

Mechanisms and models of distal enhancers of inducible gene expression

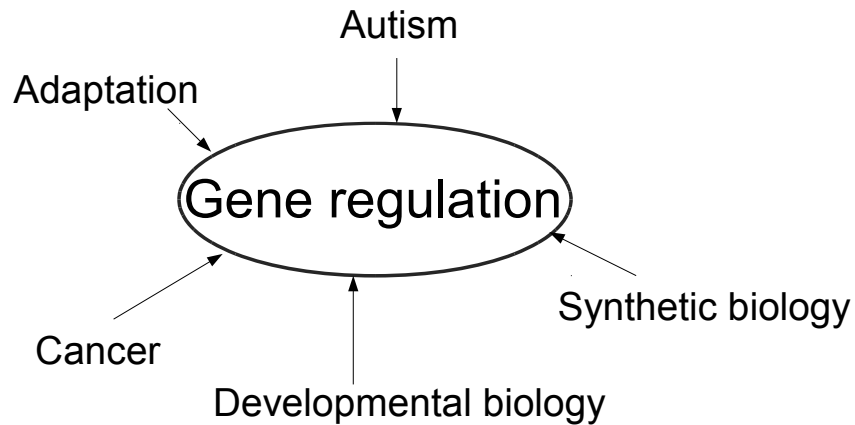
Martin Hemberg

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.

Why is gene regulation important?



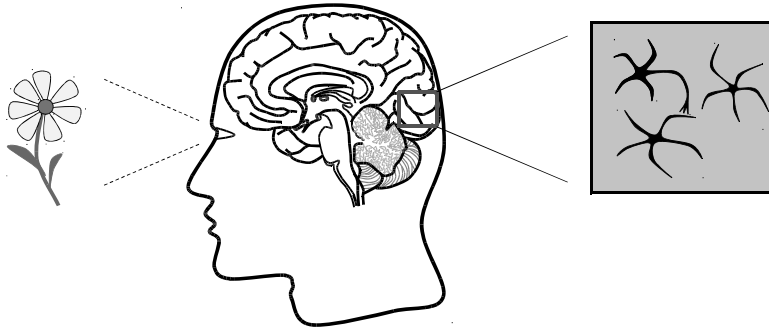
2

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

Gene regulation is central to biology and as you can see from this figure, it is involved in many aspects of biology. Yet, there are several large gaps in our understanding of this fundamental biological process. A particular shortcoming is that we mainly have a qualitative view of gene regulation today.

Synapses change in response to external environmental stimuli

- Caused by turning ~1000 genes on or off

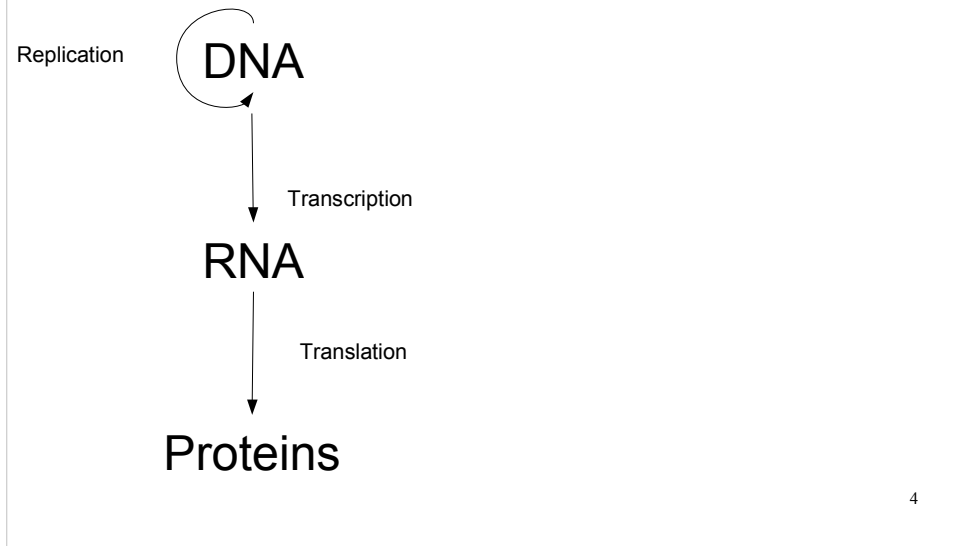


Hubel & Wiesel, 1970's

3

One important example of why gene regulation is important is brain development. During brain development, connections between neurons are influenced by sensory experience. That is, synapse formation and changes of synapse strengths respond to external environmental stimuli.

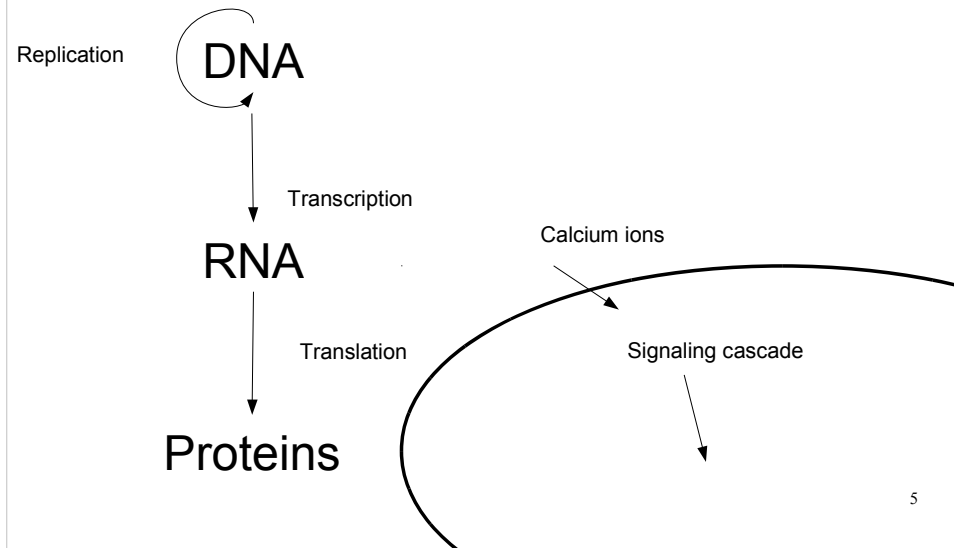
What is gene expression and gene regulation?



To understand the logic of gene expression, we need to consider this figure which illustrates the key principle in molecular biology, known as the central dogma. The central dogma states that information is stored in the DNA. The DNA can be transcribed into RNA. There are several classes of RNAs, but the most important one is called messenger RNAs and they can be translated into proteins. Proteins are considered the main work horses of the cell and when we talk about genes, we usually refer to a protein and its corresponding DNA.

When we talk about gene expression, we refer to the amount of RNA that is present in the cell. Gene regulation is the process by which gene expression is controlled.

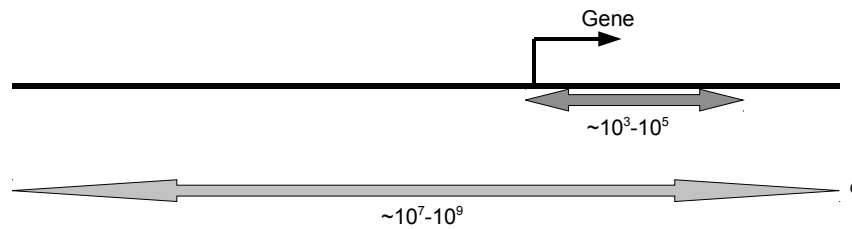
Activity dependent gene expression triggered by influx of Ca



Returning to the example with the brain, it has been shown that the changes in gene expression, known as activity dependent gene expression is triggered by the influx of calcium ions into the cell. The increased levels of calcium trigger a signaling cascade which leads to changes in gene regulation.

Mouse genome is large and has few genes

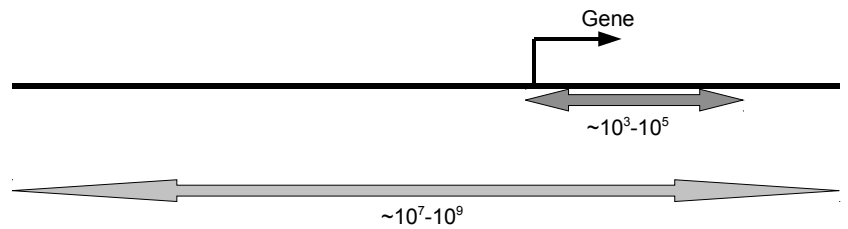
- ~25,000 genes
 - ~2% of DNA



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 ~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.

Bacterial genomes are compact

- ~25,000 genes
 - ~2% of DNA
 - Bacteria $\sim 10^6$ base pairs (**bps**)
 - 10^3 - 10^4 genes

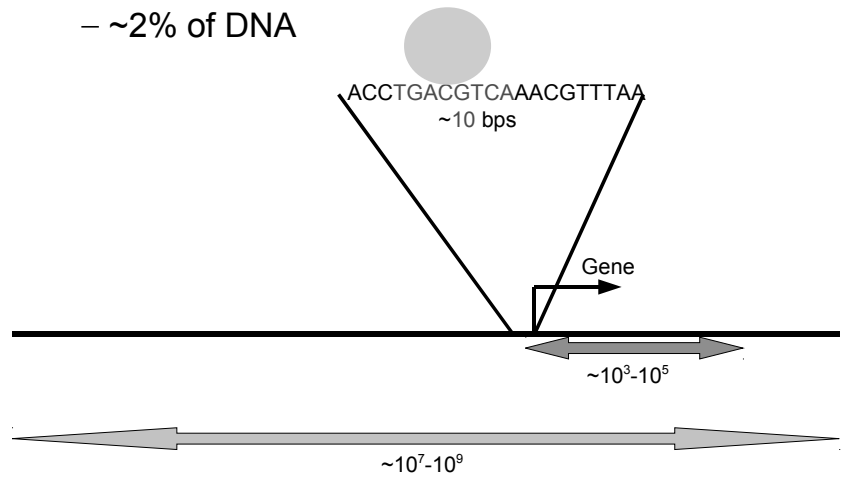


This is in stark contrast to bacteria which have been studied more widely in biophysics. Their genomes are much more compact with only about a million basepairs, with most of them coding for genes.

There are some quite successful biophysical models for bacterial gene regulation. Unfortunately, those models do not work very well for mammals, mainly because mammalian genomes have additional layers of complexity.

Transcription Factors (TFs) bind to DNA motifs

- ~25,000 genes
– ~2% of DNA



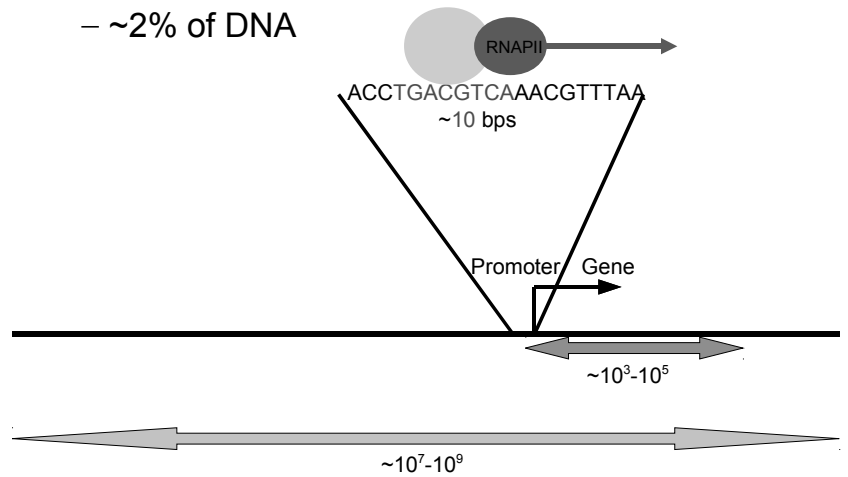
Gene regulation is a complex process and there are many different mechanisms involved. Perhaps the most important one is binding by transcription factors (Tfs).

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

Each TF has its own preferred motifs which dictates where it will bind.

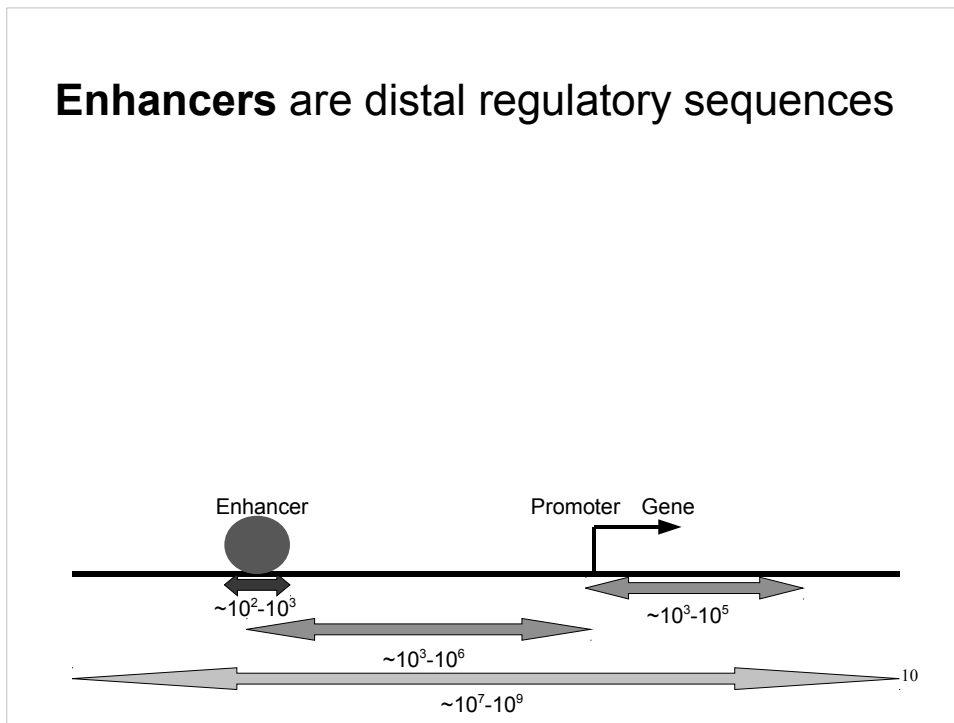
Transcription factors bind at promoter to recruit RNA Polymerase II (RNAPII)

- ~25,000 genes
– ~2% of DNA



Traditionally, Tfs are thought to bind in the promoter region, which is defined as the region near the start of the gene. In bacteria and yeast, most Tfs do bind at the promoter. 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 regulatory sequences

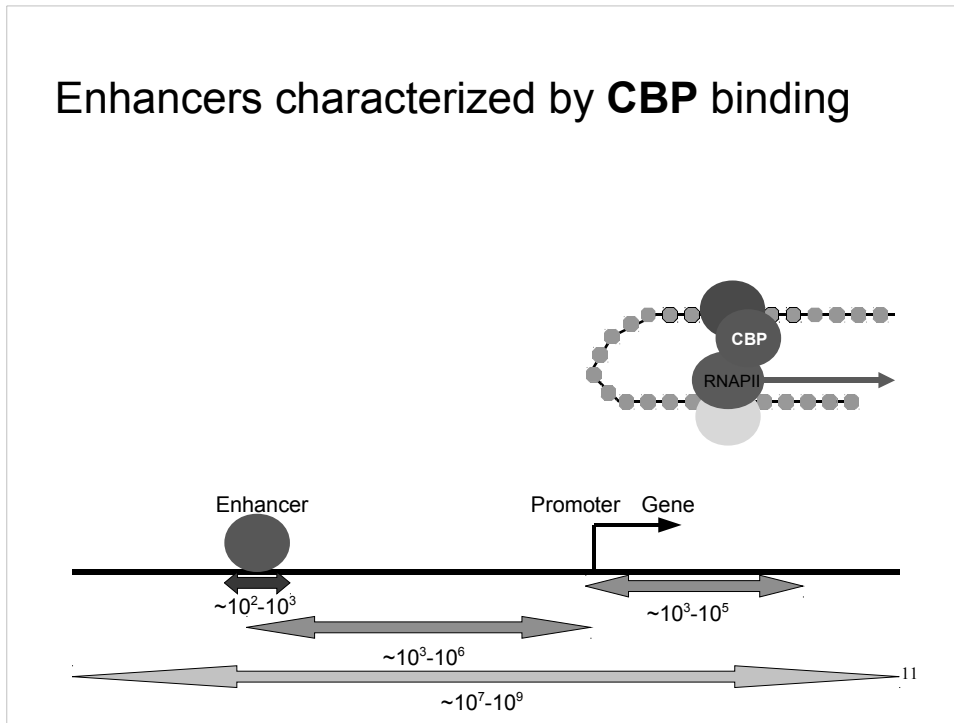


One example of a feature of gene regulation that is not present in bacteria are distal regulatory elements, which are known as enhancers. Enhancers are short and very difficult to distinguish based on their sequence alone. Enhancers are often located very far from their targets, ranging from a kilobase to a megabase.

To complicate matters even further, enhancers have been shown to be cell-type specific which means that they must be studied in a specific context.

Ideally, as a theoretical physicist, what I'd like to do here is to write down the equations for how enhancers impact gene expression, solve them for mouse neurons and then ask my collaborators to do the experiment to verify my predictions. 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.

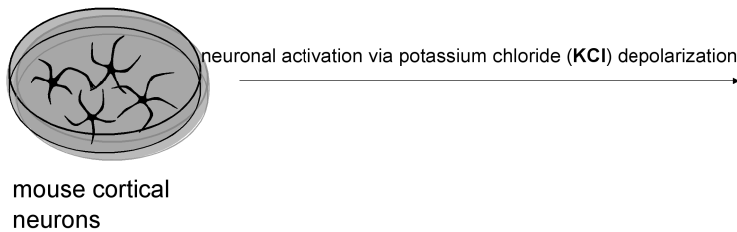
Enhancers characterized by **CBP** binding



Currently, we don't have the ability to reliably predict enhancer location based on the DNA sequence alone. However, it has recently been shown by Bing Ren and others that enhancers can be identified by the binding of the protein CBP.

The mechanisms by which enhancers work are poorly understood, but it is believed that they use a looping mechanism to interact with the promoter and thereby increase gene expression.

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

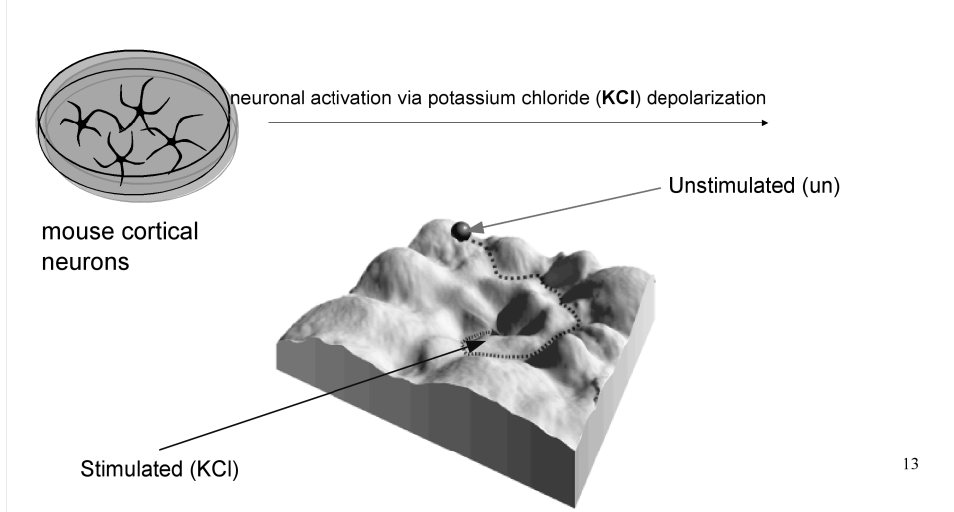


12

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.

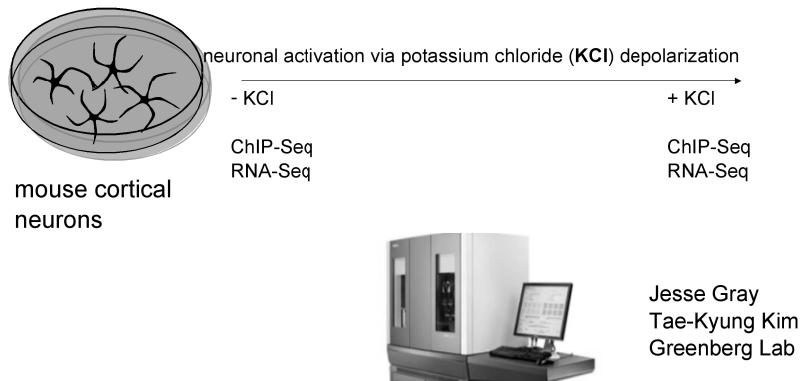
The system allows for precise control of the stimuli and the measurements in a way that would not be possible in the brain.

Potassium chloride (KCl) stimulation induces cells to change state



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



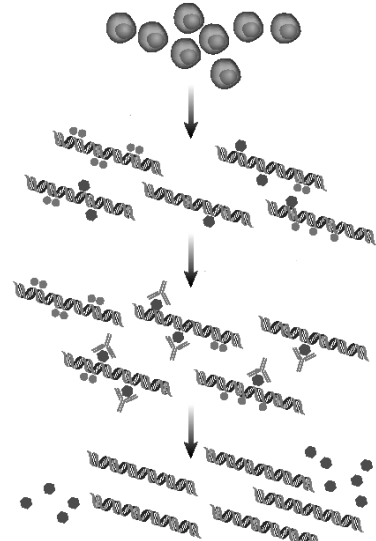
14

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*

- Short **reads** mapped to reference genome
- #reads ~ binding
- $\sim 10^6$ reads
- Unbiased



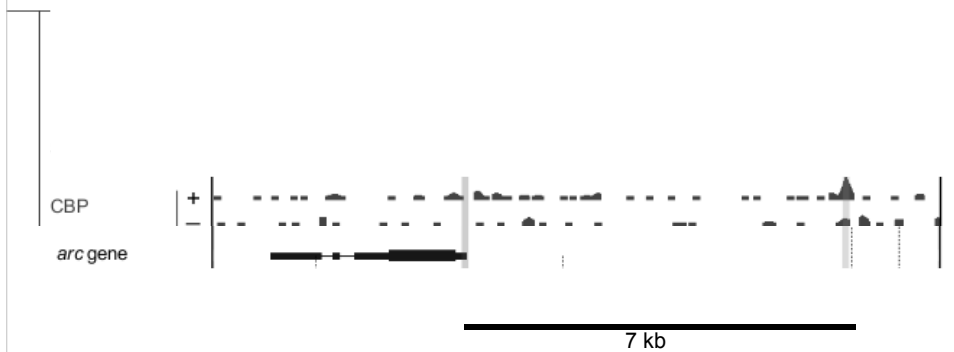
(Mardis, 2007)

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. The number of reads from a given location reflects the amount of binding and the number of reads for a typical experiment is around 10 million.

An important feature of chip-seq is that the method is unbiased and that it will provide information about binding throughout the entire genome all at once. [~ 6 min]

Inducible CBP binding at enhancers

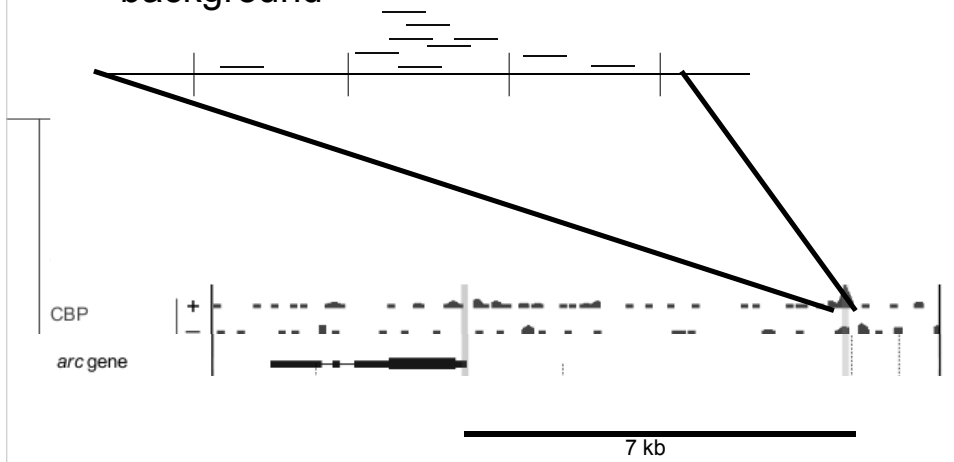


This is a screenshot from the UCSC genome browser a tool that is often used to visualize genomic data. What we have on the x-axis are genomic coordinates and this represents a small portion of the mouse genome. Over here is the *arc* gene which is one of the genes that is upregulated in response to activity. 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 in two replicate experiments

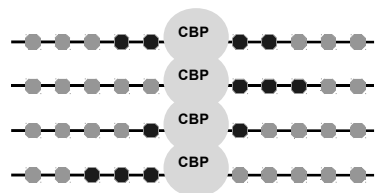
- Regions that have significantly more CBP than background



To search for CBP binding sites in a genome-wide manner, I 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.

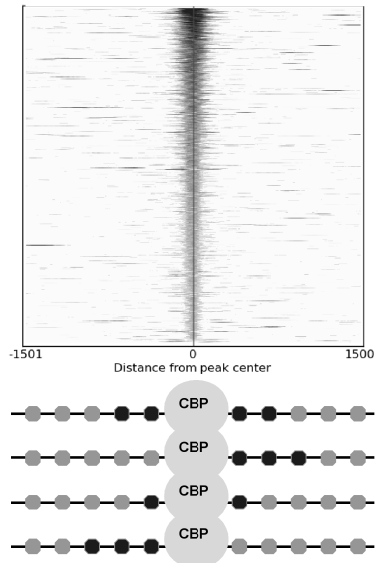
Aligning CBP peaks to calculate binding profiles



18

To give you an idea of what the data looks like, we can align all of the peaks to the center of the CBP binding as shown in this schematic.

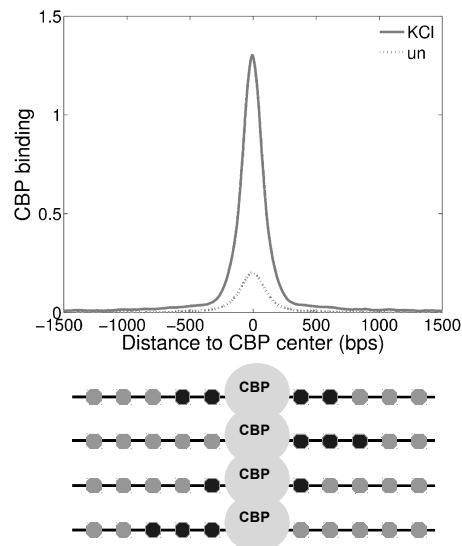
Aligning CBP peaks to calculate binding profiles



19

In reality, we have 28,000 loci and in this plot, each line corresponds to a CBP peak and the purple shade is proportional to the number of reads. The peaks have been sorted by the level of CBP binding

Average profile of CBP binding



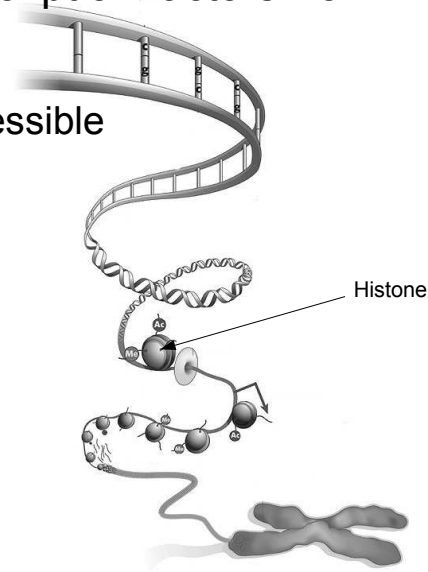
20

Another way of plotting this data is instead to calculate the average CBP level 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.

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.

Histones prevent transcription factors from binding to DNA

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



(ENCODE, 2007)

21

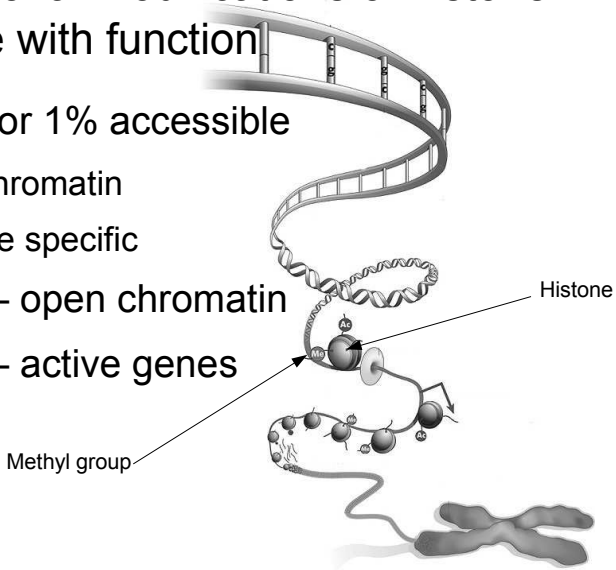
In addition to Tfs, there are many other molecules that bind to DNA and this may prevent TF from reaching its target. This competition provides another layer of regulation.

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 they are important for DNA packing. 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 main reasons why we have different cell-types, even though every cell has the same DNA..

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)

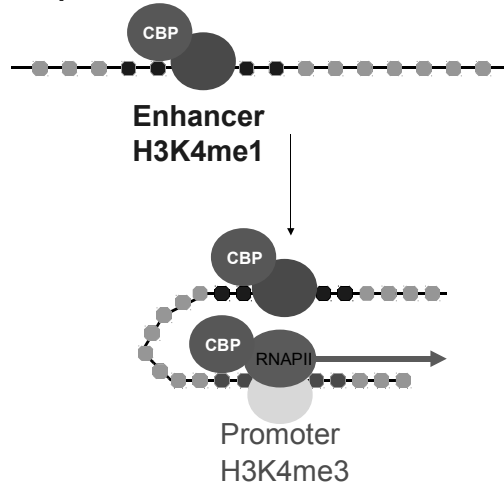
22

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 dozens of 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 putative enhancers

- **CBP** binding
- **H3K4me1** flanking
- **H3K4me3** absent
 - Many unannotated promoters in the genome



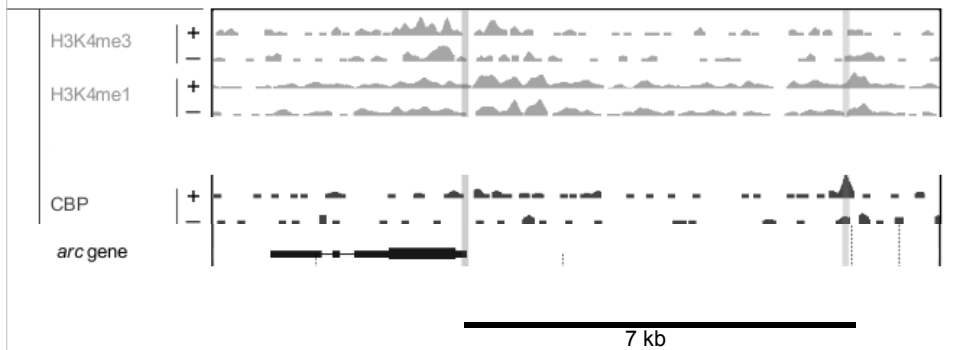
23

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 to distinguish enhancers from promoters..

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

- Click to add an outline



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.

Since there are many unannotated promoters in the genome, we must use this pattern to identify putative enhancers and we cannot rely on the annotation alone.

We identified ~12,000 activity-dependent enhancers throughout the genome

- **CBP** peak
- **High** levels of flanking **H3K4me1**
- **Low** levels of **H3K4me3**
 - Independently tested and validated 8 enhancers

25

Using these criteria, we were left with a list of ~12k putative distal enhancers.

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 bio-luminescent 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.

[~6 min, 12]

What TFs bind to enhancers?

TCGACGTAGCTAGCATGATCGATAGATC
?

- Click to add an outline

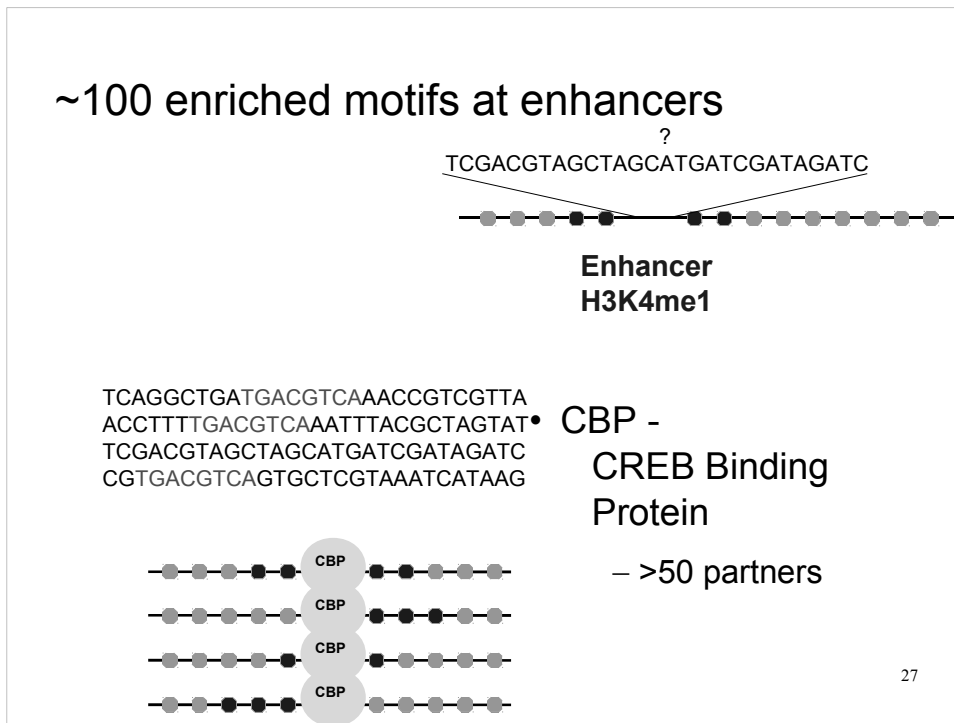
Enhancer
H3K4me1

- CBP -
CREB Binding
Protein
– >50 partners

26

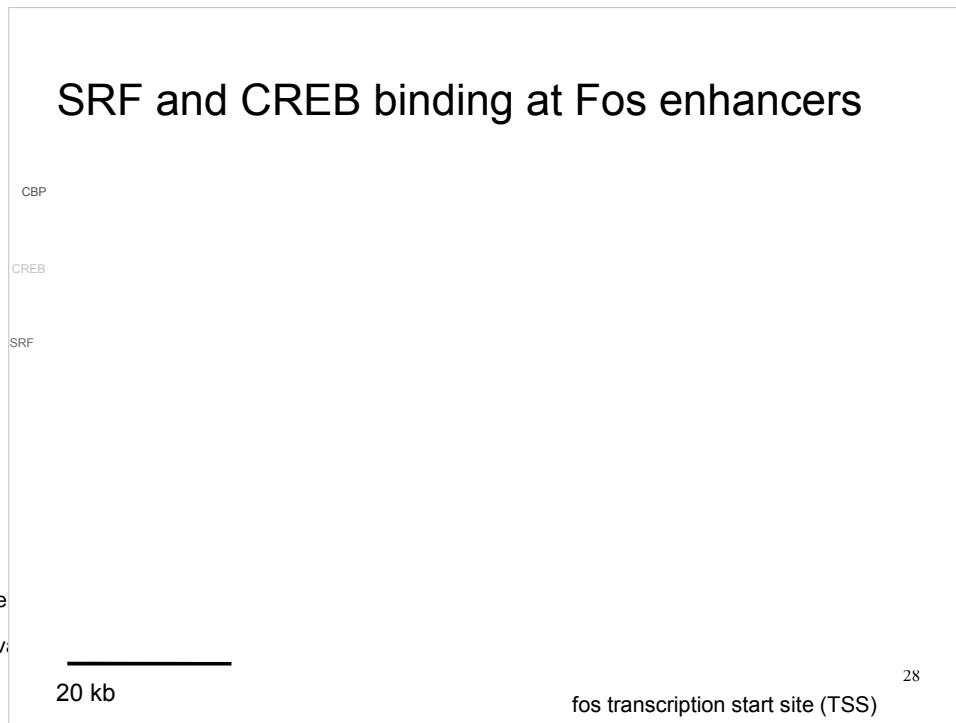
What I also need to tell you about CBP, which stands for Creb binding protein, is that it cannot bind DNA directly by itself, and as a co-activator it has at least 50 partners to which it may bind. Creb is one of those partners which does bind to DNA

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 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 at least 50 partners to which it may bind. Creb is one of those partners 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.

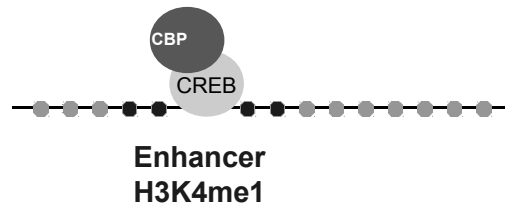


Here I am showing you another screenshot from the genome browser of the Fos locus. Fos is a gene that is known to be strongly induced by Kcl and it is located on the positive strand with the red band marking the promoter. As you can see from the CBP track up here, there is inducible binding, both at the promoter and at these enhancers here marked in blue.

As you can see here for the two Tfs CREB and SRF, they have binding sites at the promoter as well as at some of the enhancers sites, suggesting that they could be responsible for the recruitment of CBP.

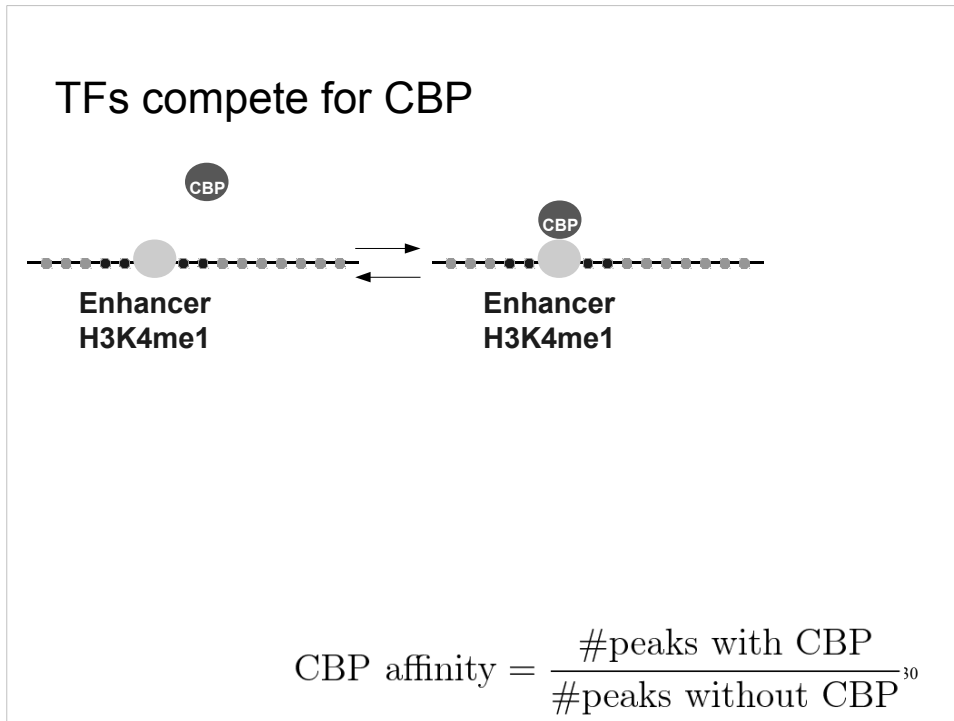
Is CBP binding determined by other TFs?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence



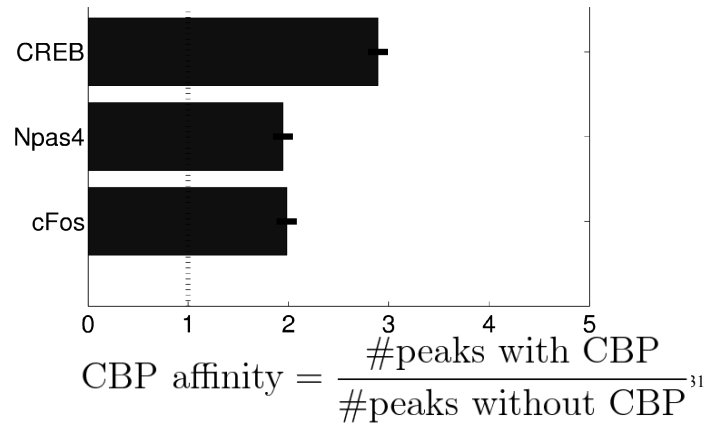
29

Now that we know where these transcription factors bind, not just at enhancers, but throughout the entire genome, we asked if we could predict CBP binding from the transcription factor binding sites.



We considered the following model where we assume that each TF has an intrinsic affinity for CBP. We then assume that each peak can be in two states: either bound to CBP or without CBP as illustrated in this cartoon and that the transition is a first order process. We assume that what we are observing is the equilibrium of this process and given the relative number of peaks with and without CBP, we can estimate the CBP affinity for each TF.

CBP binding determined by affinity of TF

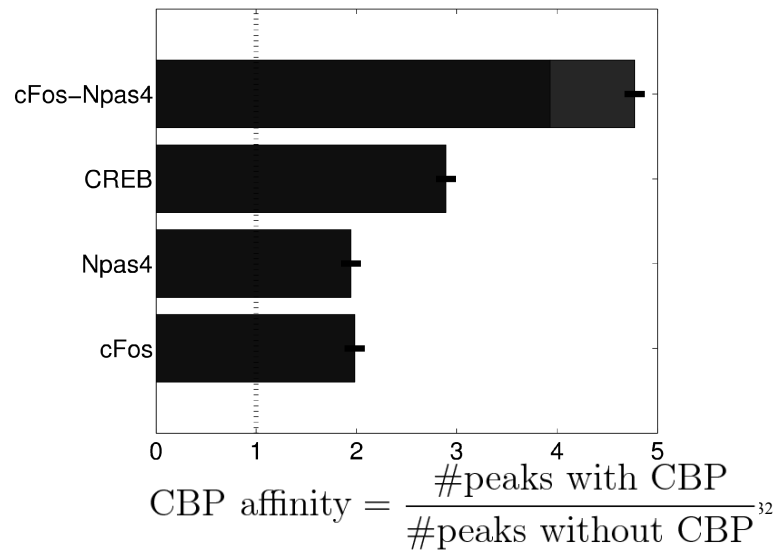


Here are the results for three of the factors in our study.

The dashed lines corresponds to no affinity above background so all factors appear to attract CBP.

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

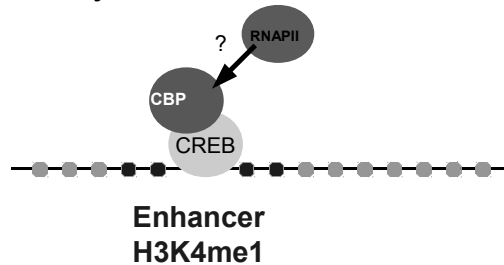


However, when we consider pairwise combinations as well, then we find that there are significant synergies. What we see here is that even though Npas4 and Fos have lower CBP affinity than CREB by themselves, when they both are present, the affinity is more than doubled and significantly larger than for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

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs



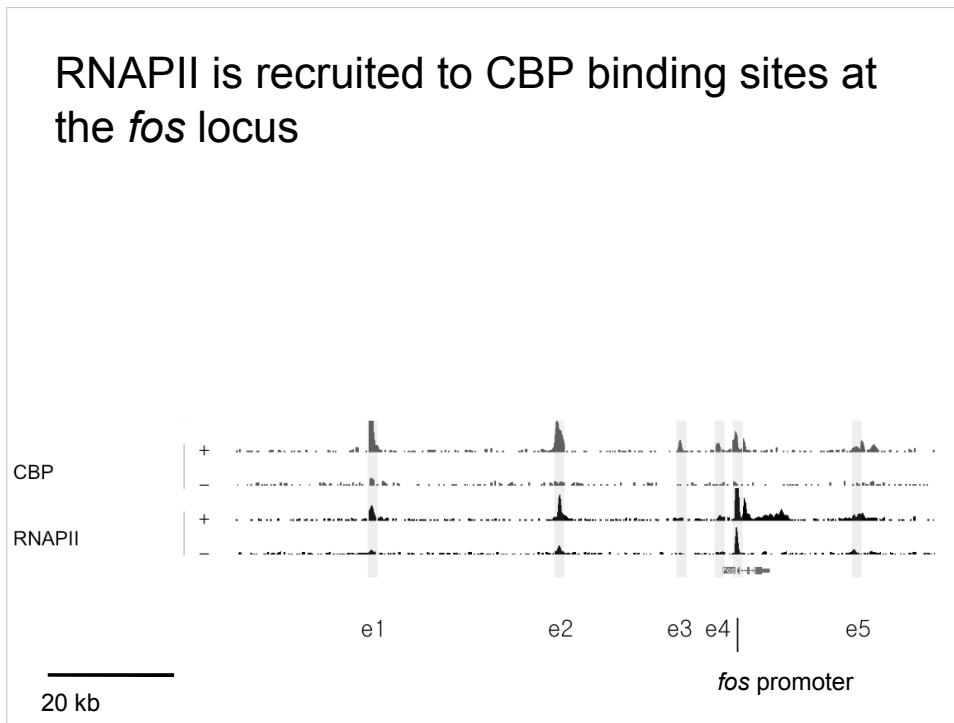
33

So what I have showed you so far is that the Tfs found at the enhancers can be predicted from the sequence and that the combination of Tfs in turn determines the CBP

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.

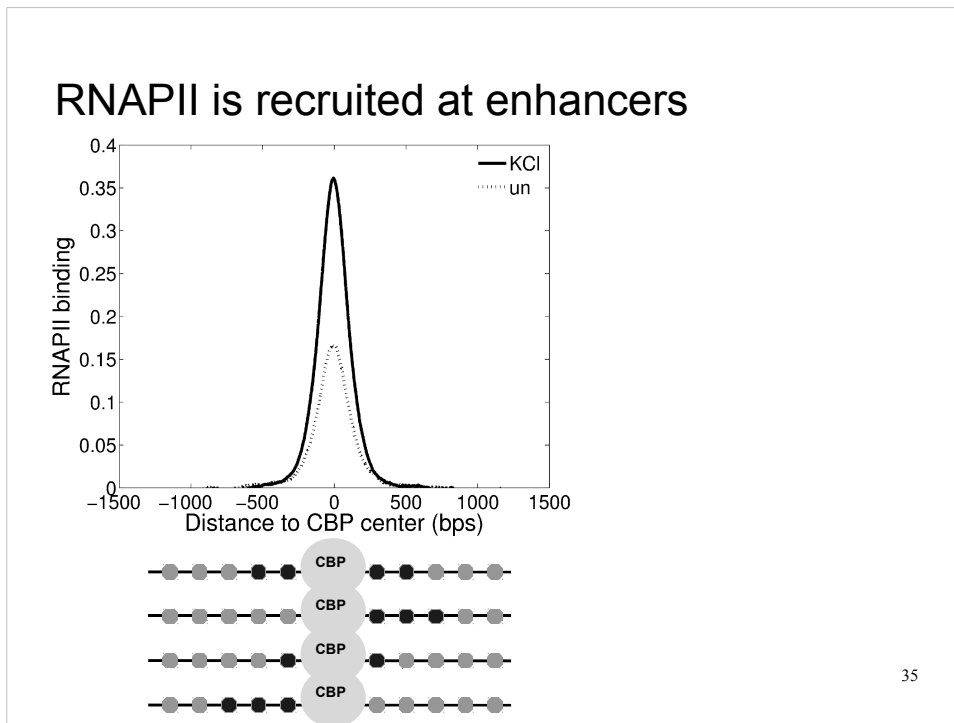
Here I should remind you that RNAPII is the enzyme that is responsible for transcription. That is, it is the molecule that reads of the information in the DNA and creates a corresponding RNA molecule and hence it is one of the most important molecules in the cell.

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



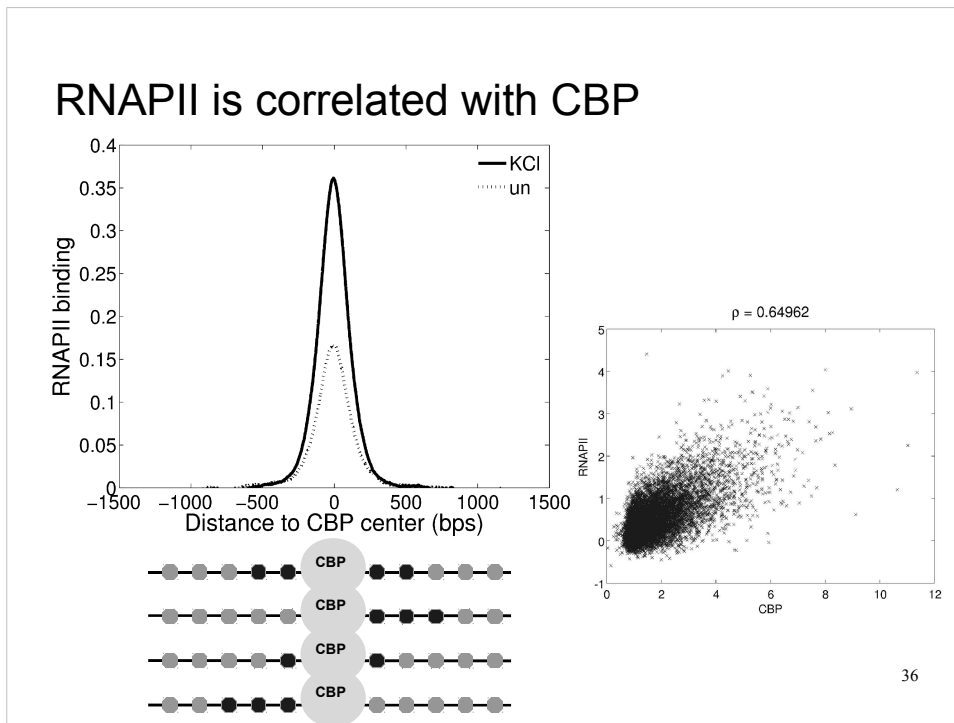
Going back to the *fos*-locus, we do indeed find that Pol2 binding shown in black overlaps with CBP and that it is also induced by the KCl-stimulation.

RNAPII is recruited at enhancers



35

Here is an average plot of the RNAPII levels. The dashed line is before KCl and we see that the Pol2 levels are doubled.



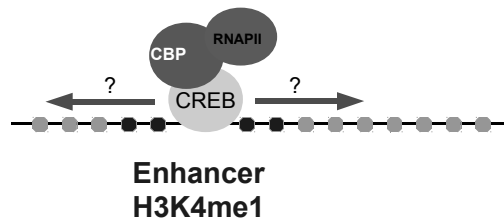
The scatter plot to the right shows that there is also a significant positive correlation between CBP and RNAPII, suggesting that CBP does indeed contribute to the recruitment of RNAPII.

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.

[I also investigated if the histones modifications or the other transcription factors could predict RNAPII binding, but additional factors have a negligible impact on the prediction accuracy.]

What is the function of RNAPII at enhancers?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII



37

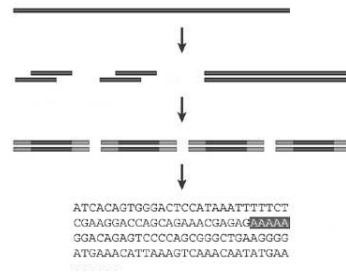
What I have shown so far is that we can use a realistic biophysical model to understand how proteins bind either directly or indirectly at enhancers.

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.

RNA-Seq finds transcribed parts of the genome

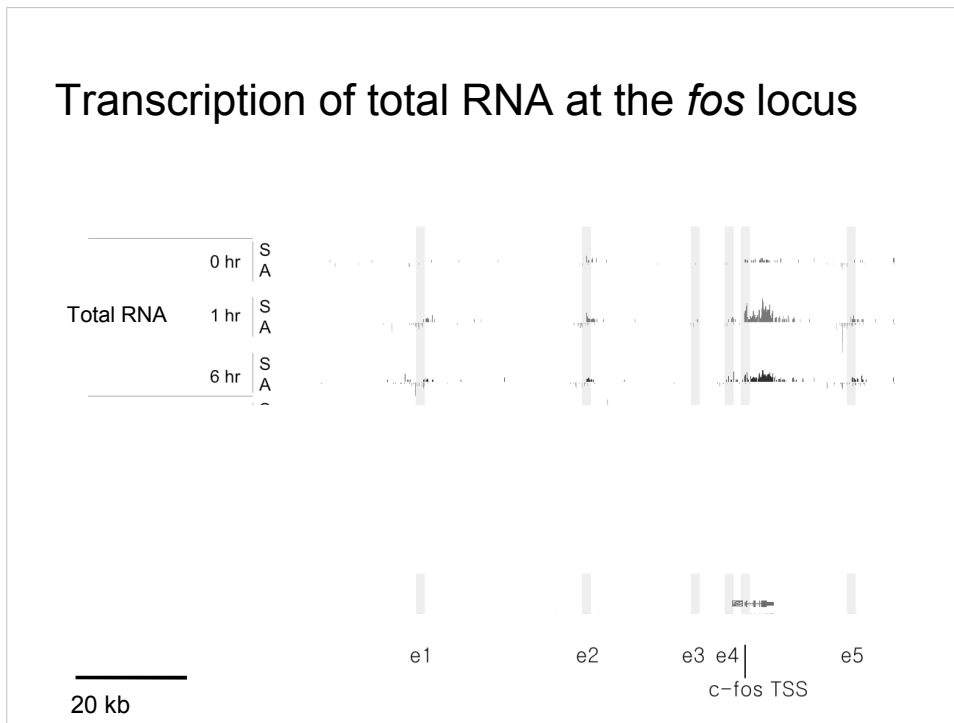
- Short **reads** mapped to reference genome
- $\sim 5 \times 10^6$ reads
- #reads \sim RNA



(Wang et al, 2009)

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

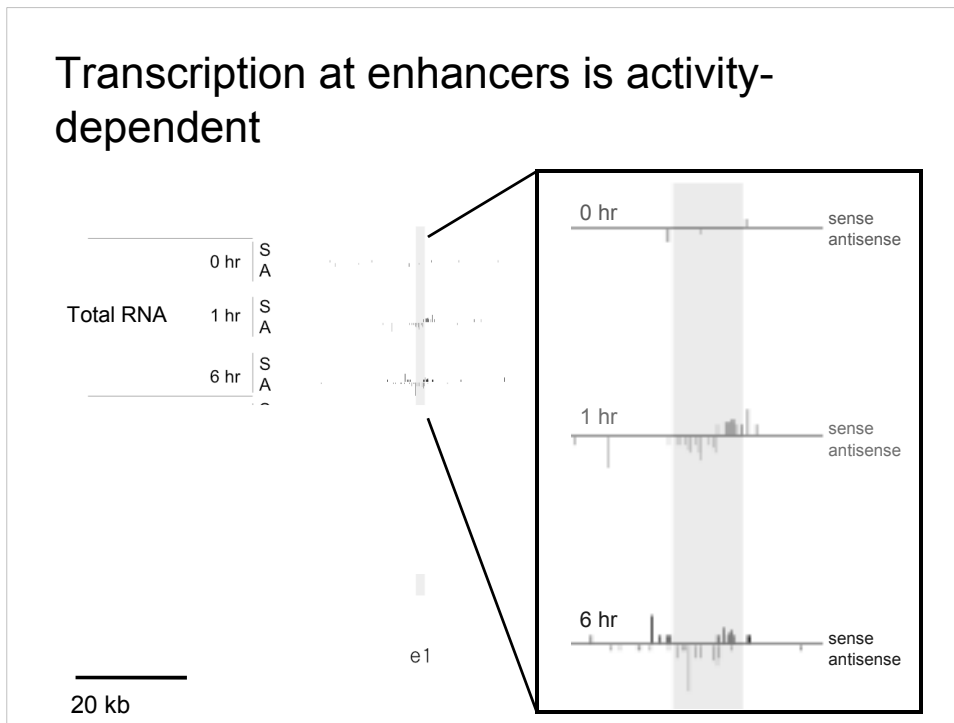
I don't have time to describe the method in detail, but the type of data that we get from the experiment is conceptually similar to ChIP-Seq. But instead of telling us where proteins bind to the DNA, it tells us which parts of the DNA that have been transcribed. Again the data consists of short reads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



Here is the *fos* locus again. DNA can be transcribed on both strands separately and we illustrate this by using upward-pointing bars for the forward strand and downward pointing bars for the negative strand. Over here is the *fos* gene and these reads correspond to mRNAs that will later be translated. As you can see, there are low levels before stimulation and the transcription then shoots up dramatically.

At the enhancers, we find a rather surprising and striking pattern of transcription. There are short transcripts at low levels that go in both directions from the center where CBP and pol2 are bound.

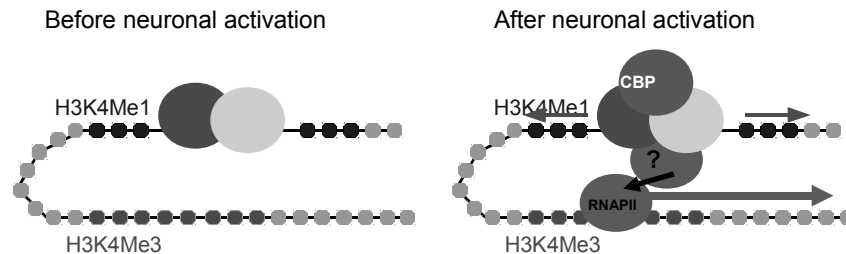
Transcription at enhancers is activity-dependent



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.

[~4 min]

Enhancer RNAs (eRNAs) novel species



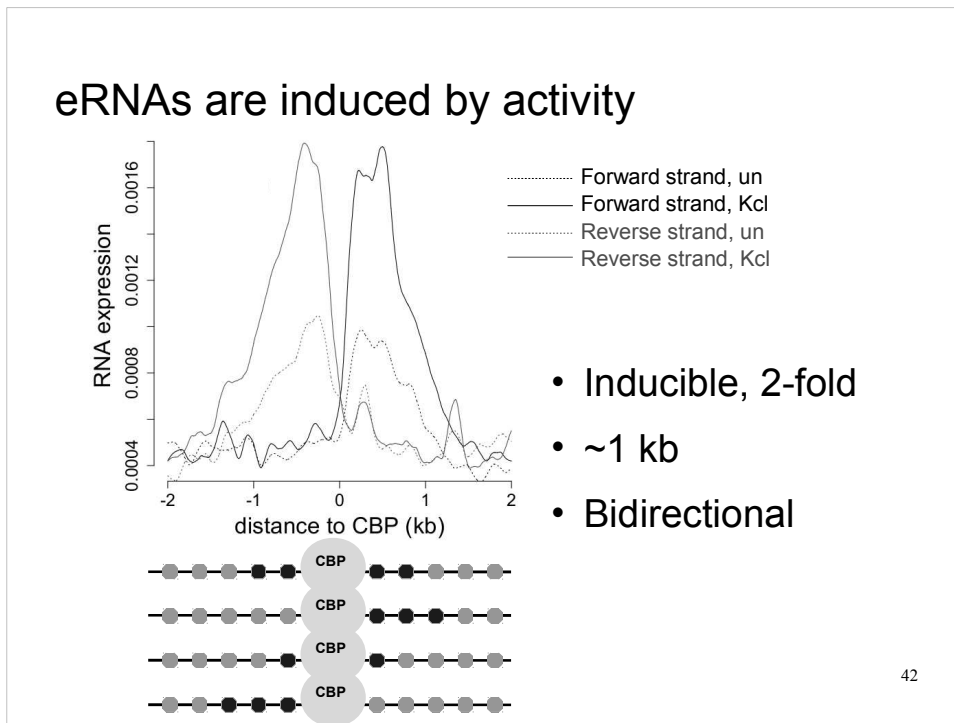
- mRNA, rRNA, tRNA, miRNA, snRNA, snoRNA, siRNA, piRNA, lncRNA, ... ?

41

Transcription of genes into mRNA is the most common type of RNA and there are around 10 more types of RNA that have been well characterized in the literature

However, this type of pattern had not been reported previously, so we were excited to have discovered a novel species 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.



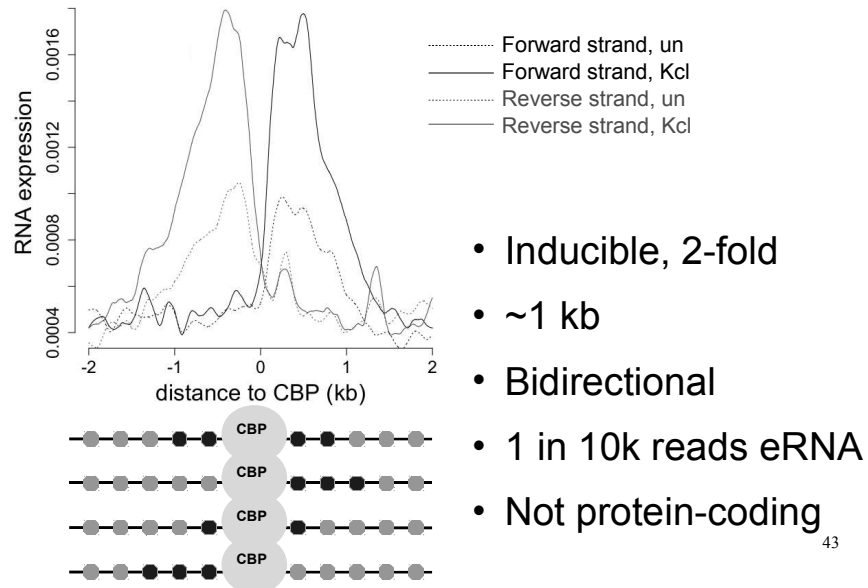
As you can see here, the pattern that we saw at the fos enhancers holds up across the genome.

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 kilobase.

Furthermore, they are induced by activity with levels almost doubling on average.

eRNAs are 100-fold lower than mRNAs

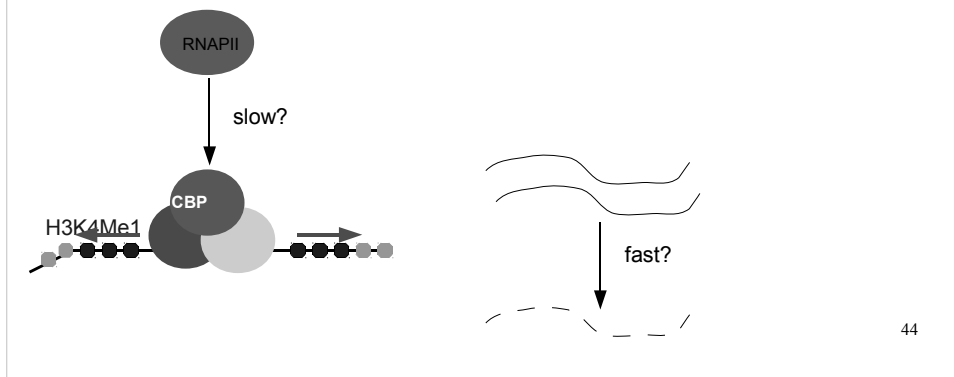


eRNAs are very lowly expressed, about 100-fold lower than typical genes. And we find that only 1 out every 10k reads appears to be an eRNA read.

[~9 min, 21]

Why do eRNAs have such low abundance?

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



There are two possibilities: either they are produced at a very low rate or they are degraded much faster than mRNAs.

A model of mRNA production and decay

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

$$\frac{dM}{dt} = \underbrace{\frac{P_M k}{L_M}}_{\text{production}} - \underbrace{\frac{M}{\tau_M}}_{\text{decay}}$$

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 the difference between a production and a decay term.

A model of mRNA production and decay

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

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

Diagram illustrating the equation above with labels and arrows:

- P_M : RNAPII
- k : Elongation rate
- L_M : Length of transcript
- τ_M : half-life
- M : mRNA

46

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.

A model of eRNA production and decay

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

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

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

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

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

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

$$\tau_E \sim 10^{-2} \times 1 \times 2 \times \tau_M \sim 2 \times 10^{-2} \times 600\text{min} = 12\text{min}$$

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 mature mRNA is about 1 kb, so this ratio is easy to estimate.

Finally, the amount of polymerase typically differs by roughly a factor of 2.

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

Estimate consistent with experiments

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

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

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

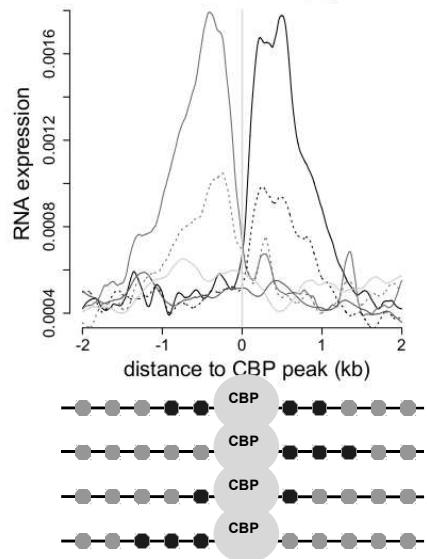
$$\frac{\tau_E}{\tau_M} = \frac{E^*}{M^*} \frac{L_E}{L_M} \frac{P_M}{P_E}$$

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.

$$\tau_E \sim 10^{-2} \times 1 \times 2 \times \tau_M \sim 2 \times 10^{-2} \times 600 \text{min} = 12 \text{min}$$

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

eRNA levels as a function of distance from center of enhancer

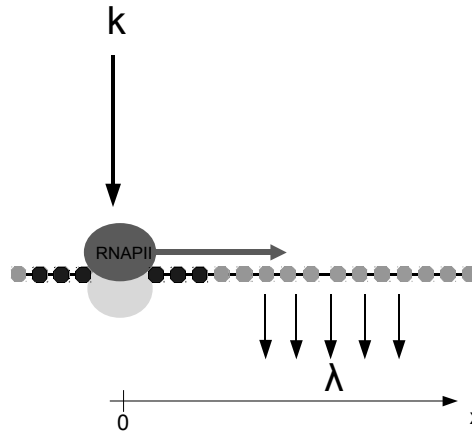


51

Next, we asked if it was possible to understand the pattern of eRNA levels using a simple mechanistic model.

RNAPII binds and falls off at a constant rate

$$\frac{dP}{dx} = k - \lambda P$$



52

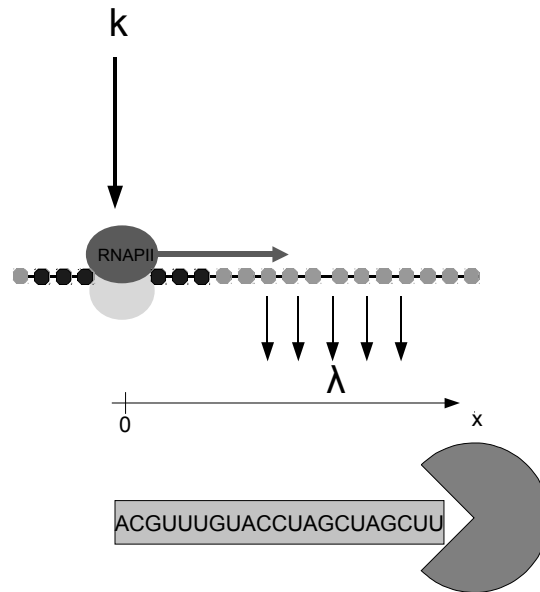
We consider the following model and what I want you to pay attention to here is the fact that I have changed variable so that I am trying to solve for the level of polymerase and eRNA as a function of the genomic position x instead of time.

For the polymerase, we assume that there is a constant rate of binding k at the center of the enhancer and a constant rate, λ , at which the polymerase will fall off the DNA

eRNA production proportional to RNAPII

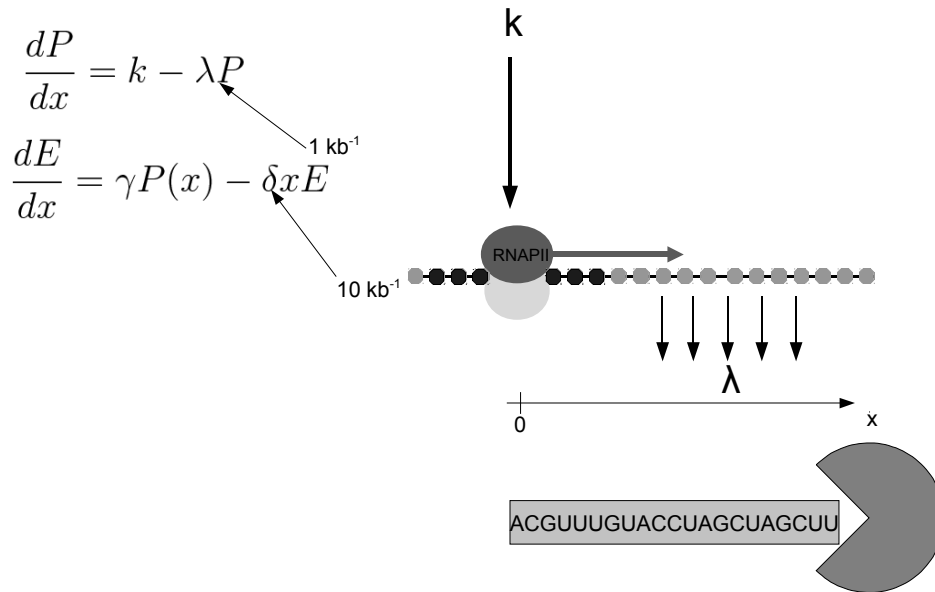
$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



In the second ODE, the synthesis rate of RNAs is proportional to the amount of polymerase. What you should pay attention to here is the x in the decay term. The reason why we have that is because of the way RNAs are degraded. The degradation process starts from the end of the transcript and it is gradually chewed up.

Parameters can be estimated from literature



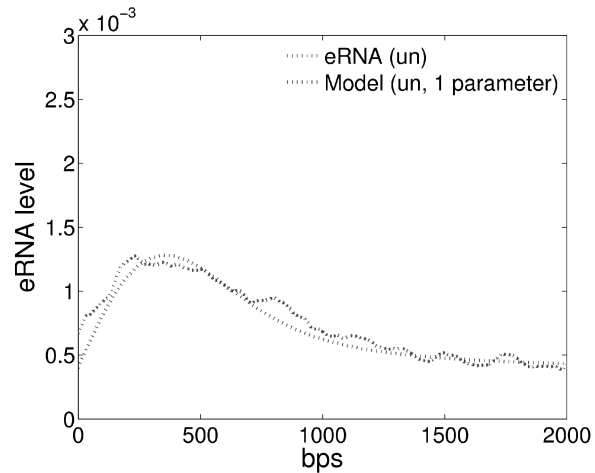
We can get estimates for two of the parameters from the literature.

The parameter k can be fit from the polymerase data and the transcription efficiency can be fit from the eRNA data.

eRNA levels can be accurately predicted

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$

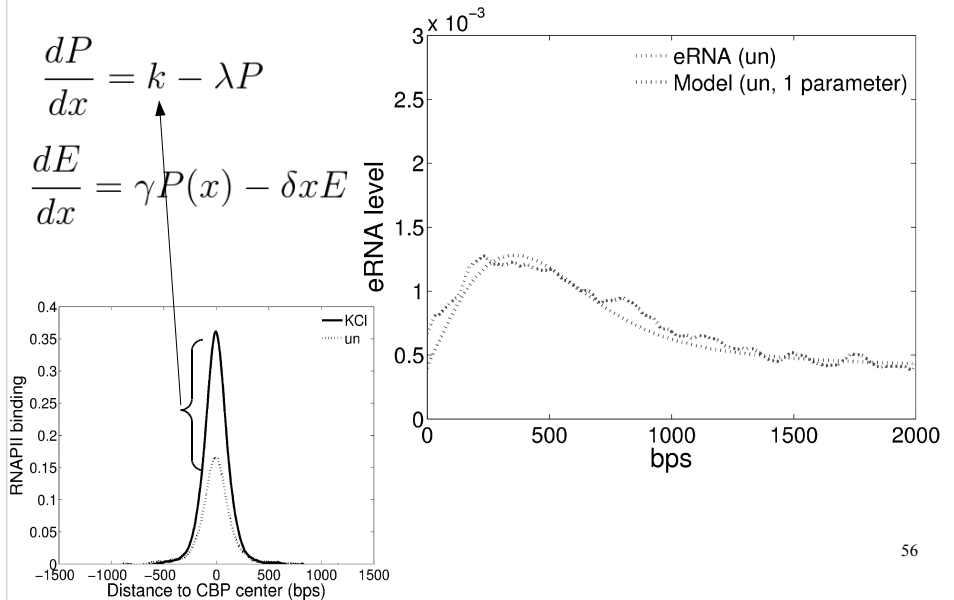


55

The fits that one obtains are very good. Here the blue line is the model where just a single parameter was fitted and the red is the data before stimulation.

Because of the symmetry, I have only plotted for one side of the enhancer.

Binding rate of RNAPII doubled after KCl

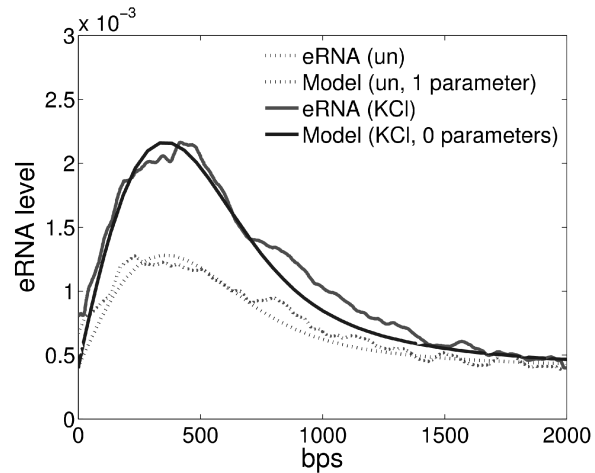


As you may recall, the polymerase levels were doubled in response to KCl stimulation, so we interpret this as doubling the parameter k in our model

No free parameters for eRNA after KCI

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



57

The fit between model in blue and the data in red is again very good and I want to emphasize here that I have not fit a single parameter to the data for the full blue line.

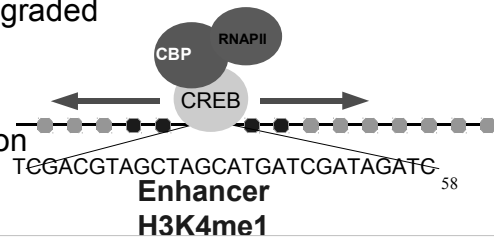
These results demonstrate that the observed patterns of transcripts at enhancers can be fully understood using a simple model of transcription and RNA decay.

Properties of activity-dependent enhancers

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII
- RNAPII synthesizes eRNAs

– eRNAs are rapidly degraded

– eRNA levels
described by
model of transcription



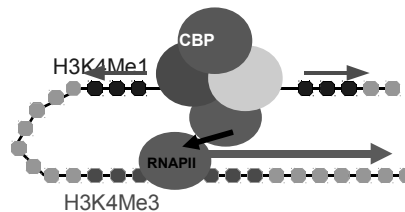
Taken together, these results suggest that the binding of RNAPII at enhancers is independent of the promoter. But that there is a mechanistic connection that is required for the synthesis of eRNAs.

Before I move on from the topic of enhancers, I would like to present a few models concerning the function of RNAPII and transcription at enhancers.

[~7 min, 28]

What is the function of RNAPII at enhancers?

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter



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

-George Bernard Shaw⁵⁹

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. The discovery of widespread binding of RNAPII and transcription at enhancers raises several important questions.

The first question is: why have Pol2 at enhancers? The most obvious answer 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.

Recruitment of RNAPII at the promoter

- 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
tau – RNAPII half life

60

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 *tau* is the half-life for the dwelling time of the polymerase at the promoter.

Difficult to estimate parameters

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

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

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

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

61

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

Unfortunately, it is quite difficult to estimate the parameters here, both from our data and the literature.

[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.]

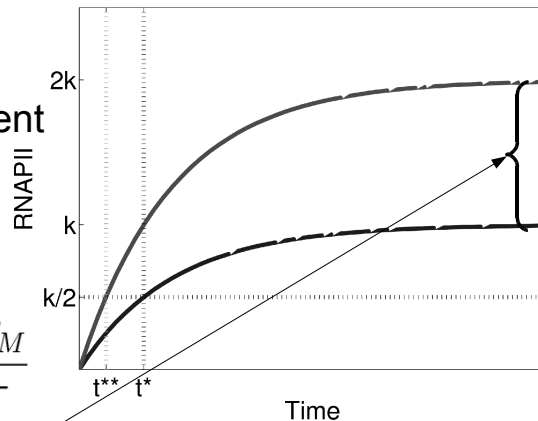
However, we may still draw some interesting qualitative conclusions using this model.

Steady state level of RNAPII is increased

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

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

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



62

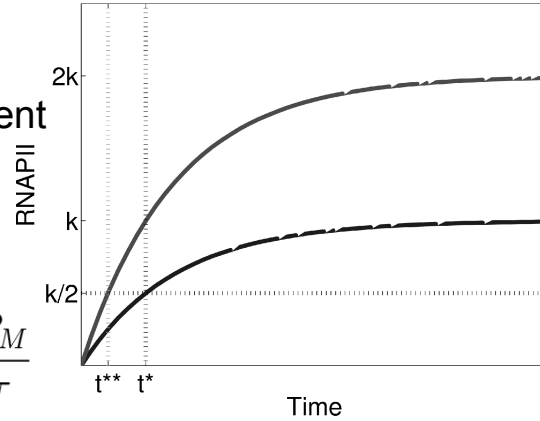
The most obvious benefit of having multiple recruitment points for Pol2 is that the maximum level at the promoter is increased by this number which is proportional to the number of enhancers. Here the blue curve is for a promoter without enhancers and the green is for one with enhancers.

[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.]

Rise time is reduced

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

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



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

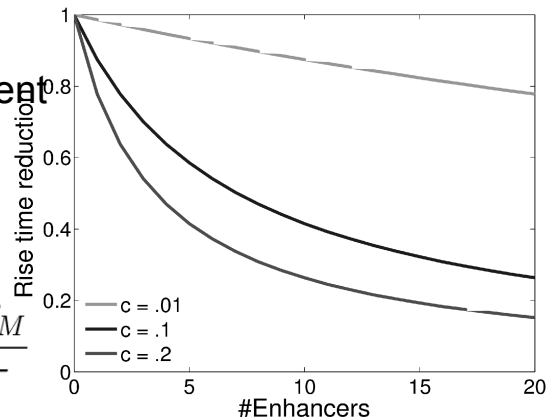
63

Another advantage that can be observed from this plot is the reduced rise times. If we assume that a certain critical level of pol2 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

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



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

64

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.

Recruitment of RNAPII is diffusion limited

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter



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

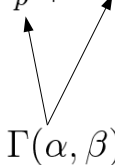
$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$

65

Inside the cell, the binding rate of any molecule is diffusion-limited, meaning that there is an upper bound to the parameter k_p here. It has been shown by Vilar and Leibler in their model of the lac operon that diffusion limited binding is important for the recruitment of transcription factors. 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.

Enhancers may reduce the noise in RNAPII

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

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


$\Gamma(\alpha, \beta)$

66

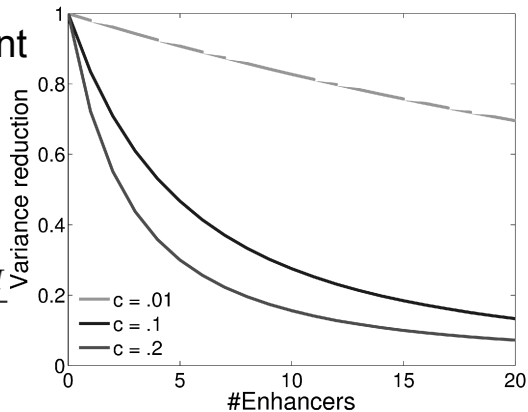
Finally, it can also be shown that by distributing the RNAPII recruitment, then we may reduce the noise in the system.

If we assume that the rate at which RNAPII binds to the DNA is governed by a diffusion process, then it can be shown that these rates will be Gamma-distributed and that the variance of the rate scales linearly with the mean.

RNAPII noise reduction proportional to number of enhancers

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

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



67

The intuition here is now that by having multiple independent recruitment points, each with a lower k_p , then the noise at each of the enhancers will be lower than for a single promoter with the same rate. Since the variance grows as n square when we increase the mean, it is clear that the variance in the system can be lowered by having n weak binding sites instead of a single strong one.

We may thus obtain the same Pol2 recruitment rate, but with a lower noise level. As this plot shows, only a few enhancers may reduce the noise levels significantly.

What is the function of eRNAs?

- What is the function of RNAPII at enhancers?
 - Increase rate of RNAPII recruitment
 - Possibly faster than diffusion limit
 - Faster rise-time
 - Reduced noise
- What is the function of eRNAs?
 - Noise
 - Transcription establishes histone modifications
 - Transcript has function

68

At this point we don't know what the function of the eRNAs is. Nevertheless, it is clear that the polymerase at the enhancers is useful even if it is the case that eRNAs are just transcriptional noise and that the transcripts themselves have no function.

At the other end of the spectrum, there is of course the possibility that the transcripts really do have an important function 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

The intermediate view is that it is the act of transcription rather than the transcripts that are important. Experiments in yeast have shown that the methylation of histones can take place as part of transcription.

[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.]

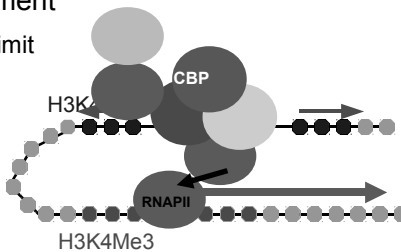
Enzymes piggyback on the polymerase

- What is the function of RNAPII at enhancers?

- Increase rate of RNAPII recruitment
 - Possibly faster than diffusion limit
- Faster rise-time
- Reduced noise

- What is the function of eRNAs?

- Noise
- Transcription establishes histone modifications
- Transcript has function



69

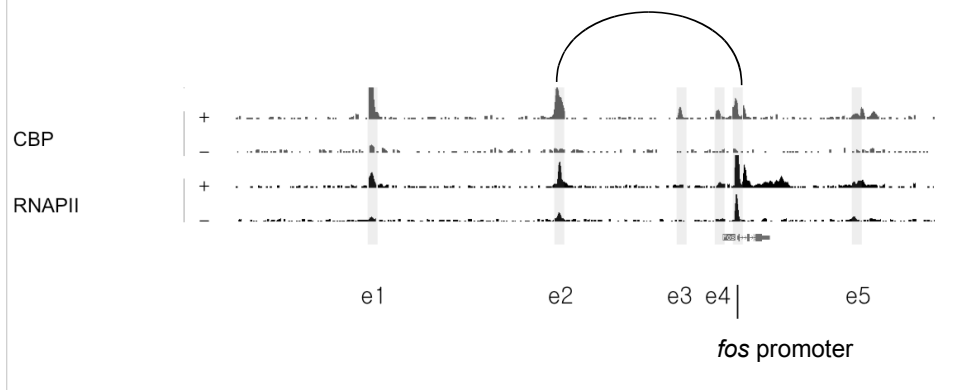
At this point we don't know what the function of the eRNAs is. Nevertheless, it is clear that the polymerase at the enhancers is useful even if it is the case that eRNAs are just transcriptional noise and that the transcripts themselves have no function.

At the other end of the spectrum, there is of course the possibility that the transcripts really do have an important function 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

The intermediate view is that it is the act of transcription rather than the transcripts that are important. Experiments in yeast have shown that the methylation of histones can take place as part of transcription.

[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.]

Pair each enhancer with nearest promoter
and compare RNAPII and RNA

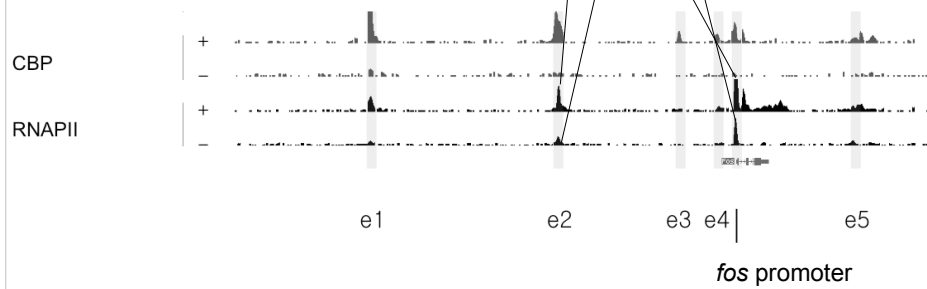


To further investigate the role of eRNAs, we investigated the relationship between eRNAs and the more well-studied mRNAs.

We did this by pairing each enhancer with its nearest promoter. For each pair we calculated the level of RNAPII binding and the level of transcription before and after stimulation.

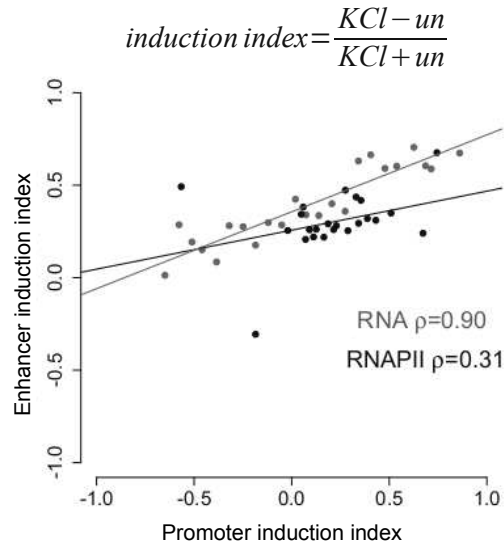
Calculate induction index for both RNAPII and transcription

$$\text{induction index} = \frac{KCl - un}{KCl + un}$$



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.

eRNA induction is correlated with induction of nearby mRNAs but not RNAPII



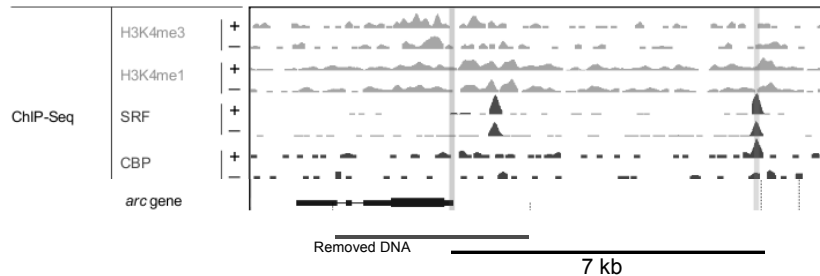
72

We then plotted the induction indexes for pairs of enhancers and promoters.

As you can see from the black dots, there is only a modest correlation for the polymerase levels. Whereas the transcript levels in red appear to be strongly correlated.

Thus, the prediction here is that the Pol2 recruitment at the enhancer is independent of the promoter while the transcription is not.

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



- RNAPII same in knock-out +/- KCl
- eRNAs not present in knock-out

73

We investigated this correlation further using a knock-out experiment. 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, both before and after stimulation.

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.

Summary

- Identified ~12k activity-dependent enhancers
- Discovered and quantified novel mechanisms
 - Identified enriched motifs and bound TFs

- Combinatorial code for CBP affinity
- Recruitment of RNAPII at enhancers
 - Faster recruitment to promoter
 - Reduce noise
- Transcription at enhancers
 - Properties of eRNA
 - Model of RNAPII and eRNA levels
 - Interaction with promoter necessary

74

To sum up: I have told you about gene regulation and in particular about distal enhancers which are difficult to identify and not much is known about how they affect gene expression. 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. This hints at additional gaps in our understanding of enhancers, but from an operational point of view it allows us to distinguish active from inactive enhancers.

eRNAs have been found in other cell types

doi:10.1038/nature09033

nature

ARTICLES

Widespread transcription at neuronal activity-regulated enhancers

Tae-Kyung Kim^{1,2*}, Martin Hemberg^{2,3*}, Jesse M. Gray^{4*}, Allen M. Costa¹, Daniel M. Bear¹, Jing Wu¹, David A. Harmin^{1,4}, Mike Laptewicz¹, Kellie Barbara-Haley¹, Scott Kuersten⁵, Eirene Markenscoff-Papadimitriou^{1,4}, 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^{1,1}, Albert W. Cheng^{1,2,3}, G. Grant Welstead⁴, Tristan Kooistra^{4,5}, Bryce W. Carey^{1,6}, Eveline J. Steine^{1,6}, Jacob Hanna^{1,6}, Michael A. Lodato^{1,6}, Garrett M. Frampton^{1,6}, Phillip A. Sharp^{1,6}, Laurie A. Boyer^{1,6}, Richard A. Young^{1,6}, 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^{1,2}, Iros Barozzi^{1,2}, Flore Mietton^{1,2}, 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

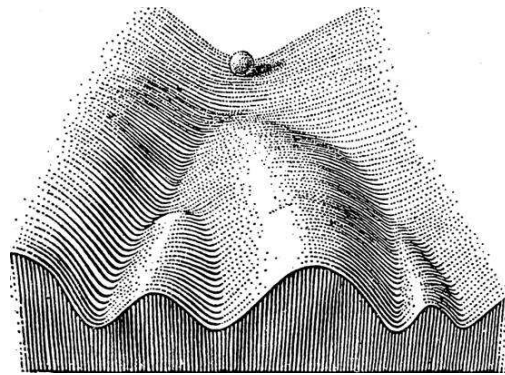
Alvaro Rada-Iglesias¹, Ruchi Bajpai¹, Tomek Swigut¹, Samantha A. Brugmann¹, Ryan A. Flynn¹ & Joanna Wysocka^{1,2}

75

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 it seems extremely likely that eRNAs are a generic feature of enhancers.

Stochastic models of gene expression

- Transitions between stable states



Waddington, 1953 ⁷⁶

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.

[~9 min, 37 min]

Master Equation (**ME**) description

- Discreteness required since ~ 10 mRNAs/cell
- Use Markov Chain Monte Carlo (MCMC)
 - Gillespie's Stochastic Simulation Algorithm (**SSA**)

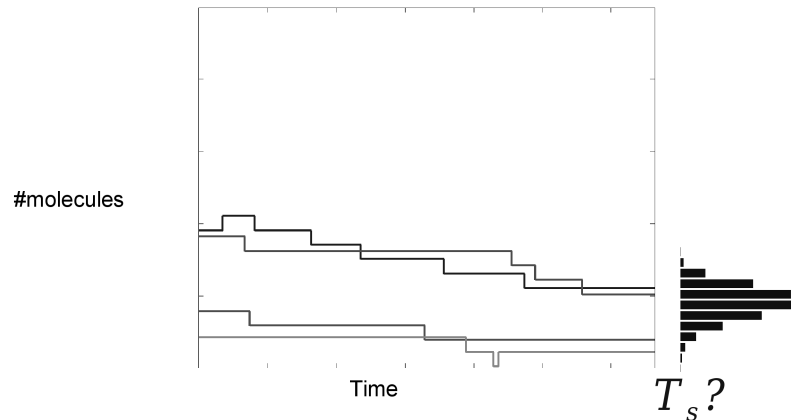
77

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 per gene and cell, 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 .

How long do we need to run MCMC?

- SSA simulates trajectories of system
 - Run repeatedly to estimate probability distribution



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 per gene and cell, 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 .

How long do we need to run MCMC?

- SSA simulates trajectories of system
 - Run repeatedly to estimate probability distribution
- 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}

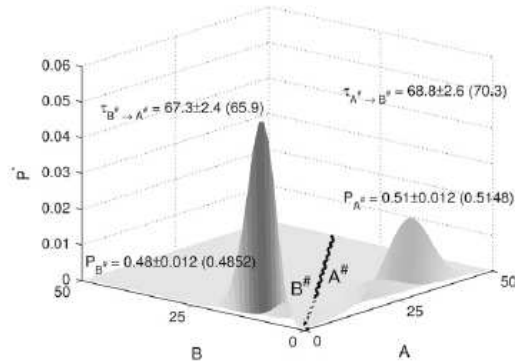
Address: ¹Department of Bioengineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK and ²Institute for Mathematical Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

79

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 per gene and cell, 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 .

Perfect sampling of transitions between steady 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

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.

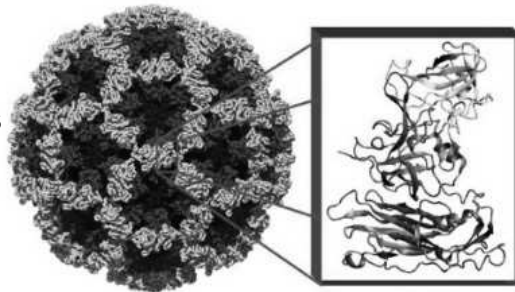
The SSA allows us to simulate trajectories and thereby sample transitions from one stable state to another.

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 this possible 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.

[~4 min]

Assembly of viral capsids

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



Biophysical Journal Volume 90 May 2006 3029–304.

Stochastic Kinetics of Viral Capsid Assembly Based on Detailed Protein Structures

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

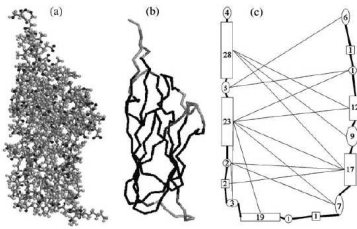
*Department of Bioengineering and [†]Department of Chemistry, Imperial College London, London, United Kingdom

I would also like to tell you briefly about another project where I combined the stochastic 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.

Coarse-grained protein model

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



82

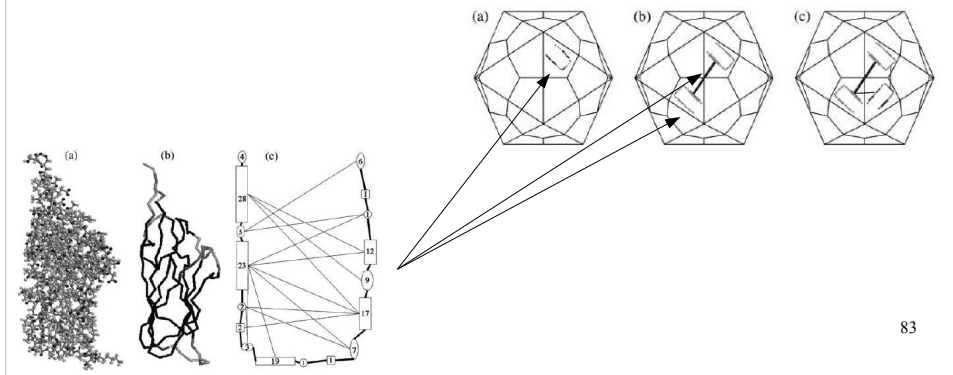
In our model, we took a bottom up approach, starting from the full atomic description as provided by the crystal structure which is shown to the left here.

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

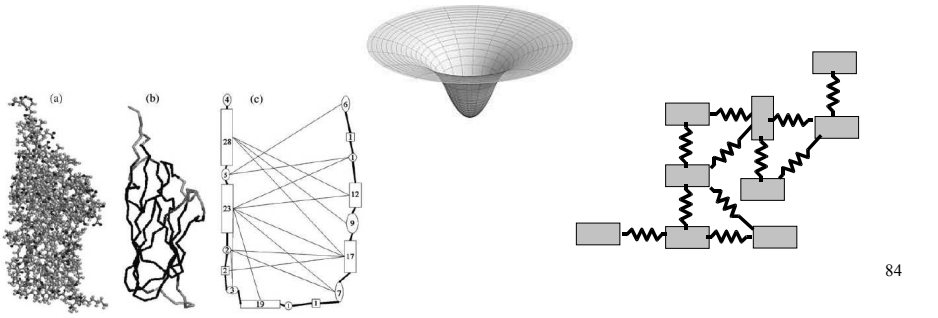


83

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 use 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

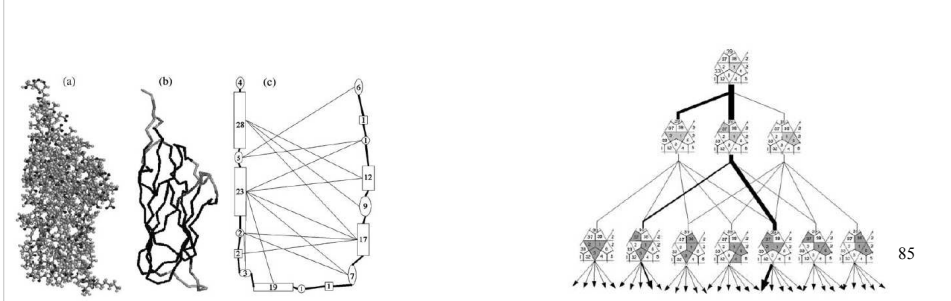


84

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 compute the rate of dissociation.

All reactions cannot be enumerated

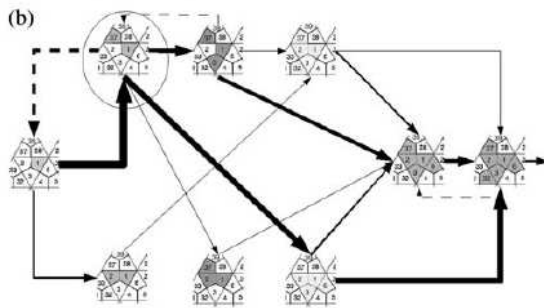
- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
 - Association restricted by diffusion
 - Dissociation escape from multi-dimensional well



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 grows 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



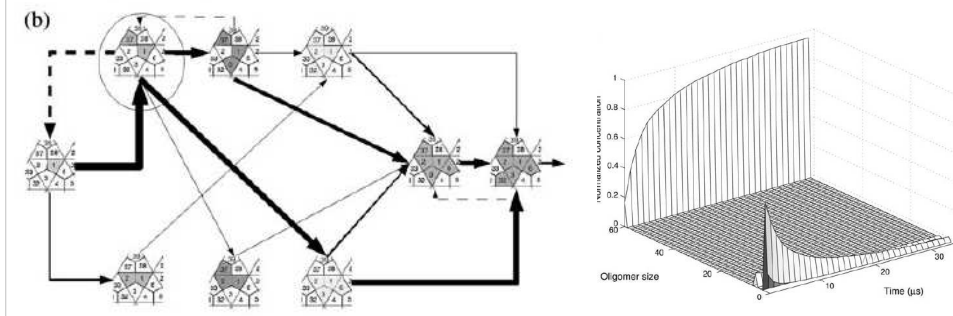
86

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.

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.

Identify stable intermediaries

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

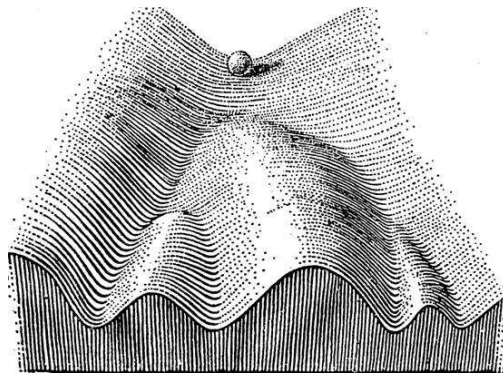


In this figure here on the right, I have compiled the information from one of these graphs. On this axis we have time, and oligomer size along this one and concentration here. As you can see, the system starts out with all monomers and these are rapidly combined to form dimers and hexamers. The hexamers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

[~4min]

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 ⁸⁸

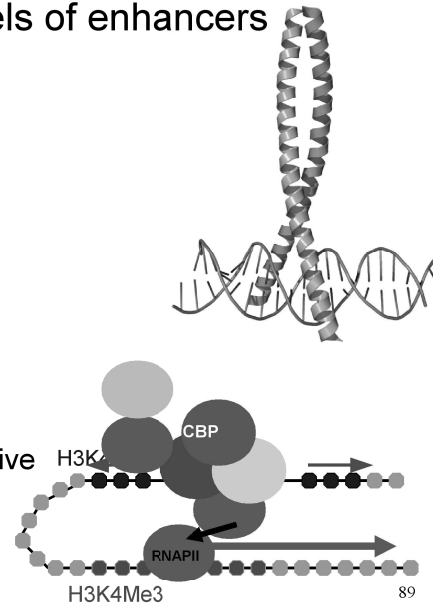
Finally, I'd like to tell you a little about my plans for future research.

To summarize as briefly as possible: My aim is to bridge the gap between our understanding of the basic laws of physics and chemistry which explain what is going on at the molecular level and Darwinian evolution which 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.

Develop biophysical models of enhancers

- TF-DNA binding
 - X-ray structures
 - ChIP-Seq binding
- DNA looping
 - Histones



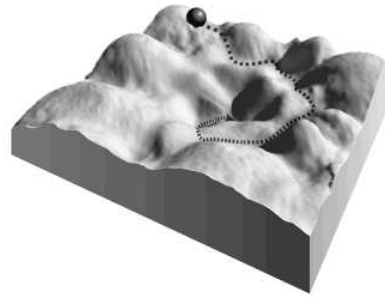
One example of such an approach would be to further our understanding of TF binding. What I'd like to do is to start from atomistic models of TF-DNA binding, similar to what I did in the virus project and then obtain not just one, but several coarse-grained models. There are crystal structures for 100s of Tfs and for many of them, there is also chip-seq data from multiple cell-types available. I should remind you, that each chip-seq experiment represents 1000s of binding events. Each coarse-grained model, can then be tested using the chip-seq data and this should will allow us to determine which one is the best model

Another aspect of gene regulation that I will study is looping of DNA. There are several models on looping of naked DNA while there is also work on the interactions of DNA and histones, it is mainly concerned with how the DNA is packaged. I plan to study how looping of DNA with histones.

Finally, I am also interested in understanding the biophysical basis of the patterns of histone modifications that we observe. For example, why is it that we have tri-methylation of lysine 4 on histone 3 at active genes and not some other mark. There has to be an explanation based on the physical properties of the methyl group and its location on the histone..

Model stochastic gene expression for entire transcriptome

- Analytical models of gene expression noise
 - Parametric robustness
 - Time-scales
- Apply to genome-wide single-cell RNA-Seq
 - Propagation in pathways
 - Global factors
 - Dimensionality



The ME work that I told you about was mainly concerned with numerical simulations. However, I have some preliminary results that have shown that it is possible to make significant progress with analytical solutions of the ME. Two aspects that I am particularly interested in are the parametric sensitivity and temporal properties of the noise. Reaction rates inside the cell are likely to fluctuate dramatically under normal physiological conditions and it is not clear how the cell is able to operate robustly in such a noisy regime. We also have a poor understanding of how noise propagates over time and over what time scales a cell can be considered ergodic.

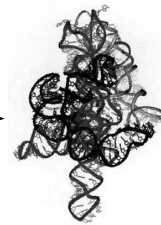
The first papers that apply RNA-Seq to single cells have already been published. I plan to extend the small-scale models of noisy gene expression that I worked on as a graduate student by taking advantage of the tools and insights for RNA-Seq data that I have developed as a post-doc.

Single-cell data will allow us to get a much better understanding of the topology and the dimensionality of the state space in which cells are operating. It will allow us to ask questions about how noise is propagated in pathways, how it varies between cell-types and to what extent it is determined by global factors.

Determine structure of RNAs

- Other species of novel non-coding RNAs
 - Identify structural motifs
- High-throughput sequencing of structure
 - PARS
 - SHAPE-Seq

.....ACGUCCAAAUUCCCUAGGCUCAAGGCAUUCGAUCGGGAUUUA..... →



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 long 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 and others at Berkeley 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 and folding of RNAs. This is important not just for understanding the function of long non-coding RNAs, but also for mRNAs since it has been shown that the translation efficiency depends on the structure of the transcript.

Acknowledgements

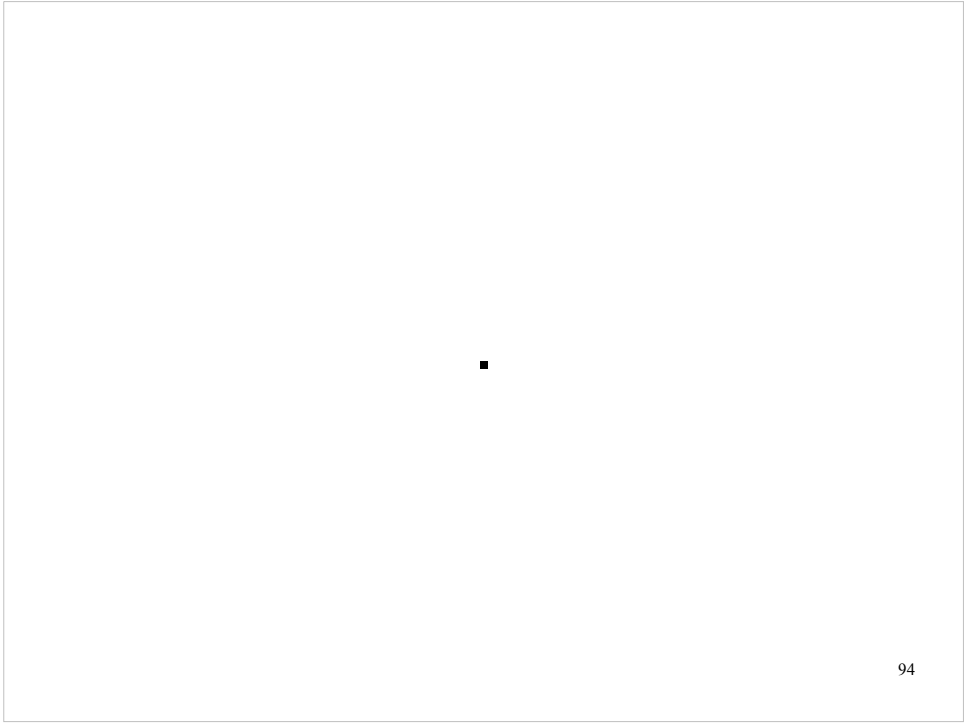
- Gabriel Kreiman
- Jesse Gray
- Tae-Kyung Kim
- Michael Greenberg
- Mauricio Barahona

Click to add title

Thank You

93

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



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

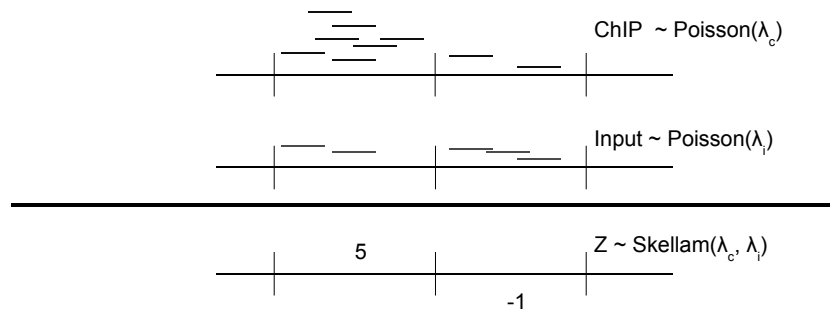
Click to add title

?

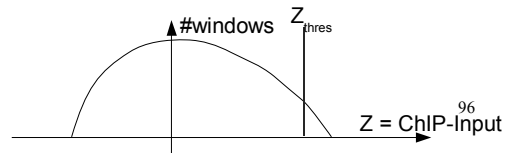
95

And I will be happy to take any questions.

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



- False Detection Rate (FDR) determine threshold



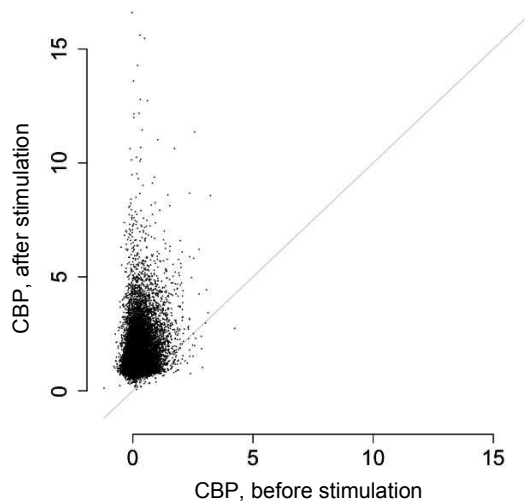
Use False Detection Ratio (FDR) to correct for multiple hypotheses

- $Z_i = \text{\#ChIP reads} - \text{\#input reads in window } i$
 - $\sim 1 \text{ read}/100 \text{ bp}$
 - 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

97

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.

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



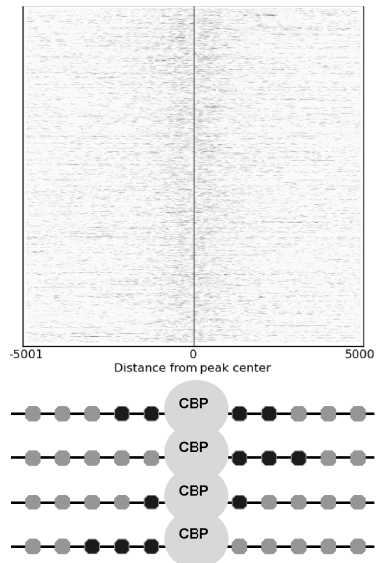
98

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 binding profiles

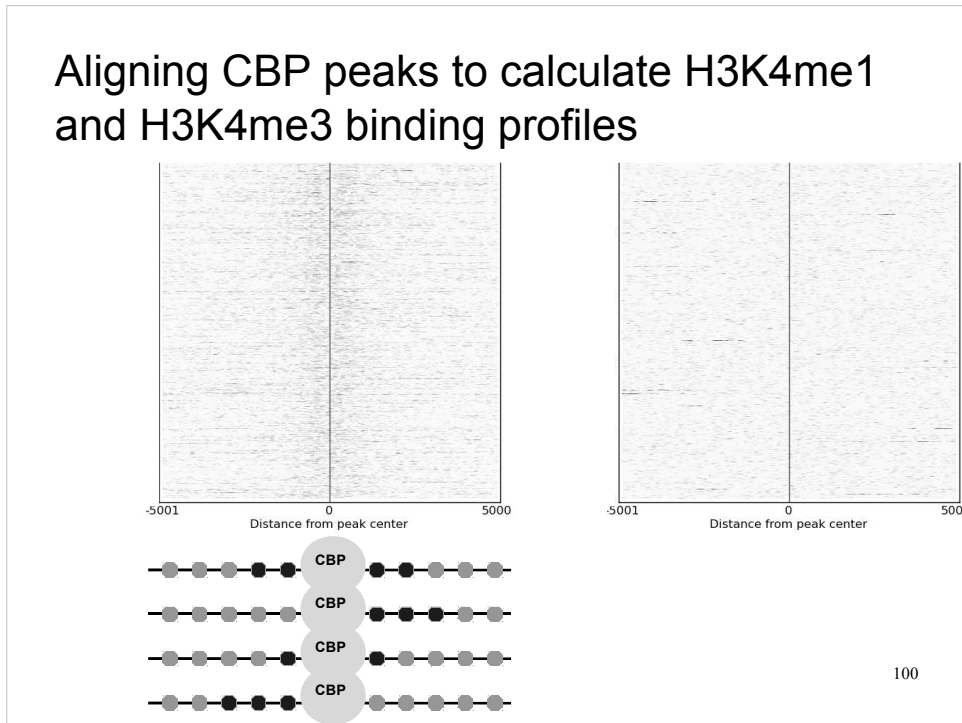


99

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.

Aligning CBP peaks to calculate H3K4me1 and H3K4me3 binding profiles

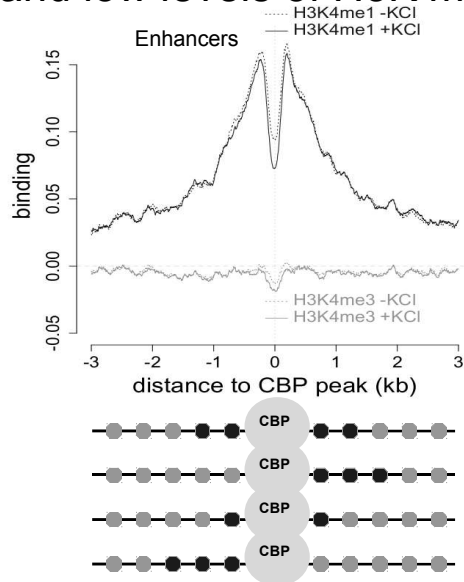


The next few slides will demonstrate why this procedure is a good one

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

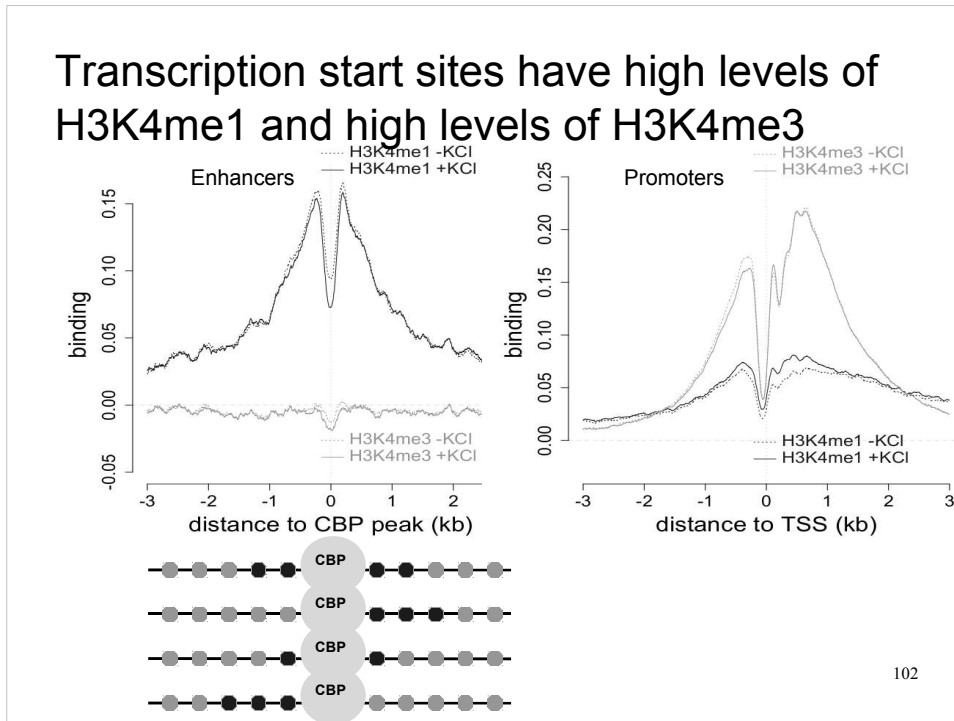
Enhancers have high levels of H3K4me1 and low levels of H3K4me3



101

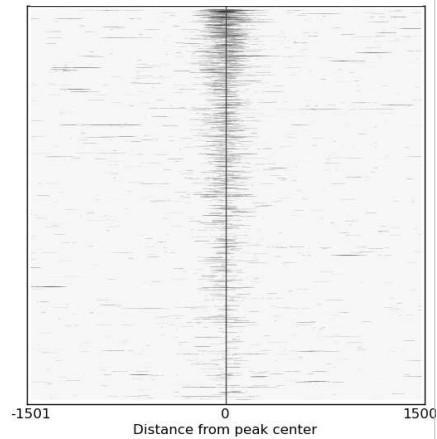
We calculated the histone as a function of distance in the same way as we did with CBP. As you can see here, there is a characteristic bimodal pattern for K4me1 in blue whereas the levels of K4me3 in green are at background.

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.

RNAPII binds at activity-dependent enhancers



103

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

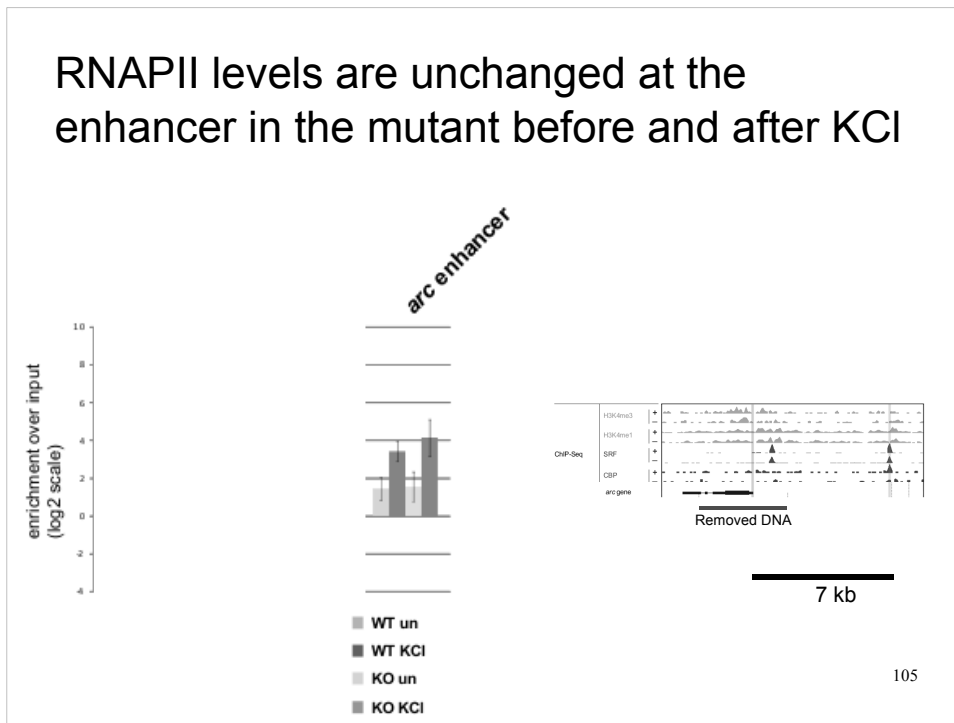
~100 enriched motifs are found

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

104

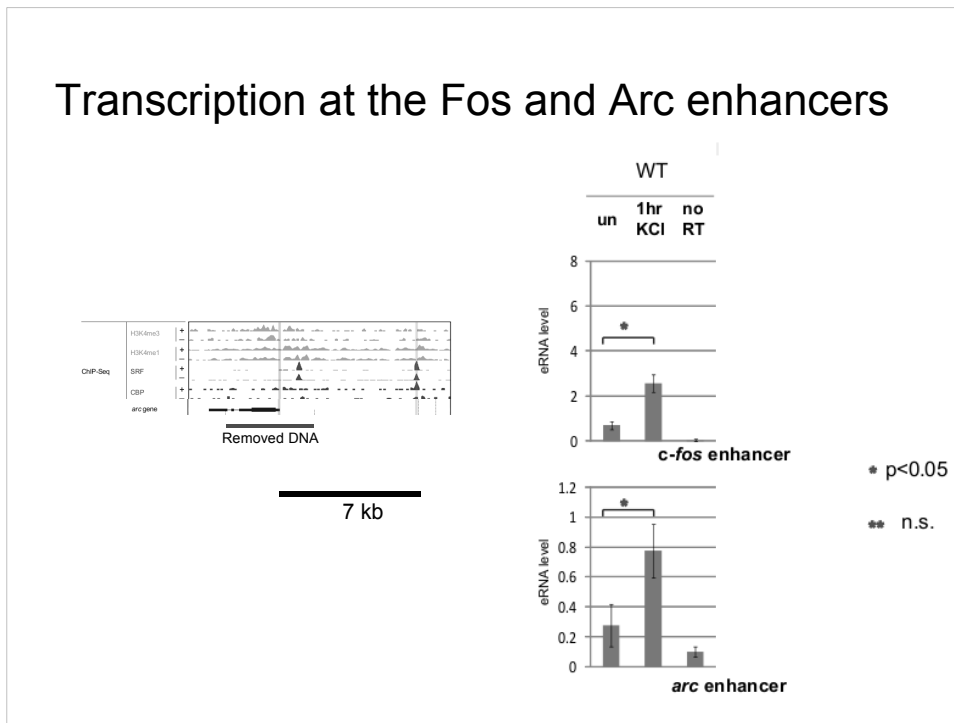
We found ~100 significantly enriched motifs in our enhancer set. Some of them, such as CREB, SRF and Npas4 corresponded to Tfs that had been identified previously, while others are similarly enriched but do not correspond to factors with known motifs.

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



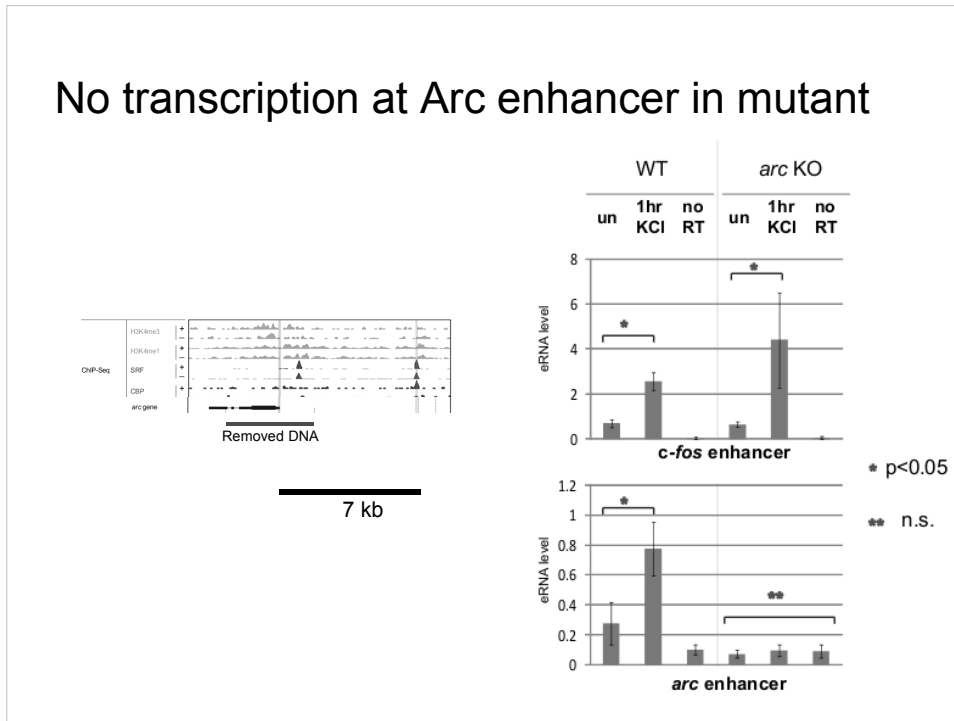
As you can see here, the levels of RNAPII at the enhancer remained the same in both the knock-out and the wild type, both before and after KCl stimulation.

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.

Estimating the production rate of eRNAs

$$\frac{dE}{dt} = kN - \frac{E}{\tau_E}$$

$$k = \frac{E^*}{N\tau_E} \sim \frac{10^3}{10^4 \times 10^{-1}\text{h}} = 1\text{h}^{-1}$$

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

Parameters for the eRNA fit

$$\lambda = \frac{k_{drop} \text{ s}^{-1}}{k_{elong} \text{ bp}^{-1} \text{ s}^{-1}} \sim \frac{2 \times 10^{-2}}{20} \text{ bp}^{-1} = 10^{-3} \text{ bp}^{-1}$$

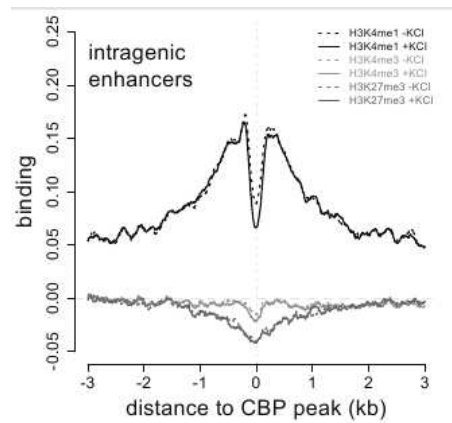
$$\tau_{decay} = \tau_{find} + \tau_{bp} L$$

$$H(x, t) = \frac{k\kappa}{\mu_x(\mu_x - \lambda)} (e^{-\lambda x} - e^{-\mu_x x}) \times e^{-\mu t}$$

$$E(x) = \sqrt{\frac{\pi}{2\lambda}} \frac{\gamma k}{\lambda} e^{-\delta^2/2\lambda - \lambda x^2/2} \left[\text{erf}\left(\frac{\delta i - \lambda i x}{\sqrt{2\pi}}\right) - \text{erf}\left(\frac{\delta i}{\sqrt{2\lambda}}\right) \right]$$

Intragenic enhancers

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

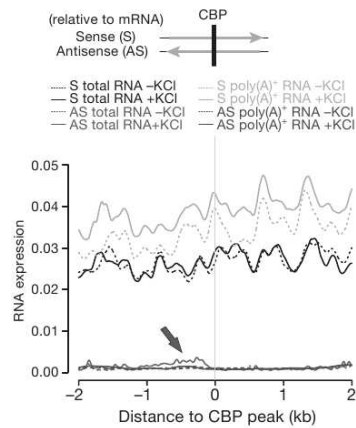


110

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



111

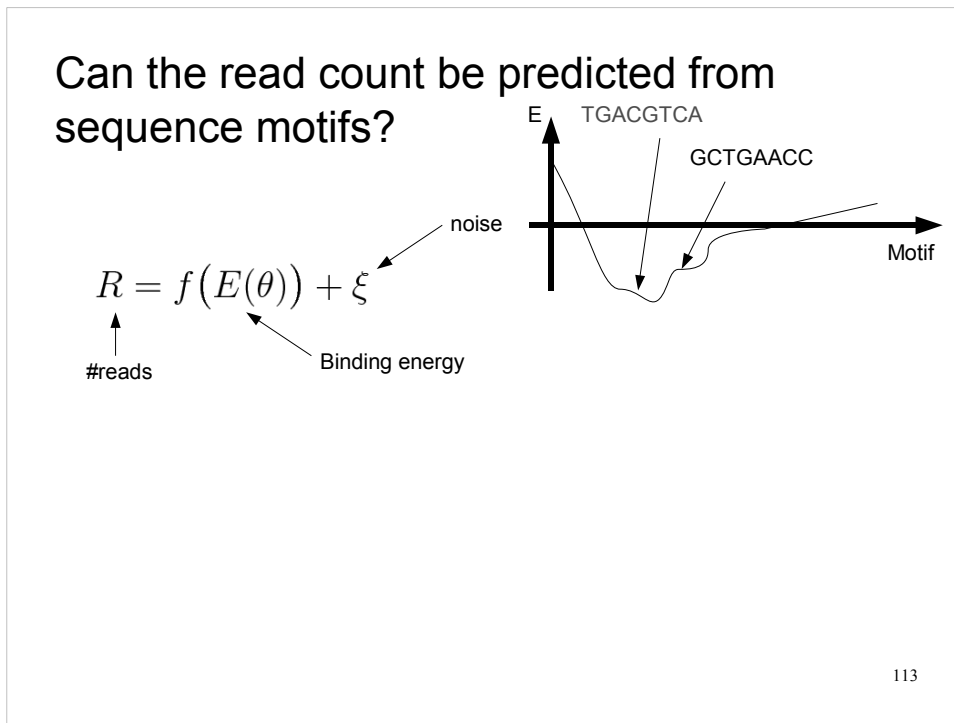
However, because of the overlapping mRNAs, we cannot expect to detect a signal on the sense strand. On the anti-sense strand, on the other hand, we do find the same characteristic signal as for the extragenic enhancers.

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

112

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.



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 there is a binding energy associated with each motif for the TF. This energy function may have parameters θ and in our case we used the position weight matrix energy model with experimentally determined parameters from the database JASPAR.

We assume that the number of reads that were observed is some unknown function f of the binding energy. 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.

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

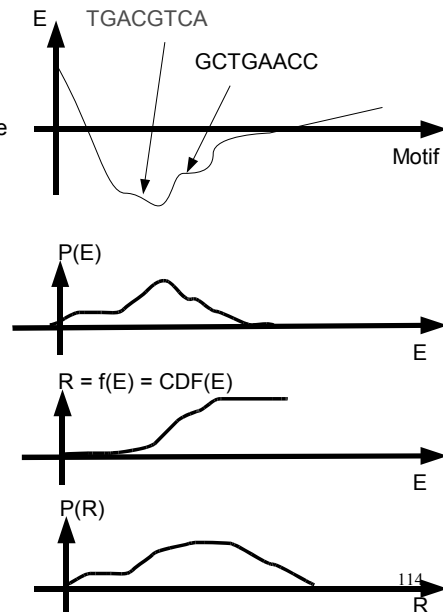
Assume transfer function maximizes mutual information

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

#reads

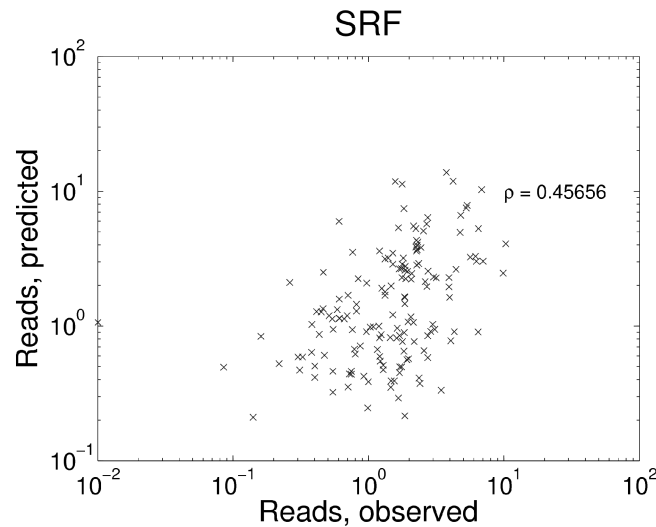
Binding energy
(from JASPAR)

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



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 using a variational approach 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. As is shown here for the enhancers where we also found SRF peaks, the model does a reasonable job at predicting the number of reads when a peak was called.

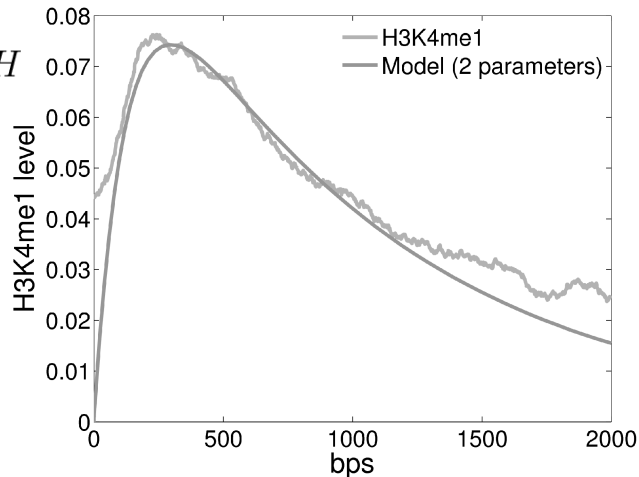
[Since the model does not take the competition with other factors or histones into consideration, it is only applicable on the condition that binding was observed, that is a peak was found in the CHIP-Seq.]

[~5 min]

Establishing H3K4me1 levels at enhancers

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dH}{dx} = \kappa P(x) - \mu H$$

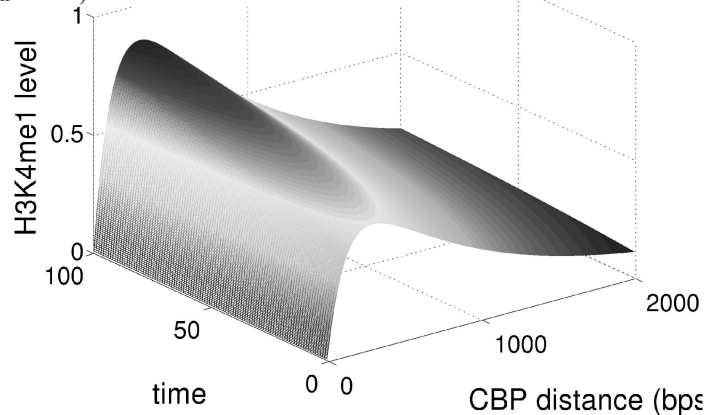


It is difficult to find values for the parameters, in the second equation, but taking the same lambda as before, we only need to fix two parameters to get a very good fit, suggesting that it is reasonable to assume that the histone marks are indeed a function of the polymerase.

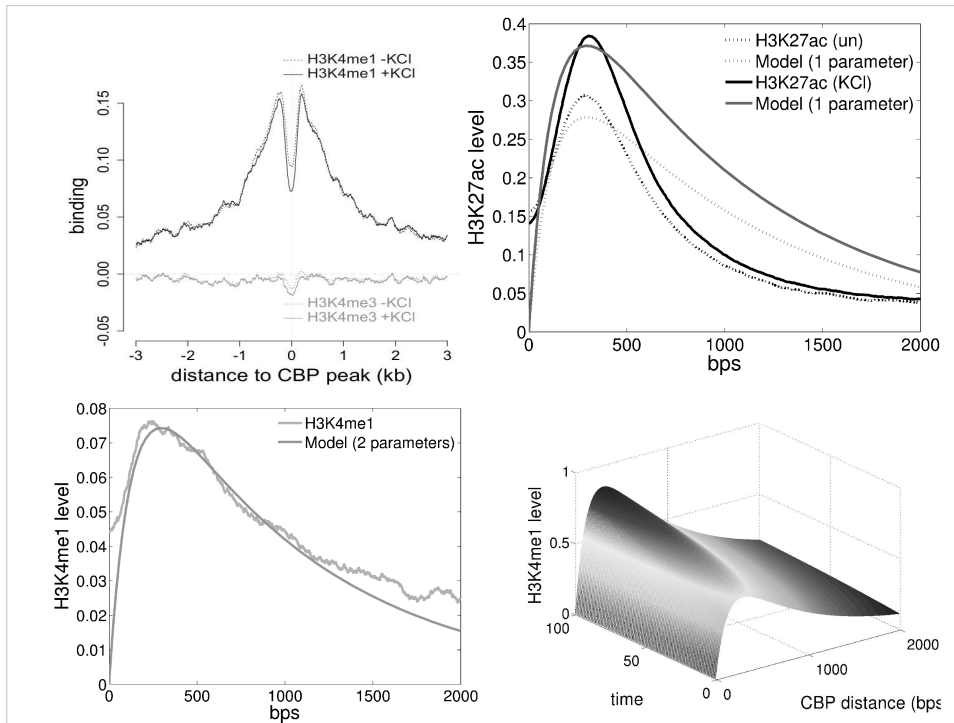
A PDE for histone levels

$$\frac{\partial H}{\partial x} + \frac{\partial H}{\partial t} = \kappa P(x, t) - \mu_x H - \mu_t H$$

$$H(x, t) = \frac{k\kappa}{\mu_x(\mu_x - \lambda)} (e^{-\lambda x} - e^{-\mu_x x}) \times e^{-\mu_t t}$$



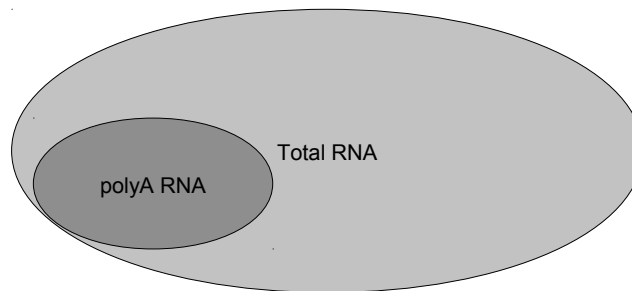
It is difficult to find values for the parameters, in the second equation, but taking the same lambda as before, we only need to fix two parameters to get a very good fit, suggesting that it is reasonable to assume that the histone marks are indeed a function of the polymerase.



It is difficult to find values for the parameters, in the second equation, but taking the same lambda as before, we only need to fix two parameters to get a very good fit, suggesting that it is reasonable to assume that the histone marks are indeed a function of the polymerase.

polyA tail is added to messenger RNAs (mRNAs)

ACGUUUGUACCUAGCUAGCUUACGAG AAAAAAAAAAAAAAAAAAAAAA

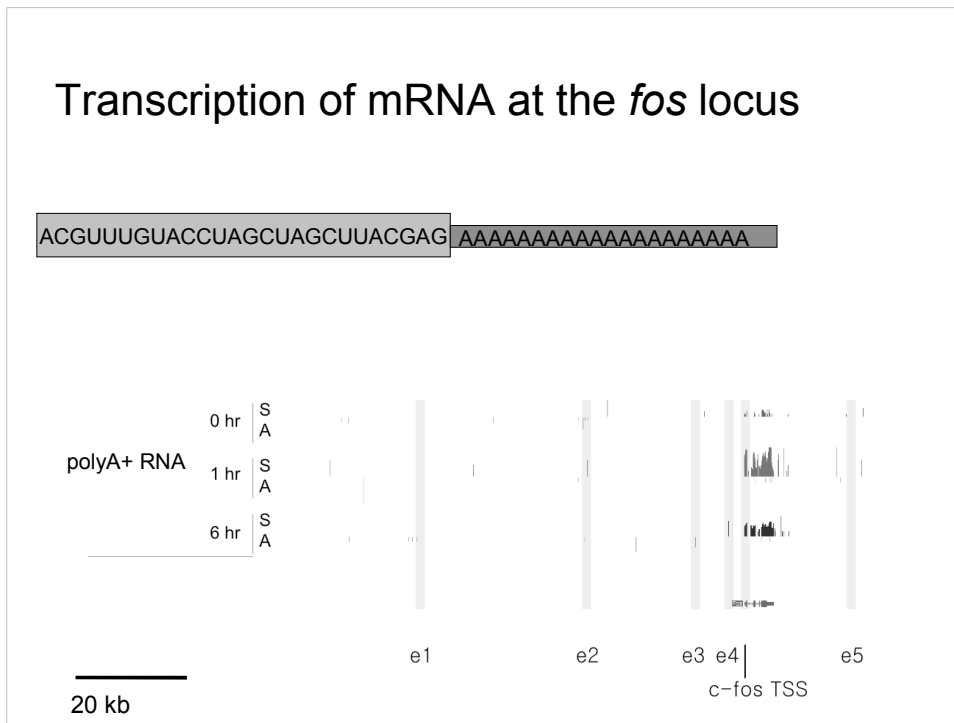


119

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.

We sequenced both the polyA fraction of the RNA as well as the entire pool of RNA in separate experiments.

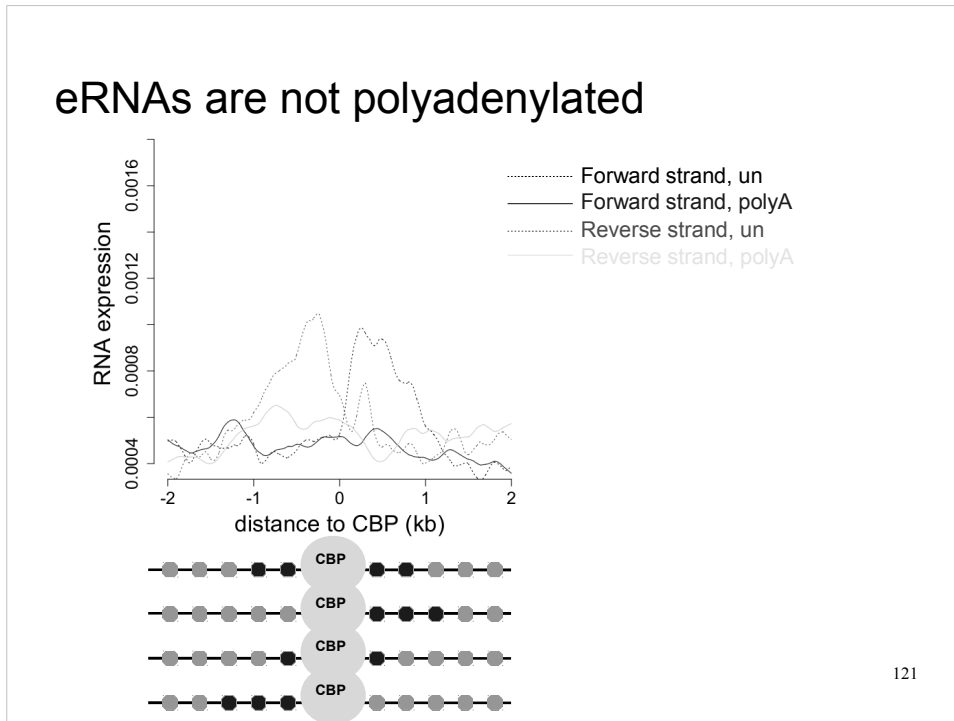
Transcription of mRNA at the *fos* locus



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.

eRNAs are not polyadenylated

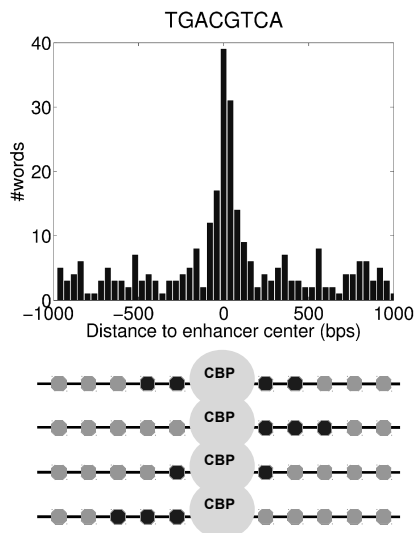


121

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.

~100 enriched motifs found at enhancers



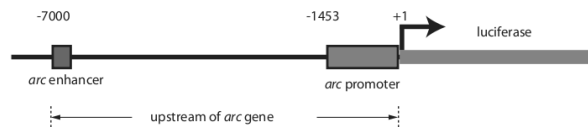
122

Here is a profile for the well-studied motif for the Creb protein. Creb has been shown to be an important transcription factor for 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. A few of these were known from before to be important, others corresponded to known transcription factors that had not been previously associated with activity-dependent gene expression and finally there were some motifs for which we do not even know which transcription factor binds there

We identified ~12,000 activity-dependent enhancers throughout the genome

- **CBP** peak
- **High** levels of flanking **H3K4me1**
- **Low** levels of **H3K4me3**
 - 8/8 tested activity-dependent enhancers were validated using a luciferase assay



123

Using these criteria, we were left with a list of ~12k putative distal enhancers.

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 bio-luminescent protein. We found that all 8 sequences were able to enhance gene expression in an activity dependent manner.

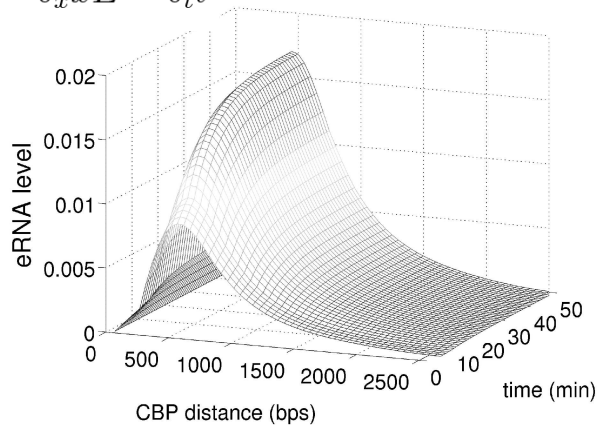
As I mentioned, it is very difficult to identify enhancers and before our study, there was only one example, the arc enhancer, of an activity dependent enhancer. Thus, finding 12k new ones is a significant achievement in itself.

[~6 min, 12]

A PDE for eRNA levels

$$\frac{\partial P}{\partial x} + \frac{\partial P}{\partial t} = k(x, t) - \lambda_x P - \lambda_t P$$

$$\frac{\partial E}{\partial x} + \frac{\partial E}{\partial t} = \gamma P(x, t) - \delta_x x E - \delta_t t$$



We may extend the model further by including the previous analysis of the eRNA degradation rate. This results in a set of PDEs instead of a set of ODEs. For the eRNAs, we have all the parameter values which allows us to plot the transcript density as a function of both position and time.

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

125

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 per gene and cell, 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.