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

<table>
<thead>
<tr>
<th>Function</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlas</td>
<td>one gene</td>
<td><em>target prediction</em> for this gene in all datasets</td>
</tr>
<tr>
<td>Recommendation</td>
<td>one dataset</td>
<td>top $n$ nearest datasets sorted by distance</td>
</tr>
<tr>
<td>Comparison</td>
<td>two or more datasets</td>
<td>matrices of <em>target prediction</em></td>
</tr>
</tbody>
</table>
## Comparison of APT and ARC

<table>
<thead>
<tr>
<th>Atom</th>
<th>APT</th>
<th>ARC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position (billions bp)</td>
<td>Gene (thousands)</td>
</tr>
<tr>
<td>Value</td>
<td>Profile or Interval</td>
<td>Target Prediction Score</td>
</tr>
<tr>
<td>Find the most</td>
<td>Visual</td>
<td>Recommendation</td>
</tr>
<tr>
<td>correlated tracks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find binding sites among datasets</td>
<td>Visual</td>
<td>Atlas</td>
</tr>
<tr>
<td>Find differences between two datasets</td>
<td>Visual</td>
<td>Comparison</td>
</tr>
</tbody>
</table>
Implementation

Target Prediction

Distance

Web

Statistics

Algorithm from Qianzi et al. (2011)
Cancer Research

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

$$D = 1 - \rho_{X,Y}$$
Three stories about ARC
Check the atlas of ESR2 gene regulated by MYC
Input a gene

Gene Symbol
esr2

ESR2: estrogen receptor 2 (ER beta)

Gene Symbol
esr2

Show atlas
Result Page

Distribution

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

Dataset list

Load to Genome Browser
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
After narrowing
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

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

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

Figure from Yusufzai et al. (2004) Molecular Cell
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

Target Score in Dataset 3514 (NANOG H1) 0 380
Details of Dataset 3514; Summary of Dataset 3514

Target Score in Dataset 3513 (CTCF H1) 0 264
Details of Dataset 3513; Summary of Dataset 3513

Type in a gene symbol:  
### Table view

<table>
<thead>
<tr>
<th>Refseq</th>
<th>Target Score in Dataset 3514 (NANOG H1)</th>
<th>Target Score in Dataset 3513 (CTCF H1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>62.1</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>43.5</td>
</tr>
<tr>
<td>5</td>
<td>50.1</td>
<td>19.3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>10.3</td>
<td>58.1</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>84</td>
<td>13.8</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>153.7</td>
</tr>
</tbody>
</table>
Graph view
After filtering by AEBP
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!