Clustering of Chromatin Signatures During Early Fly Embryogenesis Identifies GAF as a Zelda-like Pioneer Factor

ניתוח חישובי וключение חתימות כרומטין בזמן התפתחות העוברית המוקדמת בזבובי פירות מוצא את החלבון GAF כחלבון חלוץ

Arbel Moshe

Supervised by Dr. Tommy Kaplan

Thesis for Master's degree in Natural Science

18/12/2016
Abstract

The protein Zelda was shown to play a key role in early Drosophila development, binding thousands of promoters and enhancers prior to maternal-to-zygotic transition (MZT), and marking them for transcriptional activation. Recently, we showed that Zelda acts through specific chromatin patterns of histone modifications to mark developmental enhancers and active promoters. Intriguingly, some Zelda sites still maintain these chromatin patterns in Drosophila embryos lacking maternal Zelda protein. This suggests that additional Zelda-like pioneer factors may act in early fly embryos.

To test this hypothesis, we developed a computational method to analyze the chromatin landscape surrounding early Zelda peaks, based on multi channel spectral clustering. This allowed us to characterize the temporal dynamic through MZT (mitotic cycles 8-14). Specifically, we focused on H3K4me1, H3K4me3, H3K18ac, H3K27ac, and H3K27me3 and identified three different classes of chromatin signatures, matching “Enhancers”, “Promoters”, and “Transiently-bound” Zelda peaks.

We then used these chromatin patterns to further scan the genome and identify novel loci, with no Zelda binding, that show similar chromatin dynamics, resulting with >1000 putative Zelda-independent enhancers.

These regions were found to be enriched with GAGA factor (GAF, Trl), and are located near early genes and their enhancers, suggesting that GAF - together with Zelda - play an important role in activating the zygotic genome.
תקציר

מחקר קורדורום הרוא, שלחובן זלדה שין חולק חשב בהתחפשות שבירה של מחלקות של אנוסימוורפיה, Drosophila melanogaster, החלק הנוור לשולפים פרומוטוריים אנוגנורפים.

לפי תופעת התערבות העורית, MZT, של חולקב קרום ולח, המאבטת את😉 קיוור של שונות שבירה של אטומי והפרッシורים פעלים. לדמות אתchlnehmenם מאבטת את الكرום 상יכים על אטום התבנית כמותים בין תבנית מבית של בוב פיתוח שחרים וחלון זלדה. השונדנו מתחקק את התאשה, ששמונה תלבושות צלד, K4me1, K4me3, K18ac, של חולקב זלדה, ש unidad budקו החלון התאשה והזלת.

על מntag לבן את התאשה, הינו מתוח השיבה ליצוח את מבנה הקרום המסיבי אתてる קיוור מוקדמים של זלדה, הקרתם על " NST שילטון่ תלבושות תלבושות" ובעדות אנוגנורפים זה (Mazier, holcomb 1-8). בעדם התמגנת ב H3K4me1, H3K4me3, H3K18ac, H3K27ac, H3K27me3 אופייני שלוש מחלקות של זלדה, K4me1, K4me3, K18ac, K27ac, K27me3 של התיאטמוס קומפרס שחלשה המחלקות והאנוגנורפים, פרומוטורים, והיוו זלדה שונה של התיאטמוס קומפרס. שחלשה המחלקות והאנוגנורפים, פרומוטורים, והיוו,

עדימה

ללאח מכ, השמתוונת אוות התיאטמוס קומפרס, וסיקור את הגנום בקדמת על מntag חזרה אוותי נטפסים עם אוות התיאטמוס, או ללא קיוור של החלון זלדה. לא מצאנו מעל 1000 אטרים גנומיים עם מבנה קורמורום הזת, או ללא קיוור של זלדה._ATOMי, התוויות

כרצף בקדמה, התאלה נשיאים בחלובancers, GAF, וממוקמים בדומה של בניים התאשה. כלנו, GAF-1, לתד זלדה, שי חולק משמעוני, בחפ溃疡 הגנום בין תבנית כ boa10,

משיעים של-
Acknowledgements

I would like to express my dear appreciation and thank my supervisor Tommy Kaplan who accepted me as a part time student with no Computational Biology background. He was my source of knowledge in a new field I started to research, and helped me all along. I learned from him a lot throughout my Thesis, and enjoyed working with him.

I would also like to thank Melissa Harrison from the Department of Biomolecular Chemistry in the University of Wisconsin for the collaboration in the research of early development of Drosophila melanogaster.

Last, I would like to thank my family and friends, who supported me all the way.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>תקציר</td>
<td>3</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>4</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>5</td>
</tr>
<tr>
<td>Introduction</td>
<td>6</td>
</tr>
<tr>
<td>Gene Expression and Enhancers</td>
<td>6</td>
</tr>
<tr>
<td>Chromatin</td>
<td>6</td>
</tr>
<tr>
<td>Drosophila Embryonic Development</td>
<td>7</td>
</tr>
<tr>
<td>The role of Zelda in early Drosophila development</td>
<td>8</td>
</tr>
<tr>
<td>Results</td>
<td>12</td>
</tr>
<tr>
<td>Zelda Peaks Vary in their Chromatin Signatures</td>
<td>12</td>
</tr>
<tr>
<td>Chromatin-based orientation of early Zelda bindings</td>
<td>15</td>
</tr>
<tr>
<td>Chromatin-based clustering of early Zelda binding sites</td>
<td>17</td>
</tr>
<tr>
<td>Additional characteristics of the three clusters of Zelda peaks</td>
<td>19</td>
</tr>
<tr>
<td>Chromatin-based identification of Zelda like chromatin patterns</td>
<td>24</td>
</tr>
<tr>
<td>Motif Analysis</td>
<td>28</td>
</tr>
<tr>
<td>GAF Factor</td>
<td>30</td>
</tr>
<tr>
<td>Chromatin accessibility of putative enhancer (FAIRE)</td>
<td>32</td>
</tr>
<tr>
<td>Discussion</td>
<td>35</td>
</tr>
<tr>
<td>Methods</td>
<td>37</td>
</tr>
<tr>
<td>Joint Spectral Clustering followed K-means</td>
<td>37</td>
</tr>
<tr>
<td>Cross Correlating different clusters</td>
<td>39</td>
</tr>
<tr>
<td>References</td>
<td>41</td>
</tr>
</tbody>
</table>
Introduction

Gene Expression and Enhancers

Gene expression is the process by which information packed inside the DNA sequence of a gene is transcribed by RNA polymerases to produce its RNA copies. According to the central dogma of molecular biology, this step is usually followed by translation of the mRNA molecules into proteins (by the ribosome) [Crick et al. 1970].

The process of transcription is vital in all living organisms, and is tightly regulated by multiple mechanisms, including the packaging of DNA into chromatin, and the differential expression of transcription factors - regulatory proteins that bind the DNA and recruit the transcriptional machinery (RNA polymerases). There are two broad classes of regulatory DNA regions: Promoters, that are located near the Transcription Start Site (TSS) of genes; and Enhancers, that act from afar and are often found up to 1Mb away from the genes they regulate. The 3D packaging of DNA into chromatin allows enhancers to be far from the gene in a linear manner, but can be in close spatial proximity to the gene. Enhancers and promoters are basic and fundamental mechanisms that regulate transcription, and thus are of main interest.

Chromatin

In eukaryotic cells, The DNA is wrapped around special proteins that are called nucleosomes, to form chromatin. The main purpose of this packaging is to fit the 2 meter long DNA inside the cell. In addition, DNA packaging is used differentially to control in what conditions and cell types a gene is more accessible - and active - and in which conditions it is more condensed and suppressed. Each nucleosome is consisted of eight proteins (histones), to form the basic structure of packaged DNA. There are two of each of the core histones H2A, H2B, H3, H4.
Like we previously mentioned, besides transcription factors, chromatin has a main role in transcriptional regulation. The histones that form the nucleosomes, which are the basic units of chromatin, can undergo various modifications. There are different types of covalent modifications including acetylation, methylation, phosphorylation etc., which occur at different residues of the nucleosome. These histone modifications influence various processes along the DNA. For example, H3K4me1 and H3K27ac (mono-methylation of Lysine 4 and acetylation of Lysine 27 in Histone H3) are found at nucleosomes carrying active enhancer regions, while H3K27me3 is known to be repressive mark of enhancers and coding regions [Heintzman et al. 2007; Creyghton et al. 2010]. These modifications control which regulatory regions are active and which are not, and by thus influence which genes will be transcribed. But how were they formed in the first place?

**Drosophila Embryonic Development**

In mammals, early developmental stages begin with maternal proteins and RNA that control the first hours in the fertilized egg. At this stage, these proteins control the first wave of zygotic expression and direct the first mitotic divisions [Tadros, Lipshitz, 2009; Newport, Kirschner, 1982]. After that the embryo undergoes a process called Maternal-to-Zygotic Transition (MZT), in which the zygotic genome is activated and takes control of mRNA and protein production. Finally, maternal mRNA and proteins are degraded.

In the fruit fly *Drosophila melanogaster*, embryonic development is characterized by a series of 13 rapid replication cycles, occurring during the first two hours after fertilization. The division of cells slows at the 14th mitotic cycle, and zygotic transcription initiates. This marks the end of the *Drosophila melanogaster* Maternal-to-Zygotic Transition. This process is crucial for the normal development of the embryo and is tightly controlled by the strictly regulated activation, both in time and space, of a cascade of transcription factors [Li et al. 2008, MacArthur et al. 2009]. This is controlled by multiple molecular mechanisms that include chromatin, nucleosomes, DNA accessibility, steric hindrance between DNA binding proteins and more.
The role of Zelda in early Drosophila development

Previous research showed that many of the early transcribed genes in Drosophila embryos contain a specific DNA motif of 7bp (termed the TAGteam motif) CAGGTAG within their regulatory regions, including both promoters and enhancers [ten Bosch et al. 2006; Bradley et al. 2010; Harrison et al. 2010]. These were identified to be the binding site of the zinc-finger transcription factor Zelda (vielfaltig, vfl) [Liang et al. 2008]. Following studies, by us and others, showed that Zelda is present in the embryonic nucleus as early as mitotic cycle 2 and binds thousands of genomic loci, including the promoters and regulatory regions of thousands of early developmental genes [Liang et al. 2008; Li et al. 2008; Harrison et al. 2011; Nien et al. 2011].

Computational studies suggested that Zelda acts as a Pioneer Factor, by binding mostly inaccessible DNA and making the chromatin accessible for other transcription factors to bind, thus enables regulation of gene expression during these stages. Zelda plays a pioneering role in shaping the chromatin landscape and accessibility of the genome for the transcriptional machinery (at promoters) and transcription factor binding (at promoters and enhancers) [Bradely et al. 2010; Bradely et al. 2012; Zaret and Carroll 2011].
Figure 1 – **Zelda's role during the MZT of the fruit-fly**

The role of Zelda in early embryonic development in *Drosophila*. Thousands of Zelda recognition sites (CAGGTAG) along the genome are bound by the protein Zelda, followed by acquisition of a unique pattern of chromatin marks, and transcriptional or regulatory activation, leading to expression or transcription factor bindings. (Adapted from Li et al., 2014)

Zelda marks thousands of genes for activation and enables the first transcriptional program of the developing embryo. Indeed, experimental studies showed that early Zelda binding is a predictor for later transcription factor binding and open chromatin at mitotic cycle 14 [Li et al. 2009; Li et al. 2011; Liang et al. 2008]. The entire set of molecular mechanisms by which Zelda functions to access the genome and mark it for activation is yet to be discovered, as is the role of additional proteins in this crucial stage in embryonic gene expression.
Several studies mapped the chromatin landscape in *Drosophila melanogaster* [Zeitlinger et al. 2007], most of which in cells or during very broad temporal windows. Of particular interest are few studies, in which early and carefully staged Drosophila embryos were used to portray Zelda binding or chromatin marks and multiple histone marks including H3K27ac, 27me3, 4me1, 4me3, 36me3, 18ac, etc. [Harrison et al. 2011].

Moreover, it had been showed that early Zelda binding result in certain chromatin marks (enriched H3K27ac, H3K4me1, H3K27me3, etc.) [Li et al. 2014]. As we and others showed, early Zelda binding (at mitotic cycle 8) results in open chromatin in late cycle 14 (using H3 ChIP-seq, DNase, and FAIRE-seq [Liang et al. 2008]). We also showed that these accessible regions are characterized by a set of active histone modifications, including H3K27ac, H3K4me1, and H3K4me3. [Li et al. 2014]

In addition, several studies showed “developmental abnormalities” in embryos lacking zygotic Zelda, and embryos with mutations in specific Zelda binding sites near early genes [ten Bosch et al. 2006; Liang et al. 2008]. Maternal embryonic mutants showed that promoters are still accessible (H3) subject to many more processes that occur on the DNA (transcription, accessibility), while enhancers are more affected by Zelda deletion and show different chromatin marks [Li et al. 2011; Li et al. 2014].

Research also compared gene expression, both in wild-type Drosophila embryos, and in Drosophila embryos lacking maternal Zelda (zldM). Unexpectedly, although there was a developmental difference between the wild-type embryos and the zldM, many genes that show early and strong Zelda binding both at promoter and enhancer regions, remained active even when lacking maternal Zelda, and moreover - their original Zelda sites still had open chromatin at cycle 14 [Li et al. 2014; Schulz et al. 2015].

This result raises an interesting question; might there be another mechanism besides Zelda, which makes the Drosophila embryos genome accessible for transcription factors binding. Maybe there is some other factor, besides Zelda, which is responsible for activating the genome, by making chromatin accessible, and thus enables other transcription factors to bind. To answer this question, we first chose to revisit the chromatin data, and develop a computational statistical model to characterize the histone marks and their
dynamics at Zelda binding sites. At a second stage, we use this computational model to scan the genome and identify additional genomic loci - that lack Zelda binding - yet show similar chromatin patterns and dynamics. We might find other biological mechanisms that regulate transcription during the embryonic development of Drosophila melanogaster.
Results

Zelda Peaks Vary in their Chromatin Signatures

Early Zelda peaks (mitotic cycle 8) were shown to be associated with open chromatin regions and transcription factors binding by mitotic cycle 14, towards the end of the Maternal-to-Zygotic Transition [Li et al. 2008; MacArthur et al. 2009; Harrison et al. 2011]. This transition is accompanied with alteration of the chromatin marks that surround these loci. We begin by portraying the chromatin landscape surrounding these early Zelda enhancers (Fig 1). As shown in Figure 2, many Zelda peaks are accompanied with strong H3Kme1 binding, and depletion of H3K27me3 and H3K36me3 [Harrison et al. 2011]. It may seem that Zelda peaks have a certain single and unique chromatin mark.

However, a close examination of some strong early Zelda sites (Figure 3) suggests that there is more than one typical chromatin signature. Specifically, shown are two Zelda peaks, one near the schnurri gene, and the other near the bitesize gene. The former serves as an enhancer of the schnurri gene, and is characterized with high H3K27me3 and H3K4me1 around the Zelda peak. The latter falls near the TSS of the bitesize gene, and is surrounded with H3K4me1, as well as H3K4me3 on one side of the Zelda peak, unlike the previous Zelda peak we described. Intriguingly, some early Zelda peaks seem to show no chromatin pattern whatsoever, and are practically indistinguishable from the histone modification patterns at their surroundings (the third peak, inside the coding region of an untranscribed gene).
Figure 2 – **Average chromatin marks surrounding Zelda peaks.**

Average chromatin marks (at mitotic cycle 14, Li et al, 2014) surrounding early (cycle 8) Zelda peaks. Clearly visible are relative enrichment near Zelda site for the enhancer mark H3K4me1, and depletion of the repressive H3K27me3.
Maybe different Zelda peaks have different histone modifications patterns, and only when looking at the average, it seems that we have a common pattern. We believe that there are different types of Zelda peaks, and each type has a certain combination of histone modifications pattern. If we can identify the different types of Zelda peaks, we can later search the chromatin signatures, and find similar loci, which lack Zelda but act similarly during the MZT.

We therefore turned to develop a computational model that will allow us to characterize the chromatin landscape along the early Drosophila genome, and its dynamics through MZT, and will later scan the genome and identify additional loci – independent of ZLD binding – that undergo similar chromatin dynamics during MZT.

**Chromatin-based orientation of early Zelda bindings**

Prior to clustering, we wanted to check which histone modifications are symmetric on both sides of the Zelda peaks, and which are not. For this, we performed an iterative orientation method based on only the different histone modifications. Our method is based only on chromatin, is because we believe that at least some of the modification are "one-sided", and this method could
help us find the real chromatin marks around the different Zelda peaks. Another optional method would be orientation based on nearest TSS. We decided to use the method, which is based on histone modifications only, because we believe that in case of distal enhancers, the nearest TSS wouldn't always be the one that is regulated by the enhancer. While most of the histone modifications had symmetrical formation, and didn't change much after orientation, H3K4me3 and H3K36me3 are indeed not symmetrical. But rather appear only on one side of the peak.

Figure 4 - Average chromatin marks after Zelda peak orientation
A plot of average chromatin marks of early (cycle 8) ZLD peaks, after orientation. We see that H3K4me3 is not a "symmetrical" mark, in contrast to what it appears like in Figure 2.
Chromatin-based clustering of early Zelda binding sites

In order to test if the different types of Zelda peaks have different combinatorial chromatin signatures, we decided to cluster the peaks, using only chromatin marks. This way we can sort Zelda peaks into different types according to their chromatin marks, allowing for to a genome-wide scan using several characteristic chromatin signatures.

It is not enough to use a probabilistic model, based on the average signal of the histone modification. The modifications we look at are not characterized by only high signal, but rather with a certain spatial signature (as seen in Figure 4). In addition, some of the modifications are not symmetrical. We therefore want a special method, in which we can give regions with the same patterns a high score, and not rely solely on the strength of the signal. In addition we are interested on the combination of the patterns of the different modifications, and not only on each particular modification by itself.

We therefore decided to cluster the data using spectral clustering [Dhillon et al. 2004]. Our distance function would account for both the pattern (two regions with same signal average but different “shapes” would not be considered similar), and on the combinations between the different modifications. An advantage of spectral clustering is the fact that the algorithm relies only on the clusters connectivity, and not on their spatial shape. We believe the different clusters have different combinatorial histone modification patterns, and using spectral clustering, we can identify the different combinations of histone modifications.

As mentioned before, our data consists of Zelda ChIP at three different time points (cycle 8, cycle 13, and late cycle 14), and nine histone modifications at cycles 8, 11, 13, and 14 (H3K4ac, H3K4me1, H3K4me3, H3K9ac, H3K18ac, H3K27ac, H3K27me3, H3K36me3, and H3K5ac), in addition to H3.

We selected the top 2,000 early Zelda peaks (at cycle 8), and for each peak we took a window of 10,000 base pairs of significant histone modifications,
at cycles 13 and 14. The histone modifications we used for clustering were (H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K18ac). Although H3K36me3 has some significant signature, we decided not to use it for clustering, because, in contrast to the previous histone modifications, H3K36me3 is active mainly in genes and transcription starting sites, and not around the Zelda peaks, and therefore is likely to have a varying distance from Zelda peaks. The other peak patterns are related to the Zelda peaks, would therefore look similar in all peaks.

We resulted in 10 vectors - one for each of the five modifications at each of the two time points, each of length of 1000 - 1x10 downsample of a 10Kb window surrounding each ZLD peak. To calculate the distance $W_{i,j}$ between each pair of vectors (or peaks), we used the RMSE:

$$dist_{i,j} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}$$

Next we constructed an adjacency matrix, based on the distance matrix $W$. Usually, there is one distance matrix, and a kernel function is applied on the distance matrix, resulting in an adjacency matrix. We chose to use the (standard) Gaussian kernel function $e^{-\frac{dist(i,j)^2}{2\sigma^2}}$, where $\sigma$ is a parameter we can change and set differently for the different modifications, for example if some vary more than others. To find a normalization method that would assign similar importance to all modifications, we arbitrarily set $\sigma$ for each signal type as the 10th percentile in the distribution of pairwise distances (further explained in Methods section).

We then applied the Gaussian kernel function for each one of the 10 adjacency matrices (each with its matching $\sigma$), which were joint to one adjacency matrix (further explained in Methods section), and performed spectral clustering, followed by K-Means (Methods). We picked K (number of clusters) using the eigengap (further explained in Methods section). The clustering process resulted in 3 main clusters, each with a unique combinatorial chromatin signature.
The first cluster is mainly enriched in H3K4me1 (starting at cycle 13) on both sides of the Zelda peak and in H3K4me3 (also, starting at cycle 13) on one side of the peak. The second cluster is enriched in both H3K4me1 (starting at cycle 13) and H3K27me3 (starting only at cycle 14) on both sides of the peak (H3K4me1 has "short" strands while H3K27me3 has long strands), suggesting these might be poised enhancers, which are probably expressed later. The third cluster has no chromatin marks at all, and is not enriched in any of the modifications peaks. The third cluster loci might be Zelda binding sites, with no significant meaning, which have no biological meaning.

![Graphs showing characteristic chromatin signatures for different clusters](image)

**Figure 5:** Three characteristic chromatin signatures found at ZLD sites
Average plots for each of the three clusters, discovered. Shown are 5 histone marks surround ZLD peaks. Marks are from 4th time point (mitotic cycle 14), peaks are from top 2000 early (MC8) ZLD peaks. There are 1032 peaks in Cluster 1, 465 in Cluster 2, and 503 in Cluster 3.

**Additional characteristics of the three clusters of Zelda peaks**

The clustering analysis resulted in three clusters, each displaying a unique combination of histone modifications. We would like to test whether the three
clusters act differently on other biological parameters. If so, it will strengthen our assumption that they indeed have a different biological meaning.

As mentioned earlier, some Zelda peaks are transient throughout MZT, with the ChIP-seq signals dropping significantly between mitotic cycles 8 and 14. Conversely, most peaks show strong constitutive binding throughout MZT [Harrison et al. 2011]. When comparing this property against the three classes, we observe a strong bias for transient binding among peaks from the 3rd class (Figure 6A).

In addition we know that Zelda serves as a pioneer factor [Zaret, Carroll. 2011; Harrison et al. 2011]. Early Zelda peaks (cycle 8) result in open chromatin in later stages (cycle 14), following transcription [Li et al. 2011; Schulz et al. 2015]. We wanted to test whether or not the 3 clusters differ in transcription factors binding at cycle 14. Using data of genome-wide transcription factor binding at cycle 14 [Li et al. 2011], we can see that Cluster 1 and Cluster 2 peaks bind on average more than 7 transcription factors at cycle 14 on average, compared to Cluster 3 peaks that bind only ~1 transcription factor on average (Figure 6C). This suggests that, in agreement with the transient binding of Class 3 peaks, they also facilitate the binding of fewer transcription factors at later stages.

Cluster 3 peaks are characterized by temporal Zelda binding, and no chromatin effect. Therefore we decided to call cluster 3 peaks Transiently Bound.
Figure 6 - **Biological characteristics of the three clusters**

(A) - Temporal dynamics of ZLD peaks for each cluster, defined as the ratio between ZLD ChIP strength, at mitotic cycles 8 and late 14 [Li et al. 2014]. In cluster 1(left) and Cluster 2 (center), about 40% of the ZLD peaks are increasing, 40% decreasing, and 20% have minor change. In contrast, more than 80% of Cluster 3 peaks (right) are decreasing. (B) – A violin plot of the peaks distance to nearest TSS for each cluster. A distance of 0 indicates the peak is within a coding sequence. (C) – Zelda peaks for Clusters 1 and 2 show strong localized enrichment for transcription factor binding of multiple TFs ( > 7 on average) in mitotic cycle 14, compared to
Cluster 3 (right) [Li et al. 2009]. (D) – Different expression patterns of ZLD peaks in the three clusters (Maternal and Zygotic, Early/Mid Zygotic, Late Zygotic and silent). Each peak’s corresponding gene is the closest. 95% of the promoter-like peaks are active, 90% of the enhancer-like peaks are active. On the other hand, 60% of the transiently-bound peaks are silent. (E) - Classification of ZLD peaks according to their corresponding segment, by Fillion et al. (2010). According to Fillion et al. (2010) blue mark poised enhancers, yellow and red mark active genes and black marks repressive chromatin. Most of Cluster 1 peaks are either yellow or red segments. Most of Cluster 2 peaks are blue or black, and Cluster 3 are mostly black.

In addition, we wanted to further understand the location of Zelda peaks, in relation to genes, in order to understand their biological purpose. We know that Cluster 1 peaks are enriched in H3K4me3 on one side, which makes us believe they might act as promoters, and therefore fall near transcription starting sites of active genes. We also know that Cluster 2 peaks are enriched in H3K4me1 flanked by H3K27me3 on both sides, which corresponds to poised enhancers. In order to check if our hypothesis is true, we assigned each Zelda peak its distance from TSS (0 - coding region). As shown in Figure 6B, Cluster 1 peaks are typical close to TSS, with an average distance of < 1Kb, while peaks from Clusters 2 and 3 further away. To summarize, Cluster 1 peaks are closest to TSSs, and rich with H3K4me3, so we label Cluster 1 peaks “Promoter-like” peaks. Similarly, we will denote Cluster 2 peaks as “Enhancer” peaks.

Another comparison we made, was of gene expression times of nearest genes to each Zelda peak, to see if there is any difference between the 3 clusters. Previously we have mapped gene to seven different groups based on their expression patterns: Maternal, Maternal and Zygotic, four zygotic times (early, mid, late, and later) and silent. For each one of the 2000 Zelda peaks, we looked at the gene expression time of the nearest TSS. As expected, most transiently bound peaks, are closest to silent genes. Promoters-like peaks are mainly closest to MZT / Early-mid Zygotic genes, and most of them are expressed early. Enhancers are mainly closest to Zygotic genes, at different times, and as can be seen, are late zygotic. The difference between the expression times of enhancers and promoters makes sense, because the promoters-like peaks have active chromatin marks (H3K4me3), while enhancers have poised chromatin marks (H3K27me3).
We also wanted to test our results, by comparing our model to another model, which segmented the drosophila genome to different regions. Filion et al. (2010) clustered the Drosophila Genome into five segments, based on ChIP data of 53 factors. The Black segment was shown to be repressive, the Yellow was shown to be mainly in active genes, and Blue in poised enhancers. We decided to compare our models, and test the distribution of the segments in each of our clusters. *Transiently Bound* peaks fall mainly in the black segment (silent), *Enhancers* fall mainly in the Blue and Black segments (poised and silent areas), and *Promoters* fall mainly in the Yellow segment (around active genes).

We managed to characterize in high resolution and a biologically meaningful way the typical chromatin signature of Zelda peaks. The 3 main classes are:

1. Promoter-Like: Zelda peaks rich in H3K4me1 and H3K4me3. Mainly near TSSs, and correspond to early expressed genes.
2. Enhancer-Like: Zelda peaks rich in H3K4me1 and H3k27me3. Further from TSSs and correspond to later expressed genes. We assume they act as poised enhancers.
3. Transiently Bound: Zelda peaks which bound to different loci which are not followed by transcription. Correspond to silent genes.

After resulting in 2 interesting combinatorial chromatin signatures, we can scan the genome for the exact signatures, and may find new loci, with the same chromatin marks, but without Zelda binding.
Chromatin-based identification of Zelda like chromatin patterns

We know that, although Zelda plays a major role in making chromatin accessible for other transcription factors during and after the MZT of the Drosophila melanogaster, most early genes are still expressed in Zelda mutants. This suggests that there might be some other mechanisms besides Zelda, maybe some other transcription factors like Zelda.

We believe that other "Zelda like" factors may also lead to similar chromatin signatures of active and regulatory chromatin. Because of that, we hope that if we would scan the genome, for the same chromatin patterns, we might find other loci, besides loci bound by Zelda, which also act as enhancers.

To do so, we cross-correlated the genome with the chromatin patterns of the Enhancers and Promoter-Like clusters. The patterns we used were the average patterns of each modification, for each cluster, as in Figure 5. We used the same set of histone modification marks as used to cluster (H3K27ac, H3K27me3, H3K4me1, H3K4me3, and H3K18ac) at cycle 14, with the same 10,000 base pairs window.

Our first model was to separately cross-correlate each histone modification with the entire genome, and to later compute an average of all modifications. This naïve analysis did not lead to satisfying results, as the ratio between the different histone modifications was not preserved in all loci, and resulted in loci, in which, separately, each histone modification had a similar signature, but the combined signature did not look at all like it should. For example, we could result in similar patterns for each modification, but the first modification would have very high values, and the second modification would have very low values, while vectors in the cluster had a similar height for both modifications. This was addressed by concatenating the histone modifications into one long vector, and cross-correlated this vector with the genome. The results now preserved the ratio between the modifications, but were still not good enough. The reason is that the correlation doesn't take histone modification height as a parameter, but only the shape. Any solution would need to account for the height of the modification as well. The closer
the height and the correlation are to the real vector, the higher the score should be.

To solve this issue, we added a "Scaling" function normalization that penalized the score obtained for each genomic loci. The factor was estimated based on original Zelda peaks, and reflected the probability that the average height of the modification is an average height of a vector of the corresponding cluster.

This method gives each site in the genome a score. We extracted the maxima points, and eliminated peaks of a distance of less than 500 bps (we took the peak with higher score out of the two). An ROC curve of our method (Figure 8A) shows that when taking the top 2,500 enhancer-like newly found peaks we managed to locate more than 80% of original early Zelda peaks (from enhancers cluster), and when taking the top 2,500 promoter-like we managed to locate more that 50% of the early Zelda peaks (from promoter-like cluster).
Figure 8 - **Model performance**  
(A) - ROC curves of our model. X axis corresponds percentage of genome covered, and y axis corresponds to percentage of original ZLD peaks covered by our model. By covering ~2% of the genome (~2500 peaks of our correlation method) we discover ~80% of original ZLD peaks of enhancer-like cluster, and ~50% of original ZLD peaks of Promoter-Like cluster. (B) - Average histone modification patterns of enhancer-like and promoter-like newly found sites. The HM average pattern of the 2700 promoter-like peaks, and 2500 enhancer-like peaks. The center is the location of the newly found peaks.

As expected, Figure 8B shows that the average plot of the enhancer-Like regions, looks similar, to the pattern after the clustering. Our method
produced loci with similar chromatin signature, and found many new peaks, in which there is no early Zelda peak.

In addition, we want to check the average number of transcription factors bindings, to strengthen the assumption that the newly found loci are indeed biologically significant and are regulatory areas during the embryonic development.

In Figure 9 we compare the average number of transcription factors bound at putative enhancers, to the average number of transcription factors bound at original early ZLD peaks (only top 2000 Zelda peaks used to cluster and build the model).

We can clearly see that the putative enhancers are enriched in transcription factors bindings, which strengthens the fact that our method indeed identifies enhancers, based solely on the 5 different histone modifications signatures.
Figure 9 – Number of bound transcription factors
Boxplots of comparing the number of bound TFs, using data of 21 TFs bindings in 2-3 hours old embryos [Li et al, 2008; MacArthur et al. 2009]. Comparison of putative enhancers, and top 2000 ZLD peaks (cycle 8). The boxes contain percentiles 5-95, and the lines are the minimal and maximal values. The average number of TFs bound is similar; therefore our model discovered loci with biological activity similar to ZLD original sites.

Motif Analysis

Another test we performed was to check if our method of identification of embryonic regulatory regions that act as distal enhancers might find transcription factors bindings, besides Zelda. In order to do so, we ran a motif analysis by HOMER [Heinz et al. 2010] of the top 1600 peaks of the putative enhancers. Considering the enrichment of Zelda sites within our set of putative enhancers, the first motif we would expect to find is the known binding site for Zelda – CAGGTAG. The results show that CAGGTAG is
present with a p-value of $1e^{-40}$ at 10.58% of the predicted regions. This strengthens the fact that our model is capable of finding original Zelda peaks based on their induced chromatin signature. In addition we can see that 23% of these enhancers are enriched with the GAGA motif, with a p-value of $1e^{-17}$. This Motif serves as the binding site of the GAF Factor, which was also shown to act as an important transcription factor during the embryonic development of the drosophila melanogaster by activating the genome [Schultz et al. 2015].

<table>
<thead>
<tr>
<th>Motif</th>
<th>p-value</th>
<th>% of targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTA TGA</td>
<td>$1e^{-18}$</td>
<td>27.71%</td>
</tr>
<tr>
<td>CAGG TAG</td>
<td>$1e^{-40}$</td>
<td>10.58%</td>
</tr>
<tr>
<td>GAGA GAG</td>
<td>$1e^{-17}$</td>
<td>23.6%</td>
</tr>
</tbody>
</table>

Figure 10 - **Motif analysis of putative enhancers**
Motif analysis of top 1600 putative enhancer regions, identified by our model based on their chromatin signature. CAGG TAG motif (ZLD binding site) is present with a p-value of $1e^{-40}$. The GAGA motif is present in 23% of these sites with a p-value of $1e^{-17}$.
**GAF Factor**

As motif analysis showed, the GAGA binding site (the binding site of the GAGA factor, often referred to as Trl or GAF) is enriched among the putative enhancers. In addition, we showed that GAF (GAGA Factor) is present in areas of open chromatin, in *Drosophila melanogaster* embryos. It was also shown, that in areas that have both Zelda and GAF binding, and therefore accessible, in drosophila wt embryos, are also accessible in Zelda mutants, which strengthens the assumption that GAF works besides Zelda.

We first wanted to check the average GAF ChIP around the different clusters, to see if GAF is present on all three clusters, or maybe on some of them. We used data of GAF ChIP from [Negre et al. 2011] through hours 0-8 of *Drosophila melanogaster* development, and plotted the average GAF ChIP in each of the clusters. Interestingly, GAF is present on both enhancers and promoters but is not at all around *transiently bound peaks*. This strengthens both the assumption that GAF and Zelda work together as to make chromatin accessible for other transcription factors to bind, and that *transiently bound peaks* are probably random Zelda binding sites, with no significance, and that are maybe there because they might be random Zelda binding sites coded through the genome.

In addition we plotted a graph of average GAF ChIP of the "putative enhancers". We used data of GAF ChIP through hours 0-8 of *Drosophila melanogaster* development, and plotted the average GAF ChIP in each of the clusters [Negre et al. 2011]. The graph indeed shows that GAF is present in those sites. There is a clear peak of GAF ChIP (Figure 11A). This strengthens our assumption, which states that GAF might be another Zelda-like pioneer factor, and helps activating early developmental enhancers, by effects of chromatin, in order for other transcription factors to bind to these enhancers.
Figure 11 – **GAF ChIP in putative enhancers.**

(A) - Average value of 0-4 hrs GAF ChIP from Negre et al. (2011) around the putative enhancers. Y axis corresponds to GAF ChIP strength. There is a clear peaks of GAF ChIP in the putative enhancers locations.

(B) - A genome browser figure showing our correlation score. The bottom part is transcription factors binding [Li et al. 2008; MacArthur et al. 2009]. yw FAIRE is in wt embryos, and zld- FAIRE is in zld mutants. The outlined box is an example of a putative enhancer. GAF is bound to the loci, while ZLD is not present. In addition a set of TFs [Li et al. 2008; MacArthur et al. 2009] are also bound. The region is accessible (FAIRE) both in wt and zld mutants. Our assumption is that GAF acts as a pioneer factor, allowing accessibility for transcription factor binding.
In Figure 11B we can see an example of a “putative enhancer” our model discovered. The peak doesn’t have any Zelda binding, but GAF is bound to the loci. Moreover, we can see that some other transcription factors are also bound, suggesting GAF doesn't need Zelda to be bound, in order to make chromatin accessible, and it maybe can act alone, without Zelda, and makes chromatin accessible for other Transcription Factors.

**Chromatin accessibility of putative enhancer (FAIRE)**

We believe that some of the enhancers we found are transcribed, even in zld mutant embryos. In order to test this hypothesis, we use formaldehyde-assisted isolation of regulatory elements (FAIRE) performed by us [Schulz et al. 2015]. The data consists of FAIRE throughout 2-3 hours of development with wild-type levels of Zelda, and embryos depleted for maternally Zelda.

We would like to see that some of the enhancers that we found have similar FAIRE both in the wild-type and in the zld mutants. In addition we expect to see, that enhancers that have high Zelda ChiP and low GAF ChiP in the wild-type embryos, would have lower FAIRE in the zld mutants, but not vice-versa, i.e. enhancers with high GAF binding and low Zelda binding in the wild-type embryos are still expected to have high values of FAIRE in the Zelda mutants.

Figure 12A shows all the enhancers we found using our method. We divided them into 2 groups: Those that have GAF binding, and those that don’t. We can clearly see that, those who have GAF peaks have much higher FAIRE in the zld mutants, compared to those that don't have GAF peaks. This further shows that, GAF can probably act as a pioneer factor, like Zelda, and can follow open chromatin and high accessibility even with the absence of Zelda.
Figure 12B shows the accessibility of the enhancers we found, in comparison to their GAF and Zelda bindings.

We can see that in the wt embryos, loci rich in either Zelda or GAF have high accessibility (FAIRE). In addition we can see that the more enriched the loci are with Zelda / GAF, the higher their FAIRE signal is, as expected. But when we look at the zld mutants, the loci which were originally rich with Zelda but not with GAF, are barely accessible, but the loci that are rich in GAF stay accessible. This strengthens the assumption that GAF can also act as Zelda, even in the absence of Zelda.
Figure 12 – **Accessibility in zld mutants and wt**

(A) - Boxplots comparing accessibility (FAIRE, 2-3 hrs) in wild-type embryos and zld mutants. The first figure compares loci lacking GAF binding. The second figure compares the loci in which GAF is bound. In the first figure, FAIRE strengths is lower in zld mutants, while in the second figure, in contrary to the second figure. This might suggest GAF binding is the reason zld mutants still maintain high accessibility in some regions. (B) - FAIRE Heatmap in wild-type and zld mutants. The figure shows heatmaps of FAIRE in the 2500 putative enhancers. The first heatmap is FAIRE from wild-type embryos, and the second is FAIRE from zld mutants. Sites are divided into 25 groups, based on Zelda and GAF ChIP (each divided into 5 groups). Peaks with only ZLD bindings had a major impact on the accessibility in the zld mutants embryos, while those where there is also GAF bindings remained accessible.
Discussion

Early Zelda peaks had certain chromatin marks, as previous research showed. After looking at several Zelda peaks, we started to believe that there is more than one particular chromatin signature. Clustering of top early Zelda peaks, based solely on 5 histone modifications (H3K27ac, H3K27me3, H3K4me1, H3K4me3 and H3K18ac) resulted in 3 different types of Zelda peaks: enhancers, promoter-like and transiently-bound Zelda peaks.

We strengthened the fact that these are indeed unique "clusters" of Zelda peaks, by showing a different biological behavior of each cluster.

The promoter-Like fall closest to TSS, The transiently-bound peaks follow silent genes, and Zelda "leaves" most of these peaks by cycle 14. Both enhancers and promoter-like peaks are bound by seven transcription factors on average, while transiently-bound peaks are bound by less than 2.

We showed that, looking only at the chromatin marks of Zelda peaks, we can characterize the role of the peak.

An interesting question that rises up, that we didn't test during our research, is how does Zelda know to "leave" in the case of transiently-bound peaks, and to stay in the enhancers and promoter-like peaks?

In order to identify regions with similar combinatorial chromatin signature, we used a simple method based on Pearson correlation. The model cross correlates the concatenated vector of chromatin signature with the entire genome. We added a correction to model, in which we multiplied the result with the probability of the height in order to find loci that have high correlation, but also high strength.

Our model did not only find most of the original Zelda peaks, which used to build the model, but to identify loci without Zelda peak, but with the same chromatin mark.

Our model to identify areas of similar combinatorial chromatin patterns, resulted in thousands of areas similar to Zelda enhancers (Zelda peaks of the enhancers cluster). By picking the top discovered enhancers we managed to cover more than 80% of the Zelda peaks, and moreover, to find many areas
lacking Zelda binding. It means, our model was able to find the same chromatin signatures that are present in areas of high Zelda bindings, also in areas with no Zelda bindings at all. Those regions were shown to have high accessibility, and are bound to ~6 transcription factors on average. It means, those regions, lacking Zelda bindings, are still acting as regulatory regions, and are probably crucial to the development of the drosophila embryos.

A Motif analysis of these regions showed that these regions are rich in the GAGA binding site, which binds GAF (GAGA factor). GAF was shown to be related to open chromatin regions in drosophila embryos [Schulz et al. 2015]. Our model shows that the same chromatin marks that are present in Zelda peaks, are present in GAF sites as well. We were able to discover GAF peaks, by only testing the chromatin patterns of the bound loci. Moreover, we showed that Zelda mutants (drosophila embryos lacking maternal Zelda) still have high accessibility (FAIRE) in regions bound by GAF, suggesting maybe GAF acts besides Zelda, or next to Zelda, during the MZT.
Methods

Joint Spectral Clustering followed K-means

As mentioned before, Zelda attaches to a specific binding site during the MZT, and only later we see chromatin marks, followed by Gene expression. We saw that Zelda peaks bound at mitotic cycle 8 are followed by several histone modification patterns, only later, around mitotic cycle 13, and even 14. Our goal was to find more enhancers, based on HM patterns, and to do so, we first wanted to cluster the different HM patterns.

We took the top 2000 early Zelda peaks (cycle 8), and looked at the histone modifications H3K27ac, H3K27me3, H3K4me1, H3K4me3 and H3K18ac at cycles 13 and 14. We took a window of 10K base pairs for each modification. Our histone modifications data was down-sampled by 10, so a 10,000 bps window, actually is a vector of length 1,000. This resulted in a 10 X 1,000 vector for each Zelda peak. We used log of histone modifications ChIP to construct vectors, in order to lower distance between high HM ChIP data.

The general scheme for Spectral Clustering:
Given N vectors \( \{X_i\} \), that correspond the histone modification ChIP, distance function and \( \sigma \), the algorithm works as follows:

a. Construct an N X N distance matrix A, where \( A_{i,j} = \text{distance}(X_i, X_j) \).

b. Turn matrix A to adjacency matrix W using some Kernel function. In our case, we used the Gaussian kernel function

\[
W_{i,j} = e^{-\frac{\text{dist}(i,j)^2}{2\sigma^2}}
\]  

(1)

This kernel function outputs values between 0 to 1.

c. Build normalized Laplacian matrix using matrix W. We build L in the following way:
We define Matrix D as follows:
\[
    d_{i,j} = \begin{cases} 
    0 & i \neq j \\
    \sum_k w_{i,k} & i = j 
    \end{cases} \quad (2)
\]

\[L = D - W.\] We later transform L into the symmetric normalized laplacian, which is \(D^{-1/2} L D^{-1/2}\).

d. Compute K eigenvectors.
e. Run K-Means clustering on top eigenvectors to compute clusters.

In order to perform the first step, we need to define a distance function. As distance function between every couple of vectors, we used RMSE (Root Square Mean Error), which is defined

\[
    dist_{i,j} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \quad (3)
\]

This resulted, as mentioned before, in an NxN distance matrix. Because we have 10 different histone modifications, we have 10 distance matrices. To continue from here, one option is to somehow compute an "average" of these matrices, in order to have a distance matrix, as needed to continue. The problem with this approach is that modifications with larger distances would have a larger impact on the adjacency matrix, therefore would have a larger impact on the clustering.

To solve this problem, we decided to compute an adjacency matrix for each histone modification, and later compute one global matrix that is the (weighted) sum of the specific adjacency matrices. This allows us to define \(\sigma\) for each histone modification, making the adjacency matrices "normalized". We did so by looking at the histogram of distances for each modification separately, and for each modification we set \(\sigma\) (deviation parameter in the Gaussian kernel, see equation 1 above) to be the 10th percentile (as shown in Figure 13). This way, we promise that for each modification, exactly 10% of the distances, will have a value bigger than \(\exp(-\frac{1}{2}) = 0.606\) in the adjacency
matrix. This method enables us to normalize all the distance matrices, to adjacency matrices in which 10% of the values are > 0.606. Now, we are independent on the average distance in each modification, and each modification should have a similar “weight” on the clustering.

![Distance distribution - H3K4me1-c14](image)

**Figure 13 - Distance distribution of H3K4me1 at cycle 14**
The figure above show the distance distribution (purple line) of histone modification H3K4me1 at cycle 14. The x axis corresponds to the distance. The red line is the 10th percentile, in this case distance of 10.36. This value is assigned as $\sigma$. The blue line shows the gaussian function with $\sigma=10.36$. We can see the adjacency values for the different distances.

After this phase, we have 10 adjacency matrices $\{W\}$. We now need to compute the average of these matrices, so we would have a global adjacency matrix. We used RMS (root mean square) to do so, resulting in 1 adjacency matrix, that allows us to continue with the rest of the algorithm.

**Cross Correlating different clusters**

After clustering the different Zelda peaks, we wanted to search for new “Zelda-like” regions, i.e. to find other loci, with similar histone modifications patterns, which might not have Zelda.
In order to do so, we used the average histone modifications pattern of each cluster, and cross-correlated it with the genome, to find similar areas. The score of each locus is the Pearson correlation coefficient. We used a window of 10Kbs, the same as the window used for the clustering.

At first, we cross correlated each histone modification separately, and gave each locus as a score, the average of the different histone modifications scores. The problem with this method is that the ratio between the different histone modifications doesn't remain like in the cluster, and only the correlation remains. In order to solve that problem, we created a vector of all the histone modifications concatenated, and computed the correlation with each locus (where the histone modifications are concatenated the same way). Using this method we solved the problem of the ratio between the different histone modifications.

Another problem with this method is that we encountered many loci which had "similar formation", meaning the correlation was high, but the values of the histone modifications were very low, and not interesting. In order to solve this problem, we need to somehow add another factor of the "height" of the locus histone modification Pattern, and to give higher scores to "heights" which are similar to the cluster. To formalize it, we want the score of each locus to be:

\[ \text{Score} = \text{Correlation} \times P(\text{height}) \] (4)

P is the empirical probability that a peak at height \( \text{height} \) appears among the actual set of loci in the original cluster.

\[ P(\text{height}) = \frac{\text{NumOfSites}(\text{bar(height)})}{\text{TotalNumOfSites}} \] (5)

We now have a score for each locus in the genome. The next thing we did is to compute all the local maxima, and pick the top valued loci.
References


15. Li, X.Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J.A. and Biggin, M.D., 2011. The role of chromatin accessibility in


