Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome

The problem

- Different cells have different patterns of gene expression. Cells respond to their environment by changing the expression of genes.
- Enhancers are the DNA sequences used to drive the specific patterns of expression.
- People want to use a genomics approach to identify regulatory circuitry that controls patterns of gene expression.
- Once, we thought that if you knew the DNA sequence of a single enhancer that specified a pattern of expression that you could use this to identify other examples.
The problem

- The DNA sequence of an enhancer is a poor predictor of function. It only identifies a tiny subset of enhancers.
  - Canonical palindromic binding site TGACGTCA
  - Approx freq of this seq is 1/65kb (one in 4^8), genome is 3×10^9, so in theory 46,000 sites, actual count is just over 32,000
  - ChIP:Saco = A PCR/sequencing technique Genome wide 6302 sites, many are NOT canonical
  - Seq analysis id’s too many and many real ones don’t match the motif.
  - Novel = new, not predicted.

For chromosome 10
The problems

- Enhancers may not be near the promoters they regulate and so how do you identify their targets? Some experiments indicate that only ~20% of enhancers are near the promoter (~1kb).

- In addition, the presence of the DNA sequence for an enhancer does not mean that the enhancer is being used.
The problem

• The situation is a bit better for promoters but not much.
A possible solution?

- Epigenetic modifications may provide patterns that identify enhancers and promoters
- They may also provide patterns that say NOT enhancer or NOT promoter.
Now they do the entire genome but they began with data from a consortium of scientists. 30 megabases (1% of the genome) 44 genomic regions
Known genes that are ON.
Known genes that are OFF.
Known NOT genes.
The over-arching hypothesis is that the transcription machinery uses histone modifications as a first identifier of where promoters and enhancers lie AND that cells modify these histone marks as a way to regulate gene expression.

The goal of this paper is to see if histone marks can be used by the investigator to identify active genes.

To figure out if a new technique can work it is nice to know what the answer is.

Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome

Nathaniel D Heintzman1,2, Rhona K Stuart1, Gary Hon1,2, Yutao Fu3, Christina W Ching1, R David Hawkins1, Leah O Barrera1,2, Sara Van Calcar1, Chunxu Qu1, Keith A Ching1, Wei Wang3, Zhiping Weng1,6, Roland D Green7, Gregory E Crawford8 & Bing Ren1,9

First we start with a circular process:

Take KNOWN promoters and KNOWN enhancers and see if they are covered by recognizable patterns of histone modifications. Then they ask if these patterns alone can be used to identify these promoters and enhancers in the genome.

Wait! - we must have had a way to identify KNOWN promoters and enhancers because they start with them. So why bother? Because these KNOWN promoters were identified because of substantial labor and effort by a large number of people over a long period of time. We need an easier way.
Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome


Known promoters that are on versus Known promoters that are off versus non promoter

↓

1. What histone marks do they have?

↓

Are these marks a reliable way to identify them?
   Known promoters that are on
   vs
   Known promoters that are off
   vs
   non promoter

Eventually ask if the new method can discover NEW promoters!
Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome


Known enhancers that are on versus Known enhancers that are off versus non enhancer

1. What histone marks do they have?

Are these marks a reliable way to identify them?

Known enhancers that are on

vs

Known enhancers that are off

vs

non enhancer

Eventually ask if the new method can discover NEW enhancers!
Performed chromatin immunoprecipitation

Used antibodies that recognized five different histone modifications.

H4ac
H3ac
H3K4me1
H3K4me2
H3K4me3

Analyze by ChIP:Chip using an ENCODE array.
Now they do the entire genome but they began with data from a consortium of scientists.

- 30 megabases (1% of the genome)
- 44 genomic regions
- Known genes that are ON.
- Known genes that are OFF.
- Known Enhancers.
- Known NOT genes and Known Not Enhancers.
What material was used?

DNA chip

ENCODE fragments: 30 Mb database of transcriptionally active sequences in humans. 38 bp resolution

Chromatin will be prepared from HeLa cells.

10 kb blocks around each landmark. Landmarks are known promoters and known enhancers

Start with
104 TSSs (core promoters) defined as active promoters in HeLa cells

104 TSSs defined as inactive in HeLa cells
TSS (transcription start site) is synonymous with the core promoters
Conceptual explanation of simple clustering and heat maps
Align so that transcription start sites are centered.

gene 1

gene 2

gene 3

gene 4

gene 5

gene 6

gene 7

gene 8
Use genomic tiling array to determine position of each mark.

-5 kb → +5 kb

- gene 1
- gene 2
- gene 3
- gene 4
- gene 5
- gene 6
- gene 7
- gene 8
Sort the genes into self-similar groups

-5 kb  
  gene 1  
  gene 2  
  gene 3  
  gene 4  
  gene 5  
  gene 6  
  gene 7  
  gene 8  
  +5 kb

similar to one another
Sort the genes into self-similar groups
Compact so that many genes can be viewed in a small space.
Compact so that many genes can be viewed in a small space.
Compact so that many genes can be viewed in a small space.

Called a heat map:
- black = average
- red = greater than average
- green = less than average

Color scheme can vary. In papers you should double check what the colors mean.
Compact so that many genes can be viewed in a small space.

Called a heat map
black = average
red = greater than average
green = less than average

Each line is very thin. About 200 lines here. Each line is like a track in Human Epigenome Browser.

Color scheme can vary. In papers you should double check what the colors mean.
Centered on known TSSs.
How do we align enhancers?
1. Don’t use PI promoters because there is no positive signal. These are thought to be promoters that are off. *Other non-promoter regions might look like this.* Probably can eventually use an off promoter mark.

2. Shape of the peak and intensity are going to be the what we are looking for.
**Weak** H3K4me1 and **Strong** H3K4me3 is most strongly correlated with an active core promoter. The shape is also important. Notice the nucleosome-free region at +1.
The opposite is found to be most tightly correlated with enhancers. **Strong** H3K4me1 and **Weak** H3K4me3 is most strongly correlated with an active enhancer. The shape is also important. Notice that the shape is different than the promoter mark shape.
44 genomic regions - 208 well described genes

ENCODE fragments are discrete regions. This was used to generate the previous heat maps. Representing them as a line.

30 mbases of genomic DNA ENCODE fragments

Want to use this information to directly ID promoters and enhancers.

Really trying to categorize sequence as
1. promoter
2. enhancer
3. not 1 or 2.
How do we figure out what to do?

**TRAINING SET**
Calculate the average profile for a mark.

e.g. P4 H3ac could be one training set.
How do we figure out what to do?

Calculate the average profile for a mark.

Could we ID promoters with this average profile? Need something to examine but the answer must be known.
How do we figure out what to do?

**TRAINING SET**

Calculate the average profile for a mark.

For each member of the TEST SET (-) determine how closely the histone mod profile matches the shape of one of the six TRAINING SET histone mods.

Do this by determining the Euclidian distance.

**30 Megabases of genomic DNA**

TEST SET is all overlapping 10 kb windows from the same 30 Megabase region. There are more windows here that in the training set.
Euclidian distance

Plot the peak on a coordinate system. Determine how many transformations are required to convert a test plot into the average.

Done for all six histone marks.
How do we figure out what to do?

**TRAINING SET**
Calculate the average profile for a mark.

For each member of the TEST SET (-) determine how closely the histone mod profile matches the shape of one of the six TRAINING SET histone mods.

Do this by determining the Euclidian distance.

30 Megabases of genomic DNA

TEST SET is all overlapping 10 kb windows from the same 30 Megabase region. There are more windows here that in the training set.
How do we figure out what to do?

TRAINING SET
Calculate the average profile for a mark.

30 Megabases of genomic DNA

TRAINING SET

TEST SET is all overlapping 10 kb windows from the same 30 Megabase region

For each member of the TEST SET (-)
determine how closely the histone mod profile matches the shape of one of the six TRAINING SET histone mods.

Do this by determining the Euclidian distance.

Discrimination Filter
The identified member must have at least a 0.4 correlation statistical correlation with one training set.

It must also more closely match enhancer or promoter - not both.

If either test fails then it is discarded.
So what!!! This shows that a training set can identify its own members !!! Big deal. How can we determine if it can determine members that ARE NOT in the training set!!! This is what they want.

**TRAINING SET**

Calculate the average profile for a mark.

For each member of the TEST SET (-)

determine how closely the histone mod profile matches the shape of one of the six TRAINING SET histone mods.

Do this by determining the Euclidian distance.

**TEST SET**

TEST SET is all overlapping 10 kb windows from the same 30 Megabase region

Discrimination Filter

The identified member must have at least a 0.4 correlation statistical correlation with one training set.

It must also more closely match enhancer or promoter - not both.

If either test fails then it is discarded.
Demonstrate this by making training sets that are missing some members.

**TRAINING SET MINUS RANDOM 10%**
that is missing a random 10% of the promoters

Search the test set and see how well it does in retrieving the missing 10%.
Repeat this process with a different randomly selected 10%.

**TRAINING SET**

**TEST SET** is all overlapping 10 kb windows from the same 30 Megabase region

H3K4me3 is good for promoters
H3K4me1 is good for enhancers
but together they work better than either one separately.
Decide that you need to use the output from two histone modifications for best predictability.
How well does it work to identify the 10% omitted from the TRAINING SET?

Figure S6

Cross-validation to determine optimal combination of histone modifications for class P3 based on recovery of training set vs. total predictions

Number of predictions needed to achieve these results.

Each group differs by 10% of its members. Height is different because the outliers are different.

This is just for the P3 Promoter class.

Number of correct & incorrect predictions
But how can we tell if it really works?

Functionally test a predicted enhancer that was located 6 kb upstream of SLC22A5 - little is known about its transcriptional regulation.
Carnitine transporter

Test for enhancer properties in transfected HeLa cells.

Showed that this approach successfully identified KNOWN promoters and enhancers AND identified NEW enhancers. These worked as enhancers when physically tested.

“Take home: Chromatin modifications at enhancers are globally related to cell-type-specific gene expression”
• H3K4me1, H3K4me3, H3K27ac
• Surprise! Chromatin marks at promoters are similar in all cell types.
• CTCF occupancy is same across all cell types. CTCF is an insulator binding protein.
• Action is at the enhancers
  – H3K4me1 and H3K27
  – Variability is greatest between cell types
5 human immortalized cell lines

cervical carcinoma HeLa cells
immortalized lymphoblast GM06690 (GM)
leukaemia K562
embryonic stem cells (ES)
BMP4-induced ES cells (dES)

* 55,454 enhancers examined
At core promoters the marks are the same in most tissue types. This was a big surprise. We are not shown it but expression of these genes differ.
At core promoters the marks are the same in most tissue types. This was a big surprise. We are not shown it but expression of these genes differ.

But enhancer marks are diagnostic for the tissue type.

Example

TSS

Enhancers
“Chromatin modifications at enhancers are globally related to cell-type-specific gene expression”

Example
55,454 enhancers examined
5 cell line types
An enhancer mark can indicate increased responsiveness

- Interferon activates Stat1 binding.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>TF</th>
<th>% at HeLa enhancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF7</td>
<td>ER</td>
<td>32.6% ($P &lt; 1 \times 10^{-300}$)</td>
</tr>
<tr>
<td>HCT116</td>
<td>p53</td>
<td>21.4% ($P = 2.3 \times 10^{-30}$)</td>
</tr>
<tr>
<td>ME180</td>
<td>p63</td>
<td>25.8% ($P &lt; 1 \times 10^{-300}$)</td>
</tr>
<tr>
<td>HeLa-IFN-γ</td>
<td>STAT1</td>
<td>25.3% ($P &lt; 4.5 \times 10^{-142}$)</td>
</tr>
</tbody>
</table>

When you add interferon the ones with the enhancer mark (H3K4me1) turn on even both groups show STAT 1 binding.
Of course the work continues

Table 3. Distinctive Chromatin Features of Genomic Elements

<table>
<thead>
<tr>
<th>Functional Annotation</th>
<th>Histone Marks</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoters</td>
<td>H3K4me3</td>
<td>Bernstein et al., 2005; Kim et al., 2005; Pokholok et al., 2005</td>
</tr>
<tr>
<td>Bivalent/Poised Promoter</td>
<td>H3K4me3/H3K27me3</td>
<td>Bernstein et al., 2006</td>
</tr>
<tr>
<td>Transcribed Gene Body</td>
<td>H3K36me3</td>
<td>Barski et al., 2007</td>
</tr>
<tr>
<td>Enhancer (both active and poised)</td>
<td>H3K4me1</td>
<td>Heintzman et al., 2007</td>
</tr>
<tr>
<td>Poised Developmental Enhancer</td>
<td>H3K4me1/H3K27me3</td>
<td>Creyghton et al., 2010; Rada-Iglesias et al., 2011</td>
</tr>
<tr>
<td>Active Enhancer</td>
<td>H3K4me1/H3K27ac</td>
<td>Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias et al., 2011</td>
</tr>
<tr>
<td>Polycomb Repressed Regions</td>
<td>H3K27me3</td>
<td>Bernstein et al., 2006; Lee et al., 2006</td>
</tr>
<tr>
<td>Heterochromatin</td>
<td>H3K9me3</td>
<td>Mikkelsen et al., 2007</td>
</tr>
</tbody>
</table>


