison
- Prototype and 3 examples

A glimpse to traditional Genome Browser



APT
(Annotation, Position and Track)

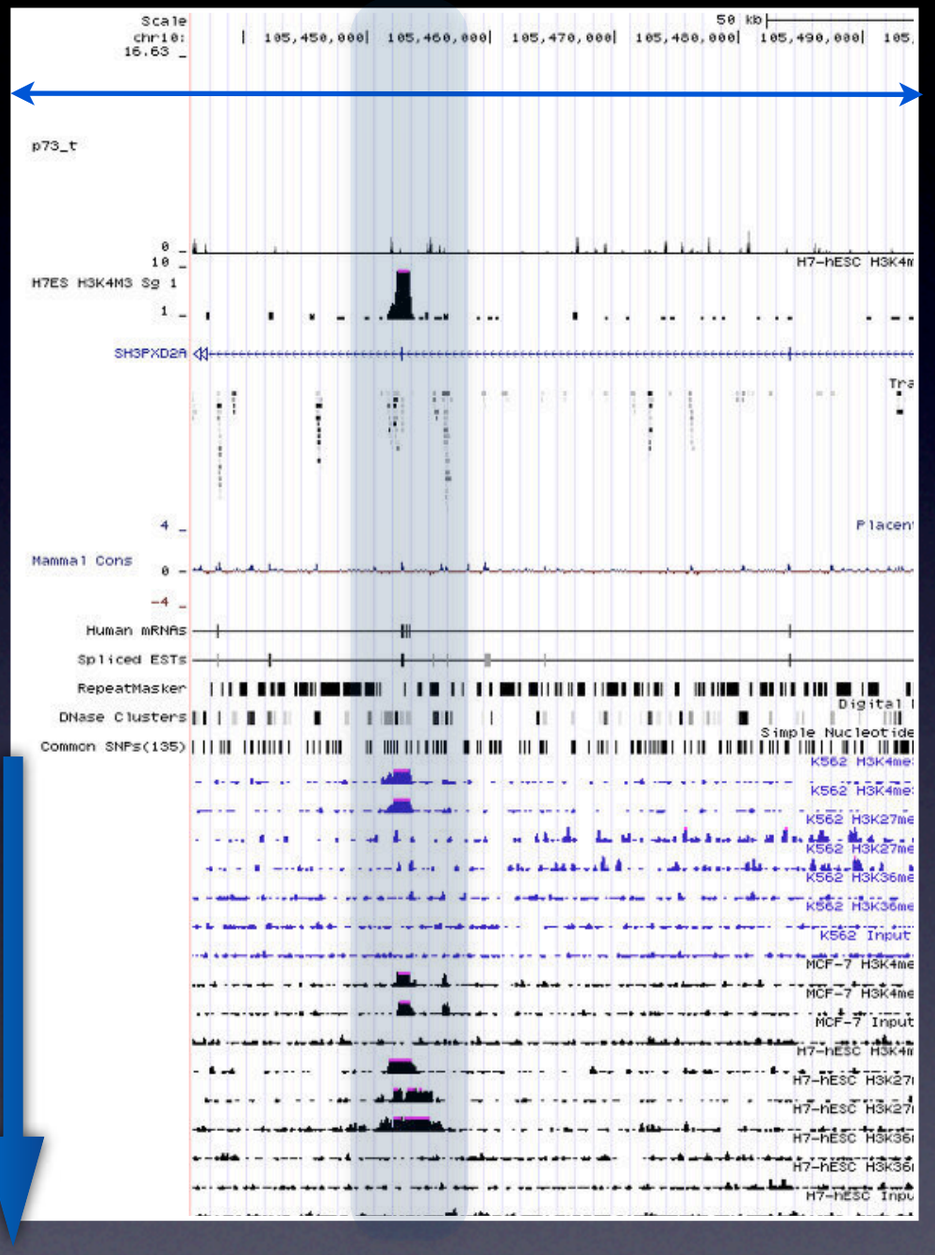
Screenshot of *UCSC Genome Browser*

Features of APT

Select regions of interest

Add tracks of relevance

Find visual correlations

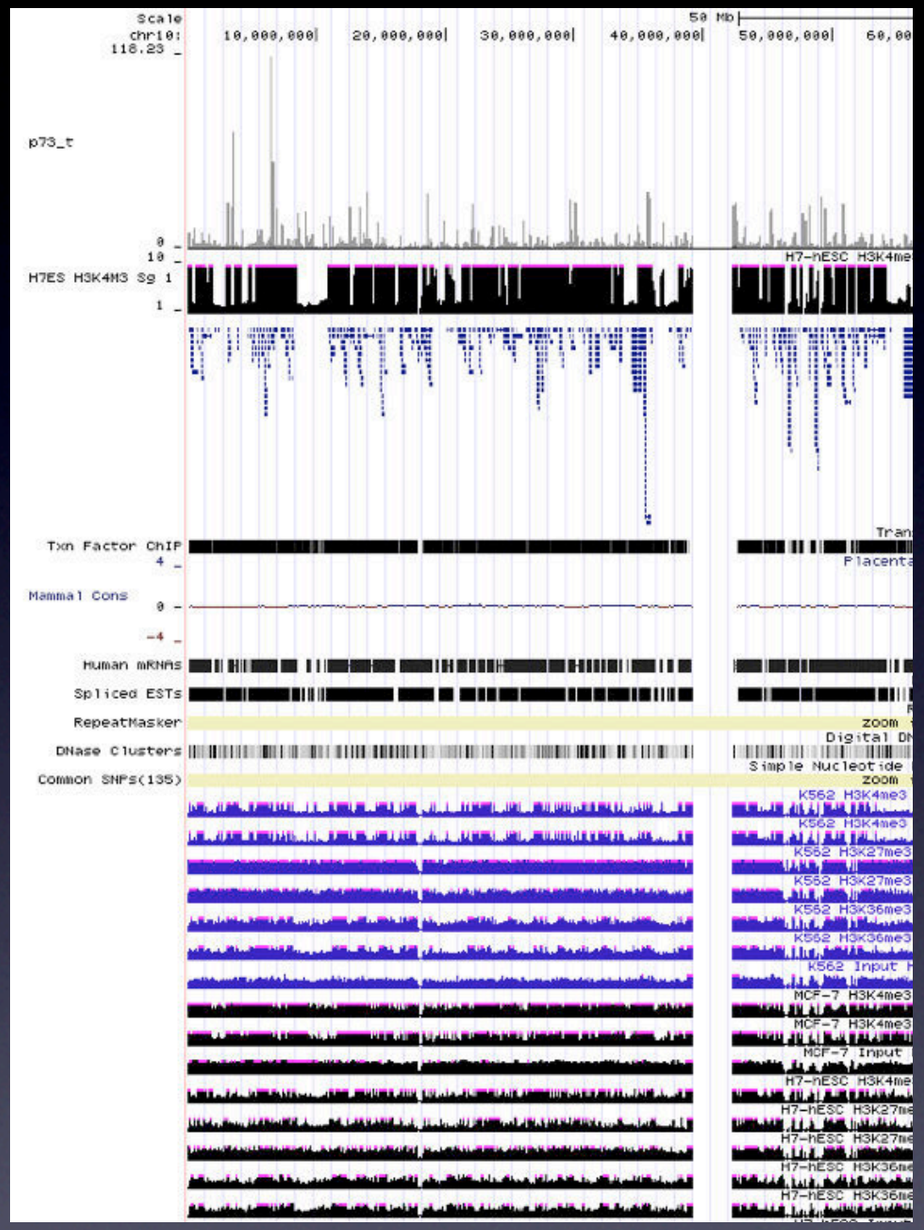


Drawbacks of APT

Which regions would be potentially interesting?

Among hundreds of tracks, which tracks would be related to current one?

How can I get an estimate of correlations in whole-genome level?



A new framework - ARC :

Atlas, Recommendation and Comparison

Presume we already have..

- target prediction score: regulatory potential for a given gene
- distance: a measurement of similarity between datasets

ARC structure

Function	Input	Output
Atlas	one gene	*target prediction* for this gene in all datasets
Recommendation	one dataset	top *n* nearest datasets sorted by *distance*
Comparison	two or more datasets	matrices of *target prediction*

Comparison of APT and ARC

	APT	ARC
Atom	Position (billions bp)	Gene (thousands)
Value	Profile or Interval	Target Prediction Score
Find the most correlated tracks	Visual	Recommendation
Find binding sites among datasets	Visual	Atlas
Find differences between two datasets	Visual	Comparison

Implementation

Target Prediction $$S_g = 100 \sum_{i=1}^{k} e^{-(0.5+4\Delta_i)}$$

Distance $$D = 1 - \rho_{X,Y}$$

Web

Pyramid™

SQLAlchemy

Statistics

RPy

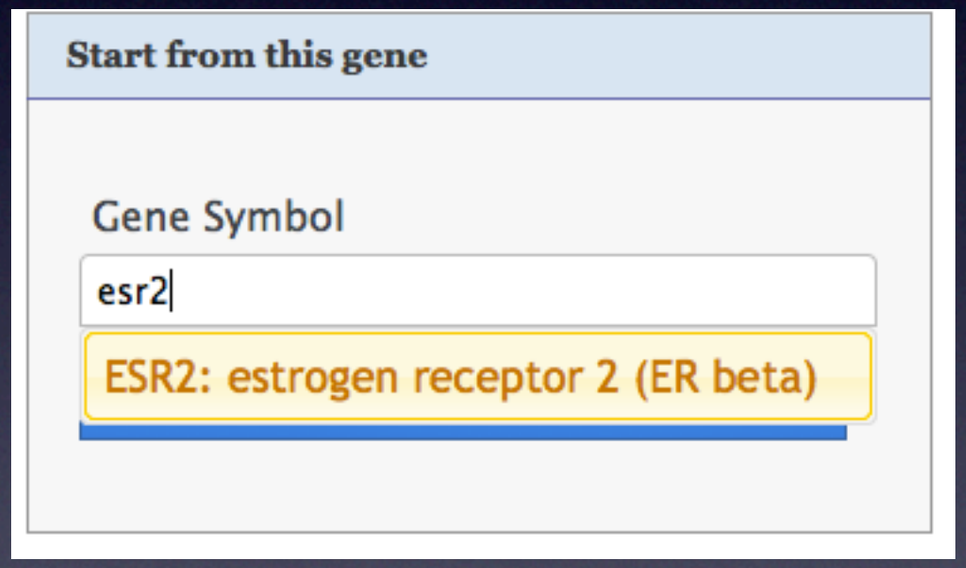
Algorithm from
Qianzi et al. (2011)
Cancer Research

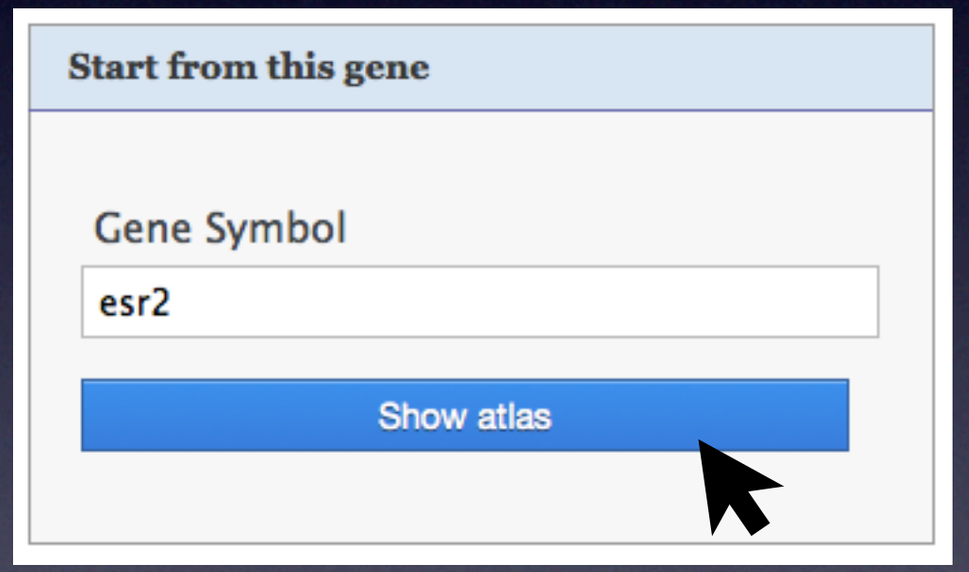
Three stories about ARC

ARC

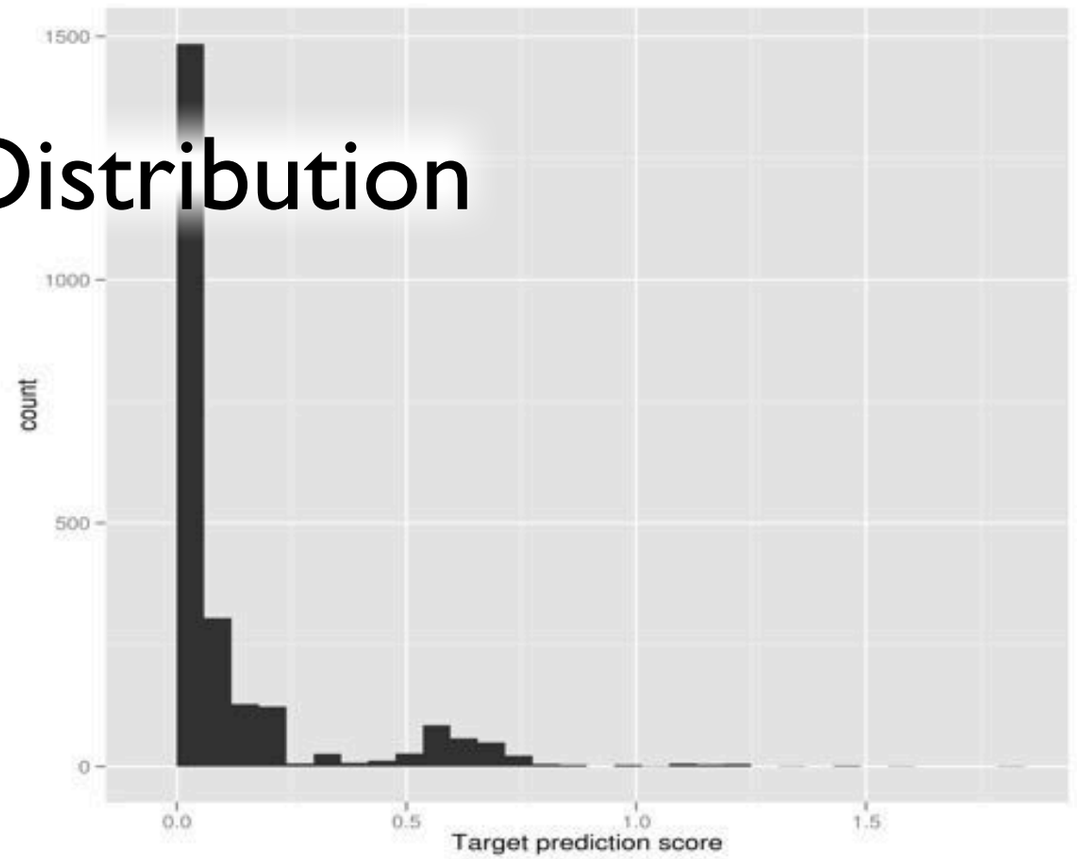
Check the atlas of ESR2 gene regulated by MYC

Input a gene


Start from this gene
Gene Symbol
esr2
ESR2: estrogen receptor 2 (ER beta)


Start from this gene
Gene Symbol
esr2
Show atlas

Result Page



Distribution of ESR2 (NM_001040275)'s target score on all datasets

Dataset ID: 5320
Target prediction for ESR2 (NM_001040275) 1.79

Factor: H3K27me3 Focus on H3K27me3
CellLine: H1 Focus on H1
CellType: Embryonic Stem
CellPop:
Tissue: Embryo
Disease: Normal

Description:
Reference Epigenome: ChIP-Seq Analysis of H3K27me3 in hESC Cells; renlab.H3K27me3.hESC-03.01
Condition:

Paper: Human DNA methylomes at base resolution show widespread epigenomic differences.
Focus on this paper
Authors: Lister R,Pelizzola M,Dowen RH,Hawkins RD,Hon G,Tonti-Filippini J,Nery JR,Lee L,Ye Z,Ngo QM,Edsall L,Antosiewicz-Bourget J,Stewart R,Ruotti V,Millar AH,Thomson JA,Ecker JR,Ren B

+

Load to Genome Browser

Dataset list

Dataset ID: 5179
Target prediction for ESR2 (NM_001040275) 1.55

Factor: MYC Focus on MYC
CellLine: BL-41 Focus on BL-41

+

Narrow your result

+

Dataset ID: 5179
Target prediction for ESR2 (NM_001040275) 1.55

Factor: MYC Focus on MYC
CellLine: BL-41 Focus on BL-41
CellType: B Lymphocyte
CellPop:
Tissue: Blood
Disease: Burkitt's Lymphoma

Description:
BL41_ChIP
Condition:

Paper: Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma.
Focus on this paper
Authors: Seitz V,Butzhammer P,Hirsch B,Hecht J,Gütgemann I,Ehlers A,Lenze D,Oker E,Sommerfeld A,von der Wall E,König C,Zinser C,Spang R,Hummel M

After narrowing

Target prediction for ESR2 (NM_001040275) 1.55

Factor: MYC Focus on MYC
CellLine: BL-41 Focus on BL-41
CellType: B Lymphocyte
CellPop:
Tissue: Blood
Disease: Burkitt's Lymphoma

Description:
BL41_ChIP
Condition:

Paper: Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma.
Focus on this paper
Authors: Seitz V,Butzhammer P,Hirsch B,Hecht J,Gütgemann I,Ehlers A,Lenze D,Oker E,Sommerfeld A,von der Wall E,König C,Zinser C,Spang R,Hummel M

Dataset ID: 5178
Target prediction for ESR2 (NM_001040275) 0.67

Factor: MYC Focus on MYC
CellLine: Ramos Focus on Ramos
CellType: B Lymphocyte
CellPop:
Tissue:
Disease: Burkitt's Lymphoma

Description:
Ramos_ChIP
Condition:

Paper: Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma.
Focus on this paper
Authors: Seitz V,Butzhammer P,Hirsch B,Hecht J,Gütgemann I,Ehlers A,Lenze D,Oker E,Sommerfeld A,von der Wall E,König C,Zinser C,Spang R,Hummel M

Dataset ID: 727
Target prediction for ESR2 (NM_001040275) 0.49

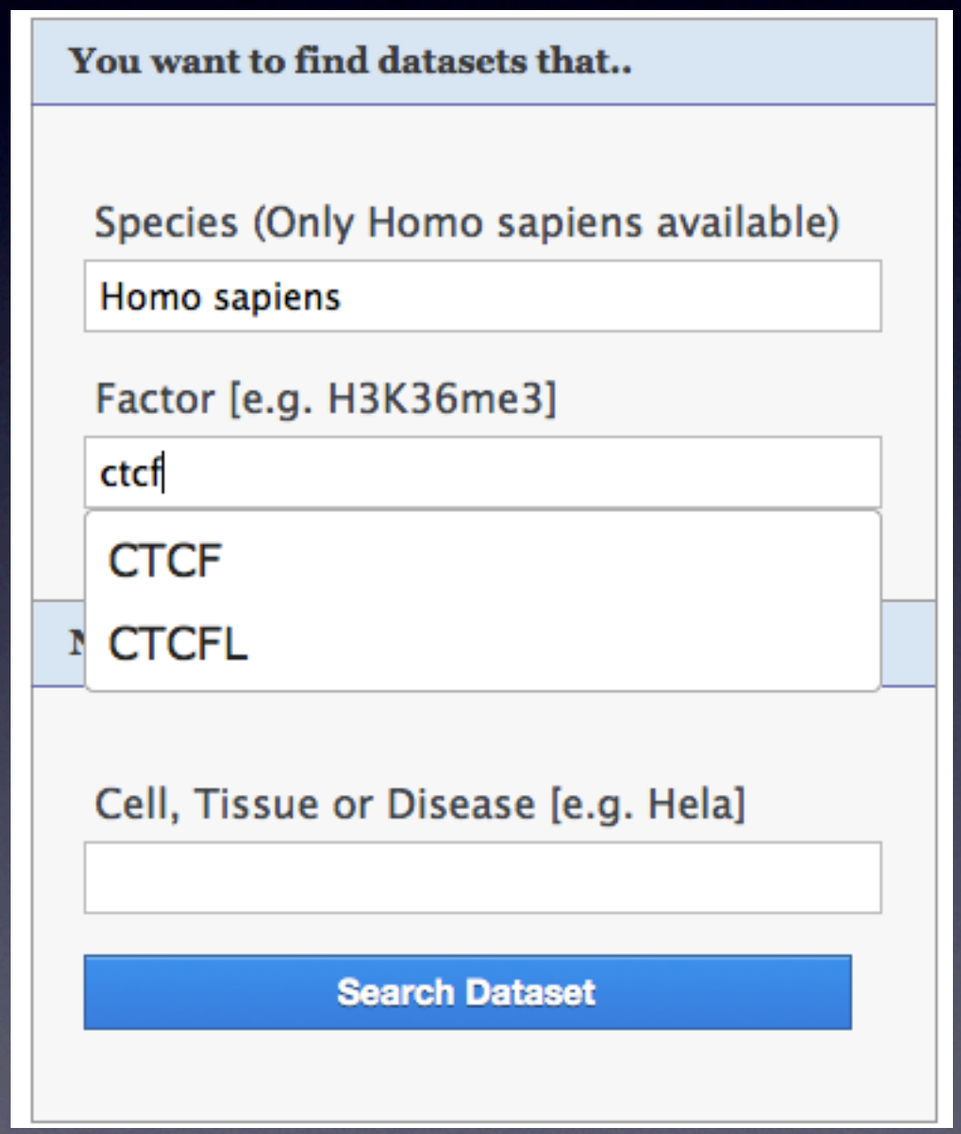
Factor: MYC Focus on MYC
CellLine: K562 Focus on K562
CellType: Erythroblast
CellPop:
Tissue: Bone Marrow
Disease: Chronic myeloid leukemia

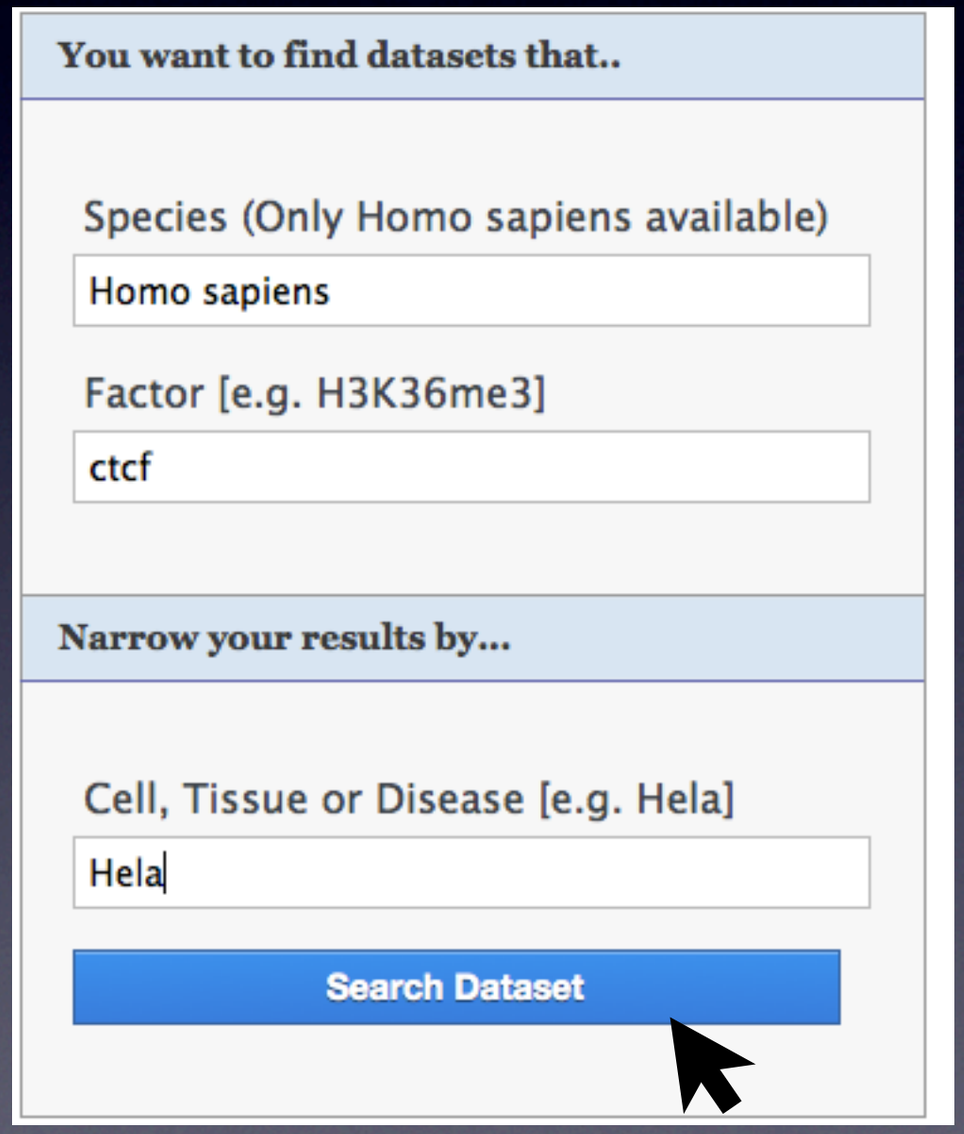
Description:

ARC

Recommend some co-binding/co-regulation transcription factors of CTCF in Hela cell

Search page


You want to find datasets that..
Species (Only Homo sapiens available)
Homo sapiens
Factor [e.g. H3K36me3]
ctcf
CTCF
CTCFL
Cell, Tissue or Disease [e.g. Hela]
Search Dataset


You want to find datasets that..
Species (Only Homo sapiens available)
Homo sapiens
Factor [e.g. H3K36me3]
ctcf
Narrow your results by...
Cell, Tissue or Disease [e.g. Hela]
Hela
Search Dataset

Rank 1: itself

Dataset ID: 975 **Similarity: 1.00** **Focus** **Diff**

Factor: CTCF Focus on CTCF
CellLine: Hela Focus on Hela
CellType: Epithelial
CellPop:
Tissue: Cervix
Disease: Cervical Adenocarcinoma

Description:
HeLa CTCF
Condition:

Paper: Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.
Focus on this paper
Authors: Cuddapah S,Jothi R,Schones DE,Roh TY,Cui K,Zhao K

Rank 2: another CTCF

Dataset ID: 1646 **Similarity: 0.73** **Focus** **Diff**

Factor: CTCF Focus on CTCF
CellLine: Hela Focus on Hela
CellType: Epithelial
CellPop:
Tissue: Cervix
Disease: Cervical Adenocarcinoma

Description:

Condition:

Paper: The ENCyclopedia Of DNA Elements (ENCODE) UTA
Focus on this paper
Authors: Iyer VR

Rank 3: an interesting co-factor?

Dataset ID: 978 **Similarity: 0.55** **Focus** **Diff**

Factor: H2AFZ Focus on H2AFZ
CellLine: Hela Focus on Hela
CellType: Epithelial
CellPop:
Tissue: Cervix
Disease: Cervical Adenocarcinoma

Description:
H2A.Z (low salt) ChIP
Condition: low salt

Paper: H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions.
Focus on this paper
Authors: Jin C,Zang C,Wei G,Cui K,Peng W,Zhao K,Felsenfeld G

The interaction validated in a previous paper

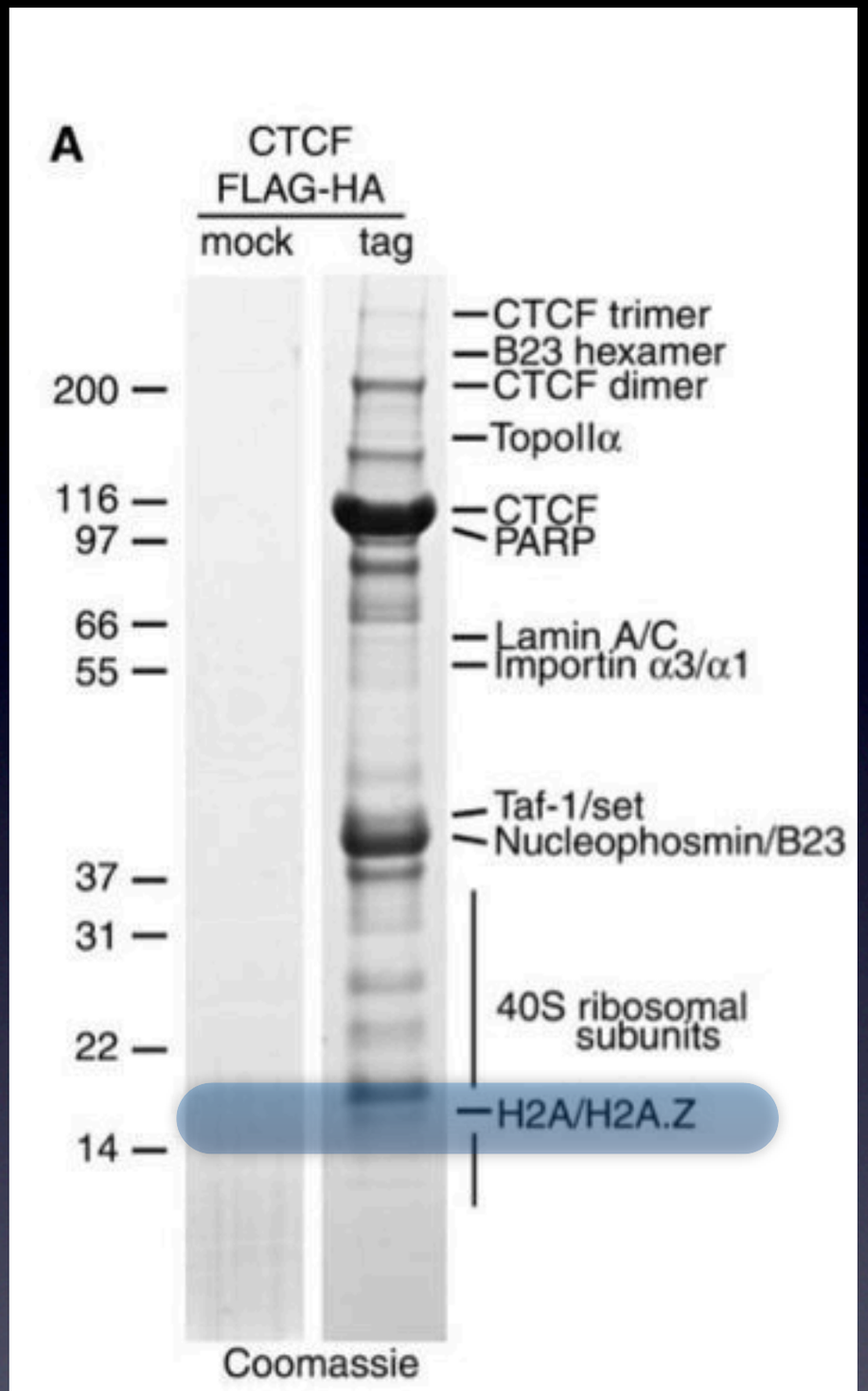


Figure 1. Purification of CTCF Complexes

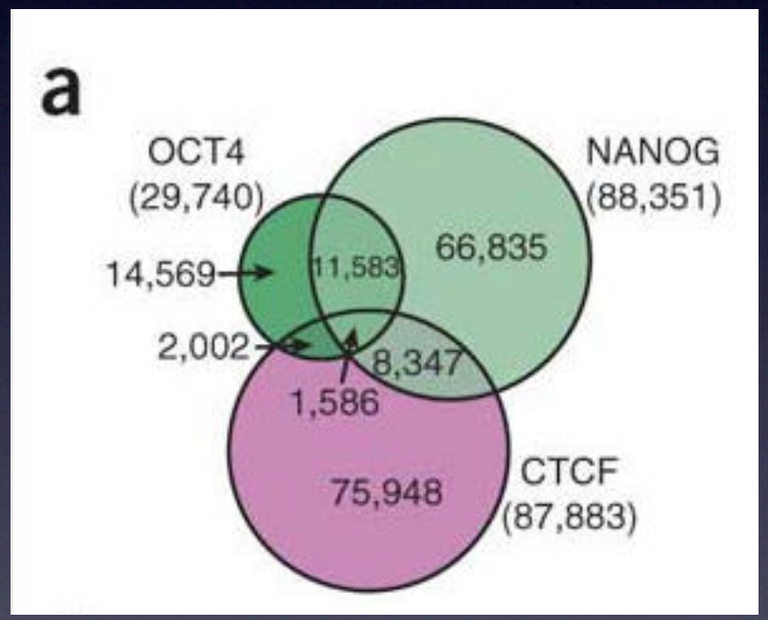
(A) FLAG-HA-tagged CTCF was purified from a stably transduced HeLa cell line,

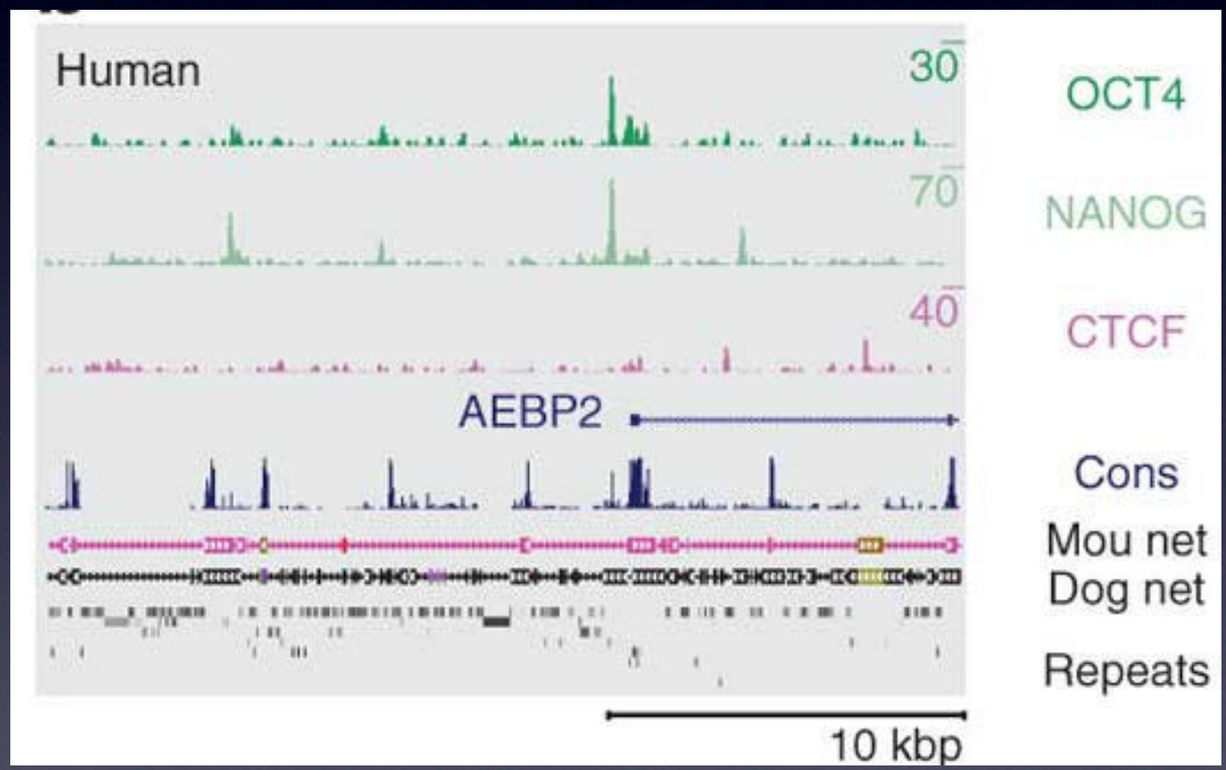
Figure from Yusufzai et al. (2004) *Molecular Cell*

ARC

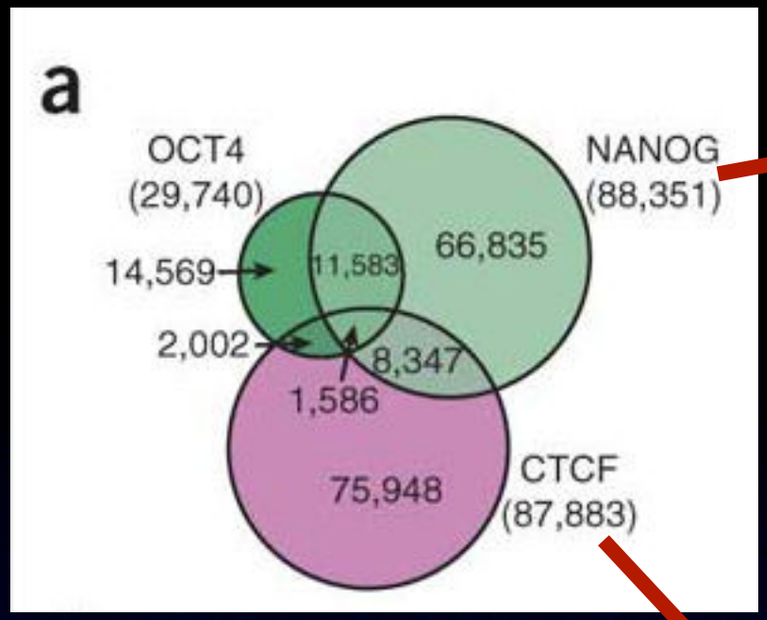
Compare 2 transcription factors in stem cell

A paper about transcription factors





Figures from:
Kunarso et al. (2010)
Nature genetics



Two datasets from that paper

Dataset ID: 3514 **Similarity: 1.00** **Focus** **Diff**

Factor: NANOG Focus on NANOG
CellLine: H1 Focus on H1
CellType: Embryonic Stem
CellPop:
Tissue: Embryo
Disease: Normal

Description:
NANOG
Condition:

Paper: Transposable elements have rewired the core regulatory network of human embryonic stem cells.
Focus on this paper
Authors: Kunarso G,Chia NY,Jeyakani J,Hwang C,Lu X,Chan YS,Ng HH,Bourque G

Dataset ID: 3513 **Similarity: 0.05** **Focus** **Diff**

Factor: CTCF Focus on CTCF
CellLine: H1 Focus on H1
CellType: Embryonic Stem
CellPop:
Tissue: Embryo
Disease: Normal

Description:
CTCF
Condition:

Paper: Transposable elements have rewired the core regulatory network of human embryonic stem cells.
Focus on this paper
Authors: Kunarso G,Chia NY,Jeyakani J,Hwang C,Lu X,Chan YS,Ng HH,Bourque G

Compare them

Dataset ID: 3513 **Similarity: 0.05** **Focus** **Diff**

Factor: CTCF Focus on CTCF
CellLine: H1 Focus on H1
CellType: Embryonic Stem
CellPop:
Tissue: Embryo
Disease: Normal

Description:
CTCF
Condition:

Paper: Transposable elements have rewired the core regulatory network of human embryonic stem cells.
Focus on this paper
Authors: Kunarso G,Chia NY,Jeyakani J,Hwang C,Lu X,Chan YS,Ng HH,Bourque G

Control Panel

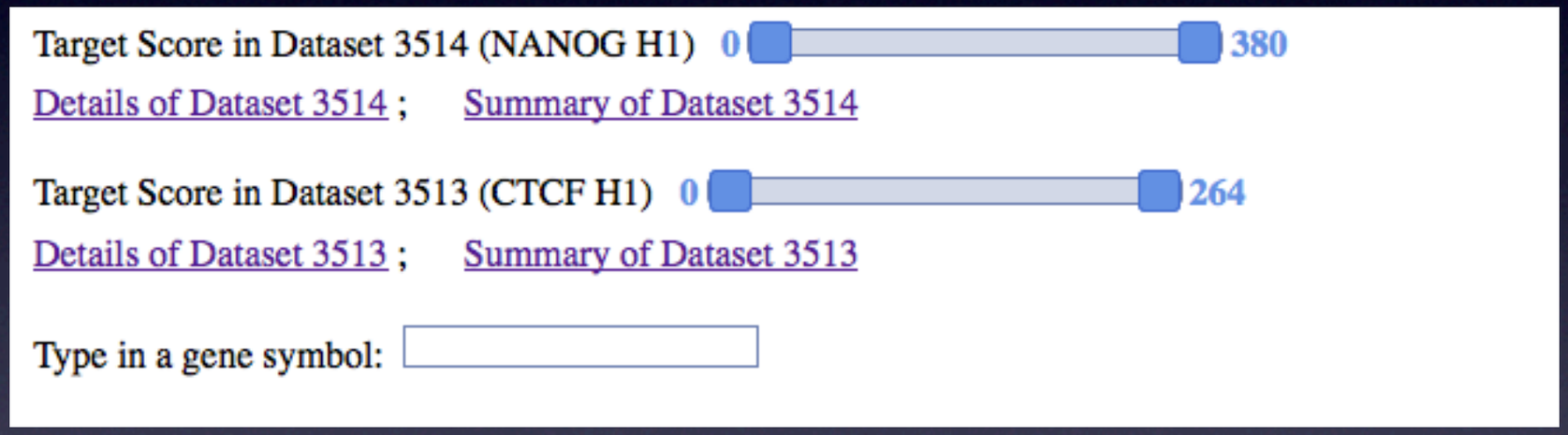


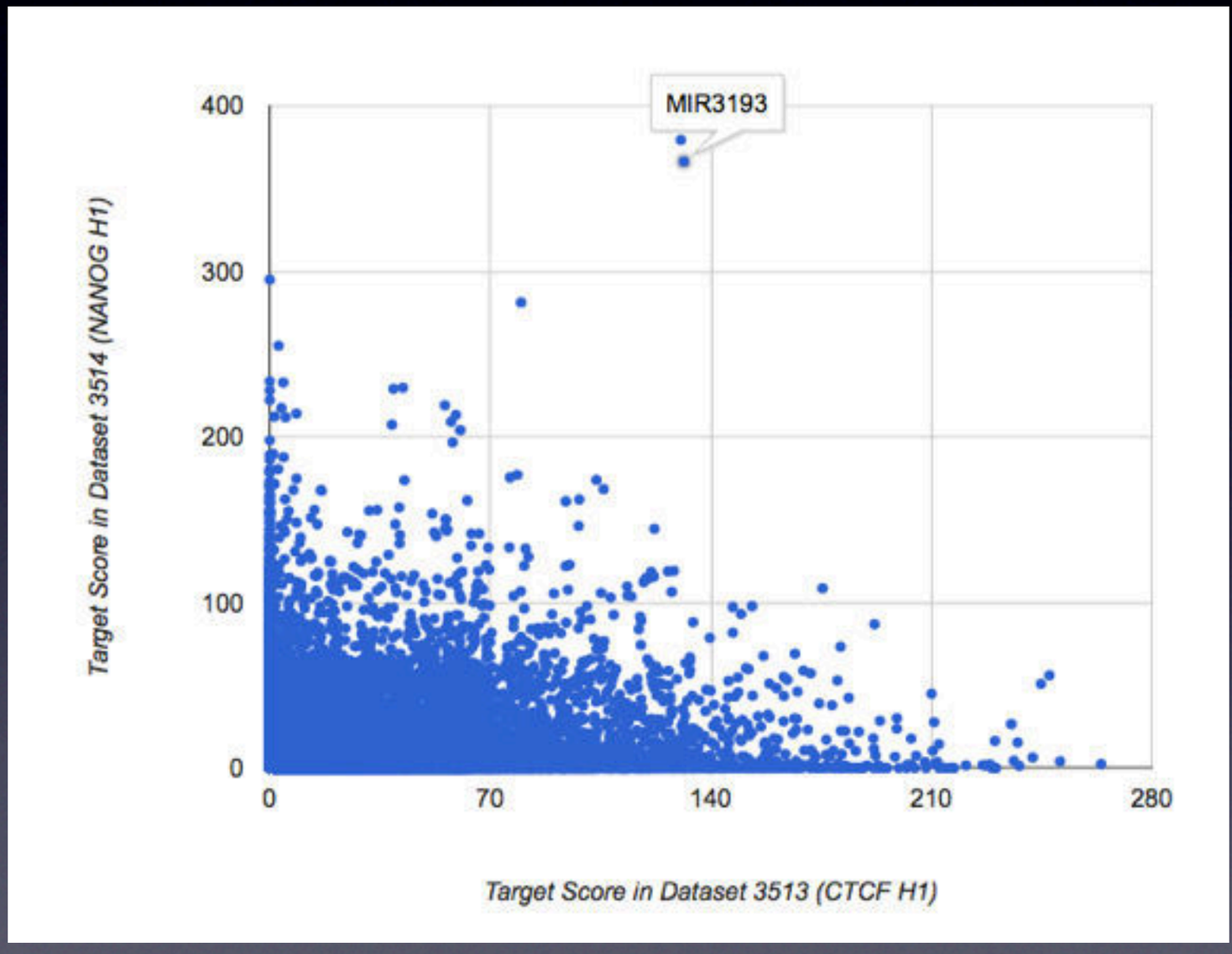
Table view

Number of rows: 10

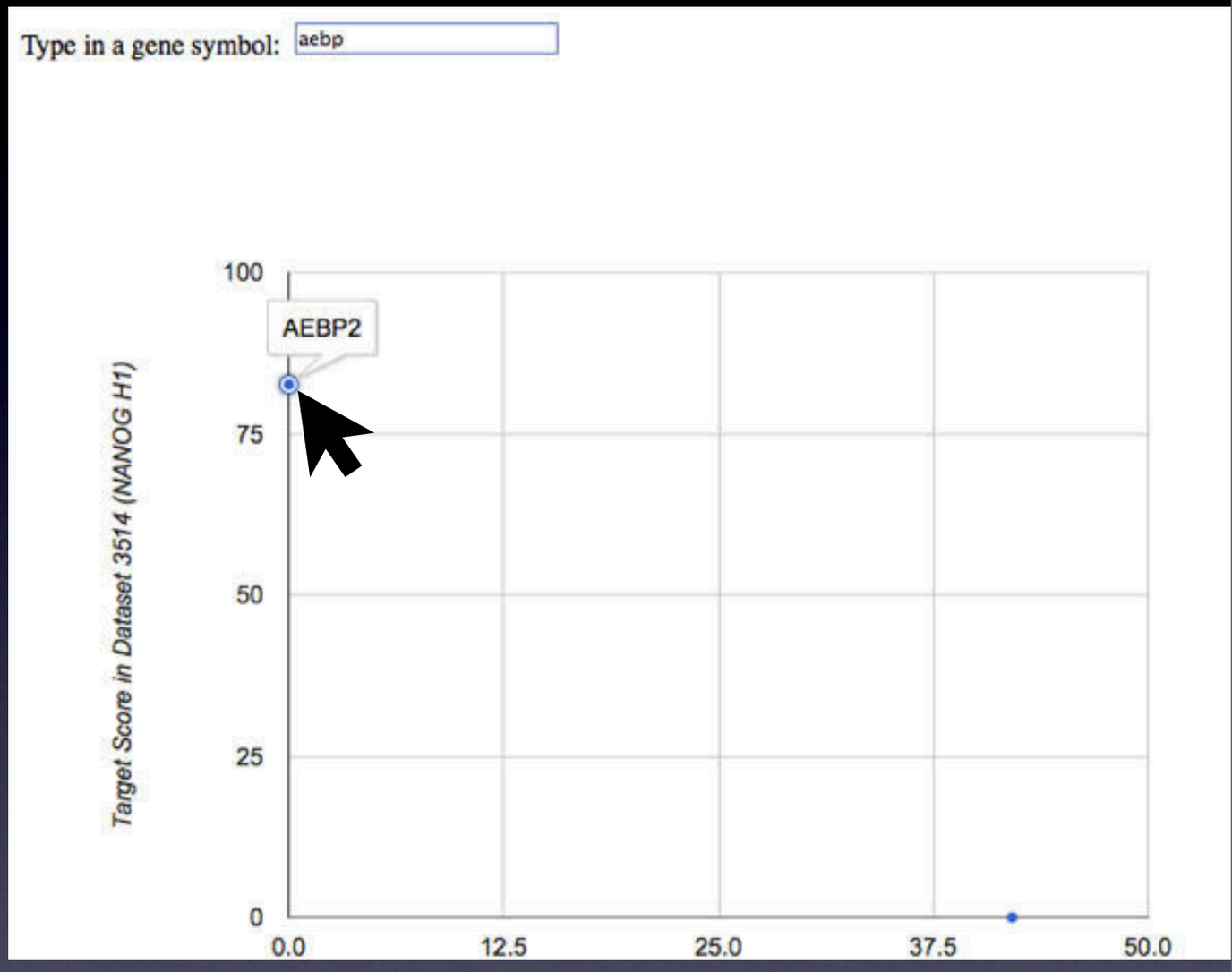
	Refseq	Target Score in Dataset 3514 (NANOG H1)	Target Score in Dataset 3513 (CTCF H1)
1	A2M	0	0
2	NAT2	0	62.1
3	ACADM	1.4	0
4	ACADS	0	43.5
5	ACADVL	50.1	19.3
6	ACAT1	0	0
7	ACVRL1	10.3	58.1
8	PSEN1	56	0
9	ADA	84	13.8
10	SGCA	0	153.7

◄ ►

Graph view


MIR3193
400
300
200
100
0
0
70
140
210
280
Target Score in Dataset 3514 (NANOG H1)
Target Score in Dataset 3513 (CTCF H1)

After filtering by AEBP



Number of rows: 10

	Refseq	Target Score in Dataset 3514 (NANOG H1)	Target Score in Dataset 3513 (CTCF H1)
1	AEBP2	82.6	0
2	AEBP1	0	42.1

Summary of ARC

- visualization
  + statistics in real time
- searching data by meta-data
  + by data
- integration with thousands of datasets or more

Acknowledgment

X. Shirley Liu lab

Tao Liu

Len Taing

Bo Qin

Myles Brown Lab

We are **recruiting** full-time bioinformatic analyst http://cfce1.dfci.harvard.edu/cfce/careers.

Zhang lab

Tongji team

Thank you for your attention!