Histone Modifications in Transcriptional Regulation

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science

by

Aharon Novogrodski

Supervised by Prof. Nir Friedman

December 2010

The School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Abstract

Chromatin is the collection of proteins that encase chromosomal DNA in eukaryotic cells. The basic unit of chromatin is a nucleosome, a protein complex around which 147 basepairs of DNA are wound. One of the most important questions about chromatin is the influence of chromatin state on gene regulation. Many evidences point towards a significant role for nucleosome positioning in transcriptional control. Beyond nucleosome positioning, patterns of covalent modifications of histone tails are correlated with transcriptionally active and silenced regions, and manifestly play a role in transcriptional regulation and in passing information in an epigenetic manner to offspring. Our goal is to untangle the causal aspect of the correlations between histone state and transcription. To do so, we followed changes in histone state and transcription of yeast genes after a sharp stress stimulation which involves rapid induction and repression of hundreds of genes. We measured the changes in the modification level at different time points following stimulation by using a high- resolution tiling microarray with single nucleosome resolution on a thousand nucleosomes.

Our findings show that the initial modification changes are affected by Pol2, while Pol2's initial response is not influenced by changes in the modification level. We also found that the behavior of most of the modifications in response to stress was consistent with the observations made about their behavior in mid-log conditions, i.e. the changes in the modification level associated with gene expression were correlated with the changes in the gene expression.

These results leads to another step in understanding the importance of histone modification for the global process of gene regulation.

Acknowledgments

I wish to thank my supervisor, Prof. Nir Friedman, who was abundantly helpful and offered invaluable assistance, support and guidance, for teaching me how to approach scientific questions, how to view problems from different perspectives, when to dig deeper and when to move on. My thanks also go to Prof. Oliver Rando at the university of. Massachusetts Medical School who always knew how to push my research forward when it seems that I had reached a dead end. I wish to thank Moran Yassour, Assaf Weiner and Ruty Rinnot, I am grateful for their help in the different parts of the work and for their guidance and persistence. I am lucky to be a part of an amazing group, the members of Prof. Friedman's lab, who made this period as enjoyable as possible.

Contents

Chapter 1

Introduction

1.1 Chromatin and Transcription Regulation

All eukaryotic genomes are packaged into a nucleoprotein complex known as chromatin. This packaging is very efficient, for example one meter of human DNA packages into a nucleus that is only a few micrometers in diameter. One of the side effects of this packaging is controlling gene regulation. Transcription is the process of production of RNA copies from the DNA, by RNA polymerase. Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. Usually when we speak about gene expression level we are referring to the amount of RNA copies of the gene, since it is the easiest way to measure it. An essential part of the gene regulation is the regulation of the transcription, i.e. how many RNA copies of the gene are synthesized. The transcription process can be regulated by many mechanisms. Most of them involve regulatory proteins, which are often referred to as "Transcription Factors" (TF). Most of the transcription factors binds to DNA sequence upstream to the Transcription Start Site (TSS), near the promoter - the DNA region where RNA polymerase associated with the DNA and transcription initiation takes place. A common assumption is that the accessibility of the gene in the chromatin affects its expression. Genes with easy access will have high expression, while genes that are hidden by the chromatin will be silenced. The critical area which need to be exposed is not the gene itself but the transcription control area upstream, to the transcription start site. The question of how the chromatin influences gene regulation is one of the key questions for understanding transcription regulation.

The primary structure of chromatin consists of a large number of repeating subunits, the nucleosomes, with almost two turns of DNA (147 bp) wrapped around each unit (Luger et al., 1997). Many evidences point towards a significant role for nucleosome positioning in transcriptional control. In general, highly expressed genes have a nucleosome free region (NFR) near the TSS, while stress genes tend to be characterized by higher nucleosome occupancy at promoters. Furthermore, as nucleosomes occlude transcription factor binding sites, changes in nucleosome position can influence gene regulatory programs (Weiner et al., 2010).

Previous researches show that most nucleosomes have a specific location in the genome, and that they can always be found around this location (Kaplan et al., 2008a). Some nucleosomes, especially around the TSS, are strongly positioned, meaning that they are always positioned exactly in the same location. Other nucleosomes are weakly positioned, which are often referred to as "fuzzy nucleosomes". Another important phenomenon is the disassembly and reassem-

bly of nucleosomes, called "nucleosome turnover" (Dion et al., 2007). These two properties make the nucleosome more dynamic and allow it to cover and uncover specific genome regions (Weiner et al., 2010). Some researches show specific examples of genes in which changes in nucleosomes position near the promoter have a large impact on the transcription (Bryant et al., 2008).

The chromatin has been proposed to carry epigenetic information throughout generations, epigenetic, for our purposes, is heritable changes caused by the activation and deactivation of genes without any change in DNA sequence. For example, the majority of cell types in the human body carry the same exact DNA, yet cell state is heritably maintained between cellular generations. The theory is that much of this heritable information is carried by chromatin, for instance in patterns of histone modifications as we shall see below (Turner, 2002; Rando, 2007).

The main goal of our research is to try to understand more about the role of chromatin and nucleosomes in the process of transcription regulation.

1.2 Histone Modifications

Histones are simple alkaline proteins combined ionically with DNA to create nucleosomes. Each nucleosome contains an octamer of core histone proteins, two of each - H2A, H2B, H3 and H4. Each histone is build from a structured domain and an unstructured amino-terminal 'tail'. These tails are substrates for a variety of chemical marks known as histone modifications (Turner, 2002). Early studies on the subject suggest that changes in the pattern of histone modifications affect chromosome functions through at least two distinct mechanisms, the first being the change in the physical and structural state of the chromatin (Hong et al., 1993; Luger et al., 1997), the second is the creation of binding sites for different proteins (Dhalluin et al., 1999). These changes have many roles during cell division and transcription, for example DNA repair during the replication process or gene regulation during the transcription. Our interest is mainly in the relations between histone modification and gene regulation.

One of the controversial hypotheses in this field is the "histone code" hypothesis, first proposed by Strahl and Allis (Strahl and Allis, 2000). This hypothesis proposes that distinct histone modifications combine to create a code read by other proteins to bring about distinct downstream events. One of the implications of this theory is that the evolutionary reason for having so many modifications is to create a combinatorial complexity of histone code, resulting in a large variety of functionally distinct nucleosome functions. While everyone agrees upon the existence of some histone combinations, the conclusions about large histone language and combinatorial complexity are questionable (Rando, 2007). One of the arguments against this theory is that all the studies that support it were done on population averages and cannot distinguish whether two of the modifications join into pattern or merely substitute each other. In addition, some modifications might have a transient effect or a serial effect and thus examining steady state would not revel these temporal relationships. Another claim against that theory is that we only see a fraction of the activation patterns, and it is not enough to prove that all the various modification were developed for so few patterns (Kurdistani et al., 2004).

In a previous study conducted in our lab, modifications with a single nucleosome resolution were examined (Liu et al., 2005). It has shown that the modifications can be divided into two categories: the first category contains modifications in which changes in the expression

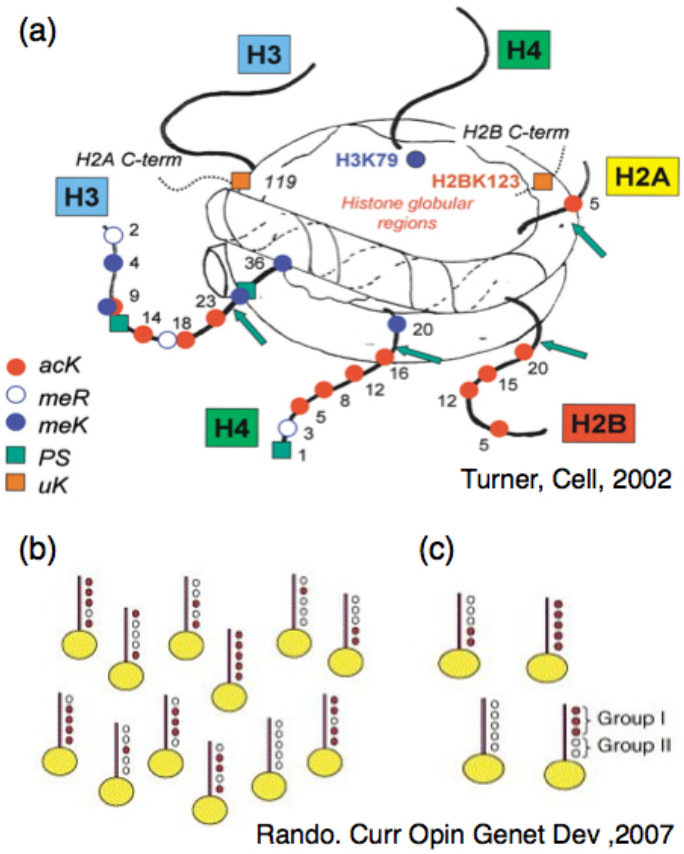

Figure 1.1: **Histone modifications**:
(a) Histone Modifications on the Nucleosome Core ParticleThe nucleosome core particle showing 6 of the 8 core histone N-terminal tail domains and 2 C-terminal tails. Sites of posttranslational modification are indicated by colored symbols that are defined in the key (lower left); acK, acetyl lysine; meR, methyl arginine; meK, methyl lysine; PS, phosphoryl serine; and uK, ubiquitinated lysine. Residue numbers are shown for each modification. **(b)** The combinatorial complexity of the histone code of nucleosome carrying five modifiable amino acids. Only a subset of the 2^5 (or 32) possible combinations is shown, to save space. **(c)** According to some researchers, all genomic studies that examine multiple potentially independent modifications have found that modifications co-occur in population averages, resulting in greatly reduced dimensionality in the complexity of patterns identified (Rando, 2007).

level causes modifications near the TSS, the second contains modifications that affects only nucleosomes in the mid coding region of the gene. One of their conclusions was that from all the combinatorial options in the data, there were only two examples, in contradiction to the histone code theory.

The conclusion of all these studies is that most of the histone modifications are related in some way to the transcription. It seems that this connection is mutual, the transcription effects the histone modifications pattern and they effect the transcription. Assuming that this mutual influence does exist, we would like to know how this influence works. For example, when specific modification is correlated with expression, we want to know whether this modification was created by the transcription in order to help it or if the modification was created by the cell in order to recruit the polymerase and to start the transcription.

The primary question in our study is the classic question of cause and effect. In this context, the question is whether a polymerase passage causes histone modifications, or whether histone modifications cause transcription. In other words, if there is a mutual influence between the two, we would like to see the timeline of the influence. Answering this question is the key to understanding whether chromatin controls gene regulation. So, if we find that histone modifications indeed change the gene expression, it will suggest that modifications play a substantial instructive role in gene regulation. On the other hand, if we find that modifications are simply a secondary effect of the transcription, their significance in epigenetic inheritance is rather questionable.

The problem is that in steady state, where the mutual influence is already exists, it is very hard learn new things about the connection between them. Therefore, we designed an experiment that checks the changes of both transcription level and modification level in response to a stress. Examination of their behavior during the timeline can teach us a lot about this relationship. Another question that we can answer using this experiment ins how the modifications level changes in response to a stress, and whether these results agree with the modification pattern in steady state. Since the behavior of the modifications in response to stress is unknown, answering this question has its own importance.

Another question that we would like to answer, is whether the assumption that the modification effected by the polymerase is accurate. Besides for a few modifications that are physically associated with Pol2, the association of most of the modifications is based on the correlation between the modification and active genes. We would like to investigate whether the assumption that these modification effected by the transcription is true. In order to answer this question we investigated the modifications pattern after we shutdown the transcription. Examination of the modification behavior without transcription can teach us about the connection between the two.

The selected modifications are those that have a lot of information about their behavior. The steady state relations between these modifications and gene expression are well known, and have been checked by a lot of researches some of which conduced in our lab (Liu et al., 2005; Kaplan et al., 2008b).

The modifications are:

1. Input (nucleosomes) - nucleosome occupancy have a major role in the yeast transcription regulation. Previous studies shows that nucleosomes influence the transcription level of a huge fraction of genes, and that the precise positioning of nucleosomes controls access to protein binding sites and thereby affects regulatory programs. In our study we are not

going to search for new knowledge about nucleosomes occupancy, just to check its effect on the modification pattern.

2. H3K4Me3 - Lysine number 4 methylation exists in three states: monomethylated (me1), dimethylated (me2), and trimethylated (me3). We examine just the last level of trimethylated H3K4. H3K4Me3 is an active chromatin mark that associated with the 5' end of the coding region(Millar and Grunstein, 2006). Although H3K4Me3 is associated with active genes, the correlation of the modification level and the transcription level is not high, which leaves us with an open question about the role of this modification. Several chromatin remodeling factors, including the chromodomain containing remodeler Chd1 are known to be recruited by H3K4Me3, especially in high level eukaryotes (Sims et al., 2007; Sims and Reinberg, 2006). This connection leads some research to connect H3K4Me3 with pre-mRNA maturation by recruiting CHD1 proteins (Sims et al., 2007). In yeast, H3K4Me3 methylation occurs by an association between methyltransferase named Set1 and a specific phosphorylated form of elnogating RNA polymerase 2(Pol2). The level of the modification is regulated by Set1 level(Santos-Rosa et al., 2002; Ng et al., 2003).

3. H3K36Me3 - Histone three trimethylated Lysine no. 36 (H3K36Me3), like H3K4Me3, is associated with actively transcribed genes. In S. *cerevisiae* H3K36 methylation is made by the methyltransferase Set2, and just like Set1 is physically associated with elongating forms of RNAP2. The connection to RNAP2 explain the correlation between H3K36Me3 to the transcription process. The difference between the two methyltransferases is that Set2 is associated with RNAP2 throughout all the gene body, unlike Set1 that associated with it just near the TSS. Therefore H3K36Me3 is generally distributed across the body of the gene, and not just near the TSS like H3K4Me3. This difference led to the hypothesis that Set2 and H3K4Me3 regulate the elongation part of gene transcription while Set1 and H3K4Me3 regulate the activation part (Martin and Zhang, 2005). As we saw before, some researchers claim that both modifications have roles in the pre-mRNA maturation, and both are important for the elongation part of the transcription(Sims et al., 2007). One of the recent finding in metazoans is that this modification has an important role in the splicing process(Sims and Reinberg, 2009), but there is no evidence that this is the case in S. *cerevisiae*, where the splicing model is less common.

4. H3K56Ac - Lysin 56 is located in H3 core domain and not in the N and C-terminal tails, like the other modifications mentioned here. Acetylation of Lysine no. 56 (H3K56Ac) is known to be associated with active genes. A recent series of studies connected H3K56Ac with DNA damage response (Masumoto et al., 2005) and with chromatin remodeling during replication (Han et al., 2007) and transcription (Schneider et al., 2006). A previous study conducted in our lab shows that H3K56Ac is not only a marker of the new nucleosomes during genomic replication, but can also mark new nucleosomes during replication independent nucleosome turnover (Kaplan et al., 2008b). As we wrote above, nucleosome turnover is essential for the transcription process, therefor this association between H3K56Ac and nucleosome turnover shows the great importance of this modification.

5. H3K14Ac - Transcription related, the more the gene is transcribed, the more acetylation near the 5' end of the gene. The acetylation of H3K14 occurs by acetylases Gcn5 that is generally recruited to the promoter regions of active genes(Pokholok et al., 2005)

6. H4K8Ac - Unticorrelated with transcription: the more the gene is transcribed, the less Acetylation. Associated with the 5' end of the gene(Liu et al., 2005).

7. H4K16Ac - Histone H4 acetylation of lysin 16 is known to have a great role in positive and negative regulations of transcription. Previous results, which part of them belongs to a study conducted in our lab, shows that H4K16Ac is unticorrelated with expression, in the 5' end of the gene (Liu et al., 2005; Kurdistani et al., 2004). The acetylation status is controlled by a specific set of HATs and HDACs and by Hst2 from the Sir2 family of NAD+ (dependent protein deacetylases)(Vaquero et al., 2006). It seem that the acetylation and deactylation influence the transcription by having a biding site for a series of proteins. These protein include proteins that effect chromatin remodeling, and both positive and negative regulators of transcription (Kurdistani et al., 2004).

1.3 The Research

All of the studies were cooperated with Prof Oliver J Rando's lab at the University of Massachusetts Medical School. All of the biological experiments described were done by his lab members. All experiments described are carried out in the model organism *Saccharomyces cerevisiae*, using homemade tiling oligonucleotide microarrays, and Agilent arrays.

To examine the relationship between histone modifications and gene expression, we conducted an experiment that examines the changes in the modifications pattern, in response to various stresses. We wanted to quantify the modification levels in each genomic loci, and use this data in order to estimate the modification changes in response to the stress. The experiment was done in a very high resolution, meaning a huge number of nucleosomes (more than 2500), and at 12 different time points, 6 of them in the first 10 minutes. The experiment used 2 different hybridization methods: one hybridization compared modifications in a specific time point compared to the occupancy at the same time; the other hybridization compared modification at a specific time with the same modification at time 0. The first reflects the absolute level of the modifications in each nucleosome; the second reflects the changes throughout the time course. These experiments were done in a two different stresses: diamide and heat-shock. All the experiments that we just described were made using tiling oligonucleotide microarrays. Another set of experiments conducted under similar conditions but with a different kind of microarrays, manufactured by Agilent company. The major advantages of the Agilent arrays are the arrays quality, and the fact that they encompass the entire yeast genome and not just small parts of it like the tiling array. The disadvantages of these arrays are the 200bp intervals between the different probes, unlike the overlapping of the tiling arrays, there are also no repeats in the Aglient arrays, and the number of time points is much lower.

The reason we chose specifically these stresses of diamide and heat shock is that diamide stress has a very fast and strong response, and heat shock allows us to shutdown the polymerase in nearly normal conditions. One of the main considerations in which parts of the genome will be includes in the tiling microarrays was to find regions where there is a strong response to diamide.

Regarding the question whether the modification is effected by the polymerase; we repeated the experiments in a different environments where the polymerase was genetically inactivated. The new environment is similar to the heat shock environment, this time with a

polymerase, which is temperature sensitive. Comparing this new experiment with the previous results will supposedly indicate which of the modifications are effected by polymerase and which are caused by stress itself independent of polymerase passage.

Testing the modifications response to exposure of diamide shows that modifications which are associated with highly expressed genes are indeed correlated with expression. These results show us that even if the steady state correlation between transcription level and modification level is weak, the response to stress is strongly correlated. It seems that for different genes there are different patterns of modification controlled by several factors, including expression level. Therefore, differences in the modifications level between different genes do not necessarily derive from differences in the expression level. On the other hand, for a specific gene, we expect that changes in the expression level will cause changes in the modification pattern, since transcription is one of the factors that effects the modification level.

Regarding the primary question of the relations between modification and the polymerase, all the modifications that we checked are related to the polymerase activity, since the response of all of them to heat-shock changed after we shutdown the polymerase. We saw also that some modifications, especially those who are known to be associated with the polymerase, had responded very slow to the stresses. Since these modifications are correlative with expression, and since their response is slower than the polymerase respond, it seems that the polymerase cause the initial response of these modifications. Other modifications respond very fast to stress, but it seems that when we shutdown the polymerase these modification response is quickly changed. Therefore it seems that these modifications also effected by the polymerase.

In conclusion, we had performed a large scale experiments in histone modifications, using computer since methods for analysis. The main goal of these experiments was to search for the role of histone modifications in the global process of gene regulation. Understanding this will help us answer the main questions of how the chromatin effects expression and of how important histones modifications to the epigenetic inheritance.

Chapter 2

Measuring Nucleosome Modification Levels

2.1 Experimental Assay

To experimentally measure the modifications levels along the Yeast DNA we first used MNase assay to create mononucleosomal DNA fragments (Figure 2.1 a). The MNase assay leaves us with fragments of approximately 150bps, each fragment containing a DNA strand that exactly covers one nucleosome. From the mononucleosomal fragment pool we immunopercipitated with antibodies to specific nucleosome modifications to enrich fragments of this modification (Figure 2.1 b). Then we labeled these fragments with a fluorescent dye and hybridized them to microarrays(Figure 2.1 c).

All of the studies were cooperated with Prof Oliver J Randos lab at the University of Massachusetts Medical School. All of the biological experiments described were done by his lab members, and especially to Chih-Long Liue. Our part in this work starts from that data analysis.

We used two different types of hybridization in our data(Figure 2.1 d):

- Type 1 - Hybridization against the pool of all the mononucleosomal fragments from the same experiment. This method compares the number of nucleosomes that were marked by specific modification against the nucleosome occupancy at the same position in the gene. The results returned by this type reflect the absolute quantity of the modification level.
- Type 2 - Hybridization against the pool of the mononucleosomal fragments that marked by the same modification at time 0. This method compare the modification level under specific conditions with the modification level in mid-log conditions. The results of this type reflect the changes in the modifications level along the time axis.

We used two types of tilling microarrays (Figure 2.1e): Homemade arrays, and Agilent arrays. These two types use different manufacturing techniques: the Agilent arrays use *in situ* synthesis where probes are built on the surface of the chip while the Homemade arrays use automated machines with pins that place previously synthesized probes onto the surface. In each one of the arrays the probe size is approximately 60bp. Another difference between the arrays is the interval between the two probes: in the Homemade arrays the probes overlap each other in about 20bp, whereas in the Agilent arrays there is interval of about 200bp between two

adjacent probes. The overlap of the Homemade array probes make the data resolution higher. For these arrays we know the value of each DNA base pair in all the regions that covered by these arrays. The problem is that the Homemade arrays cover just chromosome 3 and a few short regions from other chromosomes. Another advantage of the Homemade arrays is that we have much more experimental data, meaning more modification and more time points. The advantages of the Agilent arrays is that these arrays cover all the yeast genome, unlike the Homemade array, and the quality of the arrays. The reason that Agilent probes has about 200bp distance between each other, is because there was a search for a good probe sequence in each such window. As a result of that, and the synthesis method, the Agilent data is much better quality but it sacrifices resolution in terms of single nucleosomes. These arrays are also more reliable, since a lot of researchers already used these arrays. The Agilent data include just Type 2 data, the Homemade arrays include both Type 1 and Type 2.

In the Homemade arrays we repeat each experiment twice, to create two replications of the whole array. These replicates give us the possibility to check the quality of the assay. Since we want to check that the dye differences between the modification data and the hybridization data had no influence on the research results we swaped their dye between the two replicates.

In addition to the modifications data we also measured the expression level and the Pol2 coverage. The expression level data measured by microarrays that covered all the yeast genome. The hybridization of this data is similar to Type 2 data, meaning that the mRNA pool of a specific experiment hybridized against the mRNA pool at time 0, that is similar to mid-log growing cells. As we wrote above, the results of this experiment can demonstrate the changes in the gene expression level throughout the time. The Pol2 data is similar to the Agilent data, but instead of measuring the modification level we measured Pol2 level. The hybridization of this data is against all the genomic DNA, since Pol2, unlike modifications, does not have to be on a nucleosome.

2.2 Experimental Setup

As we mentioned above, we attempt to disentangle the causal relations between Pol2 and modification. In the first stage of our experiment we want to compare the modification response to a stress, with the transcriptional response. We already know that part of the yeast response to a stress is to change the gene expression level of large group of genes. We also know that most of the genes that respond to specific stress respond to all the stresses (Eisen et al., 1998). We want to see if the modification level is also changes in response to a stress, and if it does, whether these changes are related to the changes in the gene expression level. After knowing what is the modification response to the different stresses we will try to shutdown the polymerase and then to check how the modifications response to the stress. This experiment can give us an opportunity to check the modifications response without the polymerase influence.

We chose to work with two kinds of stress: Diamide and Heat-Shock. Both are very commonly used, so we have a lot of knowledge about their expected behavior. Diamide is an oxidative stress that is results in a very fast and wide-scale response. Heat-shock (37° Celsius) response is also fast and wide, but not as much as the former. The advantage of the heat shock response is the ability to use temperature sensitive polymerase so we can shut down the polymerase when we start our assay.

All together we performed 4 different stimulation time courses. Designations are as follows:

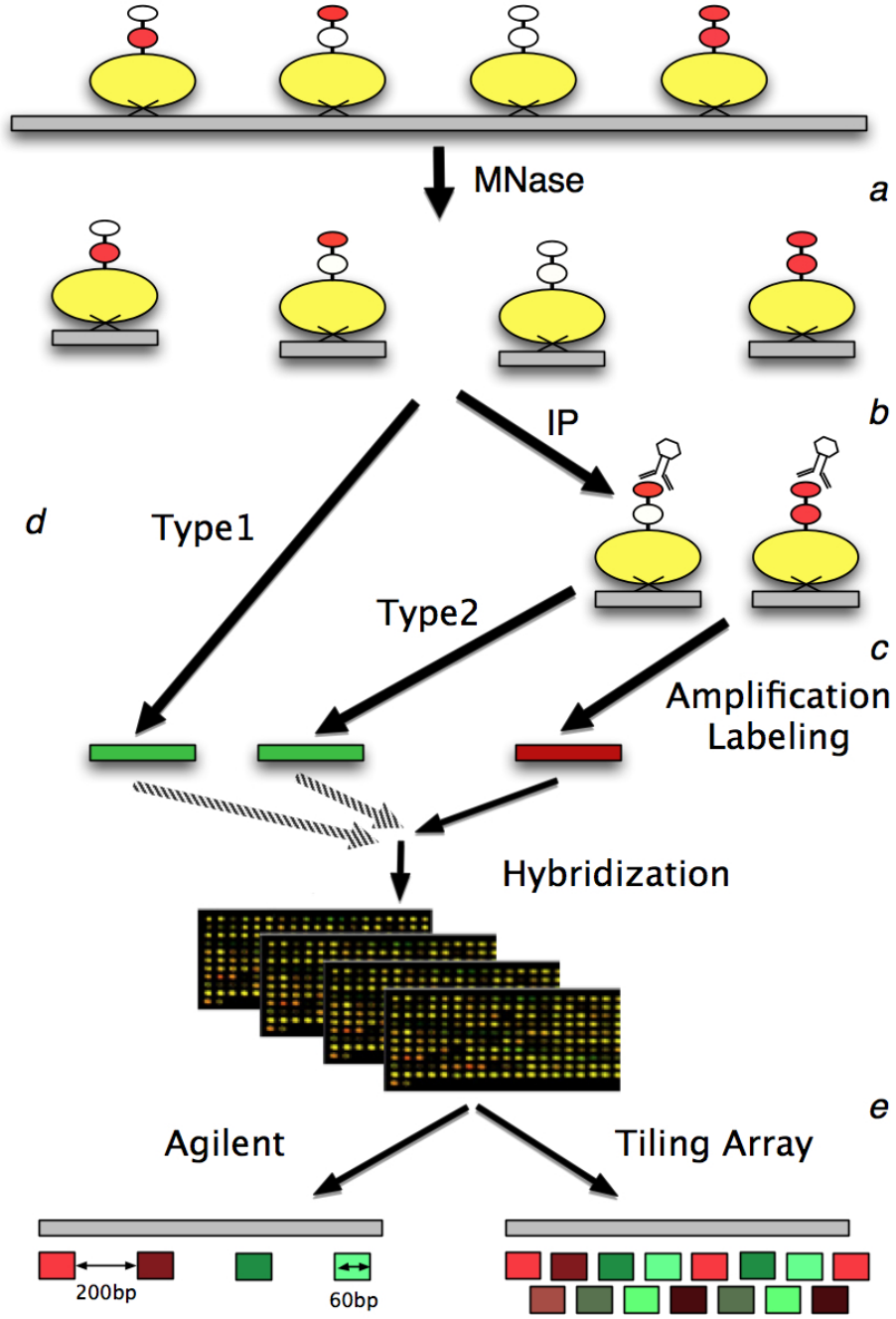

Figure 2.1: **Experimental assay**:
(a) Create mononucleosomal DNA fragments using MNase. **(b)** Pick nucleosomes that contain specific modification using specific ligands. **(c)** Labeled the fragments with fluorescent dye. **(d)** Two different types of hybridization: Type 1 - against occupancy data; Type 2 - against IP of the epitope at time 0 that suppose to be similar to mid-log growing cells. **(e)** Two types of arrays: Agilent and Homemade arrays. The Homemade arrays overlap each other in about 20bp like tiles, the Agilent have 200bp space between adjacent probes).

Table 2.1: Data dimension of the two different types of microarrays

	Coverage	Replicates	Stims	Probes	Type	Time Points	Modifications
Agilent	All genome	1	4	42,000	2	5	4 (5)
Homemade	Chr 3 +	2	4	22,000	1&2	13	7

- Stim 1 - Diamide+Pol2(no Pol2 shutdown): 1.5 mM diamide treatment at 25°C with rpb1-1 ts strain
- Stim 2 - HS(Heat-Shock)+Pol2: 37°C heat shock in BY4741 parent background
- Stim 3 - HS+NoPol2(Pol2 shutdown): 37°C Pol II shutdown with rpb1-1 ts strain
- Stim 4 - Diamide+HS+NoPol2: 1.5 mM diamide treatment with Pol II shutdown in rpb1-1 ts strain

In each one of the stimulation we checked the modification level of the 6 different modifications that we described above(1.2) and the level of the nucleosome occupancy (Table 2.1). The Agilent arrays does not contain data for all the 6 modifications, just for: H3K4Me3, H3K56Ac,H3K16Ac and occupancy data. In the first stimulation of the Agilent data we also have H3K36Me3.

The data was measured in 13 different time points: 0, 1, 2, 4, 6, 8, 10, 15, 30, 45, 60, 90 and 120 minutes. In Type 2 Homemade arrays data we have all the time points except of time 0, that used as hybridization for all the others. Type 1 of the Homemade arrays contain just 4 time points: 0, 8,30 and 90 min . The Agilent data contain 5 time points: 4,8,15,30 and 60 min, while time 0 is used as hybridization.

Altogether the data dimension for the Homemade array is:

$$2(\text{types})*4(\text{stims})*6(\text{modifications})*12(\text{time points})*\approx 22000(\text{probes}).$$

The amount of the data and its dimensions make it very hard to examine it and to search for global patterns, and therefore force us to use a lot of data mining in order to find something in this data.

2.3 Initial results

Before starting to process the data in order to look for global patterns we wanted to see some examples of the initial data. Two examples are selected, one of an induced gene, GLK1, and one of a repressed gene, HIS4. For each one of the genes we checked the behavior of three modifications under the oxidative stress of diamide.

2.3.1 H3K4Me3

As we already mentioned, trimethylation of lysine number four associated with the 5' end of active genes. In other words, this modification appears just where we find active transcription. In addition, some researchers show physical association between part of Pol2 and K4 trimethylations (Sims et al., 2007). However, we did not see a direct correlation between the

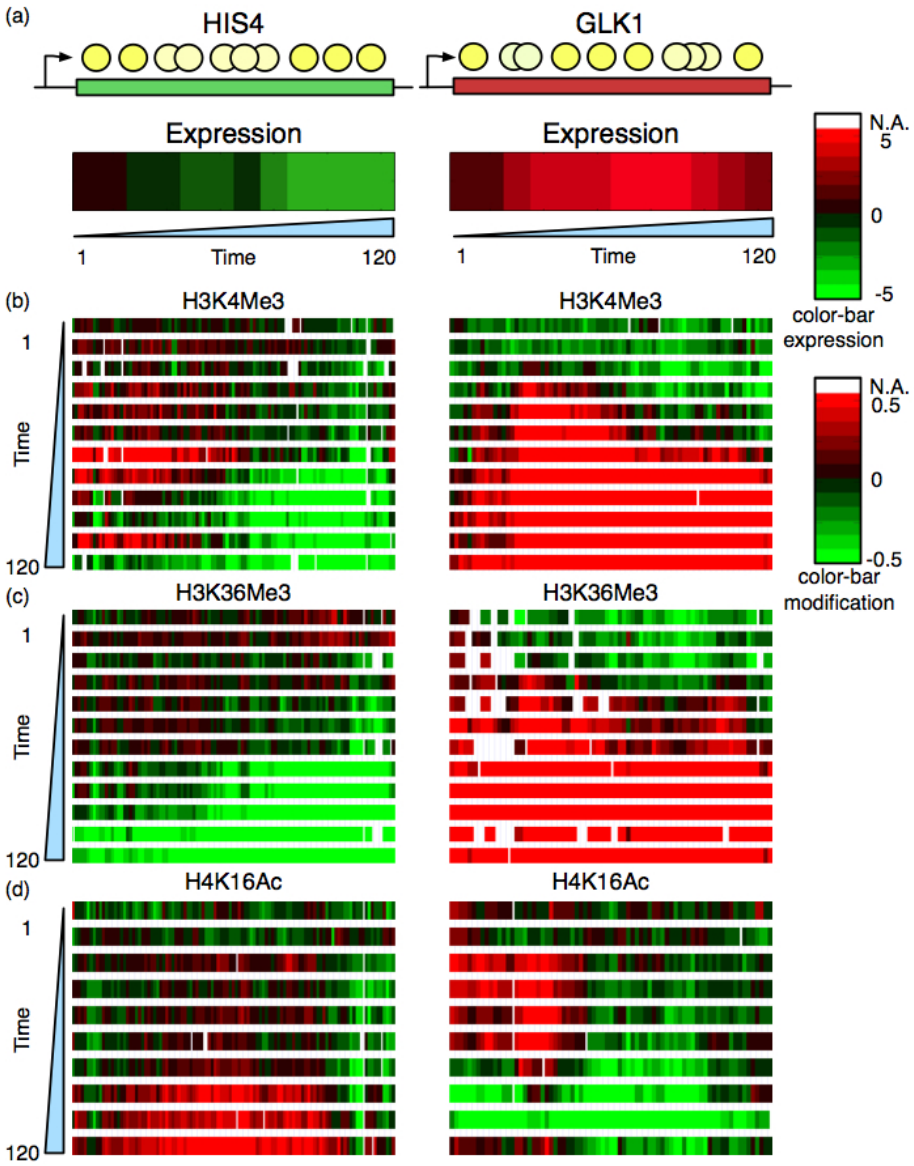

Figure 2.2: **Initial data examples**: **(a)** The two examples: GLK1 and HIS4. GLK1 expression level increase as a response to the diamide stress, while HIS4 expression level decrease. We can see also in illustration of the nucleosomes position on the genes. **(b)** H3K4Me3 results, we can see that the modification level increase for GLK1 and decrease for HIS4. We can also see that the strongest response is in the mid-coding region of the gene. **(c)** H3K36Me3 results. Here agains the modification level increase for GLK1 and decrease for HIS4. We can see how slow the response of this modification, especially in HIS4. **(d)** H4K16Ac results. Here we the modification level decrease for GLK1 and increase for HIS4. The response time here is very slow, even slower then the other two examples.

transcription level and the modification level. In our examples, of the first initial results, we can see this correlation clearly. In these results we can see how changes in the gene expression level of specific gene led to changes in the modification level on the same gene (Figure 2.2 (b)). In the first example of HIS4, the level of the transcription decrease, and we can see clearly how the modification level is also decreased. In the second case of GLK1, where the transcription level is increased, we can see how the modification level is also increased.

Interesting point is that unlike what we thought, H3K4Me3 levels change in the mid coding region of the gene and not just the 5' end (Figure 2.2 (b)). To understand this result, and other phenomenon in these examples, we need to check many more genes to see if it is a global process or merely involves these two examples. The only way to find out is to start processing all the data instead of looking on specific examples.

2.3.2 H3K36Me3

In this case we can see exactly what we expected. This modification, known to be associated with transcription, over all the genes regions. Our results show, that here again we can find a correlation between the changes in the modification level and the changes in the gene expression level. In GLK1, where the gene expression level increased, the modification level was also increased, and in HIS4, where the gene expression level decrease, the modification level decreased as well(Figure 2.2 (c)). The late response of the modification level to the stress led us to conclude that the changes in this modification level do not cause the changes in the gene expression level, since the gene expression level changed earlier. The question is whether this result is a global phenomenon or something unique to this example. If this result repeated in all the genes we would be able to conclude that H3K36Me3 was probably caused by the polymarse and not vice versa. Here again, we need to process the data before we will be able to answer this question.

2.3.3 H4K16Ac

Achetylation of lysine 16 histone 4 is known to be anticorrelated with gene expression. In our examples we can see that when the expression level decrease in HIS4 the modification level increase, and in GLK1 when the expression level increase the modification level decrease. We can see that H4K16Ac, like the previous example of H3K36Me3, responds very slowly, therefore demonstrating that the same questions can be asked here as well.

To summarize, we saw some interesting phenomena in these private cases and we would like to see these phenomena are global. To do so we have to reorganize the data and to do some analyses, as we will see in the next chapter.

Chapter 3

Data analysis

In the previous chapter we saw the pattern of some modifications for specific genes. To answer our research questions it is not enough to see a few individual cases, but rather to search for global phenomena. Since we have a tremendous amount of data, in five different dimensions, the task of finding global patterns is challenging. In this chapter we will cover some data analysis methods that should help us find these patterns. In addition, these methods should resolve several issues related to data quality, thus affecting the accuracy of the data that we will eventually work with.

3.1 Basic Data Analysis

The first step is to check the data for any quality problems. To ensure the quality of the data, we repeated each experiment of the Homemade arrays. In order to avoid the bias that is caused by the dye differences between the modifications data and the hybridization data, we swapped the dyes for the two replicates. In the first quality check we looked at the correlation between the two replicates of the same probes.

3.1.1 Reproducibility Problem

Comparing the results of the two replicates we saw that some experiments have a good correlation, while other experiments have a poor correlation. Detailed examination of the data shows that for Type 1 the correlation is generally high (Pearson correlation of about 0.8). In Type 2, however, the correlation is much lower, and in many cases it is actually very weak. Careful examination of Type 2 shows differences between the late time points where the correlation is reasonable (r=0.45), and the early time points where the correlation is weak (r=0.15). In other words, the problematic experiments that show a weak correlation usually belongs to the early time points of Type 2 data (Figure 3.1 a).

The first question that we asked was what causes the differences between Type 1 and Type 2? It seems that the dynamic range of Type 1 is higher, therefore the same differences between the replications have less effect on Type 1 data. This experimental problem is the well known tradeoff between finding the data with the highest resolution and reducing the dynamic range. Hybridization of the data with the modification level at time 0 return results with a very high resolution, but with the cost of a very small dynamic range. Detailed examination of the figures

(Figure 3.1 a) shows that the diagonal thickness in both data types is the same, i.e. the differences between the replications have the same value in both types, the only difference between them is the dynamic range. The same explanation is also true for the differences between the time points of Type 2 data. Type 2 is a relative data, therefore in the late time points where there are more changes in the modification level, the dynamic range of Type 2 is larger, and therefore the reproducibility of these time points is better.

The question is whether all the differences are caused just by the different dynamic range. Another question is why the correlation of Type 1 time 0 is weaker than the other Type 1 time points. It seems that some of the problems are caused by the variety of histone modification in mid log growing cells. This variety may be another reason for the reproducibility differences between Type 1 and Type 2, since Type 2 normalization is against mid log growing cells. It can also explain why the correlation in time 0 is weaker, because it is equivalent to mid log cells.

3.1.2 Probes Level Data Analysis

The first step in improving our data is to find and remove bad probes. We searched for probes that constantly have a bad correlation between the two replicates. To find these probes we calculated for each probe, the median of the distance between its two replicates in all of the experiments. The criterion that we set to define a probe as a "bad probe" is that its median is greater then 0.6.

Another group of probes that we removed from our data is the "linker" probes. Linker is a DNA region that is not covered by nucleosomes and since histone modification are located on nucleosomes we are not interested in these probes. The decision whether this region is nucleosome or linker is based on data from previous research about nucleosomes position (Weiner et al., 2010).

3.1.3 From Probes to Nucs

The next step was to change the observation from data that was arranged by probes into data organized by nucleosomes. For each nucleosome locus we identified the probes in that locus and set the value of each nucleosome to be the median value of all these probes (Figure 3.2 a). This reorganization of the data is natural since histones are the core proteins of the nucleosomes, and therefore modifications are process that belongs to nucleosome. Another goal for this change is to improve the quality of the tiling arrays data. In the tiling arrays data each nucleosome overlap at least 3 different probes and the value of the nucleosome is the median of these probes. The advantage of using median data is that it reduces the impact of one bad probe.

Another problem that affects the accuracy of the data is intensity depended variation. In the RI plot (Figure 3.1 c) that compares the intensity and the specificity of the data, we can see this problem in the probes data. The figure shows the intensity depended variation for the two replicates. Notice that since the dye is swapped for the two replicates, we expect that the two figures will mirror each other. To avoid this problem, after we granulized the data into nucleosome, we made "LOWESS" (LOcally WEighted Scatterplot Smoothing) normalization on the probes data (Cleveland, 1988). We can see that after doing LOWESS and granulizing the data, the correlation between the two replicates improved (Figure 3.1 b).

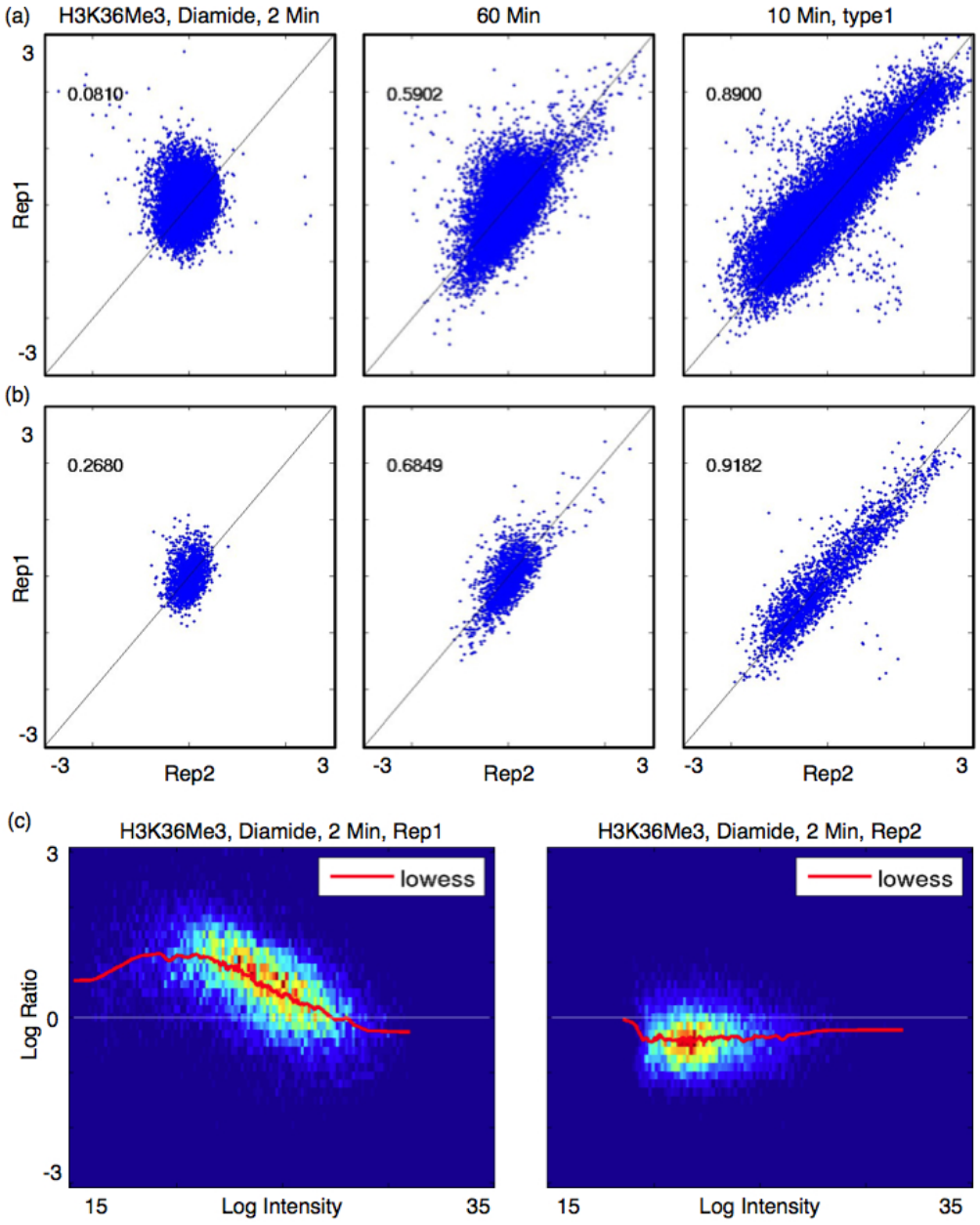

Figure 3.1: **Reproducibility of the Data** : **(a)** The correlation between the two replications of three typical and different experiments. All three experiments were made on H3K36Me3 with Diamide stress. The differences between the figures were: the experiment time and the type of the data. The numbers are the Pearson correlation between the two replicates. **(b)** The same figures for the nucleosomes data. We can see how much granulizing the data improves the correlation. **(c)** RI plot for the second result. RI plot compares the intensity and the specificity of the data. In our case, the intensity is the product of the IP and the hybridization factor, and the sensitivity is the ratio between them. The red line is the LOWESS normalization on the ratio. Notice that the two figures are supposed to be a mirror image of each other since the dye in each was swopped. We can see that for the first replication there is a bias in the small intensity values, using LOWESS normalization we can fix this bias.

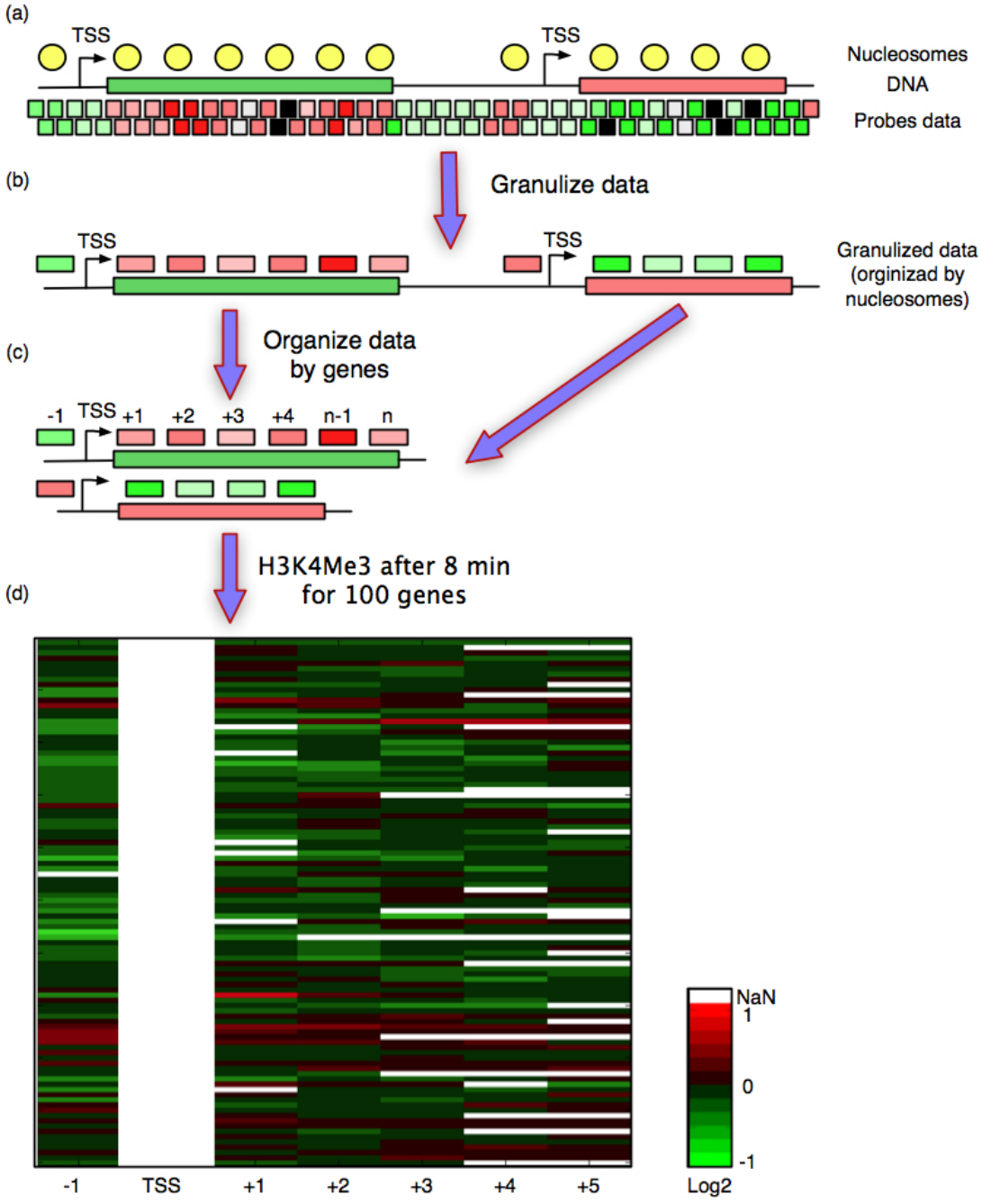

Figure 3.2: **From probes to genes**: **(a)** The initial data of tiling probes. **(b)** The data with the order changed from probes to nucleosomes. The value of each nucleosome is the median of all the probes that cover its region. **(c)** Ordering the nucleosomes according to their position at the gene. **(d)** Results of H3K4Me3 level for the nucleosomes of 100 genes with the nucleosomes ordered according to their position in the gene, as we saw in (c)

3.2 Reducing data dimensionality

3.2.1 Averaging the Data

A previous study conducted in our lab shows that the modifications level of specific nucleosomes depends on the position of these nucleosomes on the gene. According to these results, to find global patterns in our data, we need to rearrange all the nucleosomes according to their position. Another motivation to rearrange the nucleosomes according to their original genes is to allow testing of the modifications behavior for a specific genes group. To improve data quality, all the dubious and silenced genes were removed.

The procedure of rearranging the data is simple. We had information about the location of each gene's transcription start site (TSS) (REF xue). Using this information we order all the nucleosomes according to their relative position to the TSS. For example the closest upstream nucleosome will be the -1 nucleosome, the closest downstream nucleosome will be +1 and the last nucleosome in the gene will be +n (Figure 3.2 b,c).

This new organization of the data allows us to ask many new questions. For example, we can check how specific modification behave in +1 nucleosome, or divide the genes into groups and check the modifications behavior for a subgroup of genes. Another advantage of this organization is the ability to normalize all the nucleosomes of specific position. For example, if we believe that there is a deviation in the -1 nucleosomes, we can normalize just these group of nucleosomes (Figure 3.2 d).

3.2.2 Polynomial Fit - First Steps

To improve robustness and interpretability we want to reduce the data dimensionality. In addition we want to reduce the amount of data that we have from each dimension. One way to deal with the problems above is to find a polynomial fit, being a curve that best fits a series of data points. In our case we want to replace the 12 data time points with a 3-6 polynomial equation. To achieve this, we reduced the dimensions of the data to 2, by choosing a single modification, one nucleosome and a specific experiment. Now we have 2 dimensions of data where the X-axis is the time points and the Y-axis is the value of the modifications at these time points. We used this data to find the curve that best fits our data points.

There are two advantages of this method. The first is that it reduces the amount of data; instead of 12 different time points, we have 3-6 polynomial free parameters. The second advantage is having meaningful parameters. Understanding the curve parameters allows us to use them in part in relation to the question at hand. Finding a curve like that allows us to use only one parameter from this dimension and to reduce the dimensions number by one. Therefore, our goal is not just to find the equation that best fits our data but also to find a curve with the most meaningful and useful parameters.

Firstly we checked the basic methods of fitting a curve to our data. The simplest and the most common methods are regular 2-4 degree polynomial function. As expected, the fourth-degree polynomial seems to fit best. The question is whether there is an "over-fitting" of the data. To check this we removed the 8 min time point from the data and ran a train and test exam. The exam compares the median of the distance between the real 8 min time-points and the prediction of time 8 according to the polynomial function. The results where: 0.1484 for 2 degrees, 0.0792 for 3 degrees, and 0.1107 for 4 degrees. So, the 3 degrees polynomial fit is the

equation that fits our data best.

The results of the 3 degree polynomial fit were reasonable, the problem is that we did not use any knowledge about the modifications behavior. One of the implications of this problem is that the parameters that we get are not understandable, and therefore they will not help us in the future when we wish to use them. In addition, using some of our knowledge can help us better fit the time-points.

3.2.3 Two impulses model

An example of a method that uses prior information about the data is the two impulse model of responses to changes (Chechik and Koller, 2009). This method was originally developed to represent and predict changes in gene expression. In our case we would like to use it for our gene expression data and also for the modifications data. One question is whether changes in the modifications level are similar to changes in the gene expression level. Another question is whether we should use this method in our case or wether we ought to search for and use something simpler and more appropriate under the circumstances.

This method is based on some assumptions made about expression behavior. The assumptions are that for a certain set of genes, changes in environmental conditions increase the expression level for a short period. Then, after the yeast adjusts to the new environment, the expression level of this group decreases until it reaches a steady state. Genes that decrease in response to stress react in the same way, except, instead of an increasing response, we have decreasing response.

The impulse model encodes this behavior as a product of two sigmoid functions: the first represents the initial response, and the second represents the offset. The equations are:

$$f_\theta(x) = \frac{1}{h_1} * s(x, t_1, h_0, \beta) * s(x, t_2, h_2, -\beta)$$

$$s(x, t, h, \beta) = h + (h_1 - h) Sig(x, t, \beta)$$

$$Sig(x, t, \beta) = \frac{1}{1 + \exp^{-\beta(x-t)}}$$

In this equation the horizontal X-axis is the time line and the vertical Y-axis is the value of the expression. The three height parameters are: h_0 for the initial expression value, h_1 for the peak value and h_2 for the steady state value. The two width parameters are: t_1 for the time of the first transition and t_2 for the time of the second. The slope parameter β is the slope of both the first and the second transition. All together we have 6 free parameters in this curve.

Before we used this model on our expression data (Figure 3.3 b) we had to effect some changes taking into consideration the unique properties of our data. Our data is data which is relative, therefore the expression value at time zero is zero, in other words the first steady state of the model, h_0 is constantly 0. Using this property we run the model on our data, with a constraint of $h_0 = 0$. This step also reduces the degrees of the fit function from 6 to 5. The new equations are:

$$f_\theta(x) = Sig(x, t_1, -\beta) * [h_2 + (h_1 - h_2) * Sig(x, t_2, -\beta)]$$

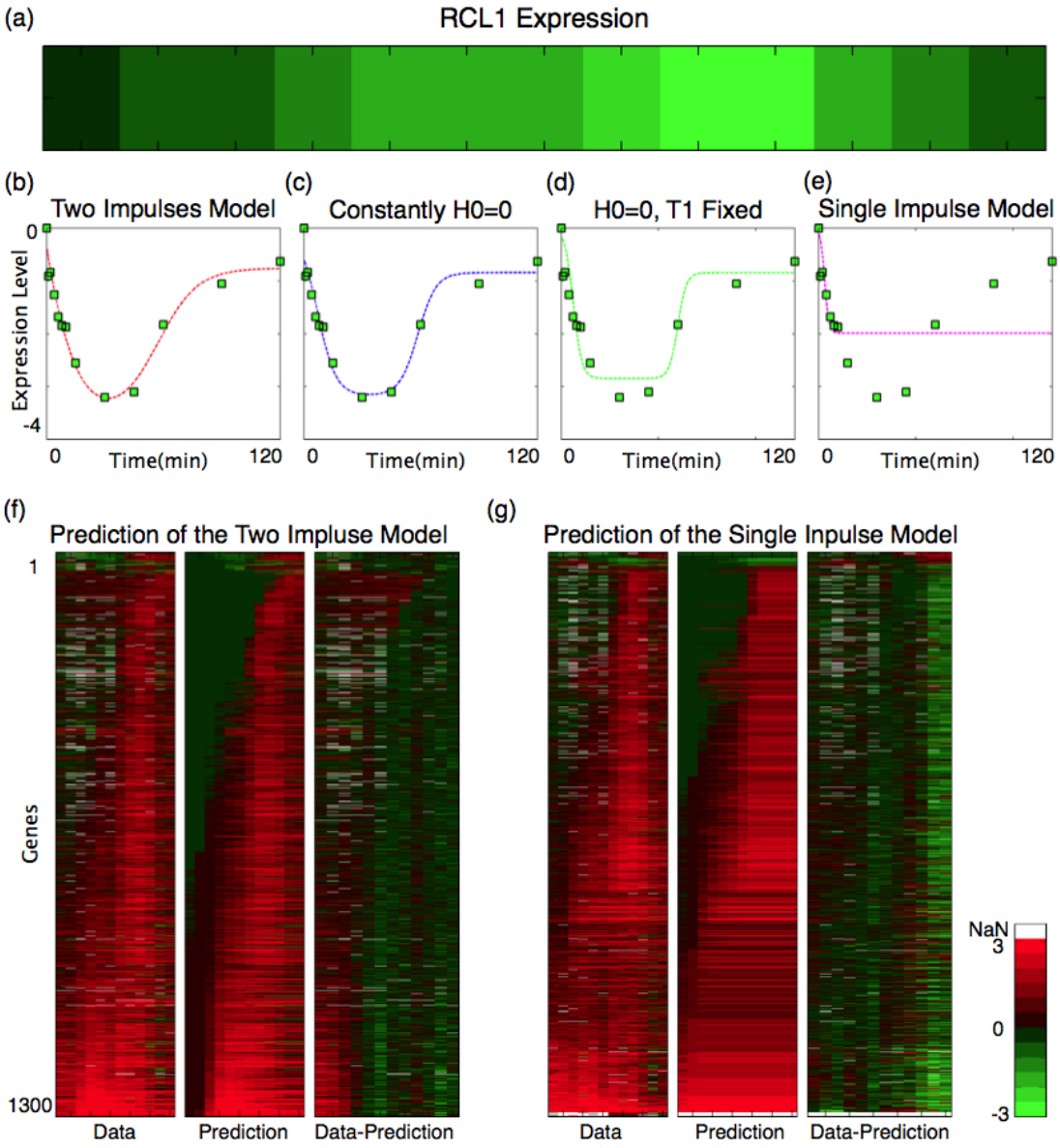

Figure 3.3: **Examples of the impulse models**: **(a)** The expression level of RCL1. **(b-e)** The results of the different impulse models (line) on RCL1 (the green squares). (b) The original two impulses function; (c) The same function with the constrain of $h_0 = 0$; (d) The same function as (c) with an additional constraint of $t_1 > 0$; and (e) Our one impulse model. **(f)** Comparison between the prediction of data using the two impules model and the real data: the first column is the real data for induced genes, the second is the prediction of those same genes and the third, is the subtraction of the prediction from the real data. **(g)** Comparison between the prediction of the data from our one impulse function and the real data.

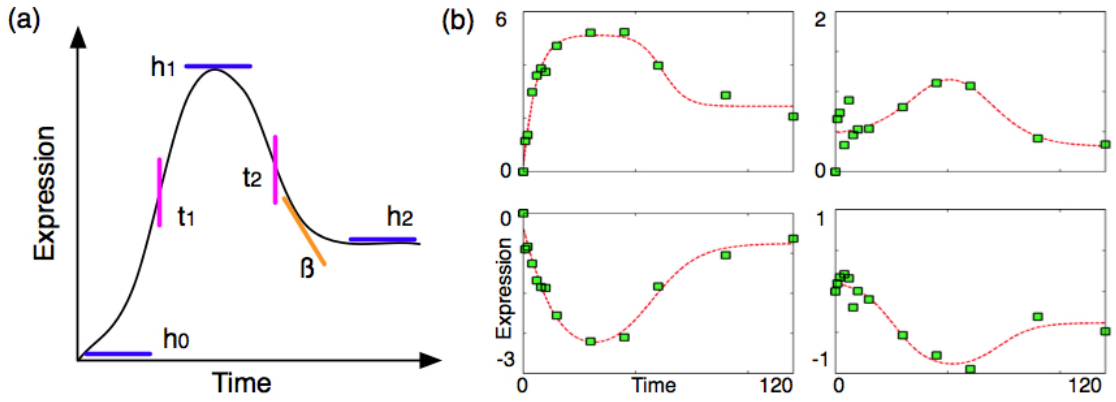

Figure 3.4: **Two impulses model**: **(a)** The six parameters of the impulse model. **(b)** Examples of impulse model fit (line) to gene expression (squares)

$$Sig(x, t, \beta) = \frac{1}{1 + \exp^{-\beta(x-t)}}$$

The result of this function (Figure 3.3 b) shows that even though we added the constraint of $h_0 = 0$, the function still looks the same. The reason is that indeed $h_0 = 0$, but the location of h_0 on the time course is not at 0 min but is on the negative side of the course. This is a result of having a very low value of β and low/negative value of t_1. As a result, at time 0 the value of the function is not 0 and therefore we were left with the same problem we had encountered earlier. To rectify this problem we made two changes: one to the function constraints and one to the data:

1. We forced t_1 value to be in the range of the data time points (between 0-120 min) to prevent h_0 from being far on the negative side of the time course.

2. We added 12 pseudo observations at time 0 with 0 value, resulting that a value other then 0 for time 0 will cost twelve times more than differences in other time points.

The results of the function with these changes are much better. We can see how the value at time 0 is almost always 0, as we expected from our relative data(Figure 3.3 d).

To check the accuracy of this model, we generated functions for each nucleosome in the data and then recreated all the data using the prediction of the results by these functions (Figure 3.3 f). Subtraction of the prediction from the real data shows that our model is very good at most time points, except for the first few minutes. The problem is that our main interest concerns the initial response of the data to the stress, therefore we need a model that is very accurate during the first few minutes, unlike Gal's model. Accordingly, we needed to make some significant changes in the model in order to get a better fit for the data in the first few minutes.

3.2.4 Single Impulse Model

As we saw in the previous part, the two impulses model poses problems for the prediction of the first few minutes and therefore we can not use this model. To resolve this issue we created a new model using 2 properties of our data:

- As we already saw, h_0 is constantly 0
- Our interest is just the first transition of the data, meaning that we needed to focus on h_1,t_1 and β and to ignore t_2 and h_2 (define $h_2 = h_1$ and $t_2 = inf$)

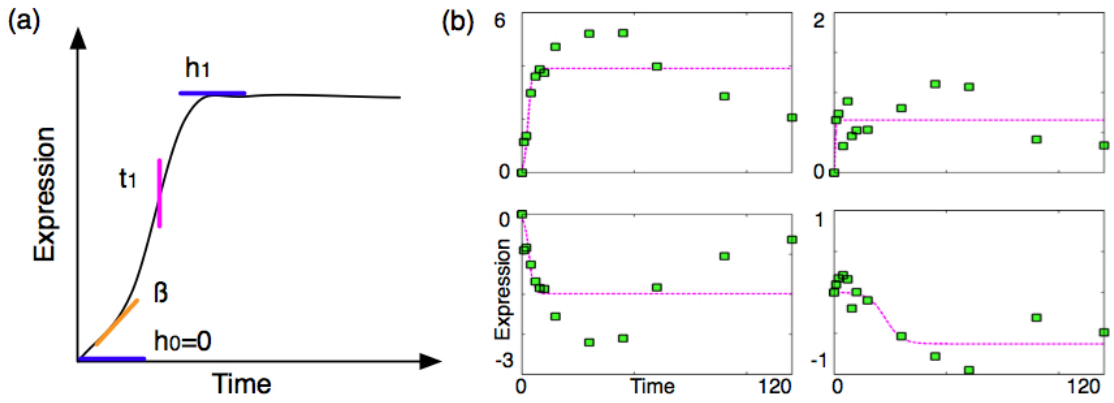

Figure 3.5: **Single Impulse Model**: **(a)** The three parameters of the one impulse model. **(b)** Examples of impulse model fit (line) to gene expression (squares)

Using these facts we can developed new equations with just 3 parameters:

$$f_\theta(x) = Sig(x, t_1, -\beta) * h_1$$

$$Sig(x, t, \beta) = \frac{1}{1 + \exp^{-\beta(x-t)}}$$

The prediction of this model (Figure 3.3 f) is very good for the first 8 time-points. The late time-points are not very accurate, as expected, since we changed the model to fit just the first transition.

Although all the examples that we have seen so far referred only to gene expression data, the same model also works for the modifications data. Moreover, this model better fits to the modifications data, especially in the late time-points. The reason for this is that unlike expression data, where we were expecting to see a second impulse a few minutes after the first one, in modification data, we were expecting just one impulse, since once a modification accrues, it does not disappear very quickly. Therefore, it seems that our new model that allows only one transition, fits precisely for the response of the modifications to a stress.

Chapter 4

Biological Results

4.1 Gene Expression and Pol2 Response

To understand the modification level response to a particular stress, and the relations of this response to the transcription process, we first need to understand the transcription response to the stress. For this purpose we measured the gene expression level throughout all the yeast genome during 12 different time points. This data, like Type 2 data, is relative, i.e. its hybridization is against the IP of the expression at time 0.

To check the quality of this data we compared it with data of Pol2 occupancy after the same stress from a different experiment. As the first step we compared Pol2 occupancy at time 0, to the expression in YPD conditions. The result of this comparison seems very good, with a Pearson correlation of more then 0.7 (Figure 4.1 a), showing that in normal conditions the two dataset are correlated.

As the next step we wanted to compare Pol2 and gene expression responses after diamide stress. To compare these datasets we first changed the Pol2 data to a relative scale, by division of 15 minutes data in 0 minutes data. After we did this, we compared the new relative Pol2 data at 15 min with the gene expression data at the next time point which is 30 minutes (Figure 4.1 a). The reason we took Pol2 data from an earlier time point is because changes in the polymerase occupancy lead to changes in the gene expression, so we expect a time lag between the two. The results of this comparison are quite good but not as good as the previous comparison like we expected, with a Pearson correlation of only 0.531.

Possible causes for the differences between the first comparison of the mid log cells and the second comparison of the response to stress can be both technical and biological. A possible technical reason is that each dataset comes from a different experiment and uses a different hybridization method. A possible biological reason is that the expression is not controlled exclusively by the polymerase and there are more factors involved.

In any case, the results are good enough to support the validity of the IP data.

To examine the relationship between transcription and modifications level in a systematic way we divided the genes into groups according to the difference in their transcription level in response to stress. This division allows us to examine the modification level behavior over genes with different transcription patterns. To obtain intuition about the correct way to perform the division we tried k-means clustering over the data with a range of values for k. We found that for k=5 there are 5 constant groups:

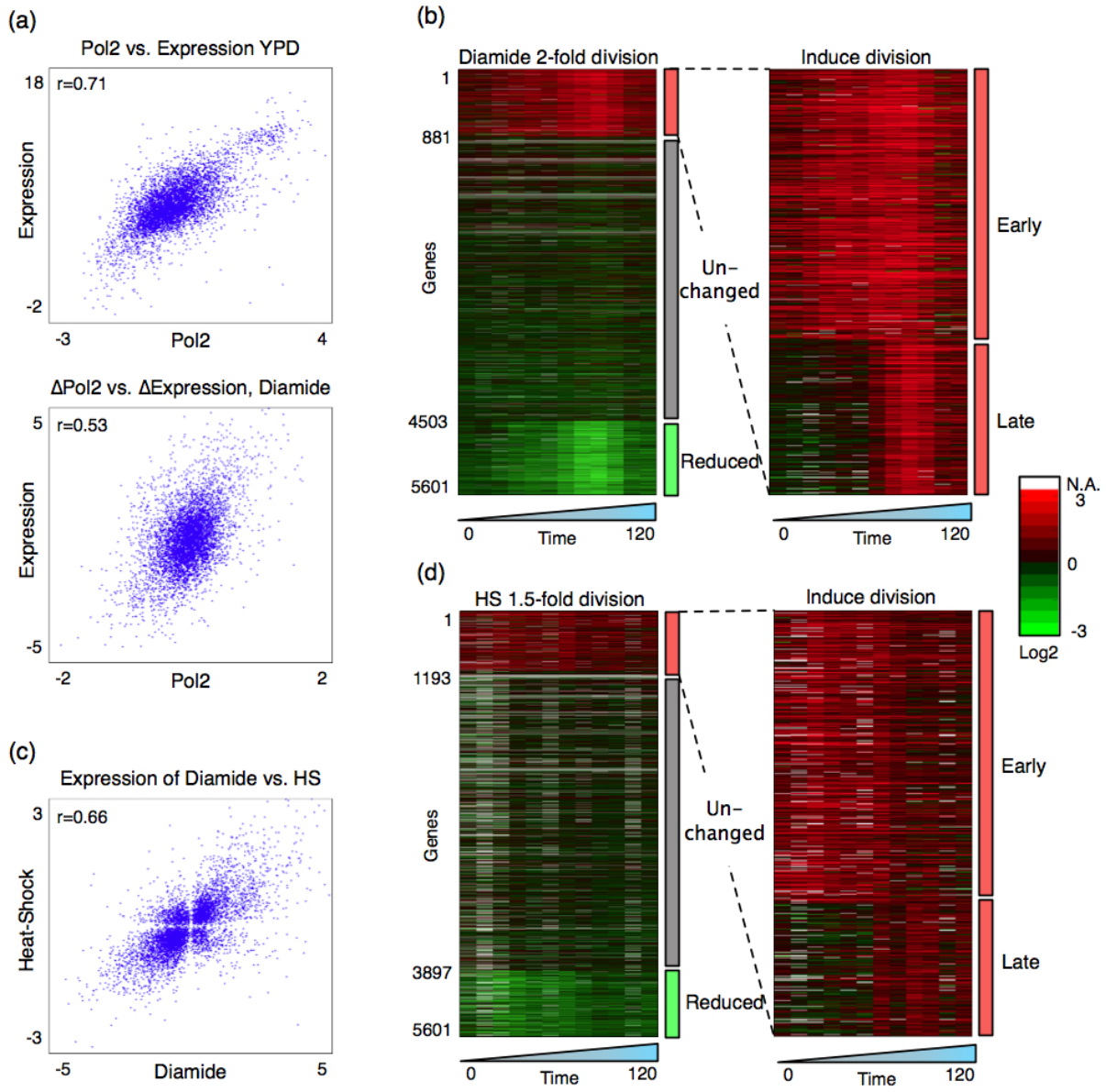

Figure 4.1: **Expression results**: **(a)** Correlation between expression level and Pol2 data. The first figure shows the correlation between expression data in YPD conditions and Pol2 data at time 0. The second shows the correlation between the responses of expression and Pol2 to diamide. 'r' is the Pearson score of the correlation. **(b)** Division of the yeast genes into four groups: reduced, unchanged, early induced and late induced. The first figure divides the data into 3 groups according to their response level, represented by $h1$ parameter. The second figure splits the induced genes into two groups according to their response time, represented by $t1$ parameter. **(c)** Correlation between the expression datasets of heat shock and diamide. Despite the good correlation between the two, we can see differences in the dynamic range. **(d)** Same graphs like (b), but this time for heat shock. Since the dynamic range of the heat shock is lower we took a lower threshold of 1.5 fold instead of 2 fold like diamide.

- Genes displaying very early induction of transcription
- Genes displaying late induction of transcription
- Genes displaying reduction of transcription
- Semi induced genes
- Semi reduced genes

Our interest is only in genes with a significant response, so we merged the groups of semi induced and semi reduced genes into one group of "unchanged genes". We will see later, that the group of genes that displaying very fast induction is problematic so we shall neglect it. Altogether we are left with 3 groups - induced, reduced and unchanged genes.

In practice we did not use the clustering results to define the threshold between the gene expression groups, instead we used a biological criteria. The reason for using biological criteria is that the border area between groups will be rich with genes and because of this will not be separated correctly according to the clustering. Another reason is the fact that we want to have the same threshold for both of the separations.

In selecting a threshold to separate the data into groups there is a tradeoff between the accuracy of the data to its size. On the one hand we want to raise the bar to ensure that only genes that for sure display induction will be in the included group. On the other hand we need to inquire a large enough groups of genes, so we will be able to trace global patterns. The threshold that we chose for the diamide stress experiments (Figure 4.1 b) is two times fold of the $h1$ parameter, which represents the peak of the gene expression (see 3.2). This threshold leaves us with a very large group of genes and ensures that these genes are a true response to the stress.

The results of the heat shock experiment are less striking. Although the correlation between the heat shock and the diamide experiments is very high, about 0.67 (Figure 4.1 c), we can still see that the response in the diamide experiment is broader, faster and stronger. Another problem is that the genome regions covered by the tiling arrays were selected according to the diamide experiments. As a result of these problems, if we will take a threshold of 2 fold for the heat shock experiment we will get a very small group of genes. So we decided to lower the threshold for the heat shock experiment to 1.5 times fold (Figure 4.1 d).

The main group where we searched for changes in the modification pattern is the induced genes group. The reason for this is that the cell has many ways to reduce gene expression level, and not all of them include reducing the transcription rate. Another reason is that the histone methylation level increases very fast, but has a long life, especially in yeast, where demethylation is a very slow process (Radman-Livaja et al., 2010). Moreover, we know that H3K4Me3 persists for a long time after the transcription, providing a molecular memory of recent transcriptional activity (Ng et al., 2003). Thus, this modification level in deactivated genes will be just a reflex of molecular memory.

In order to examine group of induced genes in a higher resolution, we split this group into two subgroups according to the genes response time: early induced genes and late induced genes. For this division we used the $t1$ parameter that represents the transition time of the response (Figure 4.1 b,d).

As we mentioned before there is a small group of genes that responds very quickly. This group responds so fast that it is hard to believe that the response of these genes is due to

the stress. The annotation of these genes includes some essential processes, unlike what we expected from genes that respond to a stress. In addition these genes have a high level of expression in YPD and a problem of reproducibility. Because of all these reasons we suspect that these genes do not respond to the stress as we thought, but are just genes with a high variability. Therefore we decided to ignore these genes and to remove them from the group of the induced genes.

4.2 Modification Response to Diamide Stress Condition

The first experiment checked the modifications response to diamide stress. In this experiment we took yeast growing on rich media and added 1.5 mM of diamide to its culture. This caused an oxidative stress by creating formation of disulfide bonds in living cells. Diamide, known as a stress with a very fast and strong response, penetrates cell membranes within seconds and also cause cell reaction very fast.

4.2.1 Nucleosome Occupancy Data

In the first stage, before we looked at the modifications results, we checked the nucleosome's occupancy level in the same conditions. To look at all the data and not at specific examples, we divided the genes into four groups, as we described above: early induced, late induced, un-changed and reduced genes. We also divided the nucleosomes into nine groups according to their position on the gene. The position of a nucleosome on the gene is determined by its relative position to the TSS. After aligning all the nucleosomes according to the gene's TSS, we rearranged all the data by the nucleosomes position (3.1.3). We chose to look at nucleosomes in nine different positions: the first nucleosome upstream to the TSS (-1), the first five nucleosomes downstream to the TSS (+1,..,+5) and at the last 3 nucleosomes of the gene (n-2,..,n). Most of the positions that we chose are near the TSS since we expected to find the most interesting phenomenon in this region. Later we used this method to analyze all the modifications in our data.

The results of the occupancy data show (Figure 4.2 (a)) a decrease of the occupancy level in +/-1 nucleosomes and an increase in the other positions. These changes were true solely for the induced genes group. In the other genes groups we could not see any significant change. The reason for the decrease of the +/-1 nucleosome's occupancy is probably because some of the nucleosome regulates the transcription, and transcription initiation requires remodeling these nucleosomes. The reason for the increase of the occupancy level in all the other positions may be the changes in the chromatin complex.Changes in chromatin complex are causing new positions to be exposed to MNase, and therefore we can see more nucleosome occupancy in these positions.

As a result of this situation, in which part of the -/+ 1 nucleosomes were removed, we ignored these two positions in our analysis of the modifications data.

4.2.2 H3K4Me3

Histone H3 trimethylated at lysine 4 (H3K4Me3) is physically associated with the 5' end of active genes. This modification performed by association of methyltransferase and a form

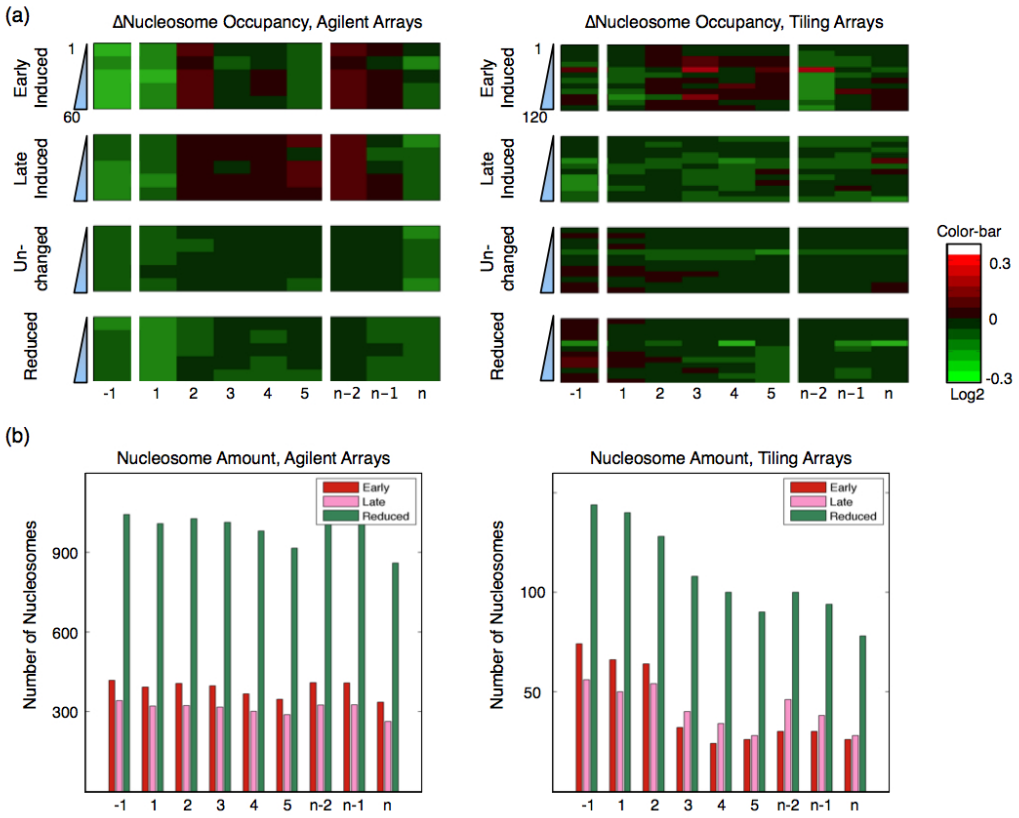

Figure 4.2: **Nucleosome Occupancy**: **(a)** Relationship between the nucleosome occupancy response to diamide and the transcription response. Genes were split into 4 groups based on their response to the stress, as we have seen above, and average data for these groups are shown. The data granulized into nucleosomes (see chapter 3), and aligned by the nucleosomes position according to the TSS. There are two graphs, one for the Agilent arrays, with four time points, and one for the tiling arrays, with twelve time points. **(b)** Nucleosomes number for each position. These nucleosomes used for the above averaging, and we are going to work with it in the next sections.

of elongating RNA polymerase 2 (Pol2) (Radman-Livaja et al., 2010). The pattern of this modification is known to positively correlate with gene activity, with a weak correlation to the gene transcription level. As a first step we examined if our data agrees with known patterns. Since H3K4Me3 is the modification that we have the most information about it, it is a good place to start our quality checks.

Most of the prior knowledge about H3K4Me3 is in steady state, therefore we need to compare it with our equivalent to steady state data, that is Type 1 data at time 0. Type 1 data (see Chapter 2) is our IP data that hybridized against the occupancy of the nucleosome at the same time. In our case it means that we have the modification level at YPD against the nucleosome occupancy at the same conditions.

To check this data we used the division of the nucleosomes according to their position (see 4.2.1), and divided the genes into three groups according to their expression level in YPD (Yassour et al., 2009). For each group of genes and for each position, we averaged the modification levels of all the nucleosomes. We observe the modification in all the groups, but is weaker amongst genes with a low expression (Figure 4.3 (a)). We also observe that the modification level is very high near the 5' end of the genes, and declines further from the TSS. These observations agree with our prior knowledge about this modification (Santos-Rosa et al., 2002; Millar and Grunstein, 2006), and support the validity of our IP data.

To examine the changes of the modification level in response to diamide, we used Type 2 data. In this data the IP data is hybridized against the IP of the same modification at time 0 (see Chapter 2). To understand the results and to search for global patterns, we divided the data into groups according to the genes response to diamide and according to the nucleosomes position (see 4.2.1), and then averaged the modification levels of each group. The results show (Figure 4.3 b) that for genes with reduced or unchanged expression level the modification level does not change. However, for the groups of the induced genes the modification level increases. We can also see that the modification level of the early induced genes changed earlier then that of the late induced genes.

Another way to look at the data is to order all the nucleosomes according to a specific parameter, and to check the gene expression trend according to this order. This parameter can be h_1 that represents the level of the response, or t_1 that represents the timing of this response (see Chapter 3). A moving average over the genes creates the gene expression trend. Using this method (Figure 4.3 (d)) we can see how the expression level increases together with the modification level. The first comparison shows that the expression level and the modification level both rise on the same genes. The second comparison shows a correlation between the transition times of the two datasets.

Altogether we can see that the expression response to the stress and the modification response are related. This relation includes the level of the response and its timing. The next step is to understand the position of the modification on the gene, in relation to the TSS.

Unlike our expectation the modification level does not increase near the 5' end of the gene, but in the mid-coding region and in the 3' end of the genes (Figure 4.3 b). A possible explanation for this is that there is no room for new modifications in the 5 end nucleosomes, since the nucleosomes are already marked. An evidence for this claim is the anti correlation between Type 1 and Type 2 in H4K4Me3. The Pearson score of their correlation is lower then -0.5, while the score of the correlation between the two types of the other modifications is between -0.1 to 0.1. To see another evidence for these explanations we need to see the Type 1 results.

To explore Type 1 data we used the same averaging on the four gene groups. We saw that

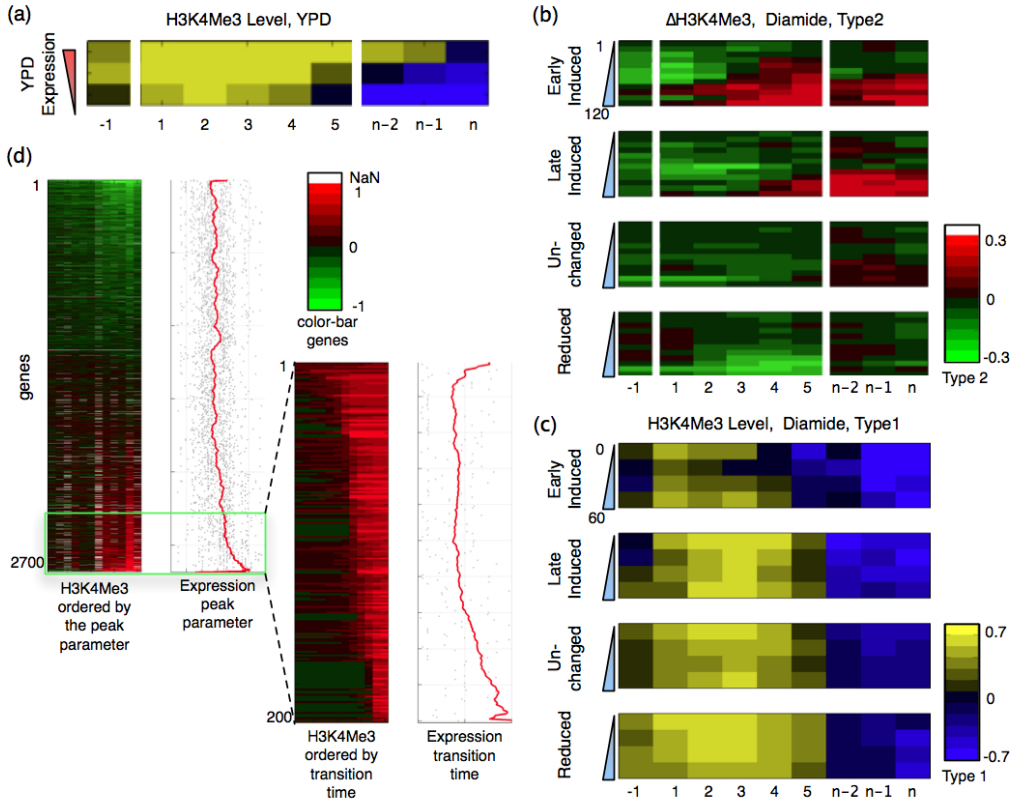

Figure 4.3: **H3K4Me3 results**: **(a)** Relationship of H3K4Me3 level at time 0 to transcription level. Genes were split into 3 groups based on their transcription rate in YPD, and average data for these groups are shown, aligned as in Figure 4.2 **(a)**. **(b)** Relationship between the modification response to diamide and the expression response. Average data, from tiling array Type 2, is shown for each group, aligned according to the TSS. The groups and the alignment as in Figure 4.2 **(a)**. **(c)** Same graph like **(b)** for tiling array Type 1 data. The graph shows the relationship between the modification level after the stress and the changes in the transcription level. **(d)** Trends of H3K4Me3 and expression in response to diamide. The first figure shows H3K4Me3 data ordered by the peak parameter $h1$ and the trend of the gene expression's $h1$ value of the same genes. The second compare the transition time of both. In the second graph we chose just the nucleosomes where H3K4Me3 levels strong increase.

the modification level changed very little throughout the time course, especially near the 5' end of the gene where it is always high (Figure 4.3 (c)). The fact that we constantly see high levels of modification near the 5 end of the gene supports the two explanations about the increases of the modification level in the mid-coding region of the gene. Interestingly, the modification level does not change even in genes whose expression level decreases. One possible reason is that H3K4Me3 is associated with active genes, and the stress just reduces the expression level of some genes but does not deactivate them. Another possible explanation is that even after a gene is deactivated it can take more than an hour until the methylation is removed. According to this our observations are consistent with a new research that shows that it goes away only on the next cell devision (Radman-Livaja et al., 2010). The fact that all the changes that we saw in Type 2 are usually do not seen in Type 1 can give us perspective regarding the level of changes that we saw in Type 2.

For summary we saw that the increase of the Pol2 presence increases the modification level in the mid-coding region. This can teach us two things:

1. H3K4Me3 can be found over all the genome body. It seems that when we have a very high presence of Pol2, and the nucleosomes near the 5' end are already occupied, the modification will spread to other regions of the gene.

2. Unlike some studies that claim that the correlation between the expression level and H3K4Me3 level is weak (REF), we obvserved a strong correlation between the two. The reason why we do not see this correlation in steady state is that for each gene the modification behaves differently. Here, where we compared the modification levels of the same gene, in different conditions, we can see this correlation very clearly.

Regarding the question of the causality between Pol2 and histone modifications, it seems that the onset of change in modification is after the change in expression (Figure 4.3 (b,d)). Therefore the polymerase may affect the modification, in consistency with our former knowledge about the recruitment of Set1 by Pol2 (Radman-Livaja et al., 2010). However, since the transcription level changes before the modification level changes, it seems impossible that changes in the modification level cause the changes in the transcription.

4.2.3 H3K56Ac

Histone H3 acetylation at lysin 56 (H3K56Ac) is known as a mark of new nucleosomes(Kaplan et al., 2008b). For example, during the replication process where there are a lot of new nucleosomes high levels of this acetylation are observed. This modification is also associated with active genes, since as part of the transcription process there is a nucleosome turnover that brings new nucleosome into the chromatin (Kaplan et al., 2008b). This modification, unlike H3K4Me3 and H3K36Me3, does not require presence of Pol2, therefore we hypothesized that this modification may comes before the polymerase and has a role in the initiation process. The methods that we used to examine this modification are the same methods that we used for H3K4Me3 (4.2.2).

The first step is to examine if the modification level of H3K56Ac in YPD conditions in our data agrees with prior knowledge. To this end we examined the modification level in Type 1 time 0. We find that this modification appears just in the group of highly activated genes (Figure 4.4 a). Thus, we observe an association between the gene expression level and H3K56Ac

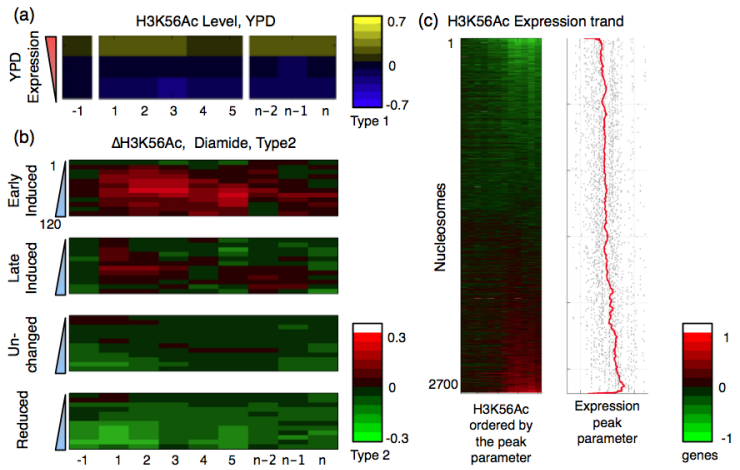

Figure 4.4: **H3K56Ac results**: **(a)** Relationship of H3K56Ac level at time 0 to transcription level. **(b)** Relationship of H3K56Ac response to diamide to the expression response. **(c)** Trends of H3K56Ac and expression in response to diamide.

level in YPD conditions. These results agree with our prior knowledge about H3K56Ac, and support the validation of our data. The next step is to see how the modification level changed in response to stress.

Examination of H3K56Ac response to the diamide stress (Figure 4.4 b) shows a correlation between the modification response to the stress and the gene expression response. We can see that the modification level decreases in the group of repressed genes, and increases in the groups of induced genes. We can also see that H3K56Ac, unlike H3K4Me3, increases over all the gene regions, especially near the 5' end of the gene. A possible reason for the strong response near the TSS is that there are more happening in this part of the transcript, and therefore more nucleosomes move, and new nucleosomes come in. Comparing the trend of the expression with K56 modification level (Figure 4.4 c), we can see how the expression and K56 increase together. Interestingly, unlike H3K4Me3 where the trends are correlated only in the induced genes, here the correlation persist for all genes.

Regarding the appearance order, it seems that the modification appear together with the transcription, and it hard to say which one comes first. Therefore the question of the causality between the polymerase and H3K56Ac is still open.

4.2.4 Other Modifications

Histone H3 trimethylation at lysin 36(H3K36Me3) is known to be correlated with expression in the mid-coding region of the gene, in YPD conditions. The methylation of lysin 36 is mediated by the methyltransferase Set2, in collaboration with elongating forms of Pol2 (Martin and Zhang, 2005). Here, like in the previous two modifications, we would like to check how this modification response to stress and what is the connection between this response and the changes in the genes expression level.

Before we answer these questions we need to valid the quality of the data by checking the response of H3K36Me3 in YPD conditions. When we compare Type 1 time 0 results with data from other researches we need to remember that in our experiment we hybridized the IP of

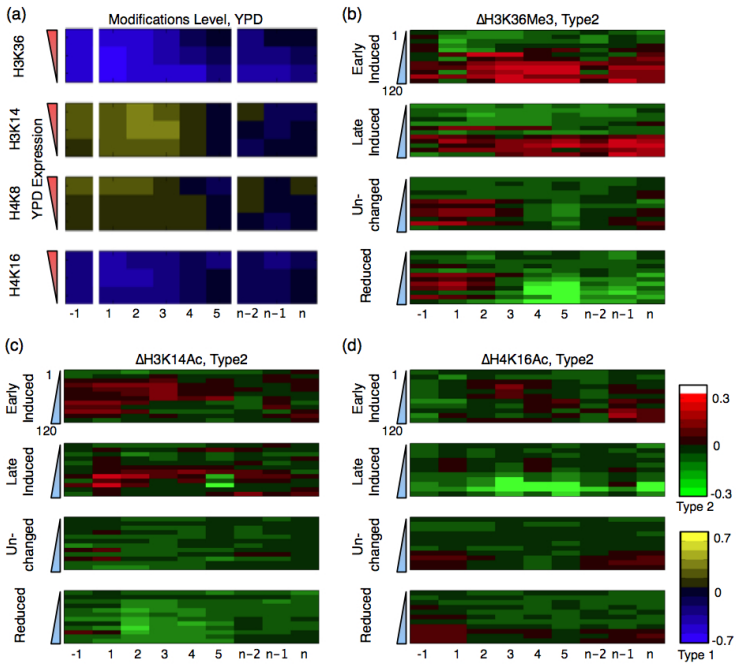

Figure 4.5: **Various modifications results**:
(a) Relationship of four different modifications level at time 0 and transcription level.
(b-d) Relationship of three modifications response to diamide to the expression response.

time 0 against the occupancy level of the same time. Therefore we can not compare the value of the modification, just the correlation with the gene expression or with the position on the gene. Looking at our results (Figure 4.5 a) we can see how the modification level increases in the highly expression group on the mid-coding region and in the 3' end of the gene. These results completely correlate with our former knowledge about this modification.

Examination of H3K36Me3 response to the diamide stress shows that the modification level of the induced genes groups rise over all the gene regions, as a result of the stress (Figure 4.5 (b)). This modification time response is slower then the transcription response and therefore it can not have any effect on Pol2 recruitment. On the other hand, since Pol2 response earlier, maybe the polymerase causes this modification. It is probably true since Pol2 is needed for the synthesis of this modification.

The acetylations of lysin 14 in histone H3 (H3K14Ac) and lysin 8 in histone H4 (H4K8Ac) does not change so much (Figure 4.5 (c)). We do see a small increase near the 5' end of both, but nothing extreme. The problem is that while H3K14Ac is sometimes associated with active genes, H4K8Ac is supposed to be anticorrelated with expression, so the increasing in the modification level is not clear, but since the increasing is very small we can't conclude anything using these results.

Histone H4 acetylation of lysin 16 (H4K16Ac) is known to be anti-correlated with expression, near the 5' end of the gene (Kurdistani et al., 2004). Our results show that this anti-correlation exists also when the expression pattern changes in response to diamide stress (Figure 4.5 (d)). Although this modification is independent in the polymerase, its response time is longer than the gene expression. Therefore, here again if there is any connection between the responses of the polymerase and the modification, the polymerase caused the modification and

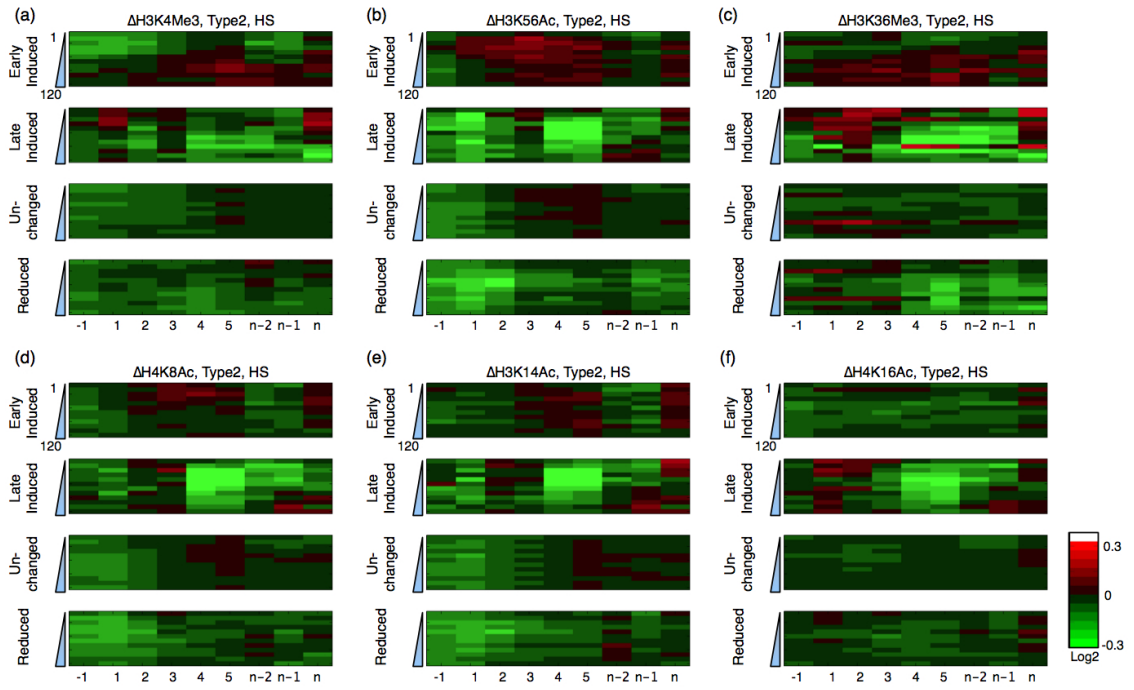

Figure 4.6: **Heat-Shock results**:**(a-f)** Changes in all six modification in response to heat shock stress.

not vise versa.

4.3 Changes of the Modification in Response to Heat Shock and Polymerase Shutdown data

The next two experiments were made under a heat shock stress (shift to 37°C). The first experiment checks the changes in the modification level in response to heat shock stress. Once we know the patterns of this response we want to repeat the same experiment with inactive polymerase. In order to shutdown the polymerase we used a temperature-sensitive rpb1-1 allele of the Rpo21 subunit of RNA polymerase (Nonet et al., 1987). This experiment returns the changes in the modification level in response to heat shock stress, without the changes that were caused by the polymerase. When we compare the two experiments we can find which changes were caused by the polymerase and which were caused by the stress. Answering this question can be very important to find the relation between the modifications and the polymerase.

4.3.1 Heat Shock

The first experiment is identical to the diamide experiment with one change; we are using a heat-shock stress instead of diamide stress. The heat-shock stress it also a very common stress that causes a wide response. The problem is that the response to heat shock is weaker and slower then the response to diamide, and therefore it is difficult to distinguish between the responses to the heat-shock stress and random changes.

Looking at the modifications response to the heat-shock stress we can see two major patterns: the first contains: H3K4Me3, H3K56Ac and H3K36Me3, and the second contains the other three. The first three modifications are usually associated with expression in steady state, and as we saw it is true also for their response to diamide. Looking at the changes of these modifications levels in response to heat shock, we can see that here also there is a correlation between the modification and gene expression. As we done earlier, we divided the genes into four groups: reduced, unchanged, early induced and late induced. The results of the first three groups are similar to the diamide experiment, meaning that the level of the three modifications: decreased in the reduced genes group, does not change in the unchanged group and increases in the early increased group. In addition, the early induced genes behave exactly like in the diamide experiment. Meaning that H356Ac responds faster than the other two modifications, and that H3K4Me3 increases in the mid-coding region of the gene, H3K56Ac increases faster near the TSS, and H3K36Me3 increases over all the gene regions. So, the only group that responses differently in the two experiments is the group of the late induced genes.

The other group contains the acetylations of: H3K14, H4K8 and H4K16. The results of this experiment, like the diamide experiment, does not look exactly like we expected from modifications that suppose to be anti-correlated with expression. Here again, if we are ignoring the group of the late induced genes the results are exactly like the diamide stress. The difference is that for these modifications, the group of the the late induced genes fits better with our expectations in the heat shock experiment.

The question is, if the results of these two experiments are correlated (4.1), and the changes in the modification level are usually similar, why is there such a big difference in the group of late induced genes? One possibility is that there are some bad arrays in this group, and therefore we can see very similar results for all six modifications. Other possibility is that as a result of the small threshold that we chose for the heat-shock stress (see 4.1) there are a lot of genes in this group that are not supposed to be there. The question is why this small threshold does not influence the early induced genes? It seems that the response of 1.5 fold during less than 4 minutes is different than the same response during 30 min. In the first case, since the response is so fast after the stress, it is probably caused by the stress; in the second case, it seems like a random change in the modification level. Therefore, reducing the threshold to 1.5 fold effects just the late induced genes group.

In summary, the changes in modifications level in response to heat-shock are similar to their changes in response to diamide. There is one problematic group, that is probably caused by the difficulty to distinguish between the response to the stress and random changes. This difficulty is caused by the fact that the genes response to heat shock is weaker.

4.3.2 Polymerase Shutdown

The following experiment repeated on the heat shock experiment with the temperature sensitive rpb1-1 mutant in Pol2. The purpose of this experiment was to distinguish between the changes in the modification pattern that was caused by Pol2 and those which were caused by the stress. Finding this distinction can help us to understand the connection between the polymerase and the modifications.In order to use this experiment we needed to compare its results with the results of the previous experiment of wild type yeast in heat-shock conditions.

Our original intentions were to look carefully at the first few minutes, where we had a lot of data with a very high resolution, and to search for changes between the two experiments.

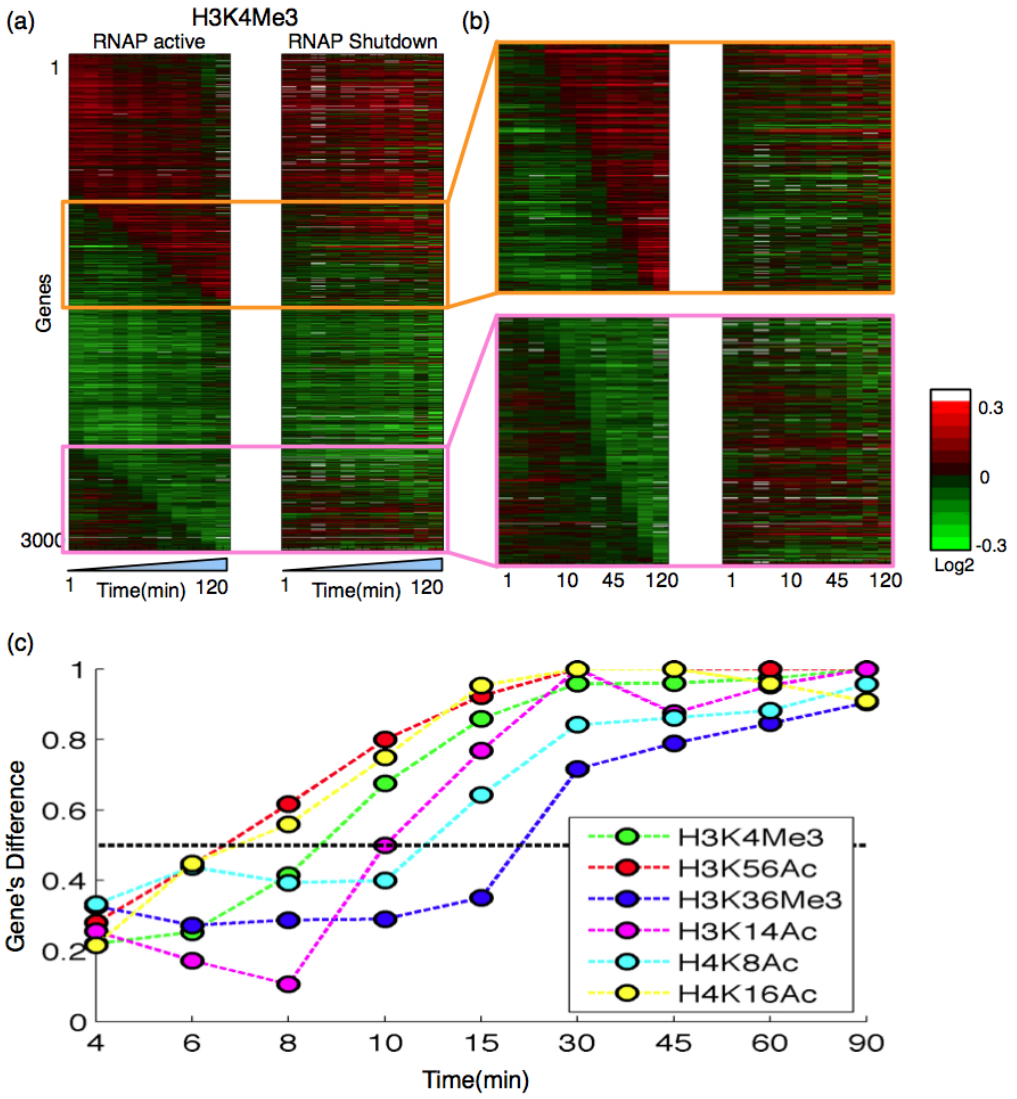

Figure 4.7: **Comparison of the heat-shock experiments**: **(a)** Compare the results of the two heat-shock experiments. The left column is the experiment with regular R NAP, the right column is the experiment that use temperature sensitive RNAP, in other words with polymerase shutdown. Both column ordered according to the $h1$ and $t1$ parameters of the left column. We can see that the column are very similar for low value of $t1$ and are not similar at all for large values of $t1$. **(b)** Focus on the regions with an high $t1$ value. The upper figure is for positive $h1$ and high $t1$ value, the bottom is for negative $h1$ and high $t1$ value. In this figure we can see that the difference between the two experiments start when $t1$ is between 10-15 min. **(c)** Shows for each modification how much genes have differences between the two experiments in each time-point. The definition of differnces between gene is change of more than 10 minutes in the $t1$ value, or change of the sign in $h1$ value.

The problem was that in a research that made by Dr. Rando's lab, in collaboration with our lab, they found some new facts about the polymerase response to the stress (Kim et al., 2010). They found that the polymerase responsed to heat-shock stress after 15 minutes and continued to associate with most of the genes for more than an hour after the heat shock. According to this, changes between the two experiments can be found after only several minutes. Therefore, comparing the first few minutes can not help us separate the responses that were caused by the stress from those that were caused by the polymerase. Since we have enough data we can still use these two experiments, but we need to consider the fact that the polymerase does not completely disappear, even after an hour.

As a first step of this comparison we looked at the results of both experiments ordered by the transition parameter ($t1$) value of the wild type experiment (Figure 4.7 (a)). We saw that for all the nucleosomes with a low $t1$ value and the modification level looked the same in both experiments. However, as the $t1$ value become higher there are more differences between the two experiments. In other words, nucleosomes, which their modification level changed quickly after the stress, behaved in the same way in both experiments, while nucleosomes with slower responses, behaved differently. A possible reason for the dependency on the $t1$ value is because of Pol2 shutdown time response. Nucleosomes that rapidly changed their modification level looked the same in both experiments, since the changes in the modification level occured before the changes in the polymerase occupancy. In contrast, nucleosomes with slower response changed their modification level after the polymerase shutdown, and therefore their modification level could be influenced by the lack of the polymerase.

This phenomenon of a correlation between the $t1$ value and the differences between the experiments, repeated in all six modifications. However, it seems that we could distinguish between the different modifications by the $t1$ value that caused differences between the two experiments. This distinction is important because a low $t1$ value indicates that this modification depended on the polymerase, and therefore it responded faster to the polymerase shutdown; while a modification with a high $t1$ value may not have been depended on the polymerase at all. In the last case, the changes in the modification pattern are results of new stress caused by the lack of the polymerase.

The next step was to find the time-point where to modification showed a significant differences between the two experiments, for each modification . In order to find this point, we needed to check how many nucleosomes are different between the two experiments for each time point. The definition of different nucleosomes is nucleosomes with a difference of more than ten minutes between their $t1$ value, or nucleosomes with different sign of the peak parameter ($h1$). Different $h1$ signs means that the modification level decreases in one experiment and increases in the other. Differences of more than 50% of the nucleosomes defined as a significant change between the two experiments. When looking at the results (Figure 4.7 (c)) we can see that H3K56Ac and H4K16Ac respond very fast, and reach the significant point in 6-8 min. In contrast, H3K36Me3 responds very slowly, and reaches to the same point after 15-30 min. The interesting fact is that H3K36Me3 known to be associated with polymerase unlike H3K56Ac and H4K16Ac. In the next chapter we will discuss the implications of these differences.

Chapter 5

Discussion

Histone modifications are, among other things, an essential part of the transcription process. The purpose of our research was to find out more about the relationship between histone modifications and transcription. We are particularly interested in the relation between RNA polymerase 2 (Pol2) and histone modifications, and how the modifications pattern changes in response to stress. Our findings show that the initial modification changes affected by Pol2, while Pol2 initial response is not affected by changes in the modification level. We also found that the behavior of most of the modifications in response to stress was consistent with the observations made about their behavior in mid-log conditions, i.e. the changes in the modification level that associated with gene expression were correlated with the changes in the gene expression.

Our study included a very large assay that checked the behavior of gene expression, Pol2 occupancy, 6 modifications and nucleosome occupancy in several conditions. Using some data analysis methods we reorganized the data in a format that is better suited to answer the research questions. We repeated this assay under 3 different conditions: (1) diamide, (2) heat shock with wild-type polymerase and (3) heat shock with Pol2 that genetically inactivated using the rpb1-1 temperature-sensitive mutation.

One major problem with histone modification data is the variability of histone modification in mid log growing cells. This variability may be one of the reasons for the differences in the understanding the modifications pattern and their goal, between the different researches (Rando, 2007). In our case, this variability adds noise to Type 2 data since it's normalization is against mid log growing cells. This is one of the reasons for the reproducibility differences between Type 1 and Type 2 and between Type 1 time 0 to the other time points (3.1.1). The question of the modifications variety in mid log is still open and waiting for a further research.

One of the goals of the reorganization of the data is to use averages to track general trends. Looking at the expression moving average trend (Figure 3.1.5 b) we can see how much of the data-points are far from the curve, meaning that the variability of the data is very high. The method of using averages is very common in this type of data, but we still need to remember that there are some problems with this method.

The first question that we asked is whether Pol2 affects all of these modifications. To answer this question we compared the results of the two heat shock experiments. The comparison showed that all the modification changed their response at a range of between 6 and 30 minutes. According to a new research the Pol2 shutdown, using the rpb1-1 temperature-sensitive mutation, is a very slow process, and it's first impact on the genes starts after more than 15 min (Kim et al., 2010). Therefore, we conclude that the modification affected by the polymerase,

otherwise, Pol2 shutdown would not affect them so fast.

Our next goal was to find more about the mutual influence between Pol2 and histone modification. To do this we checked the modification changes in response to diamide stress, and compared them with the transcription changes. The results reported that most of the changes are similar to our expectation. For example, modification that are associated with gene expression, changed in response to the stress in correlation with gene expression. Comparison of the diamide experiment with the heat-shock wild-type experiments, shows that the response to the two stresses are similar. There are some differences between the responses, but it seems that they are caused by technical problems, given to a weaker response to the heat shock experiment.

Assuming that Pol2 play a role in the synthesis of new modifications, and that histone modifications have a role in transcription regulation, we wanted to gain insight into the temporal and causal sequence of events during the initial stages of stress response. The question is whether some modifications occur before Pol2 and play a role in its recruitment, or that all modification changes as a results of Pol2 changes. Looking at the modifications changes in response the diamide stress we can see that the three modifications: H3K4Me3, H3K36Me3 and H4K16Ac had changed after the onset of transcription, and therefore it is clear that these modifications do not recruit Pol2 (Figure 5). H3K56AC, unlike the other modifications, responds very fast to stress, therefore this modification may have a role in the recruitment of Pol2, or is introduce concurrently with the initial transcription round.

Comparing the results of H3K4Me3 and H3K36Me3 we can see that H3K36Me3 response is always slower. In the diamide experiment we can see that both modification have similar response, but H3K36Me3 responds 10 minutes after H3K4Me3. The comparison of the two heat shock experiment also show that H3K36Me3 responds slower to the polymerase shutdown. These results support the association of H3K36Me3 with the elongation part of the transcription, while H3K4Me3 associated with the initiation. Since the elongation came after the intuition, all the responses of H3K36Me3 are slower. It's also agrees with the model that Set2, H3K36 methyltransferase, displaces Set1, H3K4 methyltransferase, on the polymerase.

The modifications with the fastest response to Pol2 shutdown were H3K56Ac and H4K16Ac. These two modifications were our candidates for modifications that had an initial affect on the polymerase. The fast response of these modifications to Pol2 shutdown may indicate that Pol2 had a strong affect on these modifications. Therefore, it seems unlikely that these modifications will cause the recruitment of the polymerase.

Regarding H3K56Ac, pervious research conducted in our lab, showed that its role is to mark new nucleosomes (Kaplan et al., 2008b). It seems that when the polymerase arrives to the genes it immediately causes a nucleosome turnover near the TSS, therefore this modification responds very fast to stress. The reason for this may be that the polymerase is still associated with the genome more than 15 minutes after the shutdown, but the recruitment of new polymerase to the genome stops earlier. Since this modification associated with new nucleosomes, it seems that it effected mostly by the recruitment of new polymerase, and therefore responds as fast as it does to the polymerase shutdown. The case of H4K16Ac is more questionable since these modification responds very fast to Pol2 shutdown, but on the other hand it responds very slowly to the stress.

These results provide a broad perspective about the relations between Pol2 and histone modification, and shows, for the specific modifications that we checked, that Pol2 is directly affects the modification, and that the modification changes in response to stress agrees with our former knowledge about them. We also saw how the timing of this behavior helps to understand

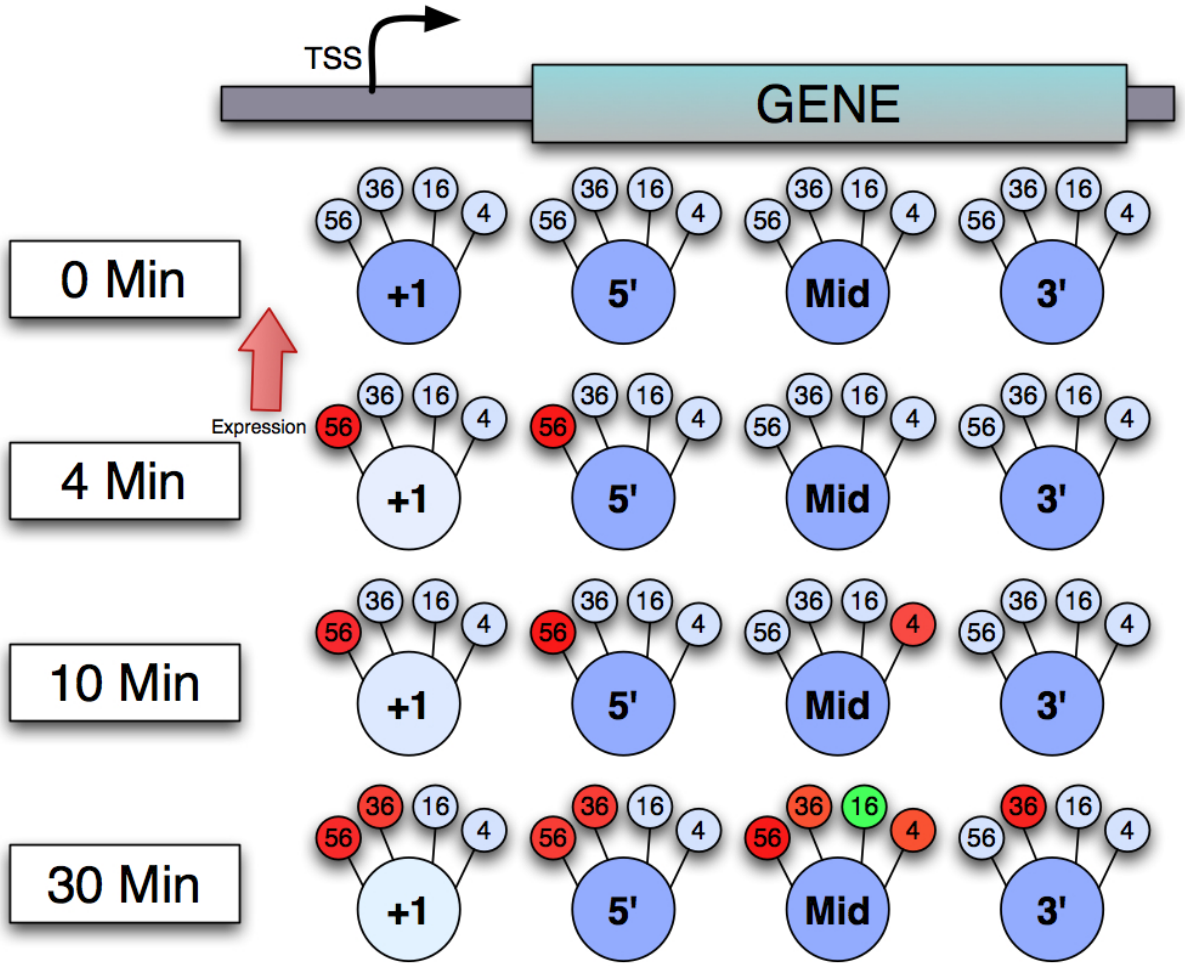

Figure 5.1: **Summary of the modifications behavior in response to stress**:
0 Min All nucleosome in basal level. **4 Min** The expression level and H3K56Ac near the 5' end are changing. The occupancy level of +1 nucleosomes decrease. **10 Min** H3K4Me3 in the mid-coding region, and H3K56Ac are changing. **30 Min** H3K36Me3 and H4K16Ac are changing.

the role of specific modifications.

An open question regarding this issue is whether the modifications affects Pol2. Here, we assumed that there is mutual influence between histone modifications and Pol2. This assumption based on a former knowledge about this specific six modifications. The question remaining is whether this assumption is true for all the modification. In addition, even for our 6 modifications, showing a direct affect of the modification on Pol2, can be very interesting.

Bibliography

Gene O Bryant, Vidya Prabhu, Monique Floer, Xin Wang, Dan Spagna, David Schreiber, and Mark Ptashne. Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS Biology*, 6(12):2928–39, 2008.

G Chechik and D Koller. Timing of gene expression responses to environmental changes. *Journal of Computational Biology*, 16(2):279–90, 2009.

W Cleveland. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, Jan 1988.

C Dhalluin, J E Carlson, L Zeng, C He, A K Aggarwal, and M M Zhou. Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, 399(6735):491–6, 1999.

Michael F Dion, Tommy Kaplan, Minkyu Kim, Stephen Buratowski, Nir Friedman, and Oliver J Rando. Dynamics of replication-independent histone turnover in budding yeast. *Science*, 315(5817):1405–8, 2007.

M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25):14863–8, 1998.

Junhong Han, Hui Zhou, Bruce Horazdovsky, Kangling Zhang, Rui-Ming Xu, and Zhiguo Zhang. Rtt109 acetylates histone h3 lysine 56 and functions in DNA replication. *Science*, 315(5812):653–5, 2007.

L Hong, G P Schroth, H R Matthews, P Yau, and E M Bradbury. Studies of the DNA binding properties of histone h4 amino terminus. thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the h4 "tail" to DNA. *J Biol Chem*, 268(1):305–14, 1993.

N Kaplan, IK Moore, Y Fondufe-Mittendorf, AJ Gossett, D Tillo, Y Field, EM LeProust, TR Hughes, JD Lieb, and J Widom. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2008a.

T Kaplan, CL Liu, JA Erkmann, J Holik, M Grunstein, PD Kaufman, N Friedman, and OJ Rando. Cell cycle–and chaperone-mediated regulation of h3k56ac incorporation in yeast. *PLoS Genetics*, 4(11), 2008b.

Tae Soo Kim, Chih Long Liu, Moran Yassour, John Holik, Nir Friedman, Stephen Buratowski, and Oliver J Rando. Rna polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast. pages 1–13, Aug 2010.

Siavash K Kurdistani, Saeed Tavazoie, and Michael Grunstein. *Cell*, 117(6):721–33, 2004.

CL Liu, T Kaplan, M Kim, S Buratowski, SL Schreiber, N Friedman, and OJ Rando. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biology*, 3(10):1753, 2005.

K Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–60, 1997.

Cyrus Martin and Yi Zhang. The diverse functions of histone lysine methylation. *Nat Rev Mol Cell Biol*, 6(11):838–49, 2005.

Hiroshi Masumoto, David Hawke, Ryuji Kobayashi, and Alain Verreault. A role for cell-cycle-regulated histone h3 lysine 56 acetylation in the DNA damage response. *Nature*, 436(7048): 294–8, 2005.

Catherine B Millar and Michael Grunstein. Genome-wide patterns of histone modifications in yeast. *Nat Rev Mol Cell Biol*, 7(9):657–66, 2006.

Huck Hui Ng, François Robert, Richard A Young, and Kevin Struhl. Targeted recruitment of set1 histone methylase by elongating Pol2 provides a localized mark and memory of recent transcriptional activity. *Mol Cell*, 11(3):709–19, 2003.

M Nonet, C Scafe, J Sexton, and R Young. Eucaryotic rna polymerase conditional mutant that rapidly ceases mrna synthesis. *Mol Cell Biol*, 7(5):1602–11, 1987.

Dmitry K Pokholok, Christopher T Harbison, Stuart Levine, Megan Cole, Nancy M Hannett, Tong Ihn Lee, George W Bell, Kimberly Walker, P Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K Gifford, and Richard A Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–27, 2005.

Marta Radman-Livaja, Chih Long Liu, Nir Friedman, Stuart L Schreiber, and Oliver J Rando. Replication and active demethylation represent partially overlapping mechanisms for erasure of h3k4me3 in budding yeast. *PLoS Genetics*, 6(2):e1000837, 2010.

Oliver J Rando. Global patterns of histone modifications. *Curr Opin Genet Dev*, 17(2):94–9, 2007.

Helena Santos-Rosa, Robert Schneider, Andrew J Bannister, Julia Sherriff, Bradley E Bernstein, N C Tolga Emre, Stuart L Schreiber, Jane Mellor, and Tony Kouzarides. Active genes are tri-methylated at k4 of histone h3. *Nature*, 419(6905):407–11, 2002.

Jessica Schneider, Pratibha Bajwa, Farley C Johnson, Sukesh R Bhaumik, and Ali Shilatifard. Rtt109 is required for proper h3k56 acetylation: a chromatin mark associated with the elongating RNA polymerase2. *J Biol Chem*, 281(49):37270–4, 2006.

Robert J Sims and Danny Reinberg. Histone h3 lys 4 methylation: caught in a bind? *Genes Dev*, 20(20):2779–86, 2006.

Robert J Sims and Danny Reinberg. Processing the h3k36me3 signature. *Nat Genet*, 41(3): 270–1, 2009.

Robert J Sims, Scott Millhouse, Chi-Fu Chen, Brian A Lewis, Hediye Erdjument-Bromage, Paul Tempst, James L Manley, and Danny Reinberg. Recognition of trimethylated histone h3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mrna splicing. *Mol Cell*, 28(4):665–76, 2007.

B D Strahl and C D Allis. The language of covalent histone modifications. *Nature*, 403(6765): 41–5, 2000.

Bryan M Turner. *Cell*, 111(3):285–91, 2002.

Alejandro Vaquero, Michael B Scher, Dong Hoon Lee, Ann Sutton, Hwei-Ling Cheng, Frederick W Alt, Lourdes Serrano, Rolf Sternglanz, and Danny Reinberg. Sirt2 is a histone deacetylase with preference for histone h4 lys 16 during mitosis. *Genes Dev*, 20(10):1256–61, 2006.

Assaf Weiner, Amanda Hughes, Moran Yassour, Oliver J Rando, and Nir Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res*, 20(1):90–100, 2010.

Moran Yassour, Tommy Kaplan, Hunter B Fraser, Joshua Z Levin, Jenna Pfiffner, Xian Adiconis, Gary Schroth, Shujun Luo, Irina Khrebtukova, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, and Aviv Regev. *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci USA*, 106(9): 3264–9, 2009.