How to make (more) sense of chromatin state data?

Wouter Meuleman
MIT & Broad Institute
ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

Affiliations | Corresponding author

Published online 28 February 2012

Unsupervised pattern discovery in human chromatin structure through genomic segmentation

Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes & William Stafford Noble

Affiliations | Contributions | Corresponding author

Received 01 July 2011 | Accepted 14 February 2012 | Published online 18 March 2012
These methods allow us to summarize our “raw” data:

A color here…

…corresponds to a state here:

```
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Emissions</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TssA</td>
<td></td>
<td>0.7%</td>
</tr>
<tr>
<td>TssAFlnk</td>
<td></td>
<td>0.5%</td>
</tr>
<tr>
<td>TxFlnk</td>
<td></td>
<td>0.1%</td>
</tr>
<tr>
<td>Tx</td>
<td></td>
<td>3.6%</td>
</tr>
<tr>
<td>TxWk</td>
<td></td>
<td>11.6%</td>
</tr>
<tr>
<td>EnhG</td>
<td></td>
<td>0.4%</td>
</tr>
<tr>
<td>Enh</td>
<td></td>
<td>2.8%</td>
</tr>
<tr>
<td>ZNF/Rpts</td>
<td></td>
<td>0.2%</td>
</tr>
<tr>
<td>Het</td>
<td></td>
<td>2.6%</td>
</tr>
<tr>
<td>TssBiv</td>
<td></td>
<td>0.1%</td>
</tr>
<tr>
<td>BivFlnk</td>
<td></td>
<td>0.1%</td>
</tr>
<tr>
<td>EnhBiv</td>
<td></td>
<td>0.1%</td>
</tr>
<tr>
<td>ReprPC</td>
<td></td>
<td>1.2%</td>
</tr>
<tr>
<td>ReprPCWk</td>
<td></td>
<td>8.3%</td>
</tr>
<tr>
<td>Quies</td>
<td></td>
<td>67.8%</td>
</tr>
</tbody>
</table>
```
Many more samples are being profiled
(cell types, disease vs. control, personal epigenomics, etc)

The genome is large
(shown here is only 0.0267% of the genome)
Chromatin state calls across many samples can be viewed as an alignment of sequences with a finite alphabet.

There are good ways of modeling such alignments: logos!

*Information content* of a region, considering background
tinyurl.com/epilogos-feedback
Kullback-Leibler divergence (relative entropy)

\[ D_{KL}(P||Q) = \sum_i p(i) \log_2 \left( \frac{p(i)}{q(i)} \right) \]

\( P \) = observed chromatin state frequencies at any single genomic location, e.g.:

\begin{array}{cccccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
100 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}

local chromatin state frequency (%)

\( Q \) = expected frequencies based on genome-wide state occurrences:

\begin{array}{cccccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
0.7 & 0.5 & 0.1 & 3.6 & 11.6 & 0.4 & 2.8 & 0.2 & 2.6 & 0.1 & 0.1 & 0.1 & 1.2 & 8.3 & 67.8 \\
\end{array}

epigenome-wide chromatin state frequency (%)
http://epilogos.broadinstitute.org
Some potential uses of epilogos

- Interactive visualization
- Consensus epigenomes
- Comparative epigenomics
Comparative epigenomics of ESC vs. “rest”

Select two subsets of epigenomes, build separate epilogos and compare them.
Comparative epigenomics of ???

Select two subsets of epigenomes, build separate *epilogos* and compare them.

Arbitrary subsets of epigenomes, such as these ones (red vs. blue)
Comparative epigenomics of male vs. female

Figure showing comparative epigenomics analysis with a focus on the XIST gene and its expression levels on the X chromosome. The graphs display the magnitude of difference in epigenetic markers across various genomic locations (Mb) for both male and female samples. The XIST gene is highlighted to show differences in expression levels between the two sexes.
Some potential uses of *epilogos*

- **Interactive visualization**
- **Use with third-party data**
- **Consensus epigenomes**
- **Comparative epigenomics**
- **Temporal pattern analysis**
- **Spatial pattern analysis**

("phyloepigenomic" modeling)

(*de novo* motif discovery & search)
input: mapped ChIP-seq reads of histone tail modifications

~10GB

Actual integration with Roadmap reference epigenomes

output: chromatin state calls according to Roadmap models

~1MB

Use with third-party data

your sample?
Some potential uses of epilogos

- Interactive visualization
- Use with third-party data
- Consensus epigenomes
- Comparative epigenomics
- Temporal pattern analysis
- Spatial pattern analysis

("phyloepigenomic" modeling)
ENCODE & Roadmap Epigenomics Consortia (data)

Soheil Feizi (information theory)

Apostolos Papadopoulos, Terrance Liang, Kevin Liu, Tiffany Chen & Miguel Medrano (web-application & more)

Ting Wang, Xin Zhou & Daofeng Li (WashU browser integration)

Luca Pinello & Nezar Abdennur (website technology)

YOU!

→ tinyurl.com/epilogos-feedback

→ epilogos.broadinstitute.org
Interpretation of large-scale (epi)genomic datasets through information-based dimensionality reduction