

Mechanisms and models of distal enhancers of inducible gene expression

Martin Hemberg

UC Berkeley
February 28, 2012

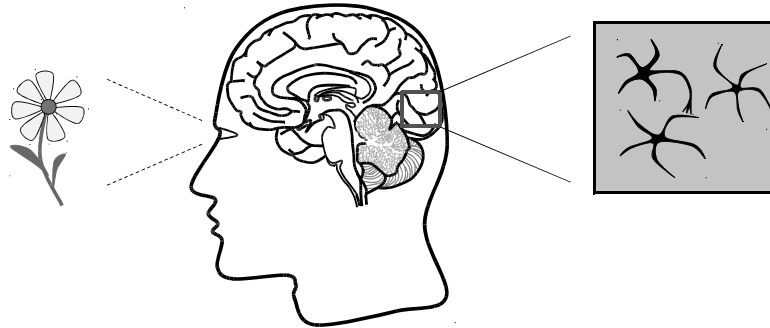


First of all, I'd like to thank the search committee for inviting me and giving me the opportunity to come here to Berkeley and present my work today.

My main scientific interest is in understanding gene regulation in a quantitative way.

[Gene regulation is a process that is central in biology since almost every biological process involves changes in gene expression. Understanding gene regulation is important from a clinical point of view since the process is involved in numerous diseases, but also from a basic science point of view, there are numerous large gaps in our understanding of this fundamental biological process.]

External stimuli change synapses



Hubel & Wiesel, 1970's

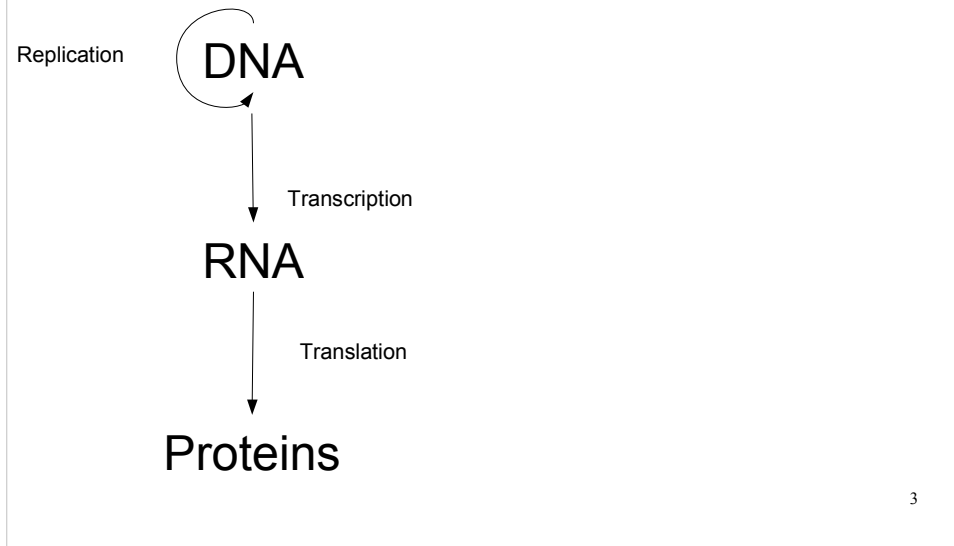
2

One of the most important properties of biological organism is their ability to adapt to the environment.

A particularly striking example of such adaptation occurs during brain development. During brain development, connections between neurons are influenced by sensory experience. That is, synapse formation and the synapse strengths respond to external environmental stimuli.

It has been shown that this and almost every other adaptive process involve changes in gene expression

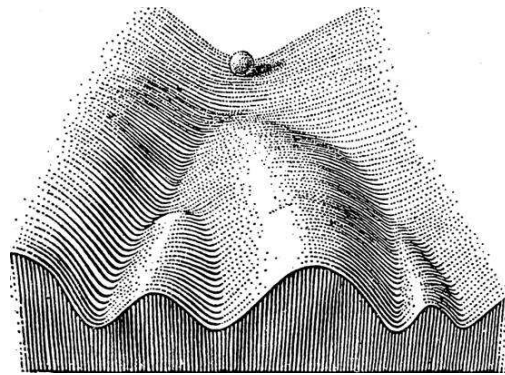
What is gene expression?



The key principle of molecular biology, known as the central dogma, is that information is stored in the DNA. Gene expression refers to the first step in this process, the act of transcribing DNA into RNA. The RNA is then translated into proteins. Proteins are considered the main workhorses of the cell and when we talk about genes, we usually refer to a protein.

What is gene regulation?

- Click to add an outline



Waddington, 1953 ⁴

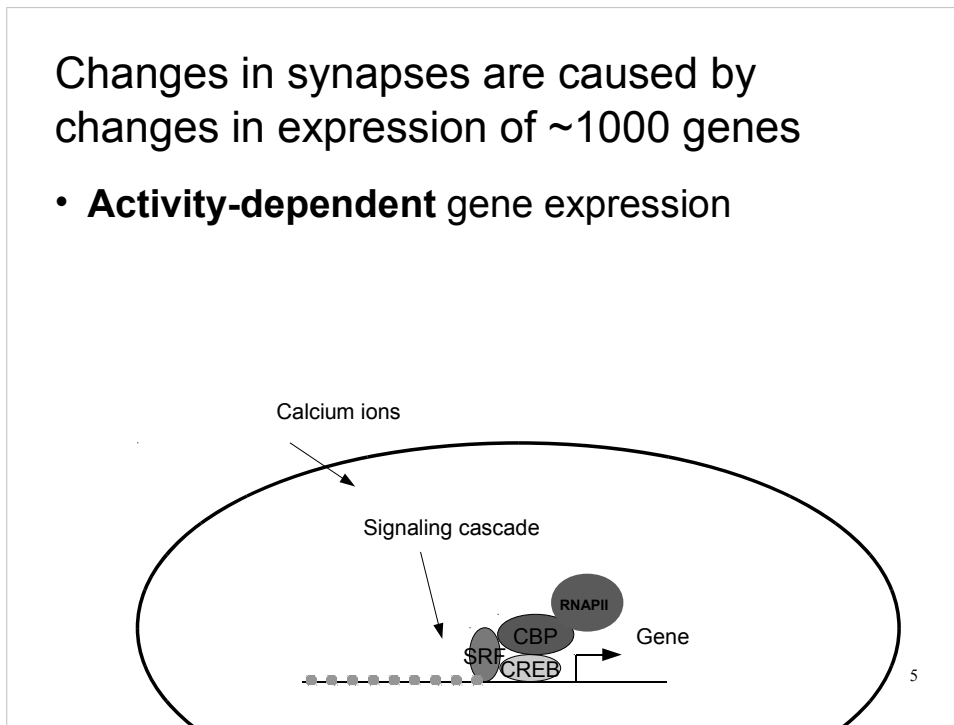
Gene regulation is the process by which the cell controls the level of transcription of each gene.

Conceptually, you may think of the cell as sitting in a high-dimensional space where each gene corresponds to one coordinate. Upon some external signal, the cell then needs to transition from one state to another.

This notion of cellular states as an epigenetic landscape goes back to Waddington in the 1950s.

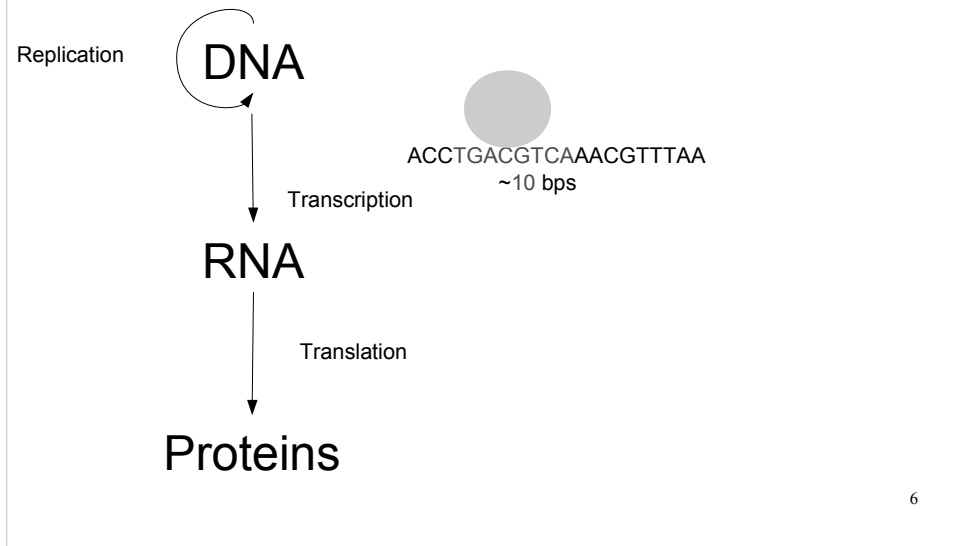
Changes in synapses are caused by changes in expression of ~1000 genes

- **Activity-dependent** gene expression



The gene expression programme that leads to changes in synapses has been studied over many years by many labs. It has been shown that the process is triggered by the influx of calcium ions into the cell. The calcium triggers a signaling cascade which in turn leads to a change in gene expression of around a thousand genes. This is known as activity-dependent gene expression.

Transcription Factors (TFs) bind to DNA motifs

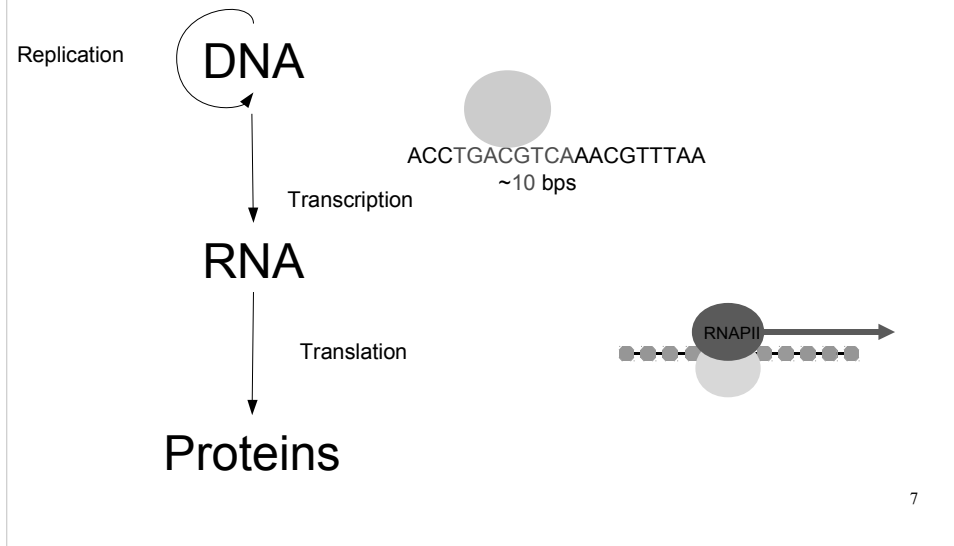


Gene regulation is a complex process and there are many different mechanisms involved. Perhaps the most important one is binding by transcription factors (Tfs).

TFs are proteins that bind to specific DNA sequences, typically ~10 base pairs, and these sequences are known as motifs.

For example, here we have our protein of interest and it happens to prefer binding to the sequence marked in red. Each TF has its own preferred motifs which dictates where it will bind.

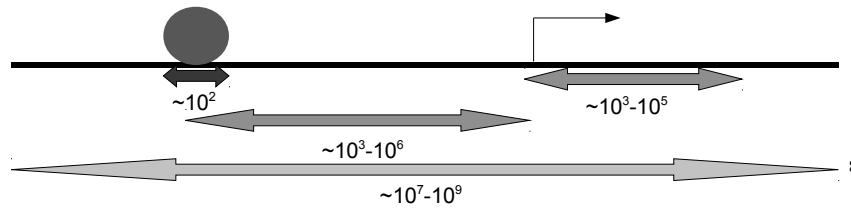
TFs bind at promoter to recruit **RNAPII**



Traditionally, Tfs are thought to bind in the promoter region, which is defined as the region near the start of the gene. Tfs at promoters serve to recruit RNAPII, which is the molecule that carries out the actual process of transcribing the DNA into RNA.

Enhancers are distal TF binding sites

- ~25,000 genes
– 2% of DNA



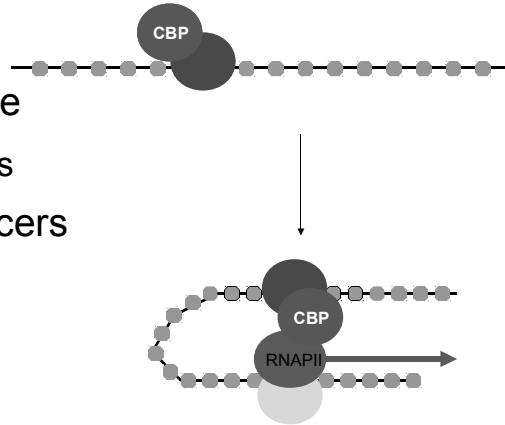
Just to give you a better sense of the numbers here, some salient features of the mouse and human genomes is that they have $\sim 3 \times 10^9$ bases. There are only about 25k genes encoded and the coding parts are no more than $\sim 2\%$ of the genomic real estate. Hence there is a huge amount of DNA for which the function is unclear, although it is believed that much of it serves a regulatory role.

Regulatory sequences far from promoters are known as enhancers and they can be very far from their target genes, ranging from a kilobase to a megabase.

Ideally, as a theoretical physicist what I'd like to do here is to write down the equations that determine where enhancers are located, solve them for the mouse genome and then ask my collaborators to do the experiment to verify my prediction. Unfortunately, biology is very complex and there are so many gaps in our knowledge of enhancers that we cannot take such an approach. Instead we have to take a much more data-driven approach and proceed by asking a series of more simple yes/no questions.

Enhancers characterized by **CBP** binding

- No universal sequence signature
 - Enriched for motifs
- CBP binds at enhancers

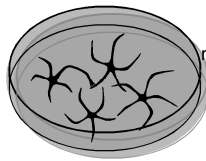


ENCODE, 2007
Heintzman et al, 2007
Roh et al, 2005
Visel et al, 2009

When we began our study, not much was known about enhancers in general and even less about enhancers in activity dependent gene expression. It was known that they are enriched for transcription factor binding and that through a poorly understood looping mechanism enhancers interact with the promoter of the target gene and help drive the expression of the target gene.

Detecting enhancers based on sequence alone is very challenging, but recently it has been shown by Bing Ren and others that enhancers can be identified by the binding of the protein CBP.

Cultured mouse cortical neurons for genome-wide study of activity dependent gene expression



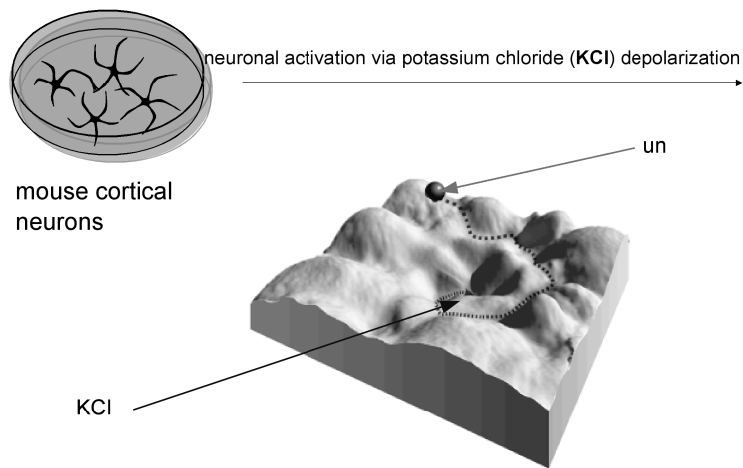
mouse cortical neurons

neuronal activation via potassium chloride (KCl) depolarization



For practical reasons, studying gene expression changes in the brain is very complicated. Instead, we used an experimental set-up involving cultured primary cortical neurons from mouse. The neurons are subjected to elevated concentration of potassium chloride or KCl. This leads to the depolarization of the membrane, triggering an influx of calcium which in turn provides a robust activation of the activity-dependent gene expression program.

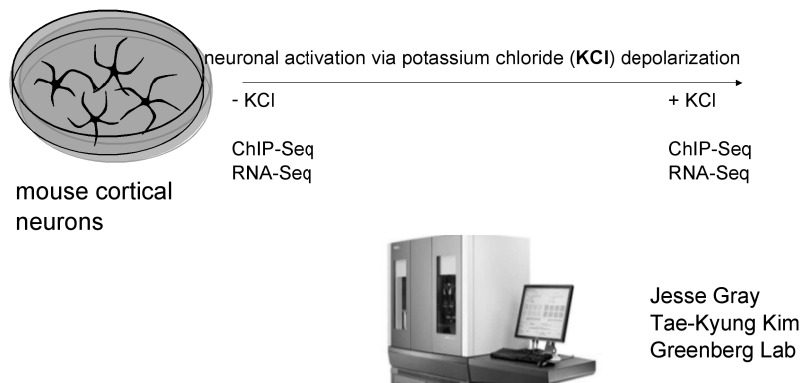
KCl stimulation induces cells to change state



11

Conceptually, you may think this experiment as an impulse to the system that pushes it away from its initial state, the unstimulated condition, and over time it moves to a new dynamical equilibrium.

Genome-wide data obtained using high-throughput sequencing

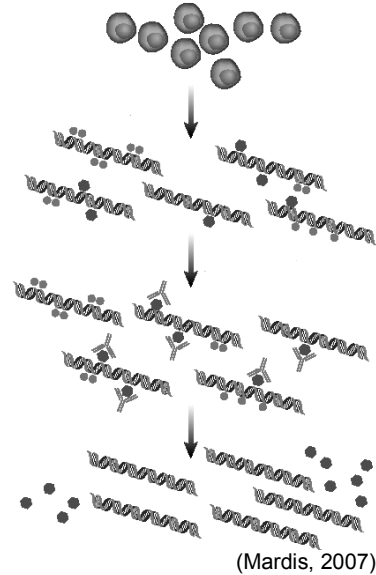


12

We monitored both the TF binding and the gene expression before and after KCl stimulation.

At this point, it is important to acknowledge the contributions of my two collaborators: all of the experiments that I will be talking about were carried out by Jesse Gray and TK Kim, and other members of the Greenberg lab at Harvard Medical School. My role was to be in charge of all the computational and theoretical analyses of the data.

Chromatin immunoprecipitation and sequencing (**ChIP-Seq**) finds protein binding sites *in vivo*

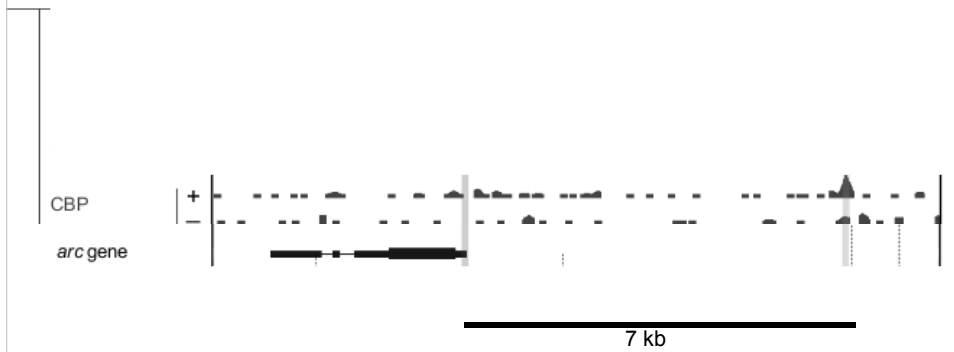


To study enhancers, we first need to know where CBP binds and for that we used a technique called chromatin immunoprecipitation combined with high throughput sequencing, or ChIP-Seq. I don't have time to describe ChIP-Seq in detail here

I will just point out that what ChIP-Seq does is that it provides us with data on the location of transcription factor binding sites throughout the entire genome. The method works by analyzing small fragments of DNA, known as reads, and the number of reads from a given location reflects the amount of binding.

An important feature of chip-seq is that the method is unbiased and that it will provide information about the entire genome all at once.

Inducible CBP binding is necessary but not sufficient for identifying enhancers

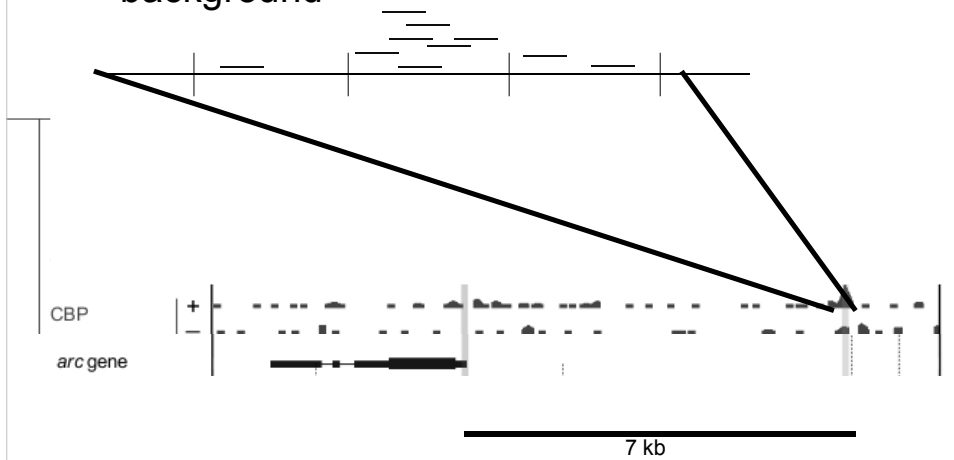


This is a screenshot from the UCSC genome browser a tool that is often used to visualize genomic data. What we have on the x-axis are genomic coordinates and this represents a small portion of the mouse genome. Over here is the *arc* gene which is one of the genes that is upregulated in response to activity. It sits on the negative strand, which means that the start is over here and it is then transcribed in this direction. Up here, I am showing you the ChIP-Seq data for CBP, before and after potassium chloride stimulation. Each blip corresponds to a read and you may think of this as a histogram where the height corresponds to the amount of CBP binding.

Over here is the only activity-dependent enhancer that was known prior to our study. It is clear from our data that there was no binding before stimulation, but high levels afterwards and this kind of peak represents a binding event.

Identifying 28,000 CBP binding sites

- Regions that have significantly more CBP than background



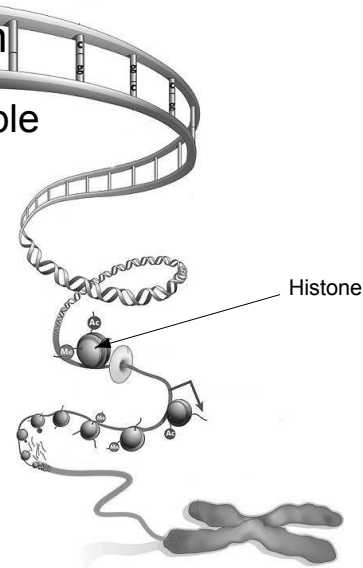
To search for CBP binding sites in a genome-wide manner, we developed a peak-calling algorithm that identifies regions of the genome where the CBP binding is significantly higher than the background.

Using a stringent threshold, we identified 28k such regions all over the genome that were replicated in two different experiments. This number of peaks is of the same order of magnitude as the number of genes.

Unfortunately, CBP binds to other places than enhancers, so what we have here is necessary, but not sufficient for identifying enhancers. So in the next couple of slides I will tell you some more about the biology of gene regulation.

Post-translational modifications of histone tails correlate with function

- ~100 k loci or 1% accessible
 - Open chromatin
 - Cell-type specific



(ENCODE, 2007)

16

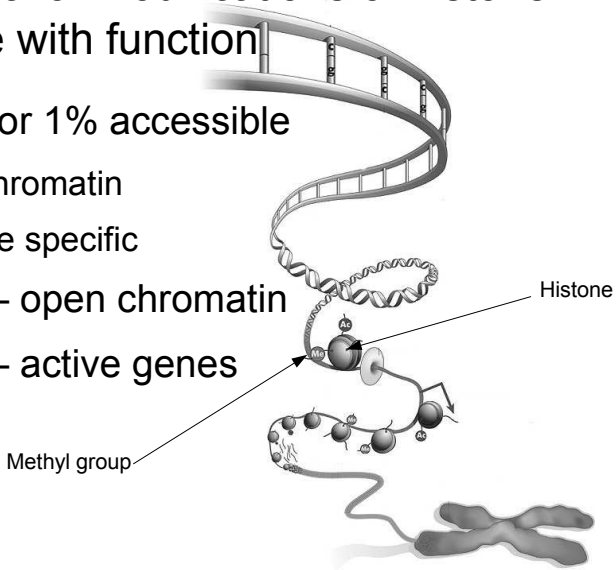
In addition to Tfs, there are many other molecules that bind to DNA and this may prevent TF from reaching its target. This competition provides another layer of regulation.

In particular we have histones. You may think of a histone as a small spool around which the DNA is wrapped. Histones are abundant and only ~1% of the genome is not bound by histones. Tfs will only bind to regions where there are no histones – these sites are known as open chromatin and it is typically found at active promoters as well as distal enhancers.

The patterns of open chromatin differs from cell-type to cell-type and it is one of the reasons why we have different cell-types.

Post-translational modifications of histone tails correlate with function

- ~100 k loci or 1% accessible
 - Open chromatin
 - Cell-type specific
- **H3K4me1** – open chromatin
- **H3K4me3** – active genes



(ENCODE, 2007)

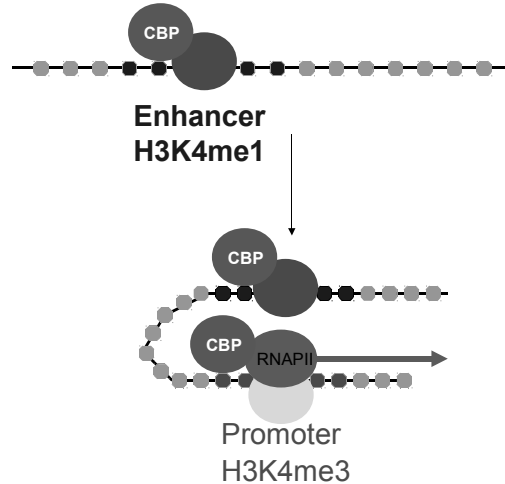
17

Additionally, histones can be biochemically modified by adding for example methyl or acetyl groups. In the 1970s it was first shown that these modifications were correlated with biological function, such as active or repressed genes.

There are more than a hundred different histone modifications but today I will only be talking about two examples: mono- and trimethylation of lysine 4 on histone 3, or H3K4Me1 and H3K4Me3. The trimethylation is associated with active genes and the monomethylation has been shown to be associated with open chromatin.

A combination of CBP and histone modifications identifies enhancers

- **CBP** binding
- **H3K4me1** flanking
- **H3K4me3** absent

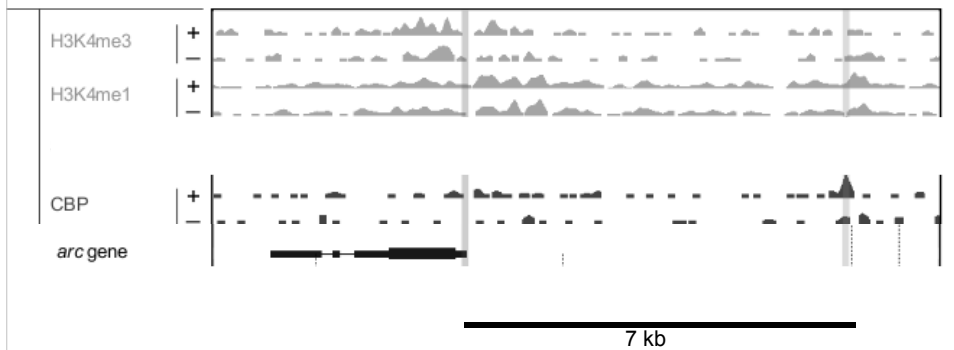


18

Going back to the problem of finding enhancers, we may thus add two additional criteria to our list: high levels of H3K4me1 flanking the CBP peak and low levels of H3K4me3.

So we carried out additional ChIP-Seq experiments for the histone modifications.

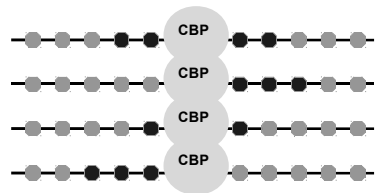
Distal CBP peaks have high levels of H3K4me1 and low levels of H3K4me3



As you can see here, the levels of H3K4me1 are high in the regions flanking both the enhancer and the promoter. This is in contrast to H3K4me3 which is high at the promoter, but not at the enhancer.

Next, I will tell you how we used this pattern in a genome-wide manner to identify enhancers.

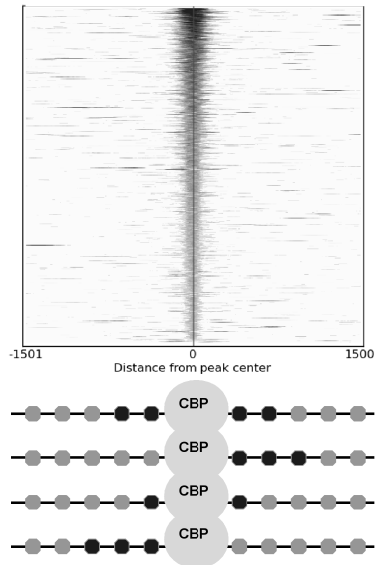
Aligning CBP peaks to calculate binding profiles



20

We first aligned the peaks to the center of the CBP binding as illustrated in this schematic.

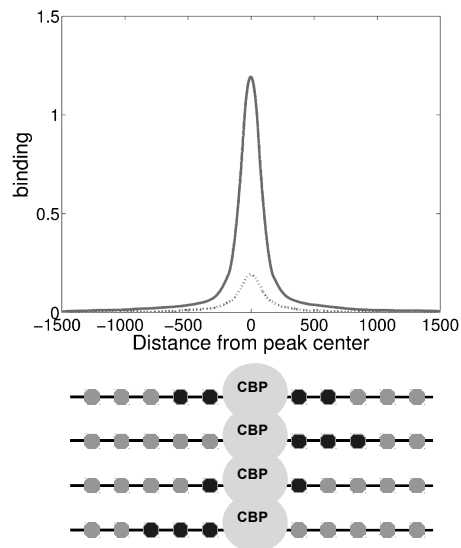
Aligning CBP peaks to calculate binding profiles



21

In reality, we have 28,000 loci and in this plot, each line corresponds to a CBP peak. They have been sorted by the level of CBP binding

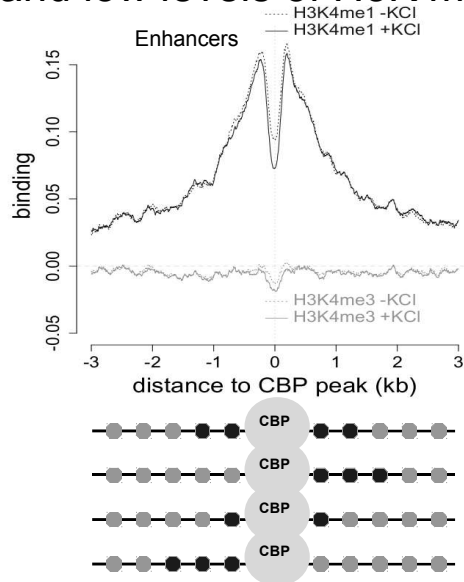
Average profile of CBP binding



22

Another way of plotting this data is instead to calculate the average as a function of the distance. As you can see here, the levels before stimulation as shown by the dashed line are very low and the binding shoots up as a response to the stimulation.

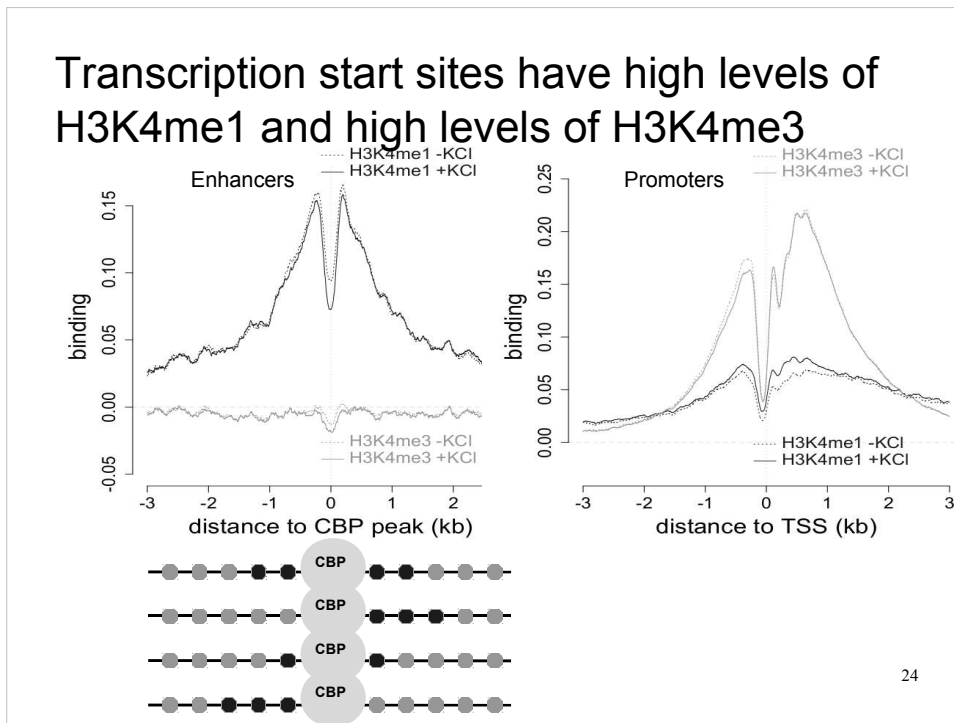
Enhancers have high levels of H3K4me1 and low levels of H3K4me3



23

We can do the same thing for the histone marks and as you see here, there is a characteristic bimodal pattern for K4me1 in blue whereas the levels of K4me3 in green are at background.

Transcription start sites have high levels of H3K4me1 and high levels of H3K4me3



This is in stark contrast to the promoters, where we have very high levels of K4me3. Hence, we can use this pattern to distinguish enhancers from promoters.

We identified 12k activity-dependent enhancers throughout the genome

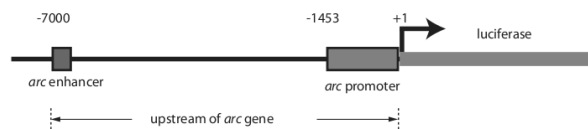
- **CBP** binding
- **H3K4me1** flanking
- **H3K4me3** absent
 - ~**5,000** extragenic enhancers
 - ~7000 intragenic enhancers

25

Using these criteria, we were left with a list of ~5k putative enhancers. We also found 7k enhancers overlapping genes.

8/8 tested activity-dependent enhancers were validated using a luciferase assay

- **CBP** peak
- **High** levels of flanking **H3K4me1**
- **Low** levels of **H3K4me3**
 - ~**5000** extragenic enhancers
 - ~**7000** intragenic enhancers




26

We tested 8 of these sequences in a luciferase assay, which is a low-throughput way of validating enhancer ability where the read-out is a fluorescent protein. We found that all 8 sequences were able to enhance gene expression in an activity dependent manner.

As I mentioned, it is very difficult to identify enhancers and before our study, there was only one example, the arc enhancer, of an activity dependent enhancer. Thus, finding 12k new ones is a significant achievement in itself.

What TFs bind to enhancers?

- CBP -
Creb Binding Protein
- 
- The diagram shows a horizontal line representing a DNA segment. Above the line, there are several small circles representing nucleosomes. A specific region of the DNA is highlighted with a thicker line and labeled 'Enhancer H3K4me1'. To the left of this region, the text 'CBP - Creb Binding Protein' is listed as a bullet point.

27

Now that we have identified thousands of enhancers throughout the mouse genome, in the next part of the talk I am going to tell you how we used theoretical models and genome-wide data to uncover new aspects of enhancer regulation.

Because of the paucity of identified enhancers, not very much is known about them. As I told you in the introduction, enhancers are generally viewed as loci far from the target gene where Tfs may bind and through some poorly understood mechanism help drive expression.

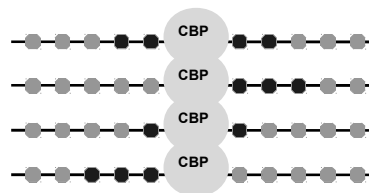
What I also need to tell you about CBP, which stands for Creb binding protein, is that it cannot bind DNA directly by itself, and as a co-activator it has multiple partners to which it may bind. Creb is another protein which does bind to DNA

The first question that we are going to ask is if it is possible to identify potential binding partners from our list of enhancers.

What motifs are enriched at enhancers?

- Calculate enrichment relative to flanks

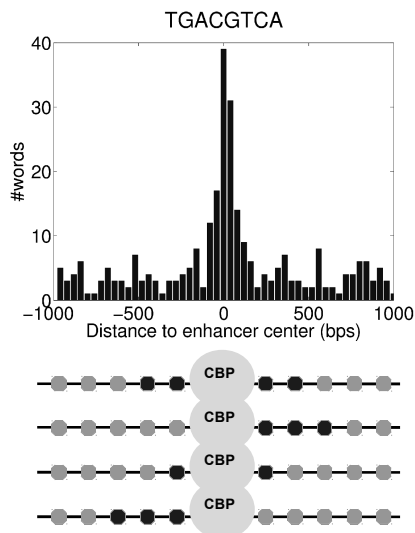
```
TCAGGCTGATGACGTCAAACCGTCGTTA
ACCTTTTGACGTCAAATTTACGCTAGTAT
TCGACGTAGCTAGCATGATCGATAGATC
CGTGACGTCAGTGCTCGTAAATCATAAG
```



28

Ultimately, what determines the function of a given part of the genome is the sequence. As you may recall, transcription factors bind to specific motifs. We searched the enhancer sequences for words that are enriched compared to the background. We did this by aligning the sequences as before and searching for the occurrences of some 2000 motifs that are known to be important for regulation.

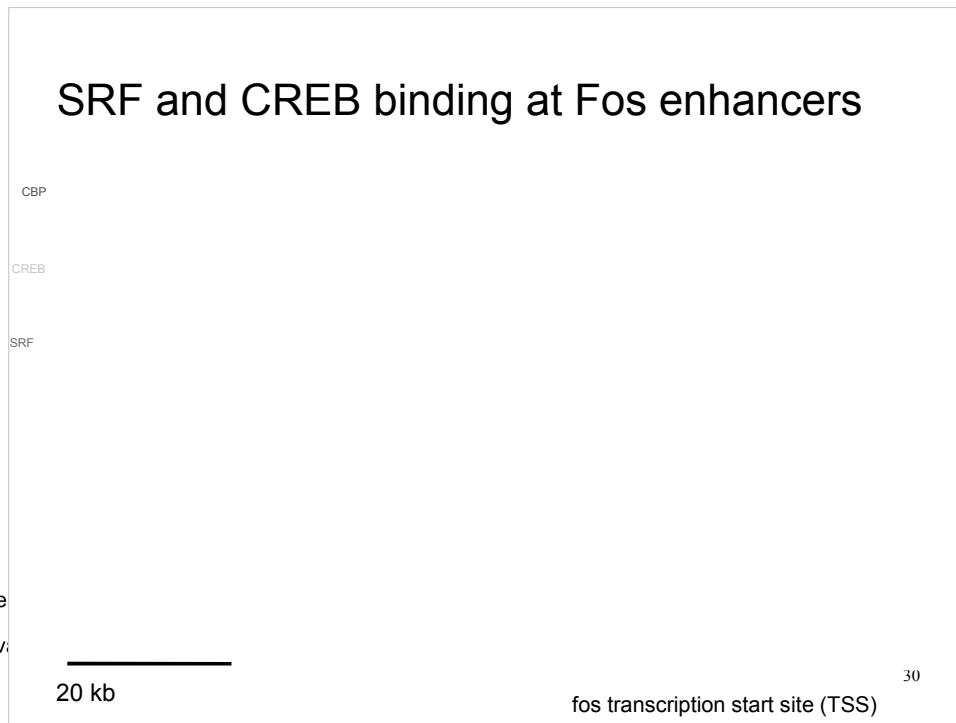
~100 enriched motifs found at enhancers



29

Here is a profile for the well-studied motif for the Creb protein. Creb has been shown to be an important component of the activity-dependent regulation and it is both reassuring and expected to see this kind of enrichment.

In total, we found around 100 words that were significantly enriched at the center of the enhancers compared to the background.



Now because of histone binding, the presence of a motif in the sequence does not necessarily mean that the factor will bind there. Similarly, factors may bind to other sites than their preferred motif. Complicating matters even further, there are typically many different Tfs that will bind to the same motif, so even if the motif is accessible from histones, there may be other factors that prevent binding.

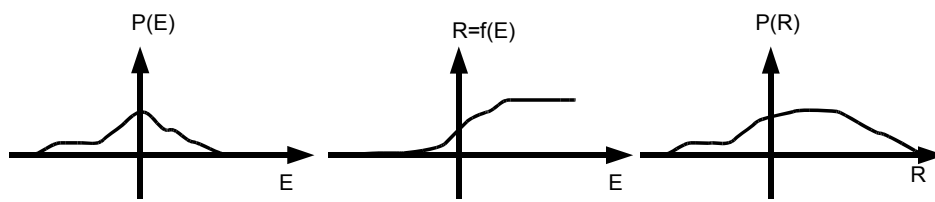
To study this further we carried out ChIP-Seq experiments for some of the most interesting factors. Here I am showing you another screenshot from the genome browser of the Fos locus. Fos is a transcription factor that is known to be strongly induced by Kcl and it is located on the positive strand with the read band marking the promoter. As you can see from the CBP track up here, there is inducible binding, both at the promoter and at these enhancers here marked in blue.

As you can see here for the two Tfs CREB and SRF, they have constitutive binding sites that overlap with some of the CBP sites, suggesting that they could be responsible for the recruitment of CBP.

Can the read count be predicted from sequence motifs?

$$R = f(E(\theta)) + \xi$$

↑ #reads
↑ Binding energy
↑ noise



31

Next, we asked if it is possible to predict the observed binding levels from the sequence features. We considered the following model of transcription factor binding.

We assume that the number of reads that we observe in our experiment is some unknown function f of the binding energy. The binding energy depends on the sequence under the peak and the energy function may also have several parameters, θ . In this study, we used position weight matrices with the experimentally defined parameters from the data base JASPAR.

We also assume that there is experimental and biological noise in our observations.

Since the experimental procedure by which the reads are obtained is very complex and cannot be easily modeled, we take a minimal approach and we try to find a transfer function that will map the energy distribution to the read distribution.

Assume that mutual information between energy and reads is maximized

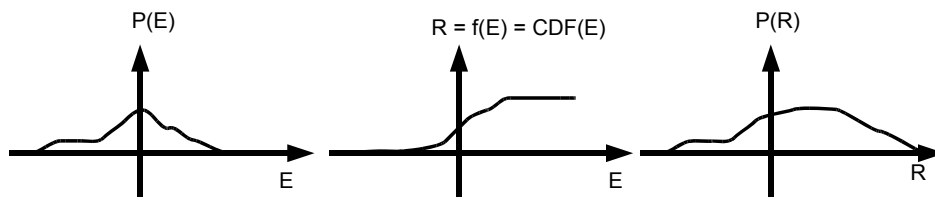
$$R = f(E(\theta)) + \xi$$

#reads

Binding energy (from JASPAR)

noise

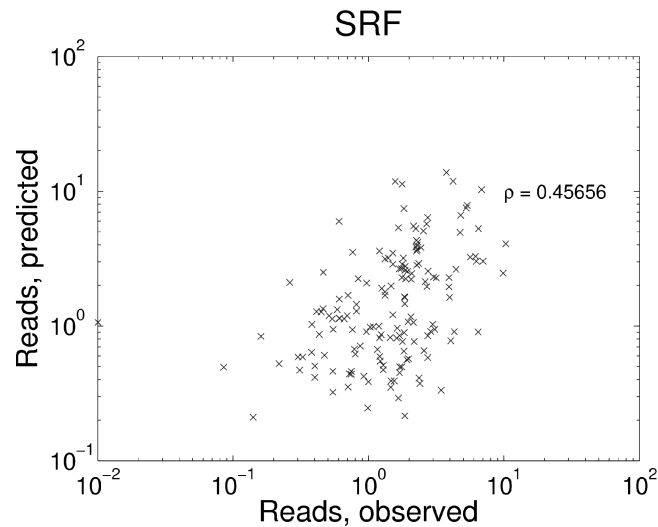
- f monotone
- $\max I(R; E)$
- Noise small and gaussian



32

If we make the rather mild assumptions that f should be monotone, that it maximizes the mutual information between the energy and the reads distribution and that the noise is small and gaussian, then it was shown by Nadal and Parga that the optimal choice for f is the cumulative distribution function of the energy distribution.

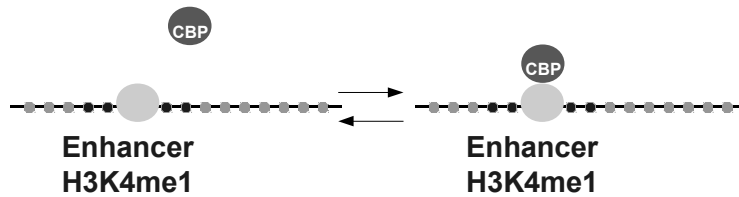
Number of reads can be predicted by binding energy



We may test this model using our data. We use the so called positional weight matrix energy model with the parameters taken from the database JASPAR. As is shown here for SRF peaks, the model does a reasonable job at predicting the observed peak sizes.

Since the model does not take the competition with other factors or histones into consideration, it is only applicable on the condition that binding was observed, that is a peak was found in the CHIP-Seq.

Is CBP binding determined by other TFs?



- TF binding sites compete for CBP

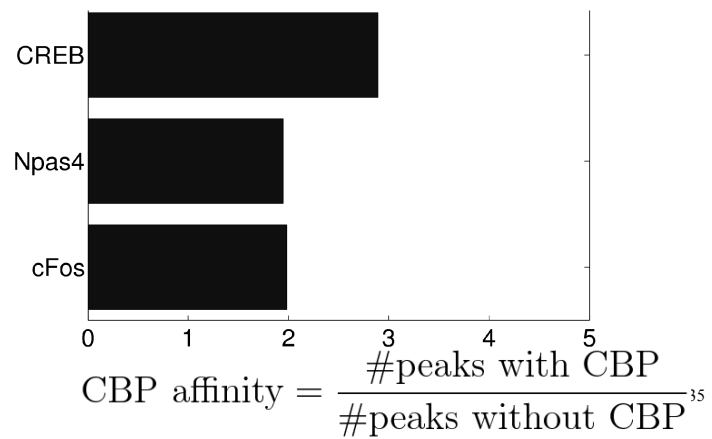
$$\text{CBP affinity} = \frac{\# \text{peaks with CBP}}{\# \text{peaks without CBP}}^{34}$$

Now that we have an idea of what Tfs are bound directly to the DNA, we asked what determined the CBP binding.

In the early 90s it was first suggested that CBP could serve as a bottleneck during gene regulation and that the choice of which enhancers will be bound by CBP is determined by the relative affinity which is in turn determined by the combination of Tfs present.

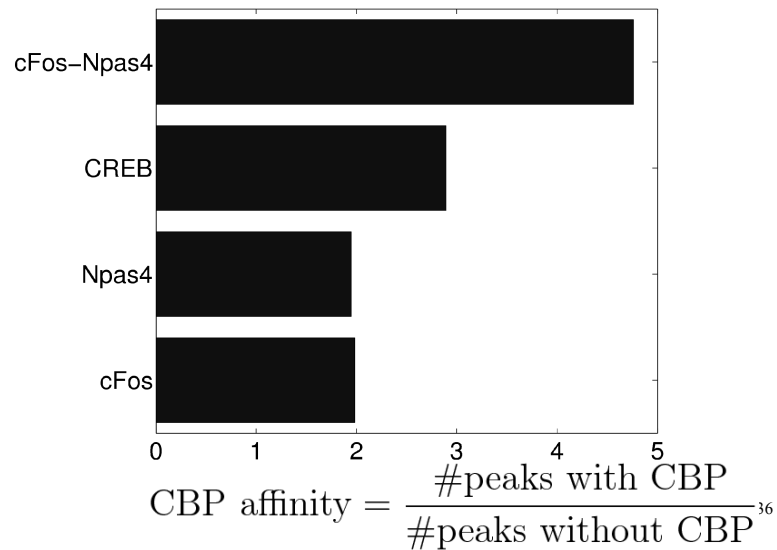
We decided to test this hypothesis using our genome-wide data. So we set up a very simple model where we assumed that for each TF there is a specific affinity for CBP and what we are observing is the equilibrium for that process. Given the relative number of peaks with and without CBP as illustrated in this cartoon, we can estimate the CBP affinity for each TF.

CBP levels determined by affinity of TF



We calculated the affinity for all of the Tfs that we studied, but here I am only showing you the data for three of them. What is re-assuring is that CREB comes out as the one that has the highest affinity for CBP, but otherwise they look quite similar.

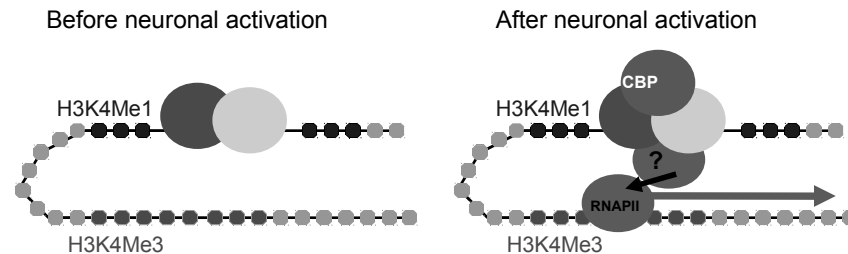
Synergistic effects for combinations of TFs



However, when we consider pairwise combinations as well, then we find that there are significant synergies. What we see here is that even though Npas4 and Fos have lower CBP affinity than CREB by themselves, when they both are present, the affinity is more than doubled and significantly larger than for CREB alone.

The idea that there are synergistic effects for combinations of Tfs at promoters where Pol2 is recruited is not new. However, for enhancers it has not been known what is the mechanism by which the synergy is manifested. Our model suggests that the mechanism is to modulate CBP affinity.

What is the role of CBP at enhancers?



- Is CBP determined by TF combinations? YES
- Does RNAPII bind at enhancers?

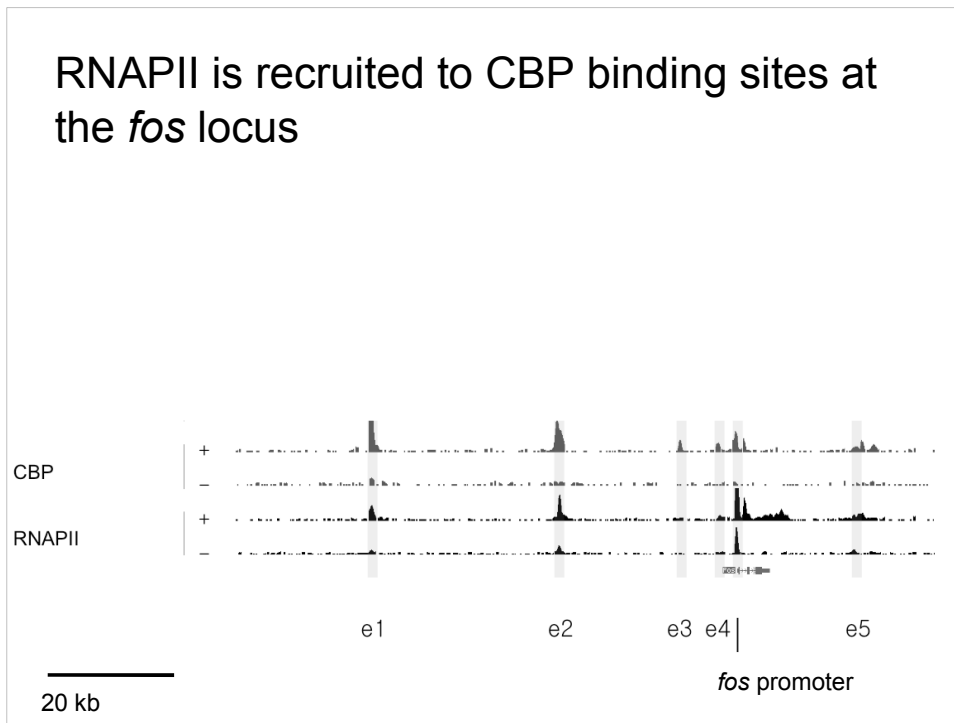
37

The next question is to ask about the functional role of CBP at enhancers. Studies at promoters have shown that CBP binding there may help to recruit Pol2.

Here I should remind you that RNAPII is the enzyme that is responsible for transcription. That is, it is the molecule that reads of the information in the DNA and creates a corresponding RNA molecule and hence it is one of the most important molecules in the cell.

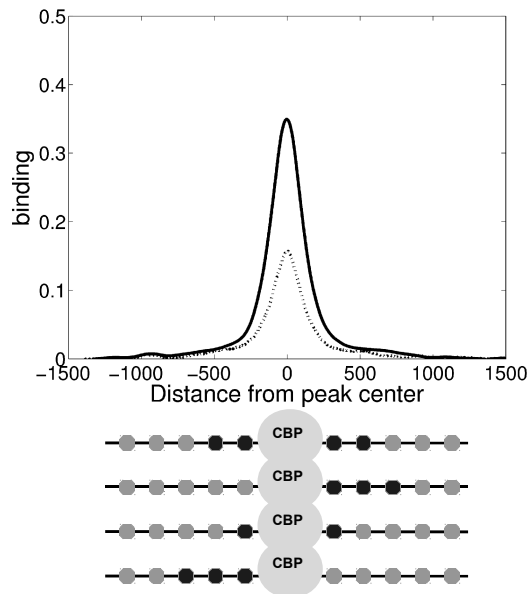
Moreover, based on studies of the beta-globin gene and a few other enhancers, it has been suggested that one of the roles of enhancers is to recruit RNAPII and transport it to the promoter. Equipped with our large list of enhancers, we wanted to find out if this is a more general phenomenon.

RNAPII is recruited to CBP binding sites at the *fos* locus



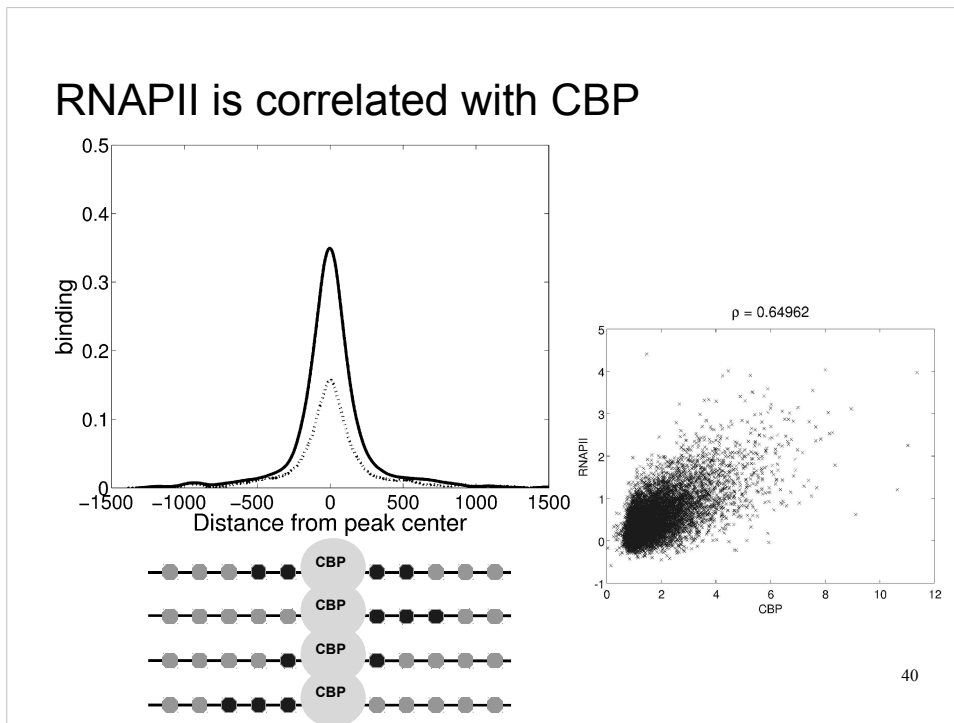
Going back to the *fos*-locus, we do indeed find that Pol2 binding shown in black overlaps with CBP and that it is also induced by the KCl-stimulation.

RNAPII is recruited at enhancers



39

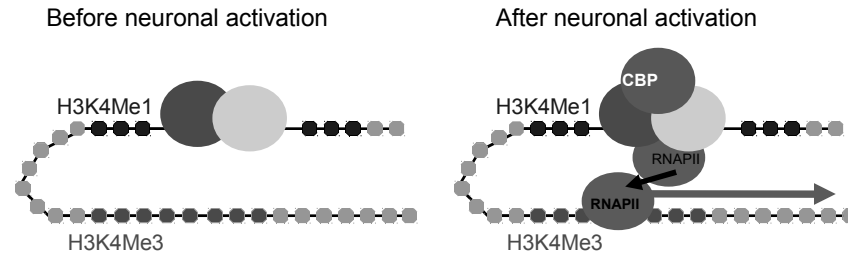
Here is an average plot of the RNAPII levels. The dashed line is before Kcl and we see that there is an average increase of Pol2 levels.



The scatter plot to the right shows that there is also a significant positive correlation between CBP and RNAPII, suggesting that CBP does indeed contribute to the recruitment of RNAPII.

I should stress here that biological data in general tends to be very noisy and observing a correlation of almost .65 is actually remarkably high.

What is the function of RNAPII at enhancers?



- Is CBP determined by TF combinations? YES
- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers?

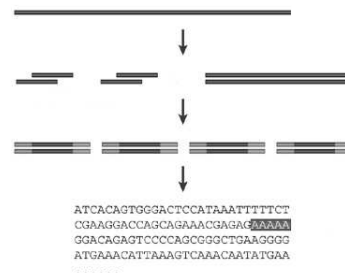
41

One possible explanation for the presence of Pol2 at enhancers is that it arrived there because of DNA looping. If enhancers are physically proximal to promoters because of a looping mechanism, then it is not unreasonable to believe that the polymerase can bind to the enhancer region which should be accessible.

Alternatively, pol2 could have an important role at enhancers.

To test for this possibility, we investigated transcription using RNA-Seq.

RNA-Seq finds transcribed parts of the genome



(Wang et al, 2009)

To study the transcriptome we use a method known as RNA-sequencing or RNA-seq.

I don't have time to describe the method in detail, but the type of data that we get from the experiment is conceptually similar to ChIP-Seq. But instead of telling us where proteins bind to the DNA, it tells us which parts of the DNA that have been transcribed. Again the data consists of short reads and the number of reads in a region is proportional to the level of transcription.

polyA tail is added to messenger RNAs (**mRNAs**)

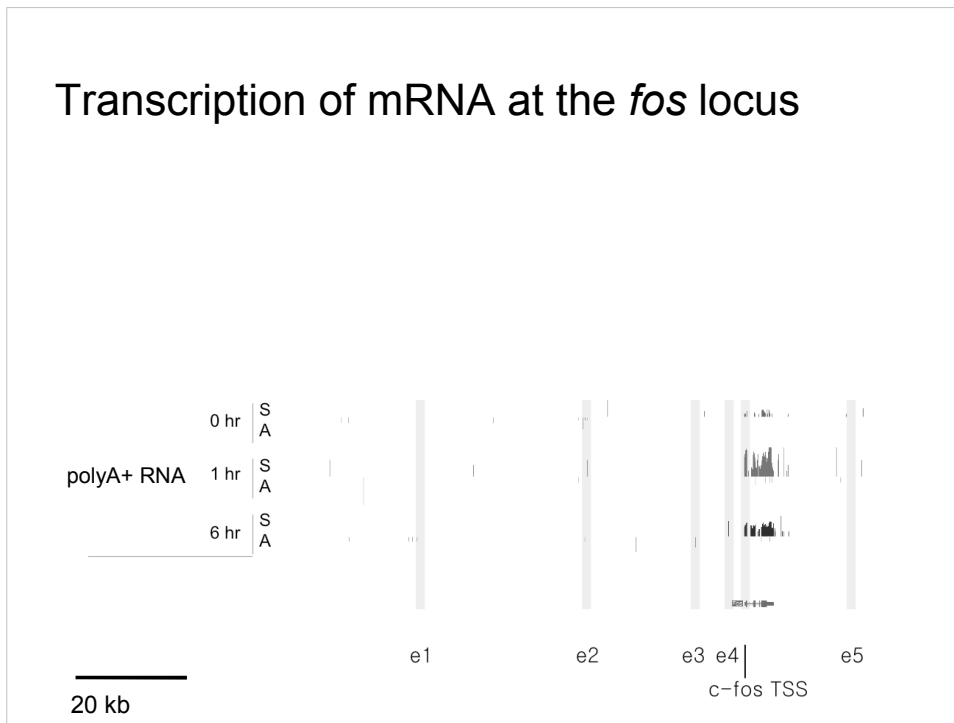
- Increases stability
- Allows transport out of nucleus

ACGUUUGUACCUAGCUAGCUUACGAGAAAAAAAAAAAAAAAAAAAAA

43

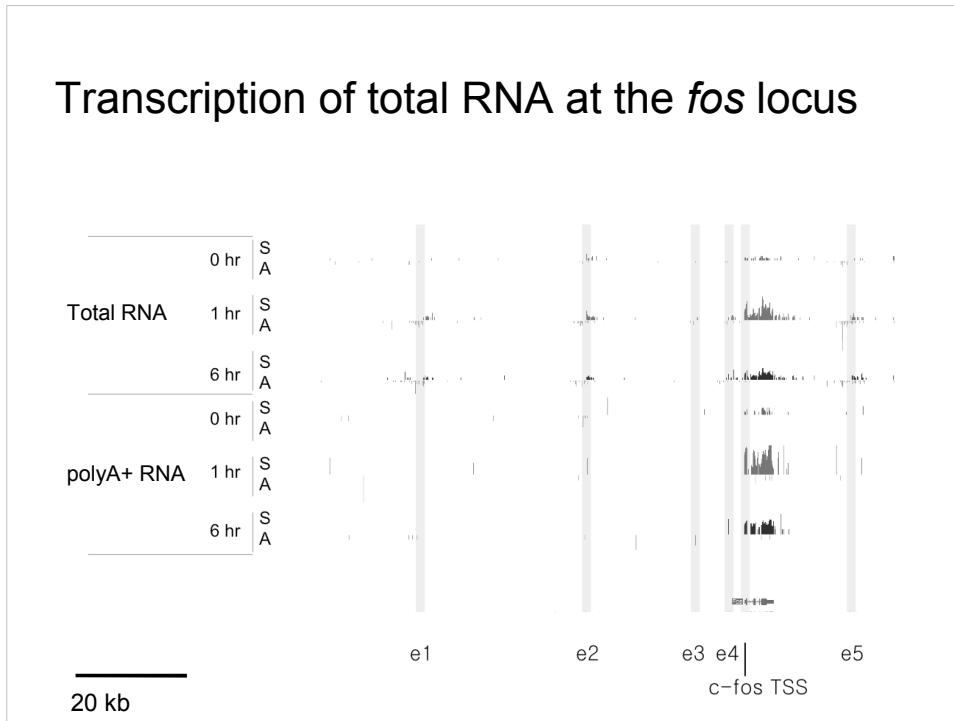
Transcription is a very complex process and there are many steps and modifications involved. For the purpose of this talk, I'd just like to highlight the fact that during transcription, the mRNAs[, the RNAs that get translated into proteins,] obtain a polyA tail which involves adding a row of adenosines at the end of the transcript. The polyA tail prevents the transcript from being degraded and it also allows it to be transported out of the nucleus.

Transcription of mRNA at the *fos* locus



Returning to the *fos*-gene, I will start by showing you the mRNA data from the polyA sample. DNA has two strands and each can be transcribed separately, but each gene is copied from only one strand. For visualization purposes, we use upward bars to indicate transcription on the forward strand and downward bars to indicate the reverse strand.

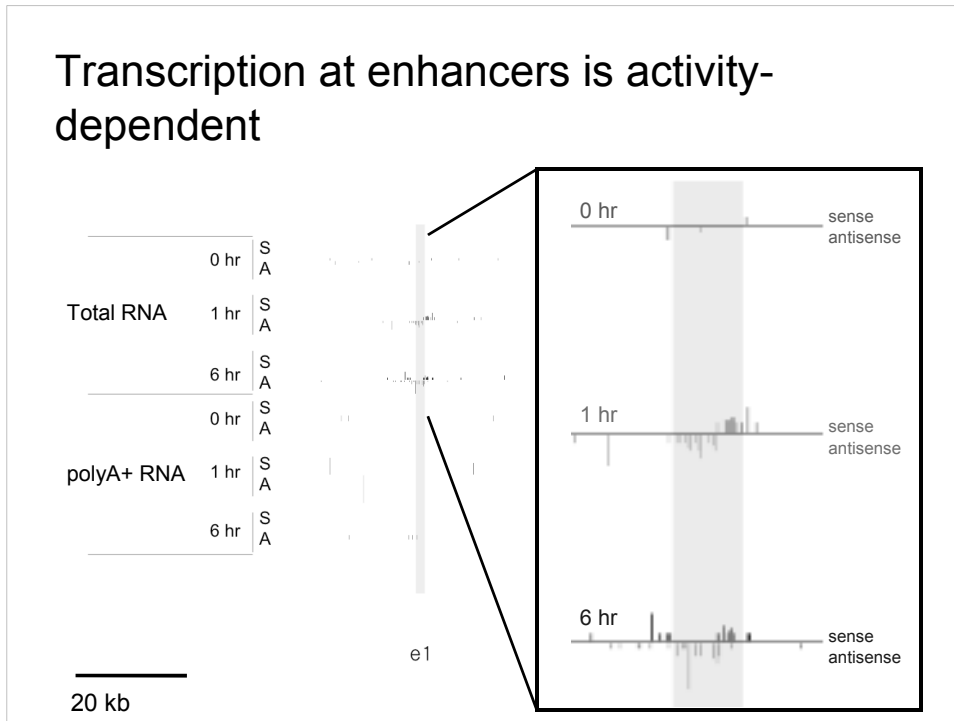
As you can see over here where the gene is located, there is little transcription before Kcl stimulation, but the activity of the gene increases significantly in response to the stimulus. Also, unlike the TF binding, there's not much going on in the extragenic regions.



However, this picture changes radically as we look at the total RNA which includes the RNA which has not been polyadenylated.

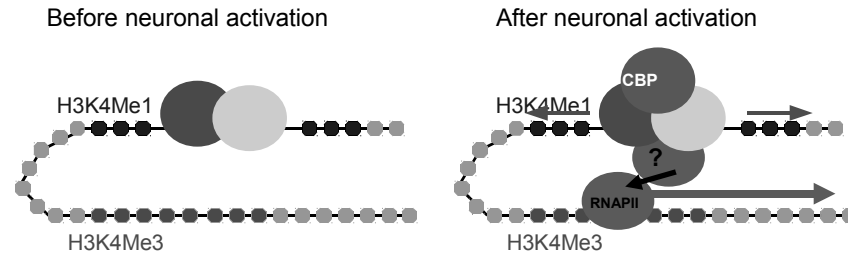
What we find is a rather surprising and striking pattern of transcription. At the enhancers we found short transcripts on both strands, originating from where CBP and pol2 is bound.

Transcription at enhancers is activity-dependent



If we zoom in on one of the enhancers, we clearly see that the transcription is activity-regulated, arguing against the possibility that it is a result of high background noise or incorrectly mapped reads.

What are the properties of enhancer RNAs (eRNAs)?

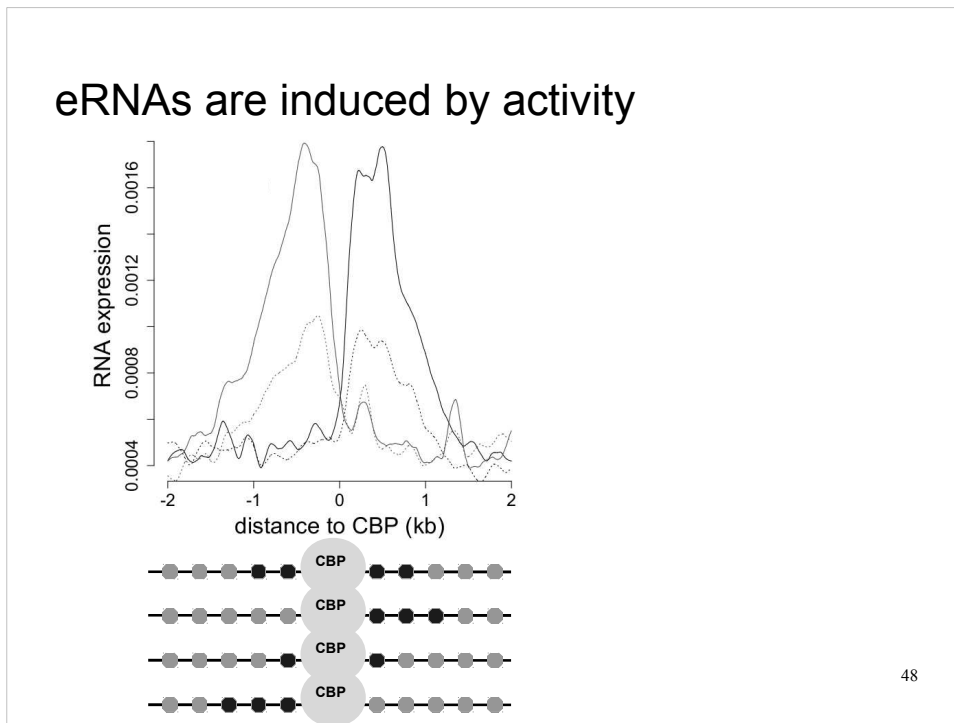


47

Transcription of genes into mRNA is the most common type of RNA, but there are also other types of RNA that are known to be important.

However, this type of pattern had not been reported previously in the literature, so we were excited to have discovered a novel type of RNA and we termed them enhancer RNAs or eRNAs for short.

Next, we set out to characterize the properties of eRNAs by using our genome-wide data.



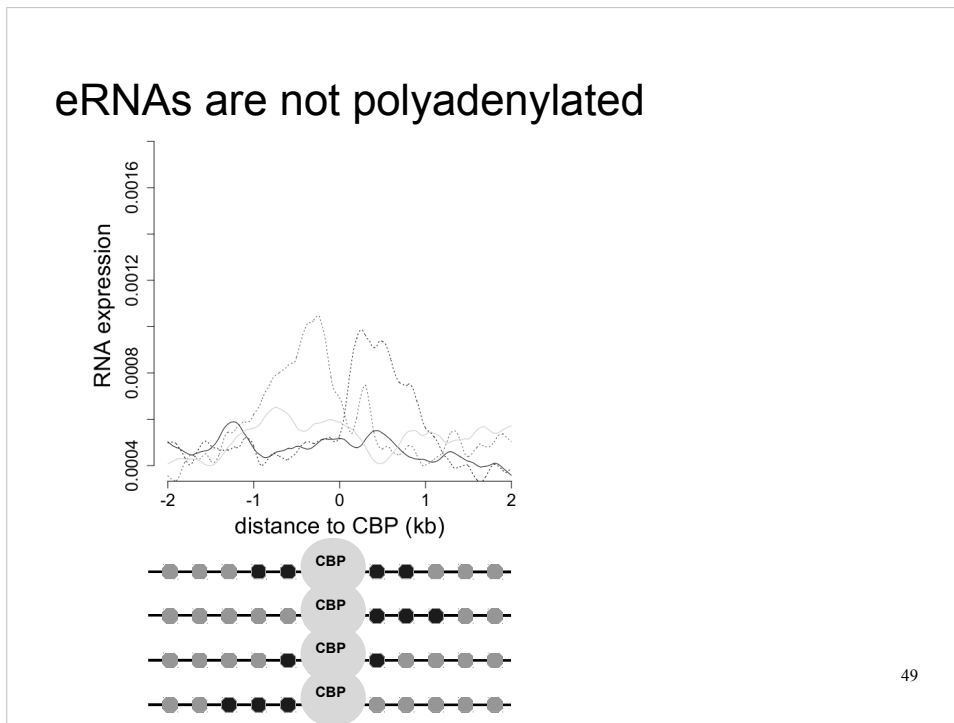
As you can see here, the pattern that we saw at the fos enhancers holds up across the genome.

The dotted lines represent the unstimulated condition and the filled lines after KCl stimulation. With red for the reverse strand and black for the forward strand.

The transcripts are bidirectional and they have a characteristic length of about 1.5 kb.

Furthermore, they are induced by activity with levels going up significantly.

eRNAs are not polyadenylated

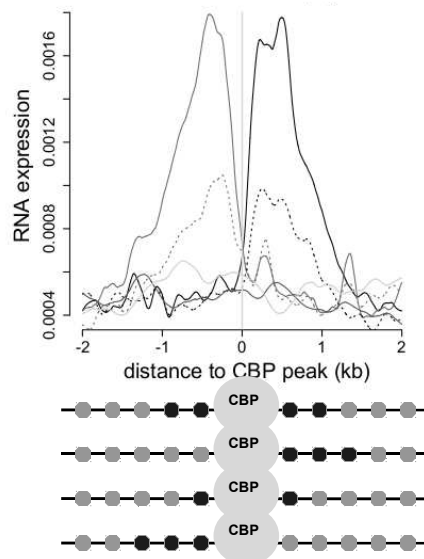


49

Looking at the polyA fraction of the RNA we see from the blue and yellow lines that there is no enrichment.

Furthermore, our computational analysis of the sequences suggests that the eRNAs have no protein coding potential.

eRNAs are 100-fold lower than mRNAs



- Inducible
- mRNA 2 orders of magnitude higher
 - 1 in 10k reads eRNA
- ~1.5 kb
- Bidirectional
- No polyA-tail
- Not protein-coding

50

From our study, we can draw the following conclusions about transcription at enhancers.

As I showed you before, the transcripts are induced by activity. Yet, eRNAs are very lowly expressed, about 100-fold lower than typical genes. And we find that only 1 out every 10k reads appears to be an eRNA read. The low expression is probably one of the reasons why they had eluded discovery up until now

Why do eRNAs have such low abundance?

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

51

A poignant question at this point is to ask why eRNAs have such low abundance. There are two possibilities: either they are produced at a very low rate or they are degraded much faster than mRNAs.

A simple model of transcription

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$

We can consider the level of RNA using the following simple ODE model. What the equation says is that the rate of change of mRNA is difference between a production and a decay term.

Parameters are straightforward to estimate

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} M - \frac{M}{\tau_M}$$

← mRNA

↑ RNAPII ↑ Length of transcript ↑ Elongation rate ↑ half-life

53

The production rate of mRNAs depends on the amount of polymerase, the rate at which the polymerase moves along the DNA and the length of the gene.

The degradation rate on the other hand is proportional to the level of mRNA and inversely proportional to the half-life.

Similar expression for eRNA levels

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} M - \frac{M}{\tau_M}$$
$$\frac{dE}{dt} = \frac{P_E k}{L_E} E - \frac{E}{\tau_E}$$

We use a similar equation for the eRNA levels, except that we allow for different parameter values.

Half life of eRNAs relative to mRNAs

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$

$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

$$\frac{\tau_E}{\tau_M} = \frac{E^*}{M^*} \frac{L_E}{L_M} \frac{P_M}{P_E}$$

At steady state, we may solve the system of equations and express the ratio of the half lives like this.

eRNAs half life is less than half an hour

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$

$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

$$\frac{\tau_E}{\tau_M} = \frac{E^* L_E P_M}{M^* L_M P_E}$$

Single ~

$$\tau_E \approx 10^{-2} \times \frac{1.5}{30} \times 5 \times \tau_M \approx 4 \times 10^{-2} \times 600\text{min} = 24\text{min}$$

Now we can use our data to estimate these quantities.

As I said before, the expression level differs by about 2 orders of magnitude.

The length of a typical mRNA is about 30 kb, so this ratio is easy to estimate.

Finally, the amount of polymerase typically differs by less than one order of magnitude.

Putting this together and using an estimate of 10 h for mRNA half-life, we find that the eRNA half-life is approximately half an hour, suggesting that they decay very rapidly.

eRNAs half life is less than half an hour

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$

$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

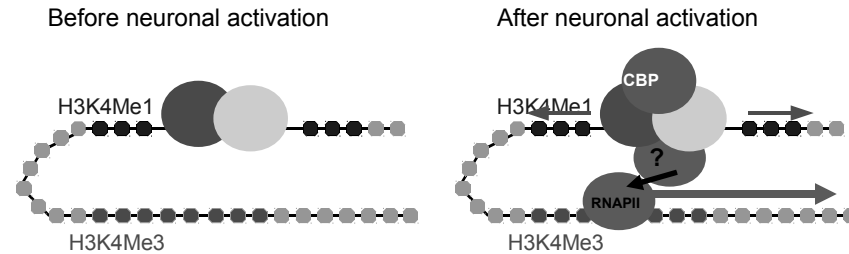
$$\frac{\tau_E}{\tau_M} = \frac{E^* L_E P_M}{M^* L_M P_E}$$

$$\tau_E \approx 10^{-2} \times \frac{1.5}{30} \times 5 \times \tau_M \approx 4 \times 10^{-2} \times 600 \text{min} = 24 \text{min}$$

Finally we measured the stability of these transcripts using an actinomycinD chase. In comparison to both the mRNAs generated by the associated protein-coding genes and some known lncRNAs (like Xist and Neat), the upstream non-coding transcripts were very unstable, being reduced by 80% to 90% after a 30 min actinomycinD treatment (indicating a half-life lower than 7.5 min) (Figure 3D and Figure S3). High instability of a subset of lncRNAs both in yeast and mammals mainly depends on degradation by the nuclear exosome [39,40] and often results in the generation of more stable short RNA products [41], which in principle might be responsible for downstream functional effects.

This estimate is very much in agreement with measurements of 7 minutes that were made by the Natoli lab after hour paper had been published.

Enhancers recruit RNAPII and produce transcripts, but does it depend on promoter?

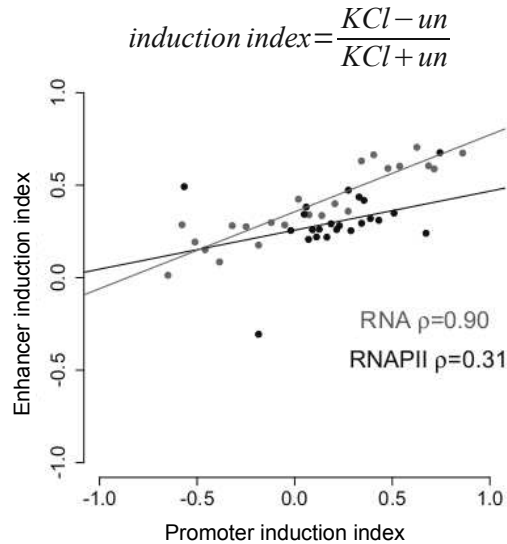


- Is CBP determined by TF combinations? YES
- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers? YES
- Is RNAPII recruitment independent?

58

Next, we investigated the relationship between eRNAs and the more well understood messenger RNAs.

eRNA induction is correlated with induction of nearby mRNAs



59

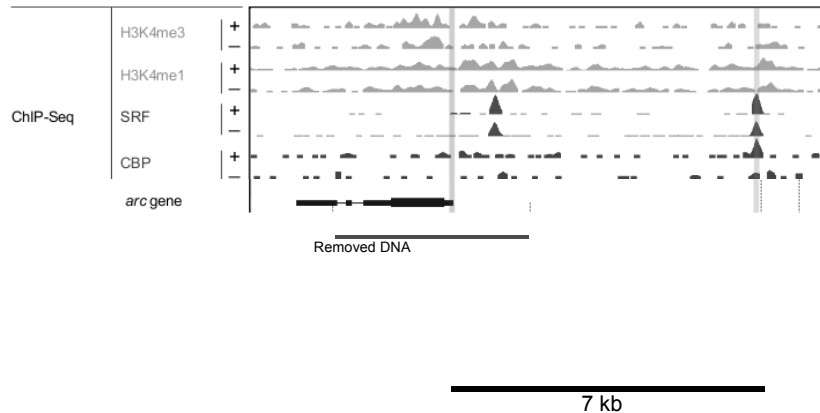
We did this by pairing each enhancer with its nearest promoter.

We used a normalized index describing the level of induction of RNAPII and gene expression at each gene and enhancer pair. The induction index captures the relative change of RNAPII or RNA at promoters and enhancers.

As you can see there is only a modest correlation for the polymerase levels. Whereas the transcript levels appear to be strongly correlated.

Thus, the prediction here is that the Pol2 recruitment at the enhancer is independent of the promoter.

Deletion of the Arc-promoter confirms that RNAPII recruitment is independent but eRNA transcription is not.



60

We investigated this correlation further using a knock-out experiment. A knock-out experiment is an experiment where you create a mouse where a part of the genome has been deleted and it is a very powerful way of investigating the function of the genome.

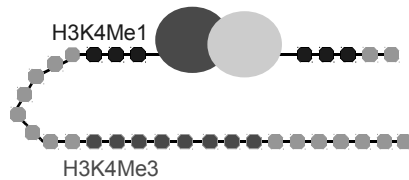
Going back to the arc-gene, we used a mouse where the promoter of the gene has been removed, but the enhancer remains intact.

The experiment showed that when the promoter was gone, RNAPII levels at the enhancer were unchanged.

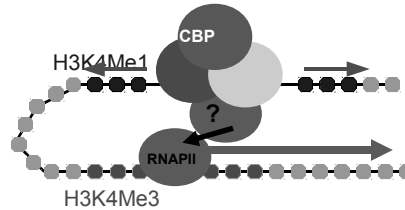
At the same time, we found that the eRNAs were not present in the mutant where the arc promoter had been removed, confirming that the interaction with the promoter is indeed required for the production of eRNAs.

Enhancers bind RNAPII independently, but the transcription is promoter-dependent

Before neuronal activation



After neuronal activation



- Is CBP determined by TF combinations? YES
- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers? YES
- Is RNAPII recruitment independent? YES
- Is eRNA production independent? NO

61

Taken together, these results suggest that the binding of RNAPII at enhancers is independent of the promoter. But that there is a mechanistic connection that governs the level of eRNAs.

Before I move on from the topic of enhancers, I would like to discuss some of my thoughts on the function of the new mechanisms that we have discovered.

What is the function of RNAPII at enhancers?

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

62

The first question is: why have Pol2 at enhancers? The most obvious possibility is of course that it has to go there in order to synthesize the eRNAs.

A second possibility, as has been suggested by Marc Groudine and others based on studies of the beta-globin locus control region, before eRNAs were discovered, is that the enhancer helps to recruit Pol2 to the promoter.

A simple model of RNAPII recruitment

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = \underbrace{k_p + Nk_e c}_{\text{Binding rate}} - \underbrace{\frac{P_M}{\tau}}_{\text{decay}}$$

P – polymerase levels
 k_p – binding rate at promoter
 k_e – binding rate at enhancer
 N – number of enhancers
 c – contact probability
 τ – RNAPII half life

63

To understand how this could work, we may write down the following equation for the amount of Pol2 at the promoter.

Here, k_p is the rate at which pol2 is bound at the promoter N_e , the number of enhancers, k_e is the binding rate at enhancers and c is the probability that the enhancer will be in contact with the promoter. P is the level of polymerase and τ is the half-life for the dwelling time of the polymerase leaves the promoter.

Steady state level of RNAPII is increased

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$
$$P_M(t) = \underbrace{(k_p + Nk_e c)}_{\text{Steady state level}} (1 - e^{-t/\tau})$$

64

This a 1st order ODE and the solution can be found as this.

Unfortunately, it is quite difficult to estimate the parameters here.

From our data, it seems likely that k_p and k_e are roughly of the same order. N is probably around 10, so these two terms probably cancel out. As for the contact probability I really do not know. Even though there are data from looping experiments, I doubt that they are that relevant since there are likely to be all kinds of molecules in vivo that help to stabilize the loops compared to the situation of naked DNA in a test tube.

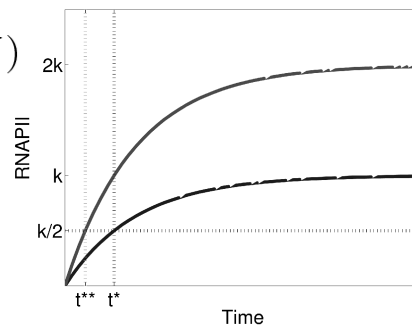
However, we may still draw some interesting qualitative conclusions using this model.

Steady state level of RNAPII is increased

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$



The most obvious benefit of having multiple recruitment points for Pol2 is that the maximum level at the promoter is increased by this number which is proportional to the number of enhancers.

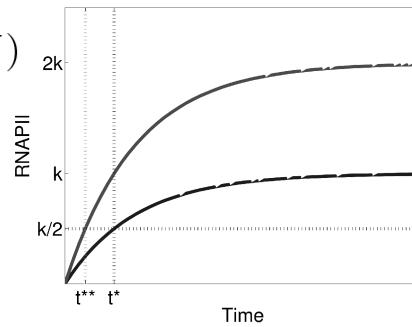
[In this plot, this factor has been set to k_p and as you can see the steady state level for the green curve is twice as high as for the blue.]

Recruitment of RNAPII is diffusion limited

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$



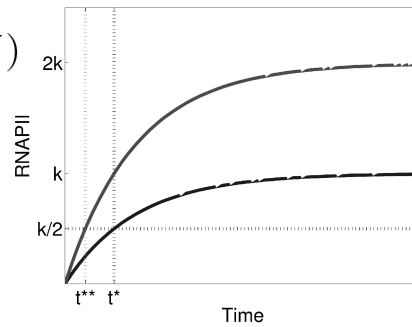
Inside the cell, the binding rate of Pol2 is diffusion-limited, meaning that there is an upper bound to the parameter k_p here. It has been shown by Vilar and Leibler that this limit can be important for gene regulation. Thus, having distributed recruitment of pol2 may allow a larger effective rate of binding than would be possible to achieve with only a single strong promoter.

Rise time is reduced

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$



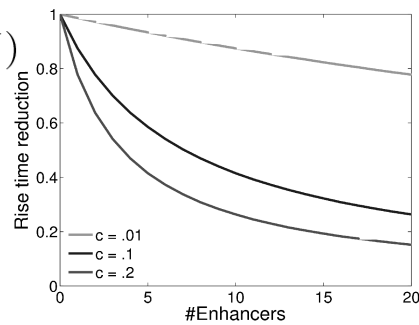
Another advantage that can be observed from this plot is the reduced rise times. If we assume that a certain critical level of pol2 level needs to be reached, then it is clear that the time to reach that threshold is reduced from t star to t double star.

Significant speed-up with ~5 enhancers

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$



We can plot the reduction in time to reach the threshold as a function of the number of enhancers for different contact probabilities.

We see that if the contact probability is 10% per enhancer, it is sufficient with only 5 enhancers to obtain an almost 50% reduction of the rise time.

Enhancers may reduce the noise in gene expression

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau} + \sigma \sqrt{P_M(t)} \xi(t)$$

69

Finally, it can also be shown that by distributing the RNAPII recruitment, then we may reduce the noise in the system. If we add a noise term to the equation here, then it can be shown that the variance of the Pol2 concentration will scale linearly with the mean.

Variance reduction proportional to number of enhancers

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau} + \sigma \sqrt{P_M(t)} \xi(t)$$

$$\frac{\text{Variance strong promoter}}{\text{Variance weak promoter with enhancers}} = \frac{\text{Var}[(1 + Nc)k]}{\text{Var}[k] + N\text{Var}[ck]} = \frac{(1 + Nc)^2 \text{Var}[k]}{(1 + Nc^2) \text{Var}[k]} \sim N$$

70

The intuition here is now that by having multiple independent recruitment points, each with a lower k_p than what would be required, then the noise at each of the enhancers will be lower. We may obtain the same Pol2 recruitment rate, but with a lower noise level. Thus, this calculation shows that we may reduce the noise level by a factor of N , where N is the number of enhancers

So to summarize, there are three immediate advantages of distributing the RNAPII recruitment: higher expression, possibly higher than what would be possible with a single promoter, faster induction and reduced noise.

What is the function of eRNAs?

Science is always wrong. It never solves a problem without creating ten more.

-George Bernard Shaw

- Noise
- Establish histone marks
- Transcript has function

71

At this point it is worth contemplating the words of George Bernard Shaw, who had some interesting opinions not just about spelling, but also about science. Perhaps the most central question that comes after the discovery of a new type of RNA is “What is the function of eRNAs?” and so far we have been unable to answer it conclusively. Nevertheless, I would like to discuss the three hypotheses that we are currently considering.

One possibility is that they are simply transcriptional noise and that they serve no biological function.

A second possibility is that the transcripts themselves do not have a function, but instead that the process of transcription is important. Experiments in yeast have shown that the methylation of histones can take place as part of transcription.

The final possibility is of course that the transcript itself actually does something useful in the cell. If this is the case, it seems likely that the transcript is used nearby where it was transcribed since the rapid degradation rate suggests that it will be difficult to transport them reliably

From an experimental point of view, the fact that the eRNA levels are induced suggests that they can be used as a read-out of the enhancer activity and distinguish active from inactive sites.

eRNAs have been found in other cell types

doi:10.1038/nature09033

nature

ARTICLES

Widespread transcription at neuronal activity-regulated enhancers

Tae-Kyung Kim^{1,†}, Martin Hemberg^{2,†}, Jesse M. Gray^{3,†}, Allen M. Costa¹, Daniel M. Bear¹, Jing Wu¹, David A. Harmin^{1,4}, Mike Laptewicz¹, Kellie Barbara-Haley¹, Scott Kuersten⁵, Eirene Markenscoff-Papadimitriou^{1,†}, Dietmar Kuhl¹, Haruhiko Bito⁶, Paul F. Worley⁷, Gabriel Kreiman² & Michael E. Greenberg¹

Histone H3K27ac separates active from poised enhancers and predicts developmental state

Menno P. Creyghton^{1,†}, Albert W. Cheng^{1,†}, G. Grant Welstead¹, Tristan Kooistra^{1,†}, Bryce W. Carey^{1,†}, Eveline J. Steine^{1,†}, Jacob Hanna^{1,†}, Michael A. Lodato^{1,†}, Garrett M. Frampton^{1,†}, Phillip A. Sharp^{1,†}, Laurie A. Boyer^{1,†}, Richard A. Young^{1,†}, and Rudolf Jaenisch^{1,2}

OPEN ACCESS Freely available online

PLoS BIOLOGY

A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers

Francesca De Santa^{1,†}, Iros Barozzi^{1,†}, Flore Mietton^{1,†}, Serena Ghisletti¹, Sara Polletti¹, Betsabeh Khoramian Tusi¹, Heiko Muller¹, Jiannis Ragoussis², Chia-Lin Wei³, Giocchino Natoli¹

LETTER

doi:10.1038/nature09692

A unique chromatin signature uncovers early developmental enhancers in humans

Alvaro Rada-Iglesias¹, Ruchi Bajpai¹, Tomek Swigut¹, Samantha A. Brugmann¹, Ryan A. Flynn¹ & Joanna Wysocka^{1,2}

72

The results presented here were published 2 years ago. Interestingly enough, several papers have appeared already where the authors found eRNAs. These studies were made in different cell types and organisms which suggests that what we observed in mouse neurons is not specific to the nervous system and I believe that eRNAs are a generic feature of enhancers.

Summary

- Identified ~12k activity-dependent enhancers
- Discovered and quantified novel mechanisms
 - Identified enriched motifs
 - Read levels explained by binding energy
 - Combinatorial affinity for CBP
 - Recruitment of RNAPII at enhancers
 - Transcription at enhancers
 - Properties of eRNA
 - Interaction with promoter necessary

73

To sum up: I have told you about gene regulation and in particular about distal enhancers. We used high-throughput sequencing data to identify 12k activity dependent enhancers.

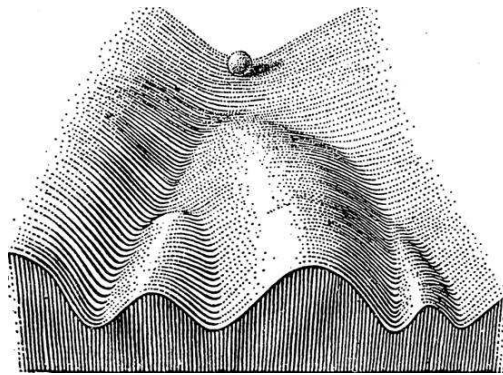
Prior to our work, the view in the field was that enhancers were simply collections of TF binding sites that somehow interacted with promoters to help drive gene expression. However, our results suggests that enhancers are much more complex than was previously thought.

So the contributions that we have made was to show that the Tfs present at an enhancer have a combinatorial affinity for CBP. The CBP is responsible for the recruitment of Pol2 at enhancers and most surprisingly of all, the polymerase produces transcripts at the enhancers.

We characterized this novel type of RNA and we were able to demonstrate that the synthesis requires the interaction with the promoter.

Stochastic models of gene expression

- Transitions between stable states
- Noise



Waddington, 1953 ⁷⁴

Now I'd like to switch topics and tell you briefly about some of the work that I did as a graduate student at imperial college london. I won't have time to go into much detail here, but I hope to be able to give you a flavor of what I worked on there.

Going back to the picture that we started with, one of the problems that I am most interested in is how cells can switch from one state to another. To understand the fluctuations and the dynamics of this process, a stochastic model of gene expression is required.

Master Equation (**ME**) description

$$\frac{dP_j}{dt} = \sum_i W_{ij} P_i(t) - W_{ji} P_j(t)$$

P_j - **Probability** of having j molecules

W_{ij} - **Transition rate** from i to j

75

There are several different methods for representing the stochastic model. The one that is often used is the Master Equation and the reason is that unlike the Langevin equation it is discrete. Since we are typically dealing with fewer than 10 mRNA molecules, using a discrete model is important.

The ME is a balance equation for the probability of finding the system in a state j. Here a state corresponds to a certain number of molecules. So on the left we have the time derivative of the probability. This is equal to the flow of probability from all of the other states i, minus the probability flow out of state j into all other states i.

MCMC required for solving ME

$$\frac{dP_j}{dt} = \sum_i W_{ij}P_i(t) - W_{ji}P_j(t)$$

P_j - **Probability** of having j molecules

W_{ij} - **Transition rate** from i to j

- Use Markov Chain Monte Carlo (MCMC)
 - Gillespie's Stochastic Simulation Algorithm (**SSA**)

76

Even though it is relatively straightforward to write down the ME for any given system, solving it, both analytically and numerically, is typically very challenging and one must often resort to MCMC methods.

The algorithm of choice is known as Gillespie's stochastic simulation algorithm and it was published in 1976.

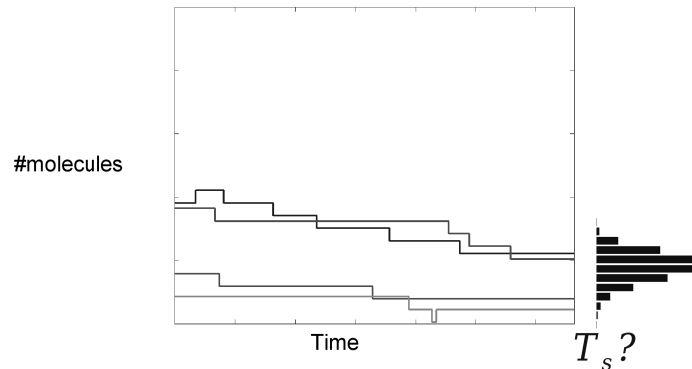
However, interest was rekindled in the late 90s following a paper by Adam Arkin, who is now here at Berkeley, that showed how the SSA could be used to model gene expression.

How long do we need to run MCMC?

$$\frac{dP_j}{dt} = \sum_i W_{ij}P_i(t) - W_{ji}P_j(t)$$

P_j - **Probability** of having j molecules

W_{ij} - **Transition rate** from i to j



When we use the SSA, what we do is to start from some initial conditions and then simulate the reaction events forward in time as illustrated here.

One of the problems which is common to all MCMC algorithms is the question about how long one needs to run the simulation.

From Markov theory we know that the stationary distribution is reached as time goes towards infinity. However, if we want to be able to publish our papers, we cannot wait that long.

Perfect sampling guarantees stationarity

$$\frac{dP_j}{dt} = \sum_i W_{ij}P_i(t) - W_{ji}P_j(t)$$

P_j - **Probability** of having j molecules

W_{ij} - **Transition rate** from i to j

- **Dominated Coupling From The Past SSA**
proven to reach stationary distribution

BMC Systems Biology



Methodology article

Open Access

A Dominated Coupling From The Past algorithm for the stochastic simulation of networks of biochemical reactions

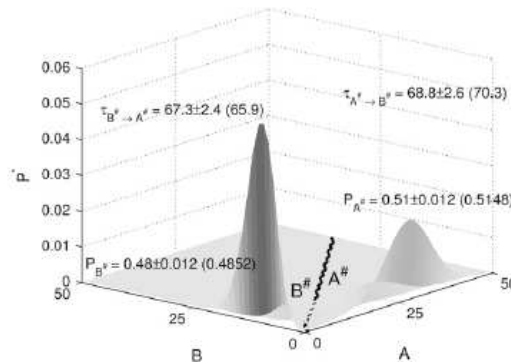
Martin Hemberg¹ and Mauricio Barahona^{*1,2}

Address: ¹Department of Bioengineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK and ²Institute for Mathematical Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

78

I combined the SSA with an algorithm known as dominated coupling from the past. The result was an extended algorithm with a very clunky acronym: DCFTP-SSA. The algorithm comes with a mathematical proof, guaranteeing that the simulation will not terminate until the stationary distribution has been reached.

Transitions between stable states



Biophysical Journal Volume 93 July 2007 401–410

401

Perfect Sampling of the Master Equation for Gene Regulatory Networks

Martin Hemberg and Mauricio Barahona
Department of Bioengineering and Institute for Mathematical Sciences, Imperial College London, London, United Kingdom

79

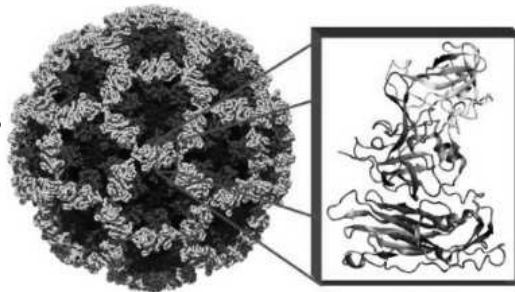
An example of a system where it is important to be able to sample from the stationary distribution is this genetic switch which is a simple example of a bistable system. Here A and B are two mutually repressing genes and the system has two stable states, either A is high and B is low, or vice versa.

The SSA allows us to simulate trajectories and thereby sample transitions from one stable state to another.

However, to get accurate estimates of the escape times, we must make sure that each simulation is started from the stationary distribution. The DCFTP-SSA makes it possible to start the simulation from the stationary distribution and hence we can be sure that the results are not tainted by transients. This allows us to accurately estimate steady state probabilities, transition times and a separatrix.

Assembly of viral capsids

- Protect viral genome
 - Self-assembly
 - Identical subunits
 - Icosahedral symmetry



Biophysical Journal Volume 90 May 2006 3029–3042

Stochastic Kinetics of Viral Capsid Assembly Based on Detailed Protein Structures

Martin Hemberg,* Sophia N. Yaliraki,[†] and Mauricio Barahona*

*Department of Bioengineering and [†]Department of Chemistry, Imperial College London, London, United Kingdom

80

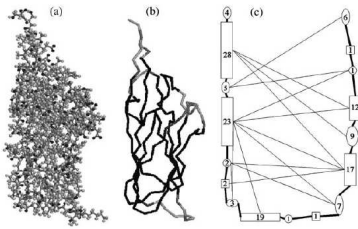
I would also like to tell you briefly about another project where I combined the analysis of a high-dimensional state space with structural analysis of proteins: the assembly of viral capsids

Viruses have a protein coat, called a capsid, which serves to protect their genomic material. The capsid has many fascinating properties, but perhaps the most intriguing one is that for a certain class of viruses they can self-assemble. That is, if you put protein monomers in a test-tube, they will automatically form perfect capsids that are indistinguishable from the ones found in the wild, without any assistance from enzymes or other molecules. Moreover, the capsids have icosahedral symmetry and they consist of only one or a handful of identical subunits.

[The assembly is very rapid, which means that it is difficult to study experimentally. However, it is a good problem to study computationally.]

Coarse-grained protein model

- Atomic-structure
- FIRST calculates rigidity of amino acids
- Identify ~20 rigid blocks



81

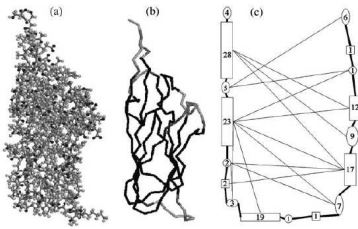
In our model, we took a bottom up approach, starting from the full atomic description as provided by the crystal structure which is shown to the left here.

We use a software called FIRST, developed by Jacobs and Thorpe at Arizona State University, which treats the protein as a structural graph and identifies rigid substructures.

The output of FIRST is further coarse-grained to produce a representation consisting of ~20 rigid blocks and their connections, represented as a graph.

Use reduced representation for aggregates

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion

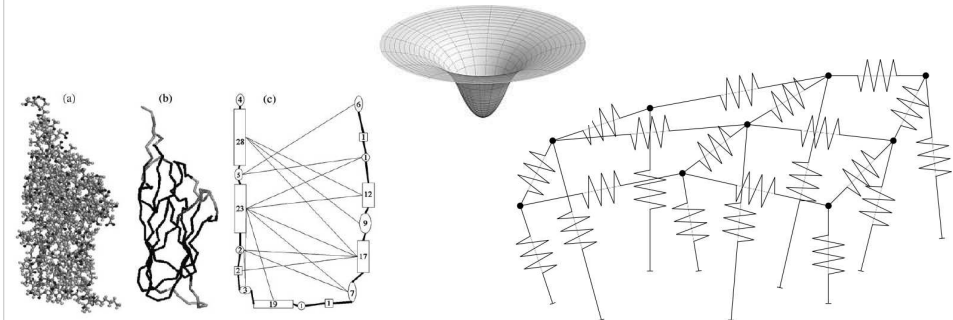


82

Having obtained a more tractable representation of individual proteins, we proceeded to consider aggregates of two or more proteins. From the crystal structure we can identify the bonds between pairs of proteins, so calculating the association rate is fairly straightforward, even though we need to take the diffusion in the surrounding medium into consideration.

Aggregates modeled as mass-spring graph

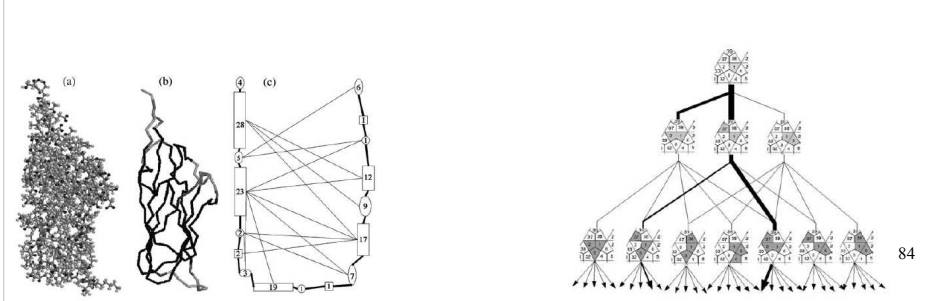
- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion
 - Dissociation escape from multi-dimensional well



However, complexes may have different stabilities and we also wanted to consider the break up or dissociation of the aggregates. We represented each complex as a mass-spring network and we calculated the most likely partition by considering the first eigenvector of the graph Laplacian. We then used Kramers' theory of escape from a multi-dimensional potential well to model the rate of dissociation.

All reactions cannot be enumerated

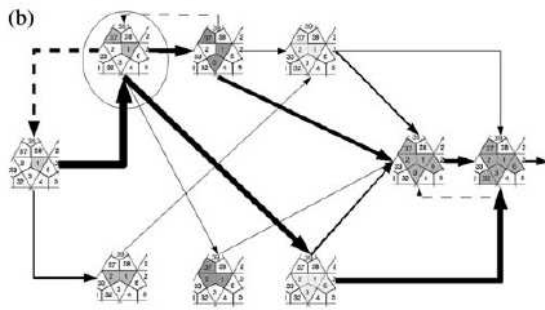
- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion
 - Dissociation escape from multi-dimensional well



Now we have a method for calculating the rate of every event in the system. However, even if we use a simplified view of the system where we only allow for the addition of one monomer at a time as represented by this tree here, it is easy to imagine that the number of possible reactions grow combinatorically and they cannot all be enumerated

Probabilistic sampling of assembly paths

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion
 - Dissociation escape from multi-dimensional well

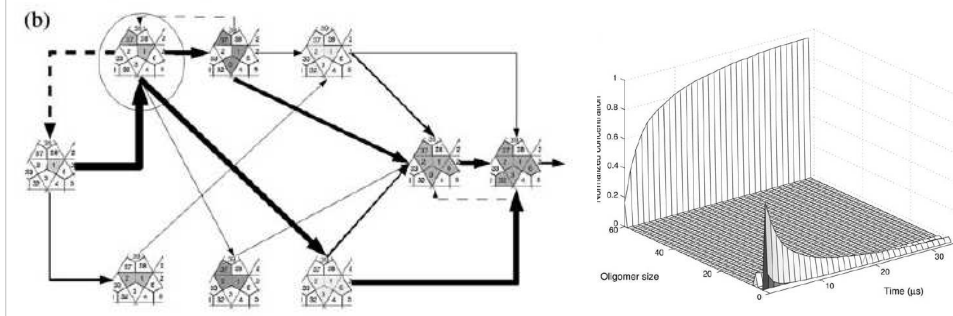


85

To overcome this issue, we used a stochastic approach which allows us to sample the most likely assembly paths. This makes the problem tractable and the method allows us to identify the stable intermediates of the process.

Identify stable intermediaries

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion
 - Dissociation escape from multi-dimensional well

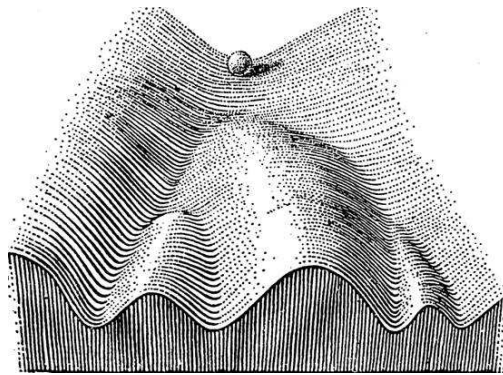


This graph shows the most likely intermediaries and the width of the edges is proportional to how frequently it is used. As you can see, the original tree has been pruned by quite a bit and it should also be emphasized that the graphs that emerge from this process are different for each virus.

In this figure here on the right, I have compiled the information from one of these graphs. On this axis we have time, and oligomer size along this one and concentration here. As you can see, the system starts out with all monomers and these are rapidly combined to form dimers and hexamers. The hexamers are the most stable unit and once larger structures are formed, the transition to the full capsid is rapid which means that no other intermediaries are observed.

Future Work: Organizing principles of the genome

- Use genome-wide data to develop systems biology and biophysical models of gene regulation and gene expression



Waddington, 1953 ⁸⁷

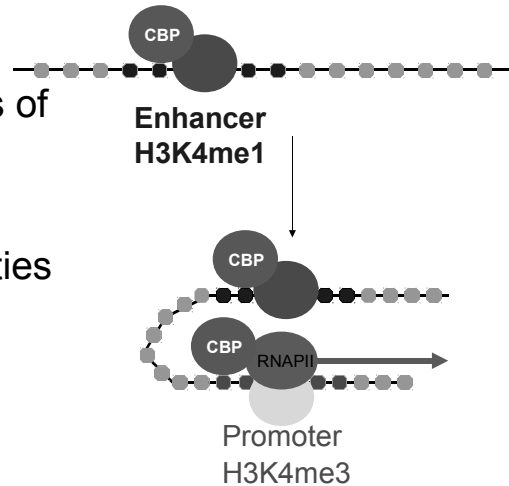
Finally, I'd like to tell you a little about my plans for future research.

To summarize as briefly as possible: My aim is to use quantitative high-throughput data to develop models that are based on biophysical theories and try to bridge the gap between our understanding of the basic laws of physics and chemistry and Darwinian evolution which conceptually explains genotypes and phenotypes.

In more practical terms, what this means is that I want to bring together the two types of research that I did as a graduate student and as a post-doc. My goal is to take advantage of the rich genome-wide data sets that are available today and use them to develop quantitative models of gene expression and gene regulation. In particular, I would like to make sure that my models explicitly account for stochasticity since that is such a central feature in biology.

Develop biophysical models of TF binding

- Use ChIP-Seq for biophysical models of TF binding
- Relate looping to biophysical properties of DNA



88

One example of such an approach would be to further our understanding of TF binding. Today our ability to predict TF binding from sequence alone is very poor and I believe that one of the reasons for this is because of our poor models of TF-DNA interactions.

What I'd like to do is to write down a handful of different models for protein-DNA interactions and then use ChIP-Seq and other high-throughput data, which constitutes 1000s of binding sites to determine which model best explains the observed binding patterns.

I'd also like to take the opposite approach and try to obtain a good energy model from a bottom-up model. Basically, my plan is to use a similar approach as in the virus project. Starting from crystal structures of TF-DNA complexes, try to obtain a coarse-grained model that can somehow be related and mapped to ChIP-Seq data.

As I mentioned, looping of the DNA is believed to be happening widely throughout the genome. I would like to get a better understanding of this process by developing models of enhancer-promoter interaction that take the biophysical properties of the DNA into consideration.

Model stochastic gene expression for entire transcriptome

- Analytical models of gene expression noise
- Apply to genome-wide single-cell RNA-Seq

89

The ME work that I told you about was mainly concerned with numerical solutions of the ME. However, I have some preliminary results that have shown that it is possible to make significant progress with analytical solutions of the ME and this is something that I'd like to explore further.

Studying noise in gene expression is challenging since it requires single cell resolution. However, the first papers that apply RNA-Seq to single cells have already been published. I plan to extend the small-scale models of noisy gene expression to the genome-wide scope and also take advantage of the tools and insights for RNA-Seq data that I have developed as a post-doc.

Determine structure of RNAs

- Many classes of novel RNAs
 - Identify structural motifs
- High-throughput sequencing of structure
 - PARS
 - SHAPE-Seq

.....ACGUCCAAAUUCCCUAGGCUCAAGGCAUUCGAUCGGGAUUUAUA..... →



In addition to eRNAs, several other novel types of RNA have been identified in recent years. Mostly their function remains unknown. However, it is very likely that the function of these non-coding RNAs is determined by their structure.

Fortunately, in the last year or so, novel methods based on high throughput sequencing, including PARS by Howard Chang and Eran Segal and SHAPE-Seq by Adam Arkin have produced large scale data sets of RNA structures.

I plan to take advantage of these data sets to develop better methods for understanding the structure of RNAs.

Acknowledgements

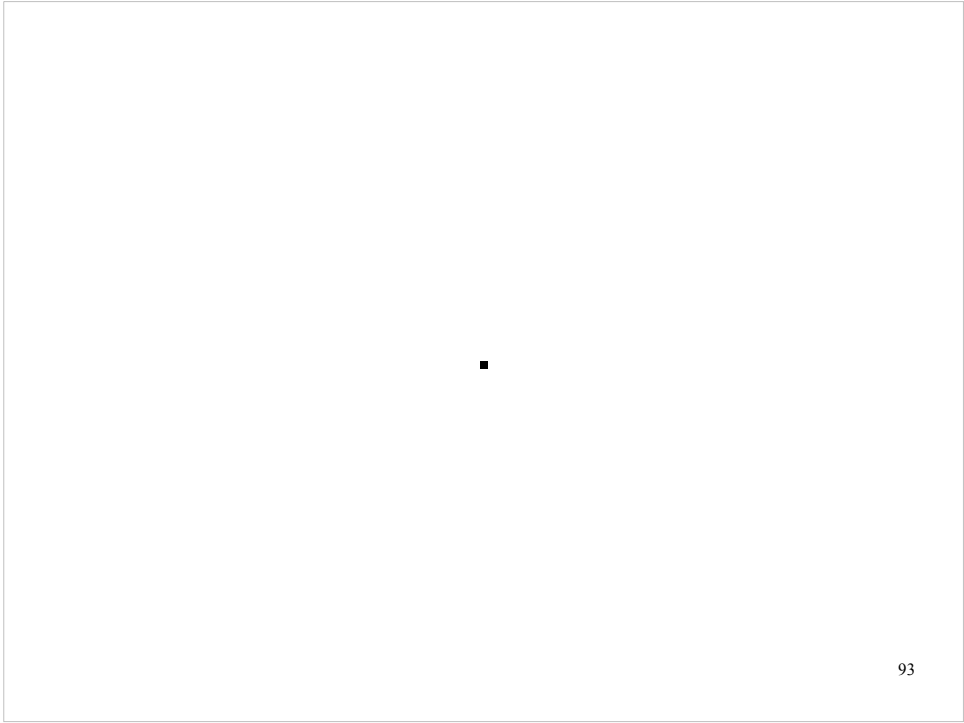
- Gabriel Kreiman
- Jesse Gray
- Tae-Kyung Kim
- Michael Greenberg
- Mauricio Barahona

Click to add title

Thank You

92

I'd also like to thank you for your attention.



With that said I'd like to make a full stop.

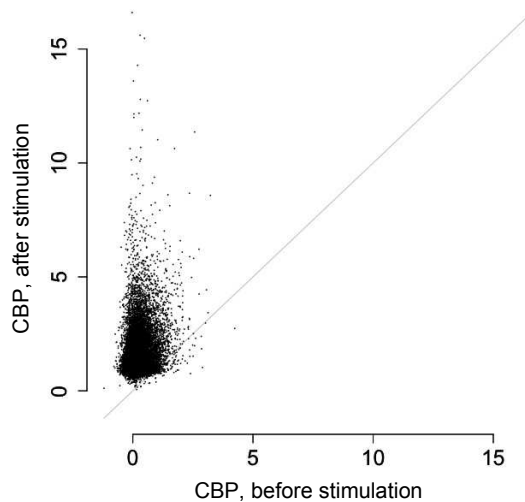
Click to add title

?

94

And I will be happy to take any questions.

CBP binds in an activity regulated manner to ~28,000 sites throughout the genome



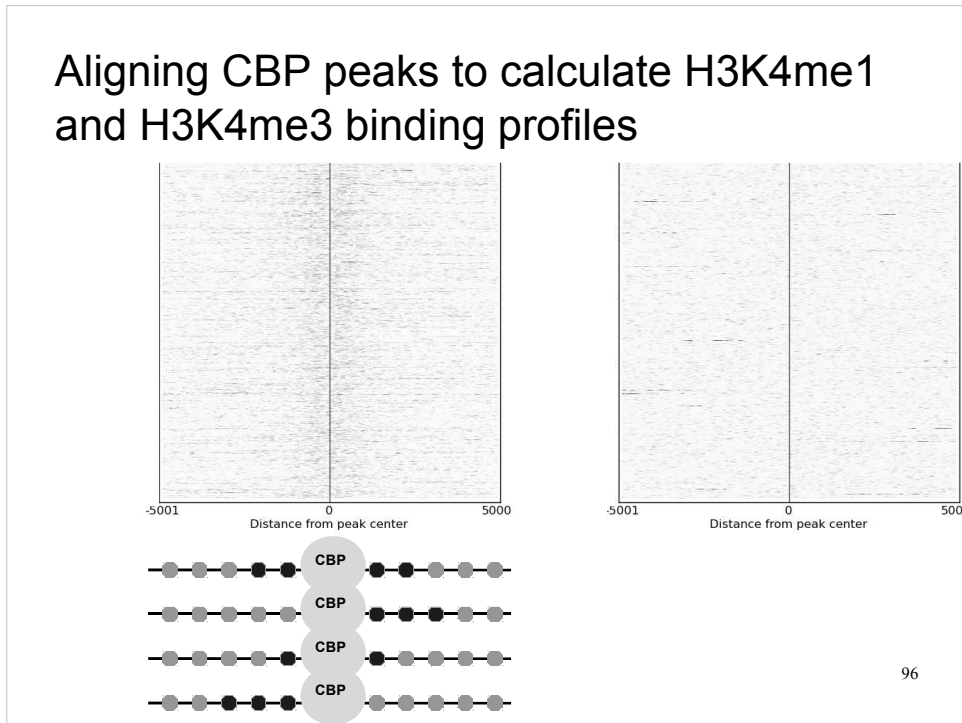
95

I am not going in to the details here, but we used a stringent statistical test to search for CBP binding in the entire mouse genome. Basically, one tries to identify regions of the genome where the histogram is higher than one would expect given a Poisson null-model

We found that CBP is bound at ~28k sites. This number of peaks is in line with other genome wide chip-seq experiments, but what is noticeable is that almost all of the CBP binding was induced by Kcl.

In this scatter-plot, each dot represents a CBP peak and on the x-axis we have the size of the peak before Kcl stimulation and on the y-axis we have the size of the peak after stimulation. As you can see, most peaks are above the diagonal and thus strongly induced.

Aligning CBP peaks to calculate H3K4me1 and H3K4me3 binding profiles



96

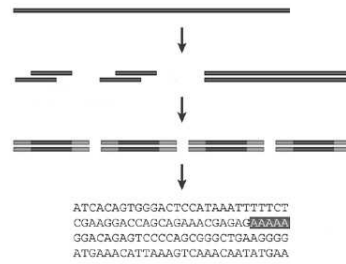
The next few slides will demonstrate why this procedure is a good one

Starting from the set of 28k CBP peaks we aligned them to the center of the CBP binding as illustrated in this schematic.

We only retained the ones that passed the histone modification requirements

RNA-Seq reveals which parts of the genome are transcribed

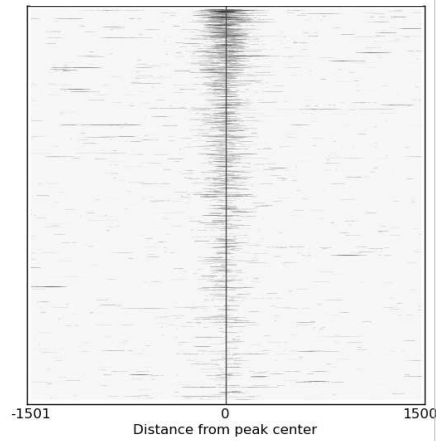
- Fragment
- RNA → cDNA
- 35 bp reads mapped to genome
 - Before and after KCl
 - Total RNA and polyA+



(Wang et al, 2009)

In our experiments, we sampled the RNA before and after the KCl stimulation. Moreover, we looked at two different fractions of RNA, namely the subset of transcripts that have a polyadenylation tail and the whole set of transcripts. The polyadenylation tail is attached to messenger RNAs and increases their stability and allows them to be transported out of the nucleus.

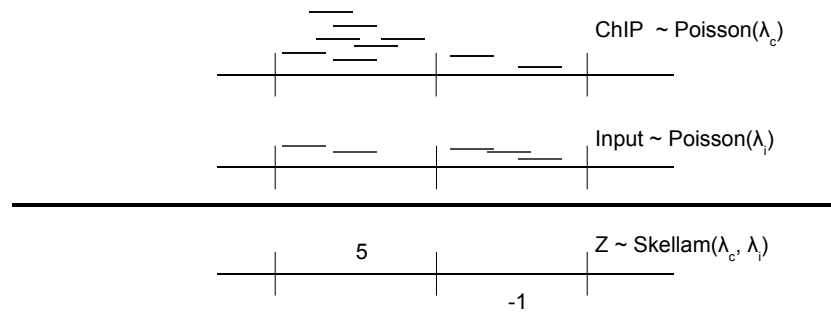
RNAPII binds at activity-dependent enhancers



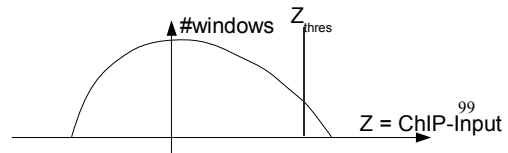
98

Looking at RNAPII at the 12k enhancers sorted by CBP binding as before we clearly see an enrichment of RNAPII for most enhancers

Identifying regions with larger than expected number of ChIP-Seq reads



- False Detection Rate (FDR) determine threshold



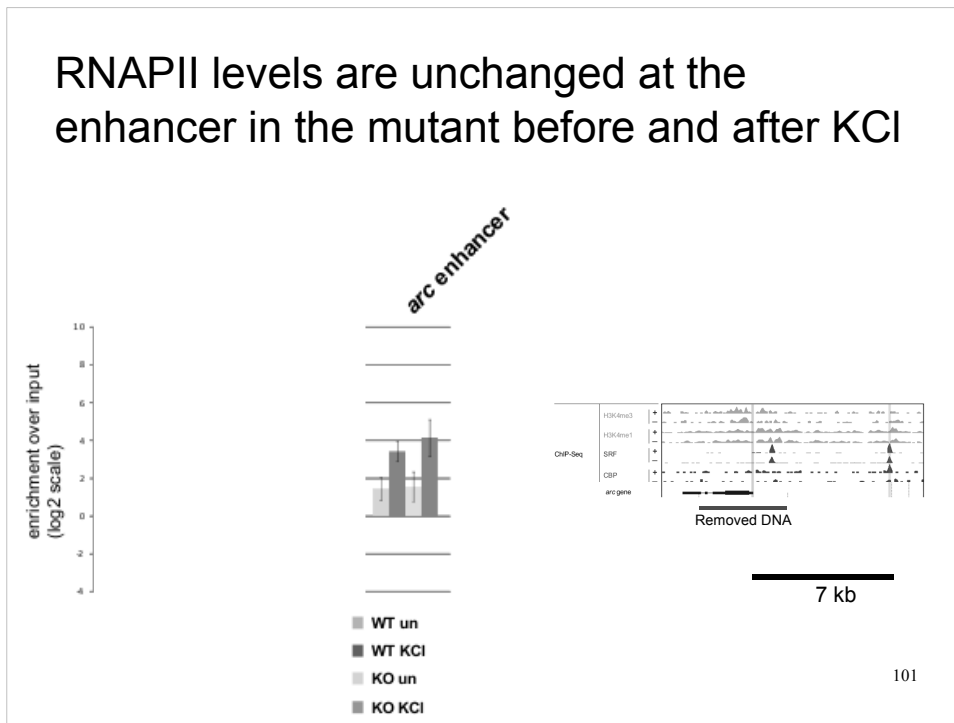
Use False Detection Ratio (FDR) to correct for multiple hypotheses

- $Z_i = \text{\#ChIP reads} - \text{\#input reads in window } i$
 - $\sim 1 \text{ read}/100 \text{ bp}$
 - Assume $\text{\#reads in window } P(k) = \lambda^k \exp(-\lambda)/k!$
 - Difference between two Poisson random variables
 - $Z_i \sim \text{Skellam}(z, \lambda_1, \lambda_2)$
- $$p(x) = e^{-(\lambda_1 + \lambda_2)} (\lambda_1 / \lambda_2)^{x/2} I_x(2\sqrt{\lambda_1 \lambda_2})$$
- Millions of windows need to be tested
 - FDR - expected fraction of false positives

100

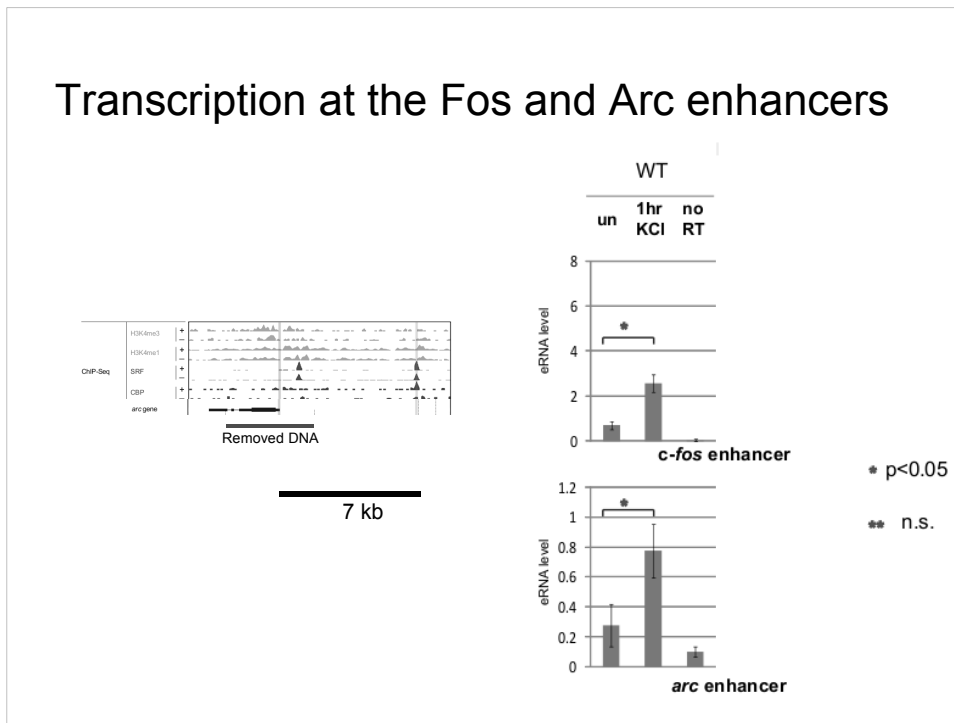
Since there are literally millions of windows that we wish to interrogate, this means that we must correct for multiple hypotheses when determining the significance levels. With such a large number of tests, Bonferroni correction is not very useful and hence we instead use a false detection ratio approach.

RNAPII levels are unchanged at the enhancer in the mutant before and after KCl



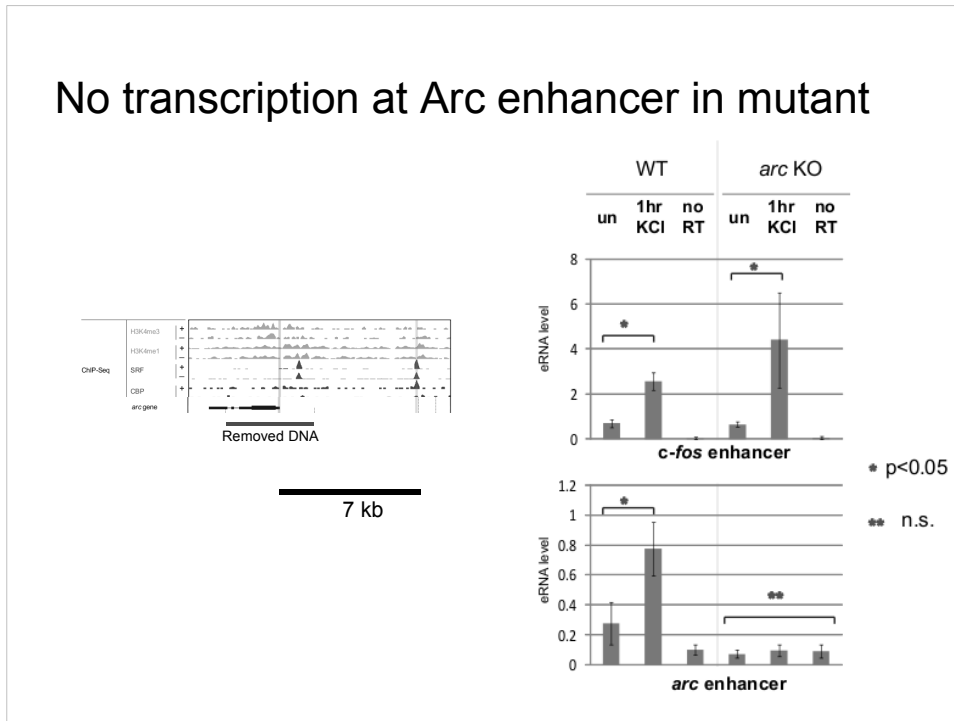
As you can see here, the levels of RNAPII at the enhancer remained the same in both the knock-out and the wild type, both before and after KCl stimulation.

Transcription at the Fos and Arc enhancers



Investigating the *arc* locus again, we monitored the transcription at the *arc* and *fos* enhancers. For the wild-type animals we found that there was strong induction in response to KCI

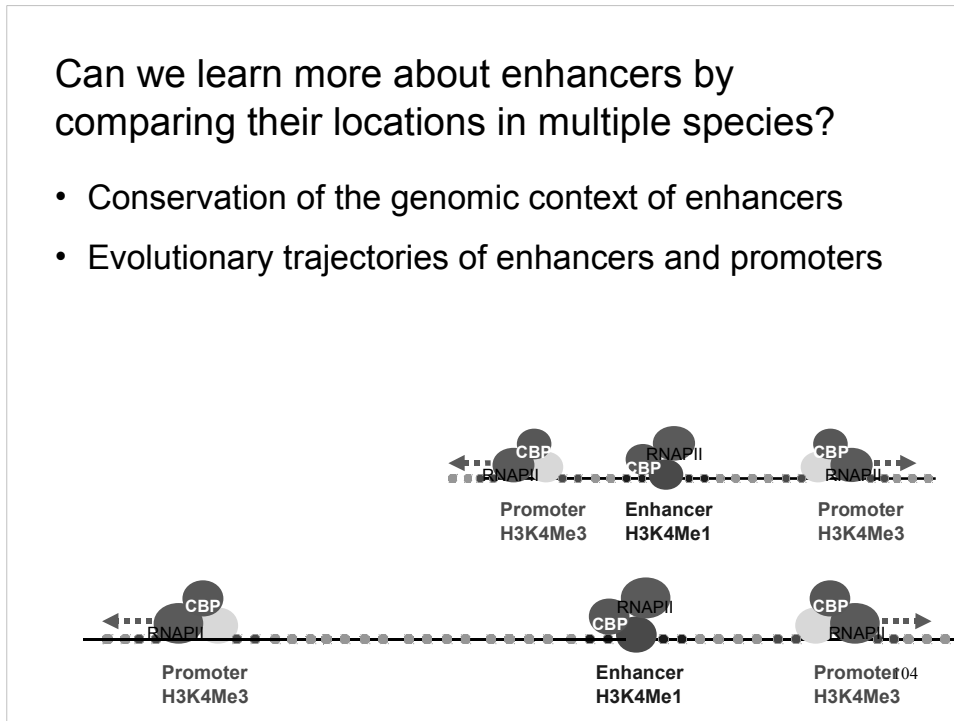
No transcription at Arc enhancer in mutant



However, for the knock-outs, we found that no transcripts were produced at the *arc*-enhancer, whereas the *fos*-enhancer which is located on a different chromosome was the same as before. This suggests that the polymerase is independently recruited at the enhancer, but that the interaction with the promoter is required to produce transcripts.

Can we learn more about enhancers by comparing their locations in multiple species?

- Conservation of the genomic context of enhancers
- Evolutionary trajectories of enhancers and promoters

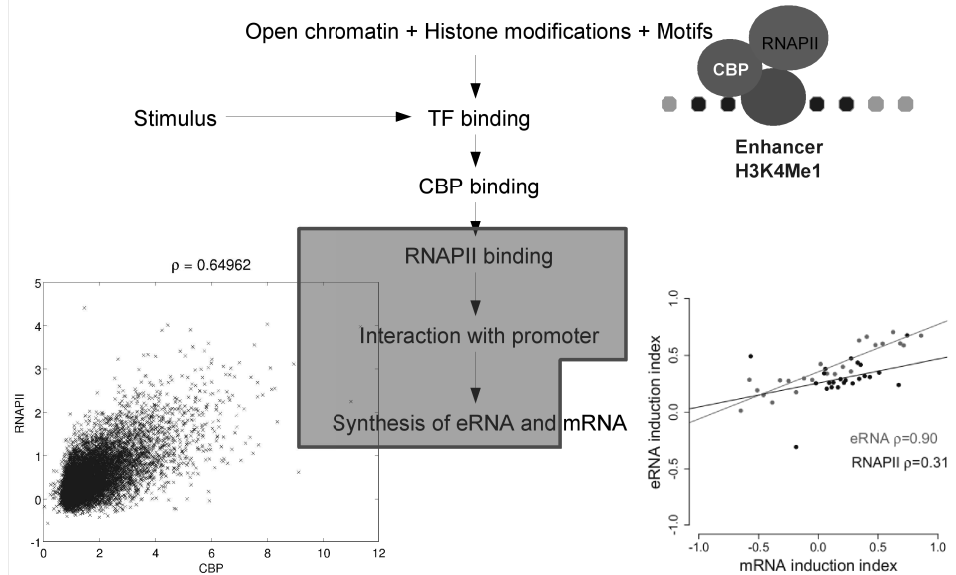


In general, the sequence found at the enhancer is often strongly conserved btwn species. However, much less is known about the position of the enhancers relative to known genes and the genomic context.

By comparing multiple species, I plan to find out to what extent the enhancers tend to remain in the same context or if they get shuffled around which might imply that they shift targets.

An important corollary would be to ask if there is any evidence of sites that act as promoters in one species but in enhancers in a different one. Our results suggest that enhancers and promoters are more similar than previously thought and I am curious as if there is an evolutionary connection as well.

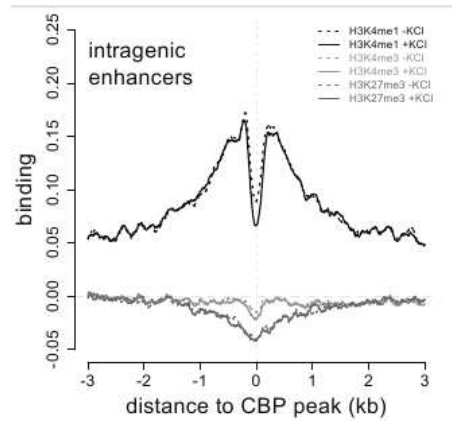
Conjectured order of events for eRNA



However, as we saw, the recruitment of RNAPII is not sufficient to produce eRNA and that through some, yet unknown, mechanism interaction with the promoter is required.

Intragenic enhancers

- ~7,000 enhancers overlapping introns
 - H3K4me1, but no H3K4me3

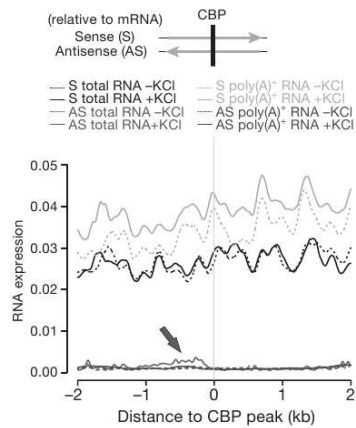


106

Moreover, they have the same characteristic bimodal H3K4me1 pattern and absence of H3K4me3 as the extragenic enhancers.

Intragenic enhancers are also transcribed

- ~7,000 enhancers overlapping introns
 - No signal detectable on sense strand
 - Significant anti-sense transcription



107

However, because of the overlapping mRNAs, we cannot expect to detect a signal on the sense strand. On the anti-sense strand, on the other hand, we do find the same characteristic signal as for the extragenic enhancers.

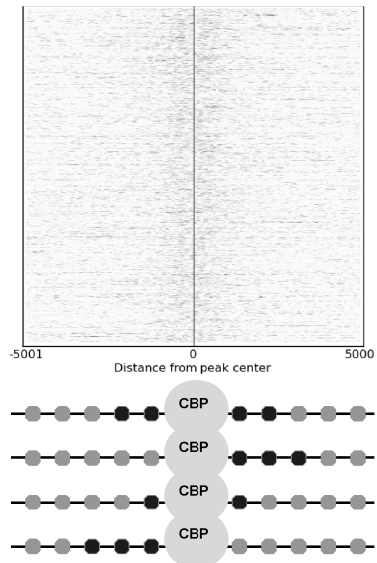
~100 enriched motifs are found

Word	Enrichment	Known TF
TGASTCA	4.74	Fos/Jun
TGACGTCA	6.41	Creb
CTAWWWATA	3.34	Srf
TCGTG	1.56	Npas4
CTGCCAAA	3.34	?

108

We found ~100 significantly enriched motifs in our enhancer set. Some of them, such as CREB, SRF and Npas4 corresponded to Tfs that had been identified previously, while others are similarly enriched but do not correspond to factors with known motifs.

Aligning CBP peaks to calculate H3K4me1 binding profiles



109

Here I am showing you the level of H3K4me1, sorted in the same order as before.

The pattern is not as clear, but if you squint, you may see that there is an enrichment on both sides of the center.

How abundant are eRNAs compared to mRNAs?

- Identify **all** transcripts in the genome
 - Wavelet-based algorithm for *de novo* detection of transcribed regions accounts for 99.8% of reads

110

As you saw from the browser screen shots, the level of eRNAs is much lower than for mRNAs.

To quantify the levels, we need to characterize the entire transcriptome. This is a significant challenge since in addition to eRNAs, there are many other types of unannotated transcripts that are discovered in an RNA-Seq experiment.

I have developed an algorithm that allows me to identify which regions are transcribed without using the annotation. I don't have time to go into the details here, but suffice to say, the algorithm is comprehensive as it accounts for more than 99% of the reads.

How abundant are eRNAs compared to mRNAs?

- Identify **all** transcripts in the genome
 - Wavelet-based algorithm for *de novo* detection of transcribed regions accounts for 99.8% of reads
 - Annotated RNAs ~ 98.3%
 - eRNAs ~ 0.02%
 - 1 in 10,000 reads is an eRNA read
 - mRNAs ~100 times more abundant

111

Once all the numbers have been crunched, it is clear that only about 1 in every 10 k read comes from an eRNA. Since mRNAs are much longer, the average expression level differs by about two orders of magnitude.