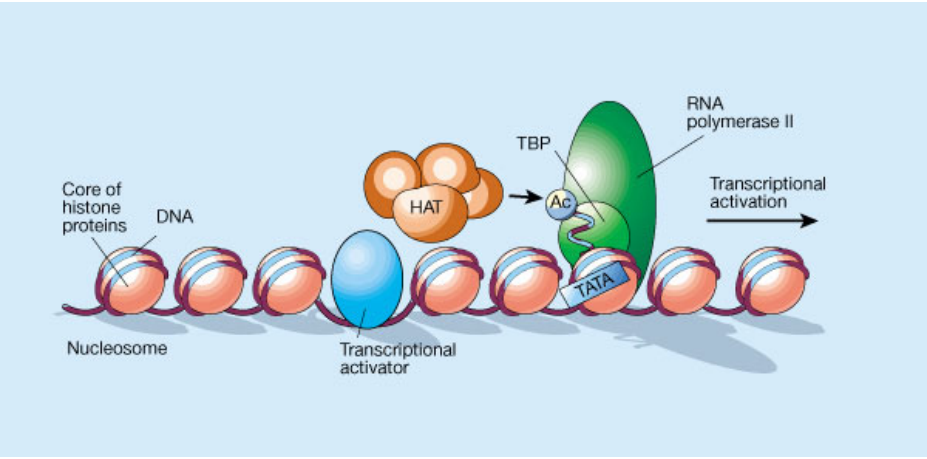
Pathway Enrichment in DNase1-Footprinting Data

David Lamparter, Aurelien Mace

February 21, 2013

Model of gene activation



- Transcription-factor binds DNA
- TF recruits factors that modify DNA to make it 'transcription-ready'.
- Polymerase is recruited and transcription is intiated.

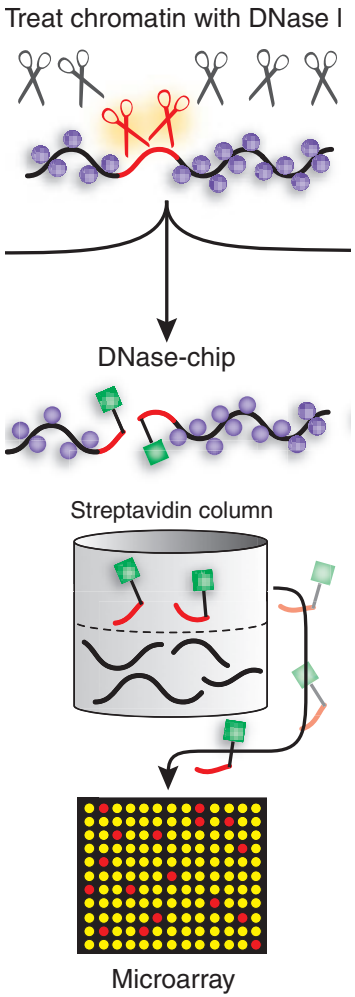
Note: Where the TF is bound, there is no nucleosome!

Finding Binding Sites

Assume you want to know where Transcription Factor bind.

- *Idea:* Exploiting Nucleosome Displacement to find binding sites.
- *Trick:* DNase1 Diggests DNA more efficiently if it isn't bound by protein.
- Use the different rate in digestion to find out binding sites.

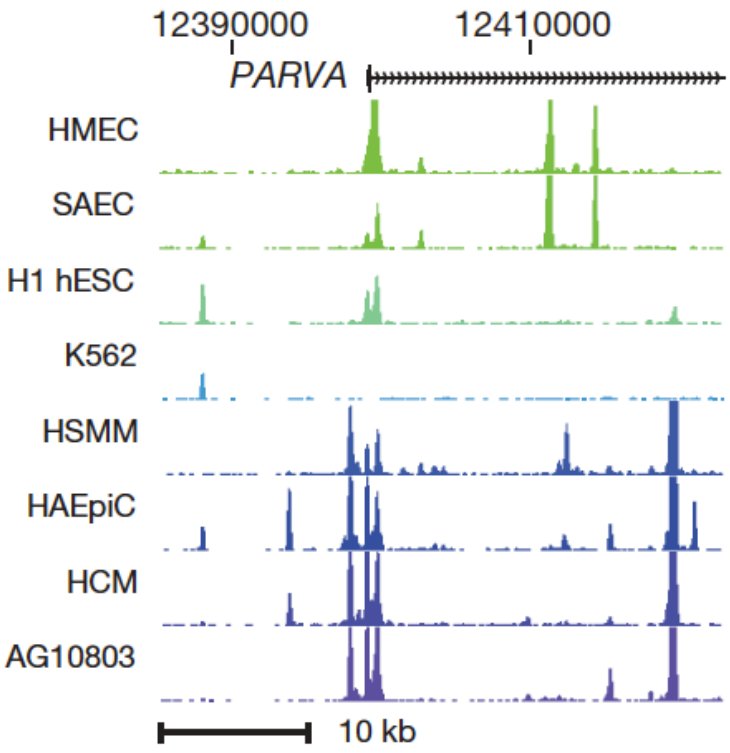
Procedures:DNAse1-seq



Steps:

- Digest (Cut) DNA with DNAse1
- Tag ends of the DNA (with streptavidine)
- further digest DNA with other nuclease to get short fragments
- isolate only the tagged fragments.
- amplify and sequence

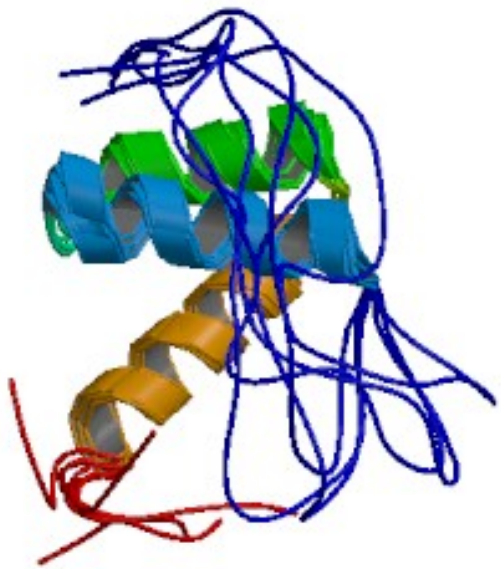
Example data for Chromatin Accessibility Assay



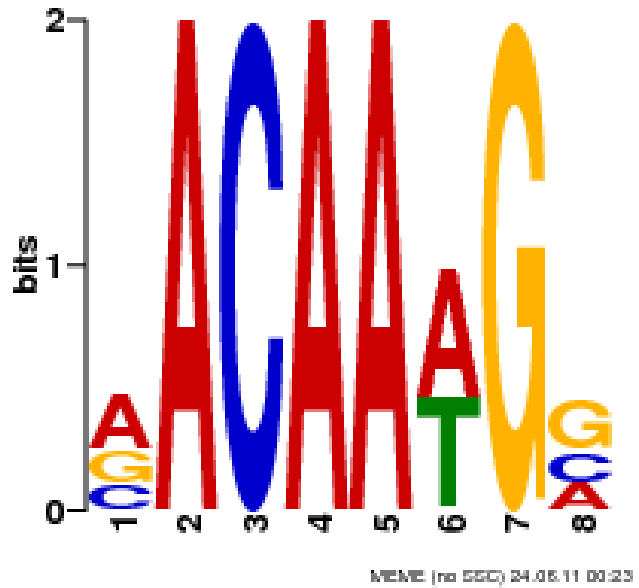
- 'Raw' data for chromatin accessibility assay.
- Peaks imply that some proteins are bound to the DNA at this positions.
- Example: Beginning of PARVA-gene has protein bound to it in some cell lines. This could imply active transcription.
- Peaks are are referred to as DNAse Hypersensitivity Sites (*DHS*).

TF-Motifs: The other side of the Story

Many Transcription Factors bind only at certain positions. Partly because they bind only certain binding motifs.

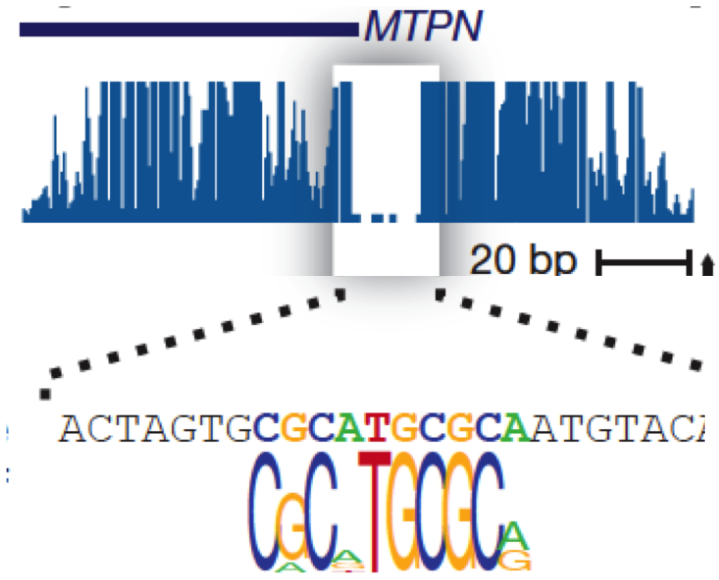


Nanog Structure



Nanog Logo

Example data for Chromatin Accessibility Assay



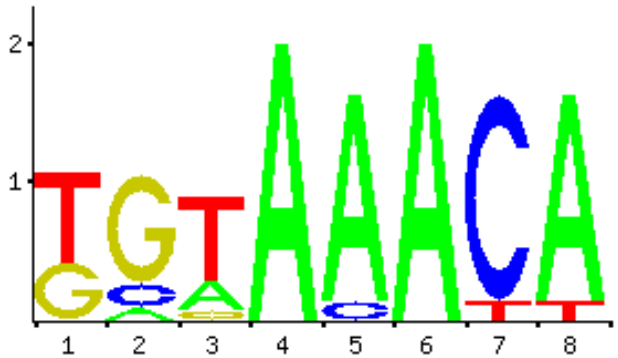
- Top: We see a DHS with a gap. This gap is called a footprint.
- Footprint is indication of precise binding postion of Transcription Factor.
- Reason: TF binding also impairs DNAse1 cutting. DNA is not freely accessible
- Bottom: Motif for Transcription Factor *Nrf1*. It sits precisely in the footprint.
- Indicates that TF *Nrf1* might be bound to DNA.

Data

- around 50 genome-wide Footprint datasets.
- a large motif library.
- a mapping from Footprint to likely genes regulated.
- Many Gene sets of genes that are involved together in a Pathway.(say genes involved in cholesterol metabolism etc)

Research Question

Main question: Can we find motifs that function preferentially in certain Pathways?
Example: **FOXO3**



FOXO3 is a well studied Transcription factor. It is involved in apoptotic cell death in the brain.

- Can we uncover this relationship from our footprinting data and the motif alone?
- do we need to restrict cell type to neural tissues?
- We will try to see how well our approach works for a handful of well studied examples

What you will learn

- Get to know the computational environment we're working in. Probably R.
- Some understanding of the biology of transcription regulation.
- Some understanding of the statistics involved. Nothing too scary.
- Some feel what a bioinformatics project can look like.