Testing for co-evolution between *eve* and Hunchback in *D. melanogaster*

Dagmar Ýr Arnardóttir

Faculty of Life- and environmental sciences
University of Iceland
2013

Testing for co-evolution between *eve* and Hunchback in *D. melanogaster*

Dagmar Ýr Arnardóttir

90 ECTS thesis submitted in partial fulfillment of a
Magister Scientiarum degree in Biology

Advisors
Arnar Pálsson
Zophonías O. Jónsson

Faculty of Life- and environmental sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, January 2013

Testing for co-evolution between *eve* and Hunchback in *D. melanogaster*
Co-evolution between *eve* and Hunchback
90 ECTS thesis submitted in partial fulfillment of a *Magister Scientiarum* degree in Biology

Faculty of Life- and environmental studies
School of Engineering and Natural Sciences
University of Iceland
Sturlugata 7
101 Reykjavik
Iceland

Telephone: 525 4000

Bibliographic information:
Dagmar Ýr Arnardóttir, 2013, *Testing for co-evolution between* eve *and Hunchback in* D. melanogaster, Master's thesis, Faculty of Life- and environmental sciences, University of Iceland, pp. 91.

Printing: Háskólaprent
Reykjavik, Iceland, January 2013

Abstract

Complex gene regulatory networks are central to the development of multicellular organisms. Trancriptional regulation is considered the most common mechanism of gene regulation. *cis* – elements are parts of DNA that proteins (transcription factors) bind to and activate and repress transcription. These regulatory elements are composed of a multiple short binding sites for more than one factor. Changes in gene regulation has been proposed to be the main contributor to evolution. Characterized binding sites within enhancers are generally well conserved and few mutations have been documented. Preliminary data show, two separate deletions of two conserved binding sites for Hunchback (HB) in an enhancer for *even-skipped* in *Drosophila melanogaster.*

To explain these deletions we proposed a model of co-evolution. We hypothesized that the concentration of HB had changed, the expression shifted or the timing of expression had changed and the deletions are a response to such a shift of the protein. We also expect to see higher incidence of mutations in HB sites genome-wide. We used population genetic and bioinformatic methods to study this, namely genome wide ChIP-chip data and sequences of ~200 inbred strains of *Drosophila melanogaster*. There is no excess of HB sites affected by SNPs compared to reference factors. We did however, see a higher fraction of deletions in HB binding sites. In sum the evidence are not conclusive on our hypothesis but confirms that a lot of predicted binding sites are affected by mutations.

Útdráttur

Stjórnraðir í erfðamenginu gegna hlutverkum við að kveikja og slökkva á genum á réttum stað og tíma, bæði í þroskun og yfir æviskeið fjölfruma lífvera. Rannsóknir sýna að stjórnraðir eru vel varðveittar í þróun, og jafnvel má finna samsvarandi raðir í fjarskyldum tegundum eins og manni og fiski. Sérstaklega eru bindiset innan stjórnraða vel varðveitt, en við þau bindast prótín sem stýra virkni gena. Oft bindast mörg mismunandi prótín á hverja stjórnröð og sem ræður tjáningu gensins, á vefjasérhæfðan hátt eða í þroskun.

Náttúrulegar stökkbreytingar í bindisetum stjórnraða eru sjaldgæfar, t.d. þegar bornir eru saman margir einstaklingar sömu tegundar. Enn er sjaldgæfara að finna úrfellingar í sömu stjórnröð, hvað þá tvær sem báðar fjarlægja skilgreind bindiset fyrir sama stjórnprótín. Vitað er um eitt slíkt tilfelli. Í einni stjórnröð *even-skipped* gensins í *Drosophila melanogaster* eru tvær náttúrulegar úrfellingar sem fjarlægja tvö bindiset fyrir Hunchback stjórnprótínið. Rannsóknin miðaði að því að kanna hvort að þessar úrfellingar væru staðbundið fyrirbæri (einungis í þessu geni) eða hvort vísbendingar væru um fleiri áþekka atburði í genamengi ávaxtaflugunnar. Markmiðið var að prófa tilgátur um samþróun stjórnprótínsins Hunchback og stjórnraða. Aðferðir lífupplýsingafræði voru notaðar til að greina stór gagnasett úr nýlegum stórum rannsóknum sem sækja má af netinu. Einnig var sameindaerfðafræði beitt til að kanna breytileika í *hunchback* geninu sjálfu.

Contents

List of figures

List of tables

Abbreviations

BS	binding site
CRM	*cis* regulatory module
DGRP	*Drosophila* Genetic Reference Panel
ENCODE	The Encyclopedia of DNA Elements
eve	*even-skipped*
GRN	gene regulatory network
HB	Hunchback
HMM	hidden Markov model
indel	insertion and deletion
kb	kilo bases
KR	Krüppel
LD	linkage disequilibrium
modENCODE	The Model Organism Encyclopedia Of DNA Elements
MSE	minimal enhancer
PWM	position weight matrix
PCR	polymerase chain reaction
SNA	Snail
SNPs	single nucleotide polymorphism
s2e	stripe 2 enhancer
s3+7e	stripes 3 and 7 enhancer
TFBS	transcription factor binding site
TF	transcription factor
UTR	untranslated region
UCSC	University of California Santa Cruz

Acknowledgements

First and foremost I would like to thank my advisor Arnar Pálsson. He has been very patient with me pursuing other dreams and having taken this long time in finishing this paper. I will certainly miss our talks in the office.

I would also like to thank Zophonías O. Jónsson for his help and advise.

For all the help in the lab and long and fun coffee breaks I wish to thank Sigrún Reynisdóttir. It was always nice to come to the lab and see and converse with her during our days in the lab.

I also wish to thank the guys under Arnar's supervision and all the programming help from Daníel Óskarsson, Jóhannes Guðbrandsson and Hlynur Bárðason.

I would also like to thank my family and friends for having endured my complaints this whole time.

I wish to thank Rannís for their contribution to this project.

1. Introduction

1.1 Gene Regulation

Most multicellular organisms develop from a single cell. That cell has the genetic instructions to guide the formation of the different tissues found in an organism. Each cell type expresses a subset of genes that varies between cell types. In humans there are over 200 different cell types that are encoded for by approximately 25.000 genes. Fruit flies have considerably fewer cell types and about half the number of genes (WATSON 2008). It has become easier to observe which genes are expressed in which tissues with microarrays and RNA sequencing. The expression profiles of cell types contains a combination of housekeeping genes that are also to be found in other the expression profiles in other genes. Each cell type also has a unique signature of expressed genes. It is remarkable how different cell types that are derived from the same progenitor cell can show such diversity in expression. What are the reasons for why a certain cell goes in one developmental direction and the adjacent cell becomes a completely different tissue? And how are cellular decisions made at the exact developmental time they are needed? Many developmental events, signaling adhesion, growth, migration at the cellular level and differential gene expression come down to gene regulation. There are several mechanisms for gene regulation such as chromatin condensation, transcriptional initiation, DNA methylation, alternative splicing of RNA, mRNA stability, translational control, intracellular trafficking, protein degradation, post-translational modification and more (WRAY *et al.* 2003). However, among the most common mechanism of regulation is transcriptional initiation. Usually there are either activators and/or inhibitors that regulate what genes are active at a particular time and place. Many of these regulators are DNA-binding proteins that recognize short sequences in the vicinity of each gene. Activators have several mechanisms to direct transcription, for instance, they can assist the binding of RNA polymerase to the promoters. The function of repressors is for example to inhibit transcription by binding to regions overlapping the promoter and makes binding of the polymerase impossible (WRAY *et al.* 2003).

Gene regulation through regulatory proteins is quite elaborate in eukaryotes. Two main categories of regulatory agents are associated with gene regulation, *cis*-elements and *trans*-factors. *cis*-elements, such as enhancers or silencers, are located close to the gene they act on, however, *trans*-factors are proteins that can sometimes be found on separate chromosomes and operate over distances (CHEUNG and SPIELMAN 2009; LATCHMAN 2010). A transcription factor (TF) affects transcription by binding to the DNA. There can be several binding sites, in the same *cis*-element, for a regulatory protein to bind to in order to turn on/off the gene. Some of these binding sites are located far from the promoter and therefore function in a different way than facilitating binding of the polymerase (CHEUNG and SPIELMAN 2009). The binding may be close to the promoter or several kilo-bases from it. There are reported cases of binding sites for activators, located many kb from the gene they regulate and their span can be considerable, up to hundreds of kb. The TF can recruit the polymerase, a mediator or other proteins associated with the transcriptional complex after the polymerase has bound to DNA. Another class of proteins recruited by activators are nucleosome modifiers. Tightly packed chromatin is inaccessible and the transcriptional machinery cannot be recruited. Possible function of theirs might be modification of the nucleosomes that make the DNA accessible (WATSON 2008). In some cases several regulatory binding sites are grouped together and form an enhancer. About 10-50 binding sites that can bind about 5-15 transcription factors are often seen in enhancer. They can often bind a few types of activators or repressors at different times which results in differential transcription of a single gene (WRAY *et al.* 2003). Enhancer function can be in both orientations in respect to the promoter, upstream or downstream. There have been two different mechanisms suggested for enhancer function. One is the direct communication with the transcriptional machinery, the other is through remodeling of chromatin. Segregating mutations in enhancers that affect single binding sites are rarely detected (ARNOSTI 2003).

It is thought that transcription factors operate over distances via DNA bending so that the binding site and promoter are close enough for the transcription factor to assist binding of the polymerase. In other cases there are proteins that mediate the binding. Some activators bind to enhancers close to more than one gene, if only one of those genes is to be activated insulators bind and insure correct regulation of genes. The insulator physically blocks communication between the promoter and the activator (WRAY *et al.* 2003). Regulatory proteins bind to the DNA. There are several different types of DNA binding domains that recognize the DNA, one such group is the zinc finger proteins. There the DNA is recognized through an α-helix placed in the grove of the DNA helix. A zinc atom has a structural role essential for recognizing the DNA. Some proteins have more than one zinc finger domain (WRAY *et al.* 2003; LATCHMAN 2010).

During development the communication between transcription factors, insulators, repressors and promoters is extremely complex where the result is different expression patterns in tissue types found in a single organism. Most of the gene regulation in development is at the level of transcription initiation or elongation (LATCHMAN 2010). There transcription factors are the key players. Three mechanisms have been identified during development that play the main role in gene regulation of two identical cells where the end result is different. One of them is maternal mRNA at different locations within the embryo. Polarity of the embryo causes mRNA to be unevenly distributed so gradients of regulatory proteins are expressed at specific locations. Another regulatory control mechanism is cell-to-cell contact. A cell can express extracellular signals that neighboring cells pick up and respond to. The third mechanism for different gene regulation during development is gradients of secreted signaling molecules (LATCHMAN 2010). Wolfram's morphogen model claims that a single cell's location within a gradient defines what that cell will develop into where different concentration produces sharp boundaries. This was called the French Flag model and was revised by Jaeger. The revised model states that it is impossible to predict dynamics and regulatory behavior of systems based on general geometric arguments. There the boundaries are a bit more vague and allow for corrections at later stages of development (JAEGER and REINITZ 2006). Factors such as morphogens affect cells differently depending in their dose. Morphogens are mobile chemical substances that form spatial gradients affecting development of inducible target cells in a concentration-dependent manner. Morphogens are able to diffuse and act over long distances in developing tissues. According to the definition of morphogen, it must also be responsible for threshold dependent responses of the affected cell. A small group of cells produce morphigenetic signals (TFs, ligands, etc.) that can act as morphogens and produce and secrete a signal molecule that forms an extracellular gradient. The cells in close proximity to the secreting cells get a high dose and develop into a certain cell type. The further away from the secreting cells, the smaller the dose and different genes are turned off and on. The processes that take place in early development form a complex cascade of activation and repression of gene expression (JAEGER and REINITZ 2006).

With emerging technologies the focus on gene regulation has increased. Numerous questions have to be answered, such as what factors are at play? How do they interact? What are the consequences of a factor wrongly expressed? In what order do things occur? Identifying the transcription factors, binding sites and all the key players is simply the first step in understanding gene regulation. With the invention of microarrays the possibilities for measuring the quantitative gene expression and the capacity for identifying general trends has multiplied (CHEUNG and SPIELMAN 2009). Without a decent picture of a gene regulatory network it is hard to predict and make assumptions on how systems are affected by tweaks and changes. One of the best resources we have to understand complex transcriptional behavior from the DNA strand alone is using *in silico* models based on actual measures *in vivo* (JAEGER and REINITZ 2006). Mapping human transcriptional networks is complex since many players are unidentified and, for ethical reasons, there are many limitations on direct measurements of concentrations. However, for the fly, worm, and yeast, such studies are being conducted and once we understand those networks better we gain more insight into the human network.

cis-regulatory modules (CRMs) are the focus of immense interest. CRMs are usually enhancers that are composed of several binding sites, usually for more than one factor. Binding sites for individual factors have been characterized and described for transcription factors, inhibitors and silencers in yeast, fly, mouse, human as well as other model organisms. One transcription factor, and corresponding binding sites, that has been well characterized is the *Drosophila melanogaster* Hunchback. Stanojevic et al. (1989) characterized 11 binding sites for HB in the enhancer of *even-skipped* for stripes 3 & 7. DNaseI assays were used to find where on the DNA a *hunchback* protein was bound. The exact 7-14 base recognition sequence was identified. There are usually some differences between the binding sequences for an individual factor, however, there is a pattern that is almost always the same. Each factor has a consensus site that refers to the "best" version of the binding site or to the sequence that captures most of the binding site matrix. For HB it is TTTTTTATG C/T (STANOJEVIC *et al.* 1989). Position weight matrixes (PWM) are a useful method to describe binding sites for specific factors, and can be used to find both novel and previously described binding sites, such as TATA box or homeodomains, in genomic sequences (BERMAN *et al.* 2002). A method based on using PWMs to find conserved CRMs has been developed and its accuracy is high in locating known enhancers (SOSINSKY *et al.* 2007). PWM for a transcription factor is built by using known active binding sites for a specific TF and create a matrix with a score for each location in the 10-14 base sequence. For each site in the sequence a score between 0 and 1 is given based on how often a particular base is found at a certain location. If it shows up in 10 out of 10 binding sites it gets the score 1, if it is only seen once it gets the score 0.1. By adding up the scores for each base in a 7-10 base sequence, each 7-10 base window in a given sequence is assigned a score that reflects the similarity of the sequence to the TF PWM. There is no optimal score for all transcription factors. The highest possible score varies depending in the transcription factor. In the case of Hunchback the highest possible score is 14 where the six T's (or A's in the reverse complement) are almost always seen and these bases have a score between 0.95 and 1 which is very high (STORMO 2000). Another way to use PWMs is by their information content. Information content takes into account the other allele, so that a base with the PWM score of 1 has the information content of 2.

Binding sites for activators and repressors sometimes overlap where there is competition for binding between activators and repressors. Some repressors act via quenching and cause repression of a gene (LUDWIG *et al.* 1998; LATCHMAN 2010). Each enhancers function differently, and in order to gain full knowledge of them each one would have to be studied. Some of the questions would be how much binding is required? What are the consequences of deletions of single binding sites? What are the effects of poorer binding of one factor as a result of mutations in binding sites? Are all binding sites within an enhancer for one factor equal, or are there some that have to be bound in order for transcription to occur? The *eve* stripe 2 enhancer is among the best studied enhancers and yet not all those questions have been answered (LUDWIG *et al.* 1998; ARNOSTI 2003). Enhancers are well conserved between species as well as the binding sites within them, however, some variation of individual binding sites seems to be tolerated (LI *et al.* 2008). Flux of binding sites in enhancers is sometimes allowed during evolution without affecting the transcriptional output. It is know that changes in single sites can disrupt regulatory output but it is hard to identify which changes affect the regulation and which are tolerated. There arises a new question of how many (or how severe) changes an enhancer can undergo until its function or gene expression output has shifted or become severely affected?

Out of the most popular model organisms a gene regulatory network (GRN) for yeast is easiest to establish. For each gene there is usually a single regulatory region located close to it. GRNs in yeast can provide useful identifying principles but since the fly is a multicellular animal many of the genetic features are shared with the human genome. Considerable understanding can come from studies with flies in the hope of uncovering human GRNs. Early development of the fly is the most studied gene network of a multicellular organism up to date but is far from complete. Even though most factors are known we can only predict with limited accuracy what actually occurs if there is a defect somewhere at a certain time-point in the fly's development. For example *eve* has at least 14 regulatory inputs which have the potential to give rise to 2^{16384} Boolean regulatory combinations and identifying which one applies at a certain moment is tricky (WILCZYNSKI and FURLONG 2010). In order to understand an entire *cis*-regulatory system many different types of data have to be collected. There has to be a comprehensive map with all *cis*-regulatory modules (CRMs) for the system. All transcription factors that bind those CRMs have to be known as well as their spatio-temporal expression pattern. Similarly the spatio-temporal output of each CRM must be known and finally, an understanding of the regulatory function of each of the CRMs. The process of understanding GRNs is quite laborious and requires many different datasets, such as information on what factors are bound, at what time-points, what is the input and output of the network. For humans, most transcription factors are uncharacterized. Once a decent GRN has been established for human development it will hopefully be useful in developing theraputic drugs or gene therapy when something goes wrong during development (WILCZYNSKI and FURLONG 2010).

Data on genetic defects influencing human disease is quite extensive. Many known disease genes have been sequenced along with their surrounding regions. Rockman and Wray (2002) searched the literature for polymorphism in regulatory regions of genes known to influence disease. They found more polymorphisms in relatively high frequency than previously was thought to exist. Another point of interest was how much the effects seemed to be from a single nucleotide polymorphism (SNP), the increase or decrease in protein output was up to 10-15 fold. The researchers were surprised to see how much polymorphism there actually is and especially that they are common pleiotropic polymorphisms. They categorized the mutational effects into 5 categories; gain and loss of activation site, gain and loss of repressor site and switching of a site. Early human genetics focused predominantly on variation in protein coding parts of genes where as modern human genetics put equal, if not more, emphasis on regulatory DNA (ROCKMAN and WRAY 2002). However, in relation to transcription factors and disease, Farnham (2009) found evidence that about 164 transcription factors contribute to 277 diseases (FARNHAM 2009).

1.2 Eukaryotic development

Eukaryotic development is a complex topic and has kept researchers busy for some time and will most likely continue to be a popular subject in the future. Among the most popular model organisms are, sea urchins, the worm *C. elegans* and the fly *D. melanogaster* (DAVIDSON 2001). The first steps of fruit fly development have been studied in great detail (GILBERT 2006). The fly is a popular model organism along with yeast, mice, frogs and *C. elegance*. The advantage of using the fly as a model is that it is small, easy to breed and has a relatively short lifespan. Another characteristic of fruit fly development is the syncytium which is a stage after fertilization where 14 synchronized nuclear divisions occur without cellularization. After 9 cell divisions most of the nuclei become localized to the periphery of the embryo surrounding the yolk sac. This occurs before cellularization takes place and gastrulation starts. In the syncytium transcription factors and morphogens can diffuse freely and local expression results in gradients with an easy access to cells and make their mark on differential transcription through gene regulation that give rise to the different segments of the fly (LATCHMAN 2010). All the mechanisms discussed above come together in the embryogenesis of the fruit fly to produce sharp seemingly on-off signals for expression of different genes (WATSON 2008). Use of computational models as well as traditional measurements of concentration has provided a detailed boundary establishment model for the gap genes of early embryogenesis (JAEGER and REINITZ 2006). In order to characterize in detail what occurs in the fly during development, four things have to be achieved: (1) formulation of mathematical modeling framework, (2) gene expression data for a number of factors in the system of study, (3) coordination of the model to the expression data and (4) biological analysis and tests of the gene circuits. This is possible in the fly because the pattern formation is a result of interaction between segmental genes only where they are known and have been studied extensively. Among the things they saw are that the effects of a transcription factors studied are potent, either positive or negative in correlation with their quantity or concentration. The interaction of genes is by reciprocal repression or activation that produces sharp segmentation of the embryo (JAEGER *et al.* 2004).

1.3 Segmentation of the fly

The genes expressed during the fly's early development have roughly been divided into 6 categories; maternal, gap, pair-rule primary, pair rule secondary, segment polarity, and hox genes (Table 1.1). These systems of development are well conserved between species, although, not equally well conserved. The hourglass model of development states that early in development there is considerable variation between species that gets narrower during mid-embryogenesis (RAFF 1996). During late embryogenesis the differences in conservation increases again. This was tested between six closely related *Drosophila* species and the finding showed that this occurs not only on the morphogenetic level but also on the gene level. In the maternal genes there is most diversity that gets progressively less, through gap, pair rule, segment polarity to the hox genes. Of the 6 gene categories of the fly's development the expression differences of the hox genes is the most between species. After that the diversity in gene expression increases again (KALINKA *et al.* 2010). One suggestion for this pattern of conservation is that during mid-embryogenesis the embryo is not in direct contact with the environment, compared to adult flies, and are therefore less likely to be subject to evolutionary forces of adaption (DOMAZET-LOSO and TAUTZ 2010). Manu et al. (2009) carried out a study to see if there is canalization of the gene expression of the gap genes in an *in silico* model. He found that there are activators that are stable against small perturbations. These activators are locally stable that affect their immediate surroundings and are responsible for trajectories of the gap gene system (MANU *et al.* 2009).

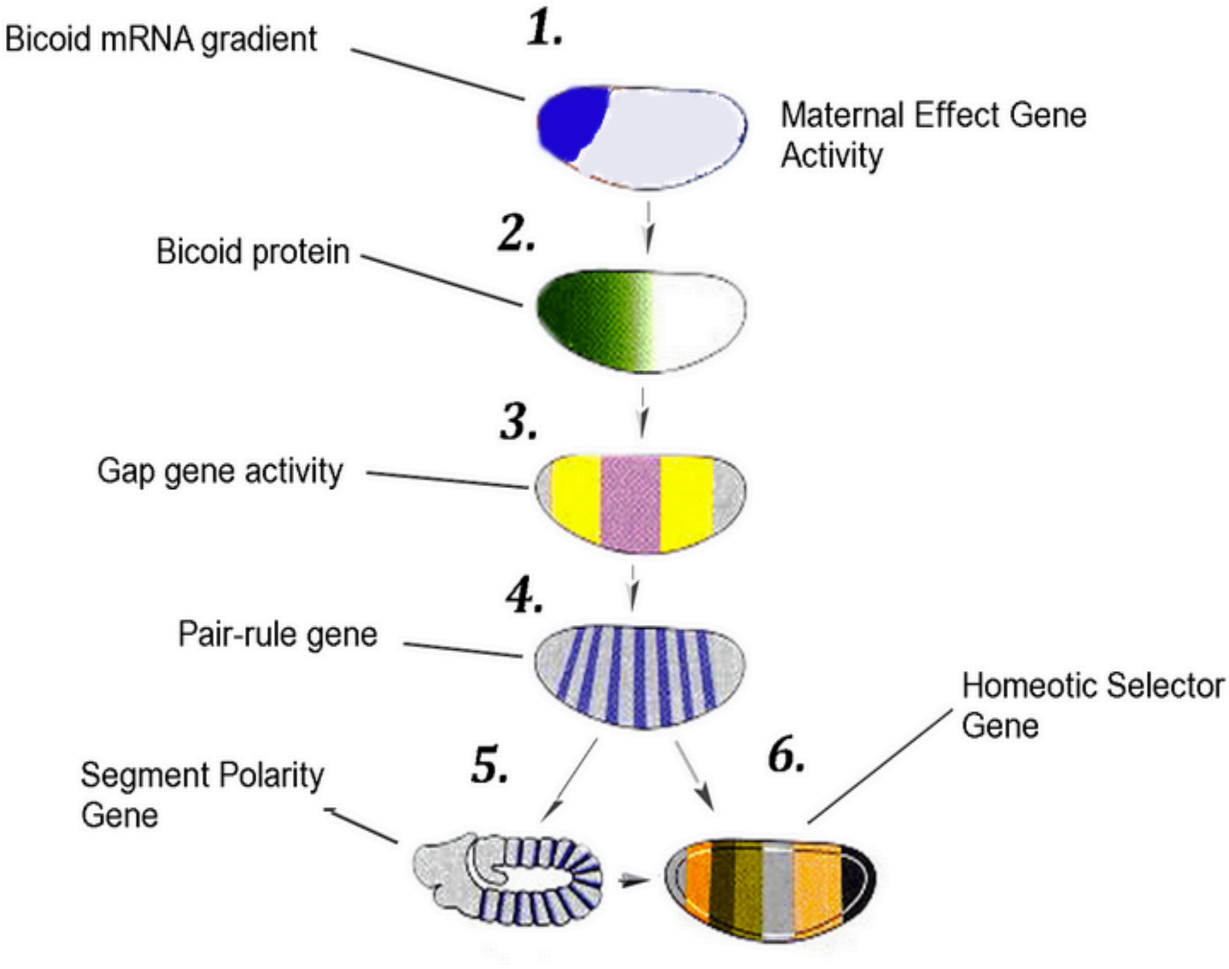


Figure 1.1: The Drosophila *developmental hierarchy (see table 1.1 for genes active at specific stage). (1) Maternal* bicoid *mRNA gradient, (2)* bicoid *protein gradient, (3) gap genes, (4) pair-rule genes, (5) segment polarity genes, (6) homeotic selector gene (BIOPAUKER 2008).*

*Table 1.1: Major genes affecting segmentation in Drosophila (*GILBERT *2006).*

Category	Gene name
Maternal genes	*bicoid* *caudal* *nanos*
Gap genes	*Krüppel* *knirps* *hunchback* *giant* *tailless* *huckebein* *buttonhead* *empty spiracles* *orthodenticle*
Pair-rule (primary)	*hairy* *even-skipped* *runt*
Pair-rule genes (secondary)	*fushi tarazu* *odd-paired* *odd-skipped* *sloppy-paired* *paired*
Segment polarity genes	*engrailed* *wingless* *cubitus interruptusD* *hedgehog* *fused* *armadillo* *patched* *gooseberry* *pangolin*
Hox genes	*labial* *proboscipedia* *deformed* *sex combs reduced* *antennapedia* *ultrabithorax* *abdominal A* *abdominal B*

1.4 Maternal mRNA determines anterior-posterior specialization

At the time of fertilization the fly has at least two distinct mRNAs from mother at each end. *bicoid* is located at the anterior pole and *oskar* is located at posterior end of the embryo. Oskar codes for a RNA-binding protein that is responsible for assembly of polar granules that result in development of posterior tissues. The polar granules are macromolecular complexes composed of many proteins and RNAs. Oskar is localized in the posterior end of the embryo. Bicoid is synthesized and diffuses after fertilization to form a gradient along the embryo toward the posterior part. *hunchback* codes for a transcription factor and is activated through high and medium concentrations of Bicoid (WATSON 2008). Hunchback is highly expressed in the anterior part of the embryo and forms a steep gradient around the center of the embryo and has a small expression domain in the posterior end (Figure 1.2). The mechanism that keeps the gradient precise is unknown. Patel and Lall (2002) checked the effects of changing the Bicoid gradient. It was both increased and decreased, however, the Hunchback gradient shows no response to the change in Bicoid indicating that Bicoid is not the only regulator of *hunchback* and that the Hunchback gradient is very precise (PATEL and LALL 2002). There is variation between individuals in the Bicoid gradient where there is little variation in the Hunchback gradient between individuals (HOUCHMANDZADEH *et al.* 2002). Hunchback affects *eve*, a pair rule gene that is responsible for the first steps in the fly's segmentation. *hunchback* has two promoters, one is activated by the Bicoid gradient in the embryo, the other regulates expression in the developing oocyte. The latter promoter leads to synthesis of *hunchback* mRNA, which is evenly distributed throughout the cytoplasm of unfertilized eggs (WATSON 2008). To activate *hunchback* translation Pumilio binds to 3' UTR of *hunchback* mRNA. In the posterior part a RNA-binding protein, Nanos, binds Pumilio and blocks the translation of *hunchback* (MURATA and WHARTON 1995). In the anterior half of the embryo Bicoid gradient activates zygotic transcription of *hunchback* and through these two separate regulators of *hunchback* a steep protein gradient is formed where the concentration is high in the anterior and almost non-existing in the posterior half, apart from a small domain at the very end (Figure 1.3.1) (LATCHMAN 2010).

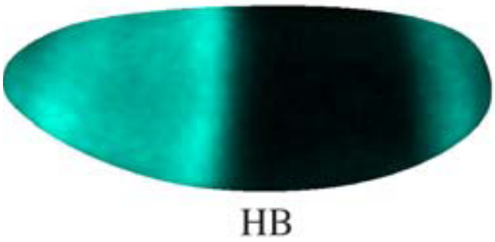


Figure 1.2: Hunchback concentration in the embryo at late stage four. Anterior to right, posterior to left, ventral is down and dorsal is up (LI et al. 2008).

Jaeger et al. (2006) presented a detailed model of how the transcription factors of the gap and segmental genes interact and work together in the embryo as a response to the maternal, morphogen Bicoid gradient. The determination of the effect Bicoid has on the gap genes and the activation/repression effects they have on each other is among the best characterized networks of gene regulation interaction during development. This was discovered using computational models based on real concentrations measured in individual cells along the embryo during a few hours in early development. This lead to characterizing the relationship between the gap genes *hunchback*, *Krüppel*, *knirps*, and *giant* (Figure 1.3) (JAEGER and REINITZ 2006).

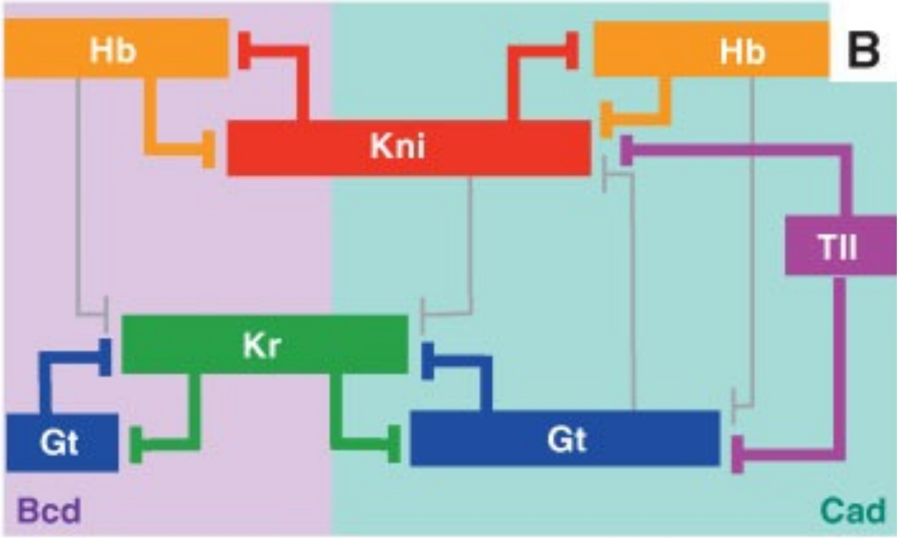


*Figure 1.3: Relationship between the gap genes, the arrows indicate repression and thickness the strength of repression (*J*AEGER and* R*EINITZ 2006).*

Hunchback functions as a transcription factor for at least three gap genes, *Krüppel*, *knirps* and *giant* and possibly other genes as well. It has dual function as both an activator of transcription but also as a repressor. High levels of HB results in repression of *Krüppel* and intermediate levels suppress *giant* and *knirps*. Hunchback and Knirps are mutual repressors, if Knirps is lacking then Hunchback extends into the posterior part of the embryo and if Hunchback is missing then Knirps stretches into the anterior part. These four genes all code for transcription factors (LATCHMAN 2010). Hunchback's gradient is not responsible for the different levels of repression, however, what is considered to be more important is the number of binding sites in the regulatory region of each gene and their affinity (weak or strong) for the protein (WATSON 2008). Along with Bicoid, Hunchback is a transcription factor that is required to bind to and activate enhancers of head-genes (GILBERT 2006). Hunchback is well conserved between species. Comparison of HB between *Drosophila* and *Tribolium* revealed that there was not much difference between the function of the two species. That was not expected prior to the experiment because of long versus short germ band embryogenesis of the two species (WOLFF *et al.* 1995).

1.5 Interaction of transcription factors

The first indication of the fly's segmentation is the expression of the pair-rule gene *even-skipped* (*eve*). *eve* is expressed in seven alternate stripes that are the precursors to the fly's segmentation. Each stripe is approximately four cells wide where *eve* is highly expressed and in the space between, also four cells wide, there is little or no expression of *eve* is found (Figure 1.4) (WATSON 2008).

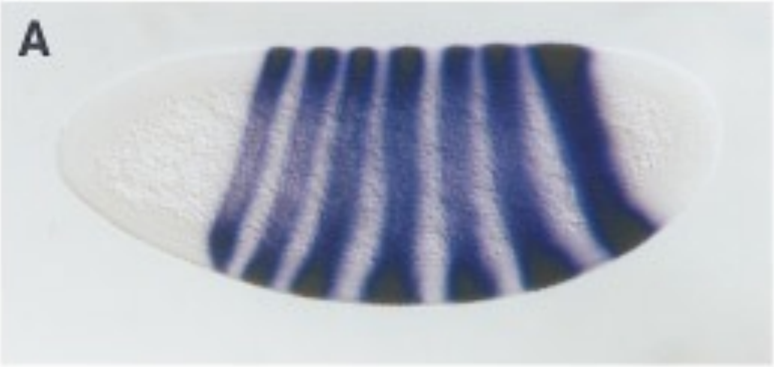


*Figure 1.4: Expression of eve in the embryo in 7 distinct stripes, each approximately 4 cells wide as well as the space between them is about 4 cells wide (*SMALL *et al. 1996).*

eve encodes a transcription factor composed of 376 amino acids. The transcribed region is rather small, only 2 kb. However, the flanking regulatory sequences surrounding *eve* is 12 kb with 4 kb upstream and 8 kb downstream of the gene. The upstream region contains the regulatory sequences for stripes 2, 3, 7 and auto regulatory element, where the regulatory sequences for stripes 1, 4, 5, 6, as well as neuronal enhancer are found within the downstream region. A total of 5 enhancers have been characterized in the 12 kb regulatory region where each enhancer affects one or two stripes (WATSON 2008). The 5 enhancers comprise 4 kbp of bound DNA and 7 transcription factors are known to bind to *eve's* regulatory region (ARNOSTI 2003).

The four gap-genes TFs, Hunchback, Krüppel, Knirps, and Giant, are involved in regulating *eve* by binding to the different enhancers (Figure 1.5). Stripe 2 is mainly regulated by an enhancer located 500 bp upstream of the protein coding part (LUDWIG *et al.* 2005). It contains five binding sites for Bicoid and one for Hunchback that serve as activators and three Krüppel and three Giant repressor sites. Stripes 3 and 7 share an enhancer with 11 characterized Hunchback binding sites as well as and 5 Knirps binding sites (STANOJEVIC *et al.* 1989). Stripe 3+7 minimal enhancer is 500 bp long and is located 3.3 kb upstream of the transcription start site. This region is described as the minimal enhancer (MSE) because if any part of the sequence is missing, the expression will be incorrect (SMALL *et al.* 1996). If any piece of a characterized MSE for a gene is missing the gene is not expressed properly, or might not be expressed at all in the designated place and time of development. When that piece of DNA is reintroduced into the genome, in proximity of the promoter and a reporter gene, the expression is rescued (LUDWIG *et al.* 2011). The individual binding sites within the minimal enhancer for stripe 3+7 were identified with DNaseI footprint assay which is a good indicator of their functionality (STANOJEVIC *et al.* 1989). Recent studies have provided new insights into the stripe 3+7 enhancer. Zelda is a transcription factor that is also necessary in order for correct expression of *eve*, no stripes are seen when *zelda* has been knocked out (STRUFFI *et al.* 2011). Among unknowns is how many binding sites have to be bound in order for transcription to be activated at any moment and the correct protein output to be produced. Hunchback and Knirps function together to form the boundaries of stripes 3, 4, 6, and 7. Hunchback demarcates the anterior border of stripe 3 and posterior border of stripe 7 where Knirps controls the expression of the posterior border of stripe 3 and anterior border of stripe 7 (WATSON 2008).

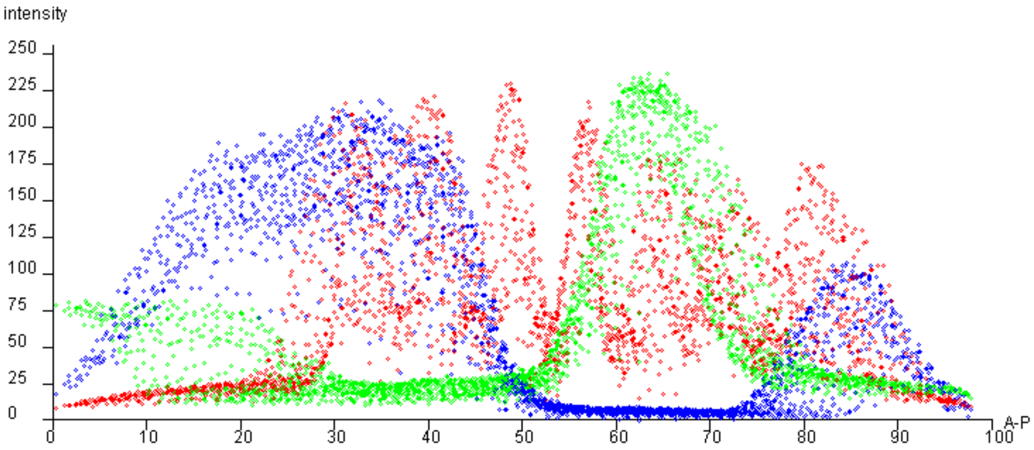


Figure 1.5: Expression of hunchback *(blue),* eve *(red), and* knirps *(green). The expression of* hb *drops after stripe 2. There is one HB binding site for stripe 2 but there are 11 for stripe 3+7 (KOSMAN 1999).*

Multiple experiments have been carried out to test the mechanism of stripe formation in the embryo. Clyde et. al. (2003) used the *snail* promoter to mis-express *hunchback* and *knirps* on the ventral side of the embryo. When *hunchback* was over-expressed, stripes 4, 5, and 6 disappeared. Stripe 3 showed anterior weakening and posterior expansion. The results show that *eve* 3+7 and *eve* 4+6 enhancers respond to different amounts of Hunchback and in order to have correct segmentation the exact spatial and temporal expression of HB and Knirps is essential (CLYDE *et al.* 2003). It has been suggested that the repressor sites are responsible for defining the boundaries of the stripes (LUDWIG *et al.* 1998). Hunchback contains two zinc-finger domains (PAPATSENKO and LEVINE 2008). The central domain mediates DNA binding and the C-terminal domain causes dimerization of HB molecules. It has been proposed that the purpose for the second DNA binding domain is to form dimers and thereby regulate *Krüppel* (HULSKAMP *et al.* 1994). With the help of computer simulations it has been shown that the binding of HB to stripe 3+7e is stronger than to stripe 4+6e. That is in concordance with the fact that the concentration of HB is considerably lower at the location of stripe 4+6e than stripe 3+7e. When the concentration of HB is low the DNA bound monomers act as activators and when there are high concentrations of HB, dimers are formed that repress transcription or block activation (PAPATSENKO and LEVINE 2008; BIELER *et al.* 2011).

1.6 Population genomics

Population genomics is the study of the genomic variation within species. *Drosophila* is a convenient model organism for such studies for instance because the genome has been fully sequenced as well as many close relatives (CLARK *et al.* 2007). With the expanding field of genomics the applications for population or evolutionary genetics are increasing as well. Population genomics needs more than one reference sequence and there have to be at least the genotypes of SNP’s scattered around the genome or ideally several fully sequenced individuals, in order to compare the differences between them. It has been shown that polymorphisms are not at fully random locations in genomes (HARTL and CLARK 2007). Different levels of polymorphism within a genome can give some indication about the history of the species or populations. Another clue might be different levels of recombination. Areas where there is very low or no polymorphism found in long stretches of DNA can give evidence of recent positive selection or purifying selection (HARTL and CLARK 2007).

Recombination rate affects polymorphism at certain locations where recombination is high or where it is low. Recombination between sites close on the chromosome are infrequent and therefore it is likely for those sites to segregate together. Forces maintaining diversity or reducing it, act within a region that will be correlated with the level of polymorphism of closely linked sites (HARTL and CLARK 2007).

There are several types of natural selection at work in genomes, natural selection, positive, negative (purifying), balancing, and stabilizing selection. Darwin in 1859 stated that organisms are i) variable, ii) the variation is heritable and therefore iii) they are differently able to survive and reproduce leading to natural selection. In genetic terms, the alleles that lead to increased fitness of individuals will increase in frequency in the population. Once that allele has become fixed in the population it is maintained through purifying selection. In stabilizing selection the individuals in the middle of the range, as opposed to the extremes, are favored (HARTL and CLARK 2007). The environment affects which attributes are advantageous. For different populations the environment can be extremely variable and therefore the signatures of selection might be totally different (BAMSHAD and WOODING 2003). Advantageous mutations in one population might be neutral in another. It has been proposed that most mutations that are segregating in populations are neutral or nearly neutral (KIMURA 1979) and that essentially all genes are subject to the force of maintaining their function (BAMSHAD and WOODING 2003). Harmful mutations are eliminated or kept at low frequency through purifying selection (HARTL and CLARK 2007).

Reduction in polymorphism can be a consequence of positive selection and to lesser extent background selection or hitchhiking. Selection leads to the fixation of favorable mutations or the elimination of harmful mutations, respectively. Selective sweep, is when an advantageous mutation is favored and becomes the predominant allele in the population. Hitchhiking is when a neutral mutation in close proximity to the advantageous one is pulled along for the ride. SNPs in the region surrounding the positively selected mutation will be in excess in the genome after, or close to fixation and the region will have low polymorphism because the recombination rate is limiting. An unusually long undisrupted haplotype on a chromosome at high frequency in the population is an indicator of positive selection where a mutation is selected for in a short amount of time (HARTL and CLARK 2007). This method is commonly used to test for recent selection. It was used for example by Sabeti et al. (2007) when searching the human genome for signs of positive selection. Another way to test for selection is by using cladograms where odd structures can indicate a selective sweep in the a population (BAKKER *et al.* 2006).

Several measures on polymorphism within species have been developed and one of those is Pi (π). Pi is the probability that nucleotide samples from two individuals will differ. It takes into account the number of and frequency of the mutation as well as the length of the sequence. Linkage disequilibrium (LD) is frequently used to see if genes are in random or in associated linkage. LD is often seen between closely linked SNPs. However, LD can also be seen between SNPs that are located far away, even on separate chromosomes. Compensatory mutations can show LD because of functional relationships (HARTL and CLARK 2007).

There are many unanswered questions regarding selection and how it operates on DNA sequences. In an attempt to answer these questions we have to study the patterns of variation within and between species. What we do know is that mutations can roughly be divided into neutral, deleterious, and advantageous. According to theories neutral and deleterious mutations are more common than advantageous ones (SELLA *et al.* 2009). We also know that random mutations occurring in functionally important locations are more likely to be deleterious and mutations in non-functional regions that do not participate in important processes are more likely to be neutral. Deleterious mutations arise often, however, they rarely become widespread in a population. They are purged by purifying selection and therefore not seen in polymorphism and almost never in divergence between species unless other factors, such as the environment, change. The opposite is true for beneficial mutations. They are often seen in divergence between species. They are very rare but are more likely to reach high frequency in the population. Polymorphism within species is often due to neutral mutations. The neutral theory states that molecular evolution occurs simply by random genetic drift, not through selection. This theory has been tested and only recently has it been questioned. Sella et al. (2009) studied polymorphism in the *Drosophila* genome. They found the fraction of deleterious new mutations in coding regions was 94%, 81% in untranslated regions, 56% in introns and 61% in intergenic regions. Also, it is estimated that about 40-50% of amino acid substitutions are adaptive. These numbers indicate that adaptive polymorphism is more common than previously believed. Because of this, researchers argue that positive selection cannot be ignored as one of the major reasons behind divergence (SELLA *et al.* 2009).

1.7 Evolution of gene regulation

Many studies have shown support for the hypothesis that non-coding regulatory DNA is what drives species divergence. Several studies have been done, both between closely related species and distantly related species such as the worm, *C. elegans* and *D. melanogaster*. Ruvinsky and Ruvkun (2003) used transgenic experiments to study enhancers of homologous genes, the transcription factors binding to the enhancers were homologous between the two species. They tested if they remained functional. In most cases they were not. However, when they moved enhancers between *C. elegans* and *C. briggsiae* the enhancers were functional despite considerable sequence difference. This finding indicates there is co-evolution between transcription factors and enhancers where the enhancers are subject to more rapid changes than the transcription factors, in other words regulatory DNA evolves faster than proteins (RUVINSKY and RUVKUN 2003).

Difference in gene regulation is one component of phenotypic divergence. The differences can arise from changes in *cis*-elements and other factors that affect transcription initiation, elongation, transcriptional rate and stability. *trans*-factors that interact with *cis*-factors can also cause changes in gene regulation. The effects of such changes are largely unknown. Only recently have researchers been able to identify changes in regulatory DNA that contribute to species divergence (WITTKOPP 2010). One argument for *cis*-factors being the main contributors to species divergence is the pleiotropy argument. It states that mutations that have less pleiotropic effects are favored over those that cause widespread effects. That is natural selection favors *cis*-regulatory mutations since they have smaller effects than coding mutations. Mutations in *cis*-elements should have more specific effects whereas a change in a protein might have more drastic affects at multiple locations (STERN and ORGOGOZO 2008).

Pigmentation in Drosophila is regulated by different *cis*-elements. The expression of yellow in different body parts is controlled by different *cis*-regulatory elements. Several transgene experiments moving enhancers between species have demonstrated different expression of the yellow gene (CARROLL *et al.* 2001). Carroll et al. (2001) claim four reasons for why *cis*-element evolution is powerful. One is that regulatory evolution enables pleiotropy of several toolkit genes, such as transcription factors. The regulatory sequence may evolve so that the response can change at different developmental stages. Second, regulatory evolution allows for developmental modularity through changes in gene regulation at the sequence level. Third, regulatory evolution is a source of rich and continuous variation both at the species level and within species. Without changing the protein sequence the effects on morphology are subtle, but changes in the regulatory sequence contribute to variation with the potential of affecting the morphology. Fourth, regulatory evolution causes novel structures to arise. New morphologies are created through novel combination of gene expression without affecting protein structures (CARROLL *et al.* 2001).

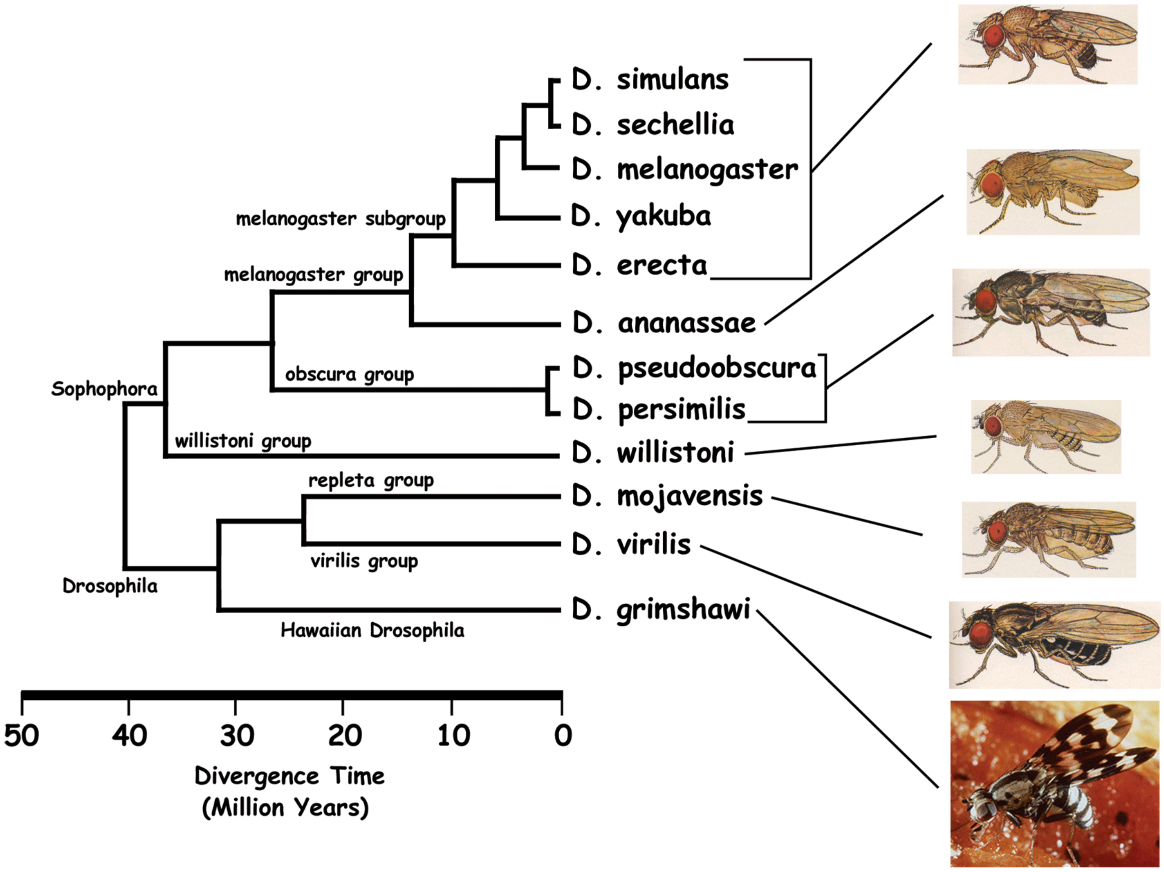


Figure 1.6: Phylogeny of 12 Drosophila *species (GILBERT 2005).*

The sequence coding for Hunchback has been moved between the species *D. virilis* and *D. melanogaster*. The proteins are fully functional within the other species (LUKOWITZ *et al.* 1994) and the same antibody can be used on both HB versions, however, the *D. virilis* protein has a few extra amino acids. The DNA binding motifs, two zinc fingers, are completely conserved between the two species. In the regions surrounding the protein coding region there are conserved blocks as well as stretches that are highly diverged (TREIER *et al.* 1989). The maternal enhancer in *D. virilis* lacks two of the 3 low affinity binding sites for Bicoid but the 3 high affinity binding sites are conserved between the species. Even though *D. virilis* has a broader stripe of *eve* expression that separates into two stripes later, the stripe in *D. melanogaster* is not affected at all (LUKOWITZ *et al.* 1994). The main function of HB in the two species, as a gap gene, is conserved where as some secondary functions seem to be somewhat diverged. Small changes that affect secondary processes are therefore considered to contribute to evolution between species (TREIER *et al.* 1989).

Among the most studied developmental enhancers is the stripe 2 enhancer of *eve*. The enhancer is well conserved between species, even though some claim the expression of *eve* has shifted slightly between the species (FOWLKES *et al.* 2011). The functional binding sites in the s2e have been characterized. It contains 12 'strong' binding sites, 6 for activation (Bicoid and Hunchback) and 6 for repression (Krüppel and Giant). Using recombination and transgenes researchers identified a minimal enhancer (MSE) for stripe 2 containing the 12 characterized binding sites that are vital for correct expression of *eve* at stripe 2 location. However, flanking regions of the MSE have additional binding sites for both Hunchback and Krüppel and those binding sites were not essential for the expression of *eve*, but when missing, effects on the fly's viability was found (LUDWIG *et al.* 2011).

Researchers were interested in testing enhancer conservation between closely related species. Stripe 2 enhancer is functionally conserved between closely related *Drosophila* species but has undergone considerable chances with regard to distances between binding sites with insertions and deletions, single nucleotide polymorphism (SNP) mutations in individual binding sites and gains and losses of whole binding sites. Comparison of regulatory output of a reporter gene, regulated by *eve* stripe 2 enhancer, between *D. melanogaster* and *D. pseudoobscura* showed that it was functionally identical in both species (LUDWIG *et al.* 1998). The same result was obtained when enhancers from four closely related *Drosophila* species were used to express a reporter gene in a *D. melanogaster* background. No differences in expression of *eve* were found, but the stripes were slightly lighter than in the wild type. Analysis on the enhancer sequence showed considerable difference. Most difference was seen in spacing of characterized binding sites. The function of the whole enhancer therefore seems to be under stabilizing selection to maintain its expression. Two binding sites, one for Bicoid and one for Hunchback were found in *D. melanogaster* but not in other related species. That suggests that they are new and are possibly a response to some change that might have occurred sometime in the *D. melanogaster* lineage, and can possibly explain why the stripes were lighter when the enhancer was placed in the *D. melanogaster* background (LUDWIG *et al.* 1998). When a chimeric enhancer, half from *D. melanogaster* and half from *D. pseudoobscura* was replaced in *D. melanogaster* the regulatory output was incorrect. This is evidence of changes in enhancer design between closely related species (ARNOSTI 2003). Ludwig et al. (2002) put out four hypotheses to explain evolutionary changes in *cis*-elements. (1) Neutral evolution of non-functional segments such as spacers between binding sites. (2) The consensus motif for binding allows for neutral or nearly neutral changes within binding sites. (3) Accumulation of new binding sites that have become functionally important and essential for the enhancer to maintain its function. (4) Functional co-evolution of the whole element and not individual binding sites (LUDWIG 2002).

Ludwig et al. (2005) expanded is studies and the same enhancer, *eve* stripe 2, was used to test evolution of *cis*-regulatory element between *D. melanogaster*, *D. pseudoobscura, D. yakuba* and *D. erecta*. The enhancer from the 3 other species was introduced as a transgene into a *D. melanogaster* background driving *eve* to see if the enhancer could rescue the expression of *eve*. When enhancer from 2 of the species was inserted into *D. melanogaster* viability was restored and the spatio-temporal expression was not affected. However, *D. erecta* enhancer was unable to rescue flies even though the sequences are orthologous. The enhancer's function was equivalent between *D. melanogaster* and *D. pseudoobscura* despite considerable changes in many binding sites. The *D. erecta* enhancer has changed, possibly with increased sensitivity. This shows co-evolution of the *cis*-regulatory elements and the dose or amount of *trans*-factors expressed. This finding is in concordance with the same structures in *D. melanogaster* are shifted posteriorly in *D. erecta* (LUDWIG *et al.* 2005). This suggests that the functional conservation of enhancers is substantial between species. However, there is still considerable structural change that might be one of the main drivers in species divergence.

Stripe 7 is regulated by Hunchback and this regulation is conserved between closely related *Drosophila* species. Expression differences of the stripe between three closely related *Drosophila* species can be linked to changes in the *cis*-regulatory element (CRE). Wunderlich et al. (2012) used transgenic lines differing only in the CRE to measure the expression differences in the HB posterior stripe. They found that subtle differences in the CRE did affect the stripe expression. The change was seen in how the regulators for *hb* alter sensitivity to the regulating transcription factors. Another thing they found was that compensatory evolution has occurred outside the CRE to maintain correct expression of *hb* (WUNDERLICH and DEPACE 2011).

Several experiments using hybrids of two closely related species has been carried out to characterize the divergence between them. When carrying out hybrid experiments *Drosophila* species are a convenient organism. *D.melanogaster* and *D.simulans* diverged about 2.5 million years ago. Hybrids of two species, can give indications of how the species diverged through changes in gene regulation. Wittkopp et al. (2004) tested 29 genes of F1 hybrids where 28 had changes in *cis*-regulatory elements. About half of these changes were enough to explain the difference in regulation where the other half showed changes in both *cis*- and *trans*-factors. The results indicate that widespread *cis*-changes, such as enhancers, are largely responsible for driving gene expression differences between species (WITTKOPP *et al.* 2004).

Landry et al. (2005) did a similar study with hybrids. There the emphasis was on testing for *cis*-*trans* co-evolution in a set of 31 genes. They generated *D. melanogaster* and *D. simulans* hybrids and looked at the expression in the parents versus an F1 hybrid. Similar allele expression differences the parents and hybrid indicates that changes in *cis*-elements are responsible. When the allele expression differs to a larger extent than in the parents, change in *trans*-factors might be the cause. When *cis*-regulatory differences of the hybrid are higher than the divergence between the parents then possibly *trans*-regulatory differences have changed to compensate for the high level of expression differences to bring it closer to that of the parents. Another way to test this experimentally is by characterization of the genetic regulatory elements and measurement of gene expression of each allele in the parental background. One example of previously characterized co-evolution of regulatory elements is the *eve* s2e between *D. melanogaster* and *D. pseudoobscura* (LUDWIG *et al.* 1998). The results from this study indicate that co-evolution of regulatory elements, regulatory DNA and transcription factors, is more common and widespread than previously thought. About 40% of the 31 genes examined, showed signs of *cis-trans* compensatory changes or co-evolution between them (LANDRY *et al.* 2005).

As discussed above, numerous studies have been carried out on individual enhancers or genes, the technology now allows researchers to carry out genome-wide scans of the effects of specific CRMs or *trans*-factors. Despite interest in the evolution of transcriptional regulation, much is unknown about the molecular mechanisms. Evidence shows that most regulatory sequences, such as *cis*-regulatory elements, are under strong purifying selection to preserve their function, possibly to maintain stable transcriptional output. Despite this, binding sites are lost and gained over time. A study where *D. melanogaster* and *D. yakuba* were compared in respect to sequence divergence, with special emphasis on binding sites of 6 factors (Bicoid, Hunchback, Krüppel, Giant, Knirps, and Caudal) responsible for A-P patterning, was carried out. They isolated DNA from whole embryos and obtained net affects of the whole genome. These species diverged approximately 25 million years ago and their genomes can be aligned. There is little difference in their spatial expression pattern. Differences in amino acid sequences of the six factors tested turned out to be small. When the binding of the factors was compared there was only about 1-5% change of the chip peaks intensity (BRADLEY *et al.* 2010), however the relative affinity was different (WITTKOPP 2010). The results show that there is a difference in binding between closely related species which can explain some of the differences in gene regulation between the two species. Binding differences for the transcription factors was explained in part by gains and losses of short sequence motifs (approximately 7 bp long), 12 separate motifs were identified for Hunchback and 10 for Krüppel (BRADLEY *et al.* 2010). An interesting observation for the current study is that the fraction of binding site gains and losses is highest for Hunchback, both for bound areas know to be significant in A-P determination and those not as significant. That indicates that HB binding sites may be in more flux in comparison to the other A-P determining factors.

Another gene regulation divergence experiment was carried out in yeast. There the objective was to quantify how much *cis*- and *trans*-factors contribute to the divergence and how much is caused by changes in both. Tirosh et al. (2009) created a hybrid from *Saccharomyces cerevisiae* and *S. paradoxus* and measured the allele-specific expression with microarrays. From this study they came to three main conclusions. First, most divergence comes from changes in *cis*-elements. Second, *trans*-factor changes are more conditional on the environment. The third conclusion is that the hybrid strains inherit compensatory mutations, which is an indicator for purifying selection that accounts for about 20% of the transcriptional differences (TIROSH *et al.* 2009).

Through several different types of studies (discussed above) researchers have characterized and come to many conclusions about the evolution of regulatory DNA. Between closely related species there are many conserved regions but also other highly diverged. Despite sequence divergence many regions are the functionally conserved. Among the things those studies have revealed is that that *cis*-elements are among the main contributors to species divergence, co-evolution between *cis*- and *trans*-factors to maintain function, enhancers allow considerable flux of binding sites, and *cis*-element changes occur at a faster rate than changes in proteins. This knowledge we have gained is only the tip of the iceberg when it comes to regulation of genes. Each enhancer works differently so therefore there is great work ahead in learning more about their operations. One of those things is learning why so many non-conserved regions are bound by transcription factors (BIRNEY *et al.* 2007).

1.8 Genomic distribution of TFs in early *Drosophila* development

Since 1974 when Sanger first developed his method for DNA sequencing, many technical advantages have occurred. New methods for sequencing that are both faster and cheaper have become available (GIBSON and MUSE 2009). Sequencing whole genomes with considerable accuracy and coverage can now be done in days as opposed to months, years or even decades. *D. melanogaster* was the first higher eukaryote to be sequenced (GIBSON and MUSE 2009). Not only have many close relatives of *D. melanogaster* been sequenced (ASHBURNER 2007) but also around 200 inbred strains, in a collaboration of Trudy Mackay and the Baylor College of Medicine (DGRP).

ChIP-chip or chromatin immunoprecipitation followed by a microarray is one method to test binding of a certain transcription factor in the genome. One transcription factor is tested each time. This method can also be used to measure the quantity of the binding as well as location of other epigenetic markers (VISEL *et al.* 2009). That allows for identification of regions in the genome bound by this factor at a specific developmental time and stage, and of sequence motifs in high frequency in bound regions that have the highest affinity for the transcription factor. This is an effective and fairly accurate method of comparing binding of specific transcription factors between species. This has been done in comparing yeast species, human and mouse, and *Drosophila* species as was mentioned above (WITTKOPP 2010). This method accurately predicts enhancers and suggests which genes are active *in vivo*. It has also been shown that if binding between transcription factors and DNA is detected then many of those locations are active in directing transcription (VISEL *et al.* 2009).

With the development of ChIP-chip and ChIP-seq detailed mapping of factors bound genome wide has become a tool for researchers interested in early development and those interested in gene regulation comparison between species. Since most players involved in *Drosophila* A-P development are known it is ideal to use the fly to find all the factors bound at a certain time-point in development. Li et al. (2008) performed ChIP-chip for six gap gene transcription factors Bicoid, Caudal, Hunchback, Knirps, Krüppel, and Giant. In the following year the study was expanded by MacArthur et al. (2009) to 15 more transcription factors known to participate in A-P determination of the embryo. They used material from whole embryos and thus the binding represents average binding of transcription factors at that time-point in development. With this study they are unable to see any tissue specific or localized binding. They saw that these factors are highly bound to all previously characterized *cis*-regulatory motifs as well as other unknown ones and weakly bound to areas spread throughout the genome. The areas of highly bound regions can roughly be split into five different groups: (1) several hundred of highly bound factors were found bound to previously characterized CRM's for genes known to be A-P genes, (2) most highly bound regions are close to genes known to have function in early development whereas poorly bound regions were found closer to housekeeping genes or metabolic enzymes, (3) majority of factors are highly bound close to genes active at this stage and poorly bound to regions not active at all or not at this state, however, a few of the factors show a different trend and are bound relatively highly at other locations, (4) some of the poorly bound regions of a few factors are found in protein coding regions of genes, and (5) those regions highly bound are more conserved than poorly bound regions, however, the individual binding sequences seem not to be more conserved than the whole bound region (LI *et al.* 2008; MACARTHUR *et al.* 2009).

The biggest problem in interpreting such results is determination of which of the bound regions are functional *in vivo*. The authors state with confidence that the highly bound regions are functional at this time in the embryo and directly influence transcription of nearby genes. Many of those regions are known to be functional in early development. These highly bound regions show higher conservation in comparison to other random, non-coding regions in the genome. The highly bound regions tend to be located in intergenic and intronic sequences. The poorly bound sequences are found closer to genes not active at this time-point of development indicating that they have other functions or none at all. For instance, HB functions as a transcription factor for development at stage 9 in development. At stage 5 embryos, binding of HB to enhancers is poorly detected. However, it is quite possible that the poorly bound regions have no function at all and represent background binding of TFs to available TF binding sites. Another possible reason for their binding is that they may have a function as buffers for the available molecules for binding to enhancers that directly regulate transcription. The poorly bound regions were often found in or near housekeeping genes or genes not transcribed in the blastoderm. Their conclusion is that developmental fates of cells are not only determined by what factors are bound, but also the quantitative differences of a set of factors that determines cellular fates (LI *et al.* 2008; MACARTHUR *et al.* 2009). Evolutionary constraints are usually correlated with function. For the regions bound by the 21 transcription factors, the sequences in the flanking non-coding regions of the genes known to be bound by transcription factors, are more conserved than other random non-coding sequences (LI *et al.* 2008; MACARTHUR *et al.* 2009). Li et al. (2008) and MacArthur et al. (2009) use PWM's for each factor within highly bound regions of each factor and found that recognition sequences for each transcription factor were enriched within an area bound by that factor. This is consistent with previous data (MOSES *et al.* 2006).

Transcription factors are bound at a quantitative level. This quantitative range is correlated with gene type, degree of gene regulation, and transcriptional state. However, most recent studies have ignored this correlation. They use either bound or not bound to classify the gene regulation relationship between factors and genes. In the study the relative level of transcription factor binding is significant in studying the complex range of regions bound by transcription factors genome wide (LI *et al.* 2008).

The findings of Li et al. (2008) and MacArthur et al. (2009) suggest that the highly bound regions are functional targets of transcriptional regulation of early A-P patterning genes. The poorly bound regions may play a role later in development or regulate housekeeping genes and many of the regions may have no function at all. The regions that are likely to have no function are possibly not preserved by selection to the same effect because they do not affect transcription. There is evidence for selection against sites that may interfere with transcription. Therefore weak binding that has little or no effect on transcription may be tolerated and possibly even preferred. Therefore there should be more emphasis on measuring and evaluating the different effects on transcription from different levels of binding (MACARTHUR *et al.* 2009).

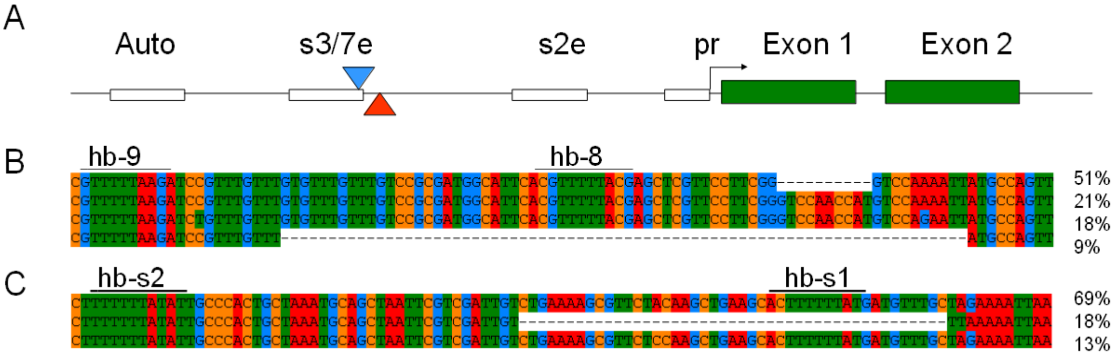
The work of both Li et al. (2008) and MacArthur et al. (2009) is typical of genomic studies they detect broad strokes and patterns. However, ENCODE (RANEY *et al.* 2011), ModENCODE (CELNIKER *et al.* 2009) and the studies of Li et al. (2008) and MacArthur et al. (2009) have given good indicators on what to expect when studying genome wide binding of transcription factors. Their data is available to smaller groups to validate or study in more detail some patterns or hypothesis. Among the things that need to be studied is why there is so much binding to locations that seem to have no transcriptional importance, what the role of the widespread binding is? The ChIP-Chip data can be used to answer this question and many more.

2. Foundations of the project and working model

2.1 Degeneration of HB binding sites

The enhancers responsible stripes 2 and 3+7 of *eve* have been studied in great details by several groups dissecting different aspects of the TF action, cooperation and interaction of these enhancers (LUDWIG *et al.* 1998; CLYDE *et al.* 2003; LUDWIG *et al.* 2011). In enhancer 3+7 11 HB binding sites have been characterized with DNaseI assay (STANOJEVIC *et al.* 1989). The current study is based on the observation that two separate deletions of individual HB sites are segregating at relatively high frequency in populations of *D. melanogaster* at different locations in the US and elsewhere (Arnar Palsson unpublished). This finding is at odds with previous results showing conservation of *cis*-regulatory elements and lack of polymorphism in enhancers (LUDWIG *et al.* 1998). The two deletions are found on separate haplotypes. The larger one is 71 bp in length and the other one is 45 bp. There is no indication that the deletions affect the fly's viability or development (Arnar Palsson unpublished). No other compensatory binding site was found within the sequenced portions of *eve*. Despite the deletions the regulatory output seems to remain the same and the loss of one binding site shows no measurable effects on the fly (Figure 2.1) (see Figure 1.6 p.14 for phylogeny) (Arnar Palsson unpublished).

A



D

	hb-14a	kni-5	stat-2	hb-10	hb-8	hb-s1
D.mel	TTTTTTGTTT	CTGCGCTAGTT	TTCCCCGAA	TTTTTTAATTC	GTTTTTACGA	TTTTTTATGA
D.sim	....	....	....	....	....	....
D.sec	....	........T..	....G....	....	....	....
D.yak	....	......C.C..	....	....	A......T..	....
D.ere	.........T	......C....	....	....	A......T..	....
D.ana	....	....	.....A...	....	N/A	N/A
D.pse	....	....	....	....	A......T..	N/A
D.per	....	....	....	....	A......T..	N/A
D.vir	N/A	....	....	....	A......TT.	N/A
D.gri	N/A	....C......	....	....	A......TT.	N/A
D.moj	N/A	....	....	....	A......TT.	N/A

Figure 2.1: The two deletions of HB binding sites in eve *and the conservation between Drosophila species. A is* eve *upstream region. B deletion of a HB and the frequency in the North Caroline sample. C the smaller deletion of HB binding site and frequency in the North Caroline sample. D the alignment of the closest relatives and characterized binding sites in stripe 3+7 enhancer for* eve *(A Palsson unpublished).*

Binding of transcription factors between closely related species has been compared. What is usually seen is when there is binding in one species, binding in the other is also documented. However, the level of binding varies between the two species (LUDWIG 2002; MOSES *et al.* 2006; BRADLEY *et al.* 2010). Conservation of *cis*-regulatory elements is higher than other non-coding regions (MACARTHUR *et al.* 2009) therefore mutations in enhancers, especially indels, are expected to be infrequent. It is quite unusual to find one deletion of well conserved binding site of moderate frequency, let alone two affected binding sites for the same transcription factor. This could be due to chance, but the most plausible explanation for these observations is positive selection (A Palsson unpublished).

Intraspecific changes in *cis*-regulatory sequence of early developmental genes in flies have been characterized before. Goering et al. (2009) found a deletion in an enhancer of *otx* and proposed two explanations. One is high local mutation rate and the other is an extreme shape of local genealogy as a result of a bottleneck (GOERING *et al.* 2009).

Several models have been proposed for enhancer function. One proposition is that enhancers operate as a unit. Drastic changes, such as loss of a binding site, would disrupt the function of the whole enhancer. However, recent experiments have indicated that this is not so. For the s2e it has been shown that the enhancer has undergone considerable rearrangement during evolution. Therefore Arnosti (2003) proposed the billboard model, which states that enhancers function as a billboard, showing information on what is bound and the basal transcriptional unit then interprets the message from enhancers (ARNOSTI 2003). Under this model, the loss of one binding site could simply be because of random drift. It is possible that only chance and nothing else is causing the two HB binding site deletions to be at such high frequency. But seeing two separate deletions in different haplotypes indicates that something is going on and the deletions are a response to that change.

He et al. (2011) set out to investigate what forces drive binding site turnovers, which is part of the questions we have proposed. They used *D. melanogaster* and *D. simulans* and studied SNPs in transcription factor binding sites (TFBS) in well characterized CRM. They only used TFBS that have been confirmed as functional with DNaseI footprint verification. One previous assumption has been that a unidirectional fitness function of SNPs in TFBS, or that selection always favors affinity-increasing mutations. Their data indicated that there is purifying selection against affinity-decreasing mutations segregating in the population as well as signs showing that positive selection both drive TFBS loss and gains. He et al. (2011) propose three reasons for why TFBS loss occurs at a fairly high rate, (1) the constrains are lost, (2) tightly linked compensatory mutation is created, and (3) positive selection drives the loss of the site (HE *et al.* 2011).

2.2 Specific questions

Based on our model I set out several specific predictions and hypotheses.

1. If positive selection is favoring deletions of HB binding sites in s3+7e we predict:
    a. More deleterious mutations in HB binding sites compared to other reference TFs.
    b. Higher frequency of mutations found in HB binding sites than in TFBS for reference TFs.
    c. On average larger effects of mutations in HB binding sites than by mutations in TFBS for reference TFs.
    d. Negative correlation between PWM information content per base and counts of mutations in each position.
    e. Deletions are more common in HB sites than reference TFs.

2. If there was a change in Hunchback activity, we would predict:
    a. Amino acid changes in the HB protein, that would change its function.
    b. It was due to increased concentration of HB and the deletions were responding to this change in concentration.
    c. The spatio-temporal expression pattern has changed or shifted because of:
        i. change (mutations) in *hb* regulatory region.
        ii. change (mutations) in 3’ UTR.
3. If Hunchback function had changed because of positive selection, we might predict
    a. evidence of selection in nearby region (with HB changing because of hitchhiking).
    b. evidence of selection in HB itself.

4. If there was on going co-evolution between HB and *eve*, we would predict:
    a. LD between variation in *eve* and *hb*.

2.3 Hypotheses and approach

We put forth several hypotheses to explain why these HB binding site deletions in *eve* have not been selected against. Binding sites and enhancers are usually well conserved between species and therefore it is very unusual to see deletions of characterized binding sites. The first prediction is that *other HB binding sites in enhancers* for other genes are also damaged and include more mutations than other transcription factors. To test this I used whole genome sequences of different strains of *D. melanogaster*. DGRP (Drosophila Genetic Reference Panel) was done by Trudy MacKay and the Baylor College of Medicine and involves sequencing of 162 inbred strains, available online are both the sequences and a SNP file from alignment of all the strains. The second piece of the puzzle, as described above, Li et al. (2008) performed ChIP-chip for HB and other gap gene factors. He identified all the regions in the genome HB binds to (LI *et al.* 2008). These two datasets together and designing algorithms that use PWMs to locate binding sites can give us a list of all possible HB binding sites known to be bound by HB and therefore likely to be functional. We will use other transcription factors to compare if mutations in HB binding sites are more frequent than mutations in other transcription factor binding sites. We expect to find more mutations and deletions in HB sites and them to be more frequent, in other words more events and more common. Positions within binding sites contribute differently to the affinity of transcription factor to the binding site. Therefore positions that are less important within the binding sites should be subject to less conserved. However, that does not suggest that each position evolves independently but rather that the whole binding sites are single evolutionary units despite single mutations. It has also been suggested that binding sites work at a level of affinity and both stronger and weaker binding sites are not preferred (KIM *et al.* 2009). Therefore it is possible that either mutations do not have considerable affect on binding or that deletion of a whole site really does not matter for the enhancer to remain functional. Combining two different kinds of data has been found especially useful when investigating, at a genome wide scale, transcriptional networks. Several articles have been published where it was feasible to combine datasets (WUNDERLICH and DEPACE 2011), similar to what we have done.

The second hypothesis is that the *hb protein itself has changed* or is undergoing change. Sequences from 12 closely related *Drosophila* species exists, so the amino acid sequence can simply be aligned and compared. If the *hb* protein in *D. melanogaster* has undergone any changes the alignment will reveal them. It is most likely that the DNA binding domain will be changing between them. Single amino acid substitution can result in the protein being a bit more stable or having slightly higher affinity for the DNA. That affects the degradation of the protein and perhaps leads to a lower or increased equilibrium concentration. Therefore each amino acid change is subject to its context. A slightly deleterious mutation in one background can be slightly advantageous in another (SAWYER *et al.* 2007). Another possibility for the deletions is that the *spatio-temporal expression of HB has shifted* or *the level increased* and mutations in HB sites are preferred to counter the shift.

The deletions might be preferred by selection as a response to increased concentration of HB at the exact time and location where stripe 3 element is biologically active (Figure 2.2).

The reason for increased concentration of HB could be because *the* hb *gene was positively selected* for. There could also be that selection is acting on a gene close by and *hb* is hitchhiking along. Positive selection acting on a region would result in a long haplotype with low polymorphism (HARTL and CLARK 2007). To test this we sequenced the area surrounding *hb*.

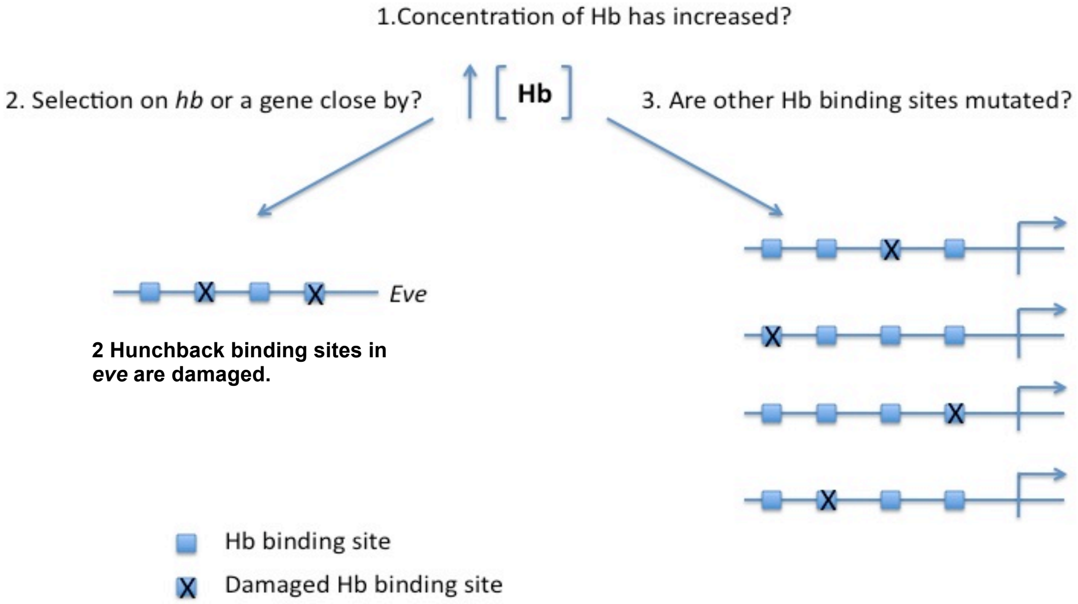

1.Concentration of Hb has increased?
Hb
2. Selection on hb or a gene close by?
3. Are other Hb binding sites mutated?
X
Eve
2 Hunchback binding sites in eve are damaged.
Hb binding site
X Damaged Hb binding site

Figure 2.2: Model of co-evolution of hb *and its target genes. In this version, we postulate a change in HB concentration.*

If there is functional interplay between HB and eve then possibly LD between *hb* and *eve* can be seen. Mutations in Pumilio binding sites affect the binding of Pumilio to the *hb* mRNA (MURATA and WHARTON 1995) and therefore transcription of the protein. Another possibility might be that the Pumilio binding site is changing in *D. melanogaster*. Alignment of the 3' UTRs of the species can show if this is the case.

3. Materials and Methods

3.1 Screen for polymorphism in *hunchback* region

3.1.1 Flies and isolation

Flies were kindly provided by Ian Dworkin. They were caught in the summer of 2004 in North Carolina (see Goering et al. 2009 for description). The 32 lines had been bred to isogenicity through 15-20 generations of full sib-mating. Flies were kept on 12hr light–dark cycles in vials with 10 mL standard cornmeal medium supplemented with yeast. The strain was sustained for 3 years until they were put in ethanol and kept at -20°C. A single fly's DNA was isolated from each strain, one individual per strain, using Sigma DNA isolation Kit (see Appendix A).

3.1.2 PCR to DNA sequencing

All sequences were obtained by direct sequencing of PCR products amplified from genomic DNA of a single fly. Primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3/ (ROZEN and SKALETSKY 2000) and the fly's genome sequence, version 3, from the University of California Santa Cruz Genome Browser (http://genome.ucsc.edu/ (KENT *et al.* 2002)). Primer pairs were designed in total to sample diversity around *hb*, spanning a 72,268 base pair region (see Appendix B, Table B.1). Each amplicon was designed to be 400-600 bases in length. The same PCR recipe was used for the amplification and the same PCR program (see Appendix A, Table A.1).

Each product was run on a 1% agarose gel to evaluate the PCR amplification success. The PCR product was then cleaned using Exonuclease enzyme and sap (see Appendix A, Table A.3). The next step was BigDye sequencing reaction (see Appendix A, Table A.4) followed by ethanol precipitation (see Appendix A, Table A.5). After that the pellet was dissolved in HiDi (see Appendix A) and run on an Applied Biosystems 3500xL Genetic Analyzer (Hitachi). Most fragments were sequenced unidirectionally (forward). Those that turned out to be problematic were also sequenced in reverse for verification. (Sequence data will be submitted to genebank).

3.1.3 Editing and sequence analysis

Raw sequencing data was base-called by Sequencing Analysis Software v5.4 with KB TM Basecaller v1.41 (Applied Biosystems). Phred, Phrap and Consed were used to edit the sequences (EWING and GREEN 1998; GORDON 2003). The primer sequences were removed and the ends trimmed. After editing, insertion of ambiguity codes for heterozygous bases, and alignment using Clustal W (LARKIN *et al.* 2007), sequences were analyzed with Tassel (BRADBURY *et al.* 2007). Tassel calculates population genetic statistics for sequences like Pi.

To further test if there are any signs of positive selection around *hb* I used published sequence data of 162 lines of different *Drosophila* strains (retrieved from: http://www.hgsc.bcm.tmc.edu/project-species-i-Drosophila_genRefPanel.hgsc, Drosophila Genetic Reference Panel (DGRP)) (MASSOURAS *et al.* 2012). I downloaded the raw sequence data set and used a python algorithm to extract the sequences for the same areas as I sequenced. The data has not been manually edited and therefore it is likely that it contains false SNPs, especially surrounding areas that are difficult to sequence and are represented with N at the low end of the quality spectrum. I therefore wrote an algorithm to extend all N areas of seven N's in each direction. By doing this, many false positive SNPs will be excluded. Next I ran a sliding window for Pi over the area using Tassel (BRADBURY *et al.* 2007). I used a window of 1000 bases sliding every 200 bases.

3.2 TF binding in the genome

3.2.1 ChIP-chip and SNP Datasets

For this part I used two separate datasets available online, one is the DRPG data previously described (MASSOURAS *et al.* 2012) and the ChIP-chip data from Li et al. (2008) (Hunchback and Krüppel) and MacArthur et al. (2009) (Snail). From the ChIP-chip I got coordinates of locations that HB, KR and SNA bind to in the genome (Li et al. 2008), Table S.1 downloaded from Plos website in January 2010 (MACARTHUR *et al.* 2009). There were 1762 chip-bound regions for HB, 3028 chip-bound regions for Krüppel and 595 chip-bound regions for Snail. Krüppel and Snail are used as reference transcription factors because they all have the same DNA binding domain as HB namely a C2H2 zinc finger. I then wrote a Python algorithm that extracts these exact locations from the sequencing data (see Appendix D Algorithm D.1). The chip areas span from 1 kb and up to 10 kb. I downloaded the genomic locations of those regions for each transcription factor from the *Drosophila* reference genome, release dm3, from the UCSC genome browser (http://genome.ucsc.edu/ (KENT *et al.* 2002)). Some of the regions overlap and some span more than two regions from a region bound by another factor. Therefore the total number of chip-regions is higher than when each factor is considered individually (Table 3.1).

Table 3.1: Overlap between ChIP-chip bound regions.

	Hunchback	**Krüppel**	**Snail**
HB only	470		
HB and Kr	1004	1004	
HB and Sna	13		13
HB, KR and Sna	319	319	319
KR only		1735	
SNA only			589
KR and Sna		104	104
Total	**1806**	**3162**	**1025**

The other large dataset used were sequencing data from the collaboration of Trudy MacKay and Baylor College of Medicine called DGRP (MASSOURAS *et al.* 2012). 162 separate *Drosophila* inbred lines from a single population in North Carolina were sequenced with an Illumina shotgun method to reveal the SNPs in the genomes of all the strains. The DGRP team then used algorithms to align and make contigs for each strain.

They then made available SNP files split up by chromosome arm (see example file Appendix C Figure C.6) from alignment of all the strains. The file, Final_Variants_2L, Final_Variants_2R, Final_Variants_3L, Final_Variants_3R and Final_Variants_X was downloaded from the Baylor College website. However, many of these SNPs are singletons and possibly due to sequencing errors but the data is too extensive to be manually verified. The data also includes the frequencies of SNPs in the sample and how many reads are behind each SNP. Thus all the singletons can be removed with appropriate filtering. I extracted all the SNPs that were located within the regions bound by each factor (see in table example Appendix C Figure C.1).

3.2.2 Finding binding sites using PWM's

Next I identified all putative binding sites for each of the three factors (HB, KR and Sn) in the areas bound by the three factors, a total of 9 different sets of possible binding sites. The reference genome sequence was used for finding all possible binding sites. I used position weight matrixes (PWM) to predict and calculate a score for each TFBS (downloaded from: http://www.danielpollard.com/bergman2004_matrices.html) (Table 3.2). Using PWM's to predict change in affinity to the DNA has shown to be quite accurate, especially if the change is not close to zero (Li *et al.* 2008; Struffi *et al.* 2011). A cutoff of 7 was used for the predicted binding sites. If the score for a site is higher than 7, then I deemed it a putative binding site and it is reported in a table (see Example of reporting Appendix C Figure C.6). There are three separate tables, one for each transcription factor. I found all binding sites, both from major and minor alleles. That enables me to also find most of the binding sites segregating in the population not only those that are present in the published genome sequence.

Table 3.2: The motifs and PWM scores used to score and predict binding sites.

	Positions									
Hb	1	2	3	4	5	6	7	8	9	10
A	0.32	0.17	0	0.01	0	0	0.01	0.57	0.19	0.04
C	0.25	0.13	0	0.05	0.01	0	0.04	0.07	0.17	0.16
G	0.31	0.07	0.03	0.01	0	0	0	0.28	0.14	0.48
T	0.12	0.64	0.97	0.93	0.99	1	0.95	0.08	0.5	0.31
Kr										
A	0.84	0.75	0.45	0.05	0.09	0.14	0.02	0.18	0.64	0.36
C	0	0.14	0.25	0	0.02	0.07	0.07	0.09	0.14	0.18
G	0.09	0.02	0.2	0.91	0.84	0.8	0.09	0.09	0.11	0.11
T	0.07	0.09	0.09	0.05	0.05	0	0.82	0.64	0.11	0.34
Snail										
A	0	0.36	0.91	0.09	0.09	0	0	0.09	0	0.27
C	0.64	0.45	0.09	0.91	0.45	0	0	0.55	0.18	0.18
G	0.27	0.09	0	0	0	0	0.91	0	0.18	0.27
T	0.09	0.09	0	0	0.45	1	0.09	0.36	0.64	0.27

3.2.3 SNPs within binding sites and deletions

Next step was to find SNPs within binding sites for each factor. I wrote a Python algorithm that takes the SNP location for each chromosome and goes through all the binding sites. The first criterion is to match the correct chromosome arm (2L, 2R, 3L, 3R, and X) and if a

SNP is found in the 10 base long binding site sequence it is reported. I first did a simple count where all the SNPs within a binding site for each transcription factor are located. There are a total of 45 (5 chr. arms x 3 TF bound regions x 3 TFBS within each regions) runs. Possible binding sites for the 3 TF are found in each of the three chip-bound regions. This is only a count of all the SNPs within the binding sites but gives no information about where each SNP lands within the site or its frequency.

We wanted to be able to remove all singletons since they are likely to be sequencing errors and not true SNPs, or do analysis with and without singletons. To do that I designed a new algorithm that reports each SNP found in a binding site in a table. That table lists the location of each SNP, major and minor alleles and the frequency of minor allele, the score of binding site, the coverage of the sequence and the sequence of the binding site itself (see example of output in Appendix C Figure C.6). Once the tables listing all the SNPs within binding sites for each factor of separate chromosomes has been written, all the singletons can be removed. That gives a more accurate count for SNPs found within binding sites.

In solving the problem of finding deletions of HB TFBS we collaborated with Thomas Zichner at EMBL (ZICHNER *et al.* 2012). He wrote an algorithm that finds indels and deletions in the DGRP data. He both scanned all our ChIP-chip bound areas as well as *eve*, including the 12 kb flanking regulatory region, and found all insertions and deletions. We provided him with the same files we had been working with. The data files Zichner ran for us were 3 ChIP-chip data files, or HB predicted binding sites within HB bound regions, KR predicted binding sites in KR-bound regions and SNA predicted binding sites in SNA-bound regions.

3.2.4 SNPs in binding sites with function in anterior-posterior patterning

I selected 21 developmental genes from a list of 54 listed in Bradley et al. (2010). I downloaded the sequences from UCSC genome browser dm3 version (http://genome.ucsc.edu/ (KENT *et al.* 2002)), the genomic locations for 21 anterior-posterior (A-P) determining genes (Table 3.2.1) and looked for all possible binding sites for the three factors using appropriate PWM as before. The protein coding part of the gene and non-coding sequence around it was studied. The surrounding sequences are included where known enhancers for the genes are located. I then used this dataset as an additional dataset and processed it the same way as the other larger datasets. These developmental genes are used because they are known to be regulated by those factors during A-P development of the fly.

Table 3.3: List of developmental genes and their chromosomal location according to dm3 in genome browser.

Chr. arm	Start	Stop	Name	Name abbr.
2L	3767904	3789896	*brother of odd with entrails limited*	*bowl*
2L	20767027	20786675	*caudal*	*cad*
2L	2432599	2464944	*decapentaplegic*	*dpp*
2L	3814824	3863077	*sloppy paired 1+2*	*slp1+2*
2L	15470168	15483951	*snail*	*sna*
2R	5859697	5878050	*even-skipped*	*eve*
2R	21101483	21127593	*Krüppel*	*kr*
2R	18929970	18940833	*twist*	*twi*
3L	14161159	14179469	*dachs*	*D*
3L	20677356	20695169	*knirps*	*kni*
3R	4506583	4532807	*hunchback*	*hb*
3R	165720	180579	*huckebein*	*hkb*
3R	2669573	2715812	*fushi tarazu*	*ftz*
3R	658543	697678	*odd paired*	*opa*
3R	26673989	26684347	*tailless*	*tll*
X	7189206	7217933	*brinker*	*brk*
X	9577781	9596312	*buttonhead*	*btd*
X	2314004	2338741	*giant*	*gt*
X	20548045	20573726	*runt*	*run*
X	15493496	15543292	*short gastrulation*	*sog*
X	468578	502624	*ventral nervous system defective*	*vnd*

3.2.5 Effects and location of SNPs within binding site

I next ran an algorithm (see Appendix D Algorithm D.2) that takes each SNP and calculates a new score including the mutated SNP and subtracts the old score of the binding site. The delta score change is then reported in a table that lists all the SNPs (see example table Appendix C Figure C.6). We noted for each SNP if they were a major allele SNP or minor allele. The effects of changing a single base in the binding site will either strengthen the binding site or weaken it. In some cases a binding site was created that had a PWM score higher than 7. That binding site was then reported in the table as well. Strengthening or weakening of a site will be reflected in the score calculated from the PWM. The score for the mutated binding site was calculated and compared to the old score. The SNPs that increased and decreased the score were counted.

3.2.6 Statistical analysis using R

The data was statistically analyzed with R (www.R-project.org version 2.15.2). The first thing I did was to throw out all SNPs where the count of sequences was below 30 because where only 30 strains had sequencing data out of 162 strains indicates an unreliable sequencing at that area (Table 3.4).

Table 3.4: How many SNPs are removed by removing those with a count below 30.

TF	Total	Removed	Used for analysis	% removed
Hb	30558	181	30377	0.006
Kr	63569	321	63248	0.005
Sn	11531	59	11472	0.005

I then checked where within binding sites the SNPs resided for each factor, HB, KR and SNA. I calculated a Chi-square value to see if the SNPs were found randomly in a binding site or if the constraints are different depending on where in the 10 base long sequence they are located. I also checked if the signal was different depending on the strength of a binding site by using different PWM scores as a cut off and noting if the pattern was different. The binding site scores ranged from 7-14. I also split up the dataset by reverse and forward sequences. That simply reflects the orientation of the genes, not a biological difference, and serves as two independent data sets. To test whether the strength of TFBS-prediction (measured with the PWM score) affected the evolutionary constraints I calculated *Pearson* correlation coefficient in R (*cor* and *cor.test*) for counts of SNPs in each location within the TFBS, in 5 different TFBS strength categories (sites with scores above 12, between 10 and 12, 8 and 10 etc). I used R program to plot up the mean and median frequency for the SNPs at each location within the binding sites and also permuted the data using bootstrap to test for outliers (data not shown).

In order to better understand our data we did further analysis. We wanted to identify what factors could possibly explain the frequency of major and minor allele of the SNPs within the binding site. We ran generalized linear model (*glm*) in R on different factors, as well as the interaction between factors. Our dataset is very large, therefore we have to split up our data table in half for R to be able to run analyses we are interested in. We ran the data files that have HB binding sites in HB bound areas, KR binding sites in KR bound areas and SNA binding sites in SNA bound areas. We used a binomial response variable that reflects the frequency of major and minor allele, and studied them with increasingly complex linear models. The equation had the general form:

Frequency = area + delta + sequence + score + position + error

We tested most of the variables that we imagined could provide signal. The four factors (see equation above) were those that could best explain the frequency. The area that was bound by a TF, the change in score when a new score with the mutation was calculated, the sequence of the binding site, the original score of the binding site and the position of the SNP within the 10 bp binding site. We are aware that possibly some of those variables are confounded, like the score and the delta score. To evaluate the additional contribution of each variable (or interaction of variables), we used Chi-square test of full vs. reduced models, and ascertained appropriate p-values.

Another idea we tested with R was how many or what fraction of predicted binding sites were found on the minor allele as opposed to the major allele. We scanned each TF-bound area and the predicted binding sites in each of them.

3.3 3+7 enhancer mutations

I looked more closely at the *eve* enhancer region in the 162 sequenced strains. I aligned the *eve* area, using Genedoc (NICHOLAS *et al.* 1997), where the five enhancers for the stripes in the embryo are found. I compared every sequence at every Hunchback binding site to see if there are any SNPs or other deletions found in them. Research shows that enhancers are well conserved (LUDWIG 2002) and therefore it would be interesting to see if mutations in hb binding sites are found that are at a different haplotype than where the deletions are found.

3.4 Comparison of *hb* protein between species

One of the hypotheses was that the *hb* protein itself has changed between *D. melanogaster* and *D. simulans*. The amino acid sequence for the protein and the 12 closely related *Drosophila* species was retrieved through UCSC genome browser (http://genome.ucsc.edu/ (KENT *et al.* 2002)) and aligned for comparison. Pfam (http://pfam.sanger.ac.uk/ (FINN *et al.* 2010)), a protein database, was used to find the DNA binding domains. Then I could check specifically if any amino acid changes are found in the DNA binding domains. The reference TFs Snail and Krüppel were also studied with Pfam. Hunchback and Snail have three DNA binding domains and Krüppel has five.

3.5 Binding sites for Pumilio

Pumilio is a protein that binds to *hb* mRNA in the posterior portion of the embryo and inhibits translation of *hb* (MURATA and WHARTON 1995). To see if there are any mutations of the Pumilio binding sites I downloaded the 3’ UTR region of HB and looked at the alignment for the 4 closest relatives of *D. melanogaster*. The exact Pumilio binding sites have been characterized (MURATA and WHARTON 1995).

3.6 *hb* promoter between and within species

The zygotic and maternal promoter locations were downloaded from UCSC genome browser (http://genome.ucsc.edu/ (KENT *et al.* 2002)). I then used the locations to extract the sequences from DGRP data for the zygotic and maternal promoters to see if any polymorphisms are seen in them. I then used genome browser to find the sequences of individual binding sites previously characterized, to see if any SNPs are seen in them. I also downloaded the sequences of the closest relatives to *D. melanogaster* and aligned the promoter sequences to see if any of the binding sites for either Hunchback itself or Bicoid had changed between *Drosophila* species.

4. Results

4.1 Other HB binding sites are deleted/mutated

The main hypothesis of the thesis was that if the concentration of Hunchback had increased then the deletions observed in *eve* were a response to that. Testing for a change in the concentration of a transcription factor this early in development within a population or between species is beyond the scope of this project, but we can pursue this hypothesis in other ways. It is possible that the concentration of HB had increased (for instance through changes in the regulation of the gene or the product stability). We hypothesized that other Hunchback binding sites in developmental genes would be affected as well, either by severe mutations, reflected in a negative change in the PWM score, or even abolished binding site, or more deletions of areas containing binding sites for Hunchback genome wide. However, deletions and point mutations can be expected by chance (GOERING *et al.* 2009; HE *et al.* 2011), then we need to establish a point of reference. Specifically, is a higher fraction of HB sites damaged, than binding sites for other TFs? Are more severe mutations found in HB compared to sites for other factors? When testing if other HB binding sites are damaged I used two transcription factors that have the same type of DNA binding domains, zinc finger domains (C2H2), for reference (Snail and Krüppel). In order to test this I used SNPs from the DGRP data and ChIP-chip data (LI *et al.* 2008; MACARTHUR *et al.* 2009; MASSOURAS *et al.* 2012) and extracted the sequence of the regions where these factors bind in the Drosophila genome. I then ran an algorithm to search for all possible binding sites that each factor can potentially bind to according to the PWM. I calculated the PWM scores for each site, for both the minor and major allele, to estimate the effects of a single mutation on the site. The changes in PWM were consolidated into 3 categories (Figure 4.1), positive (blue), mildly harmful (green), extremely harmful (red). If our prediction was true, one might expect a higher fraction of harmful mutations in HB sites, than in binding sites for the other two TFs (KR and Snail). The results do not corroborate our hypotheses, the ratio of severely mutated HB binding sites seems not to be higher than for the comparison factors on a genome wide scale (Figure 4.1). According to the data the fraction of mutations that increased the score was highest in HB binding sites. The highest ratio of mutations with severe effects on the score was seen for KR. Snail binding sites showed the highest ratio of slightly negative effects on the PWM score. These findings are not in concordance with our hypothesis, quite the contrary actually. The high effects on KR binding sites indicates that possibly something is affecting KR. It could also be that the PWM for KR is not accurate or that changes in KR binding sites do in general have detrimental effects on the score of the site.

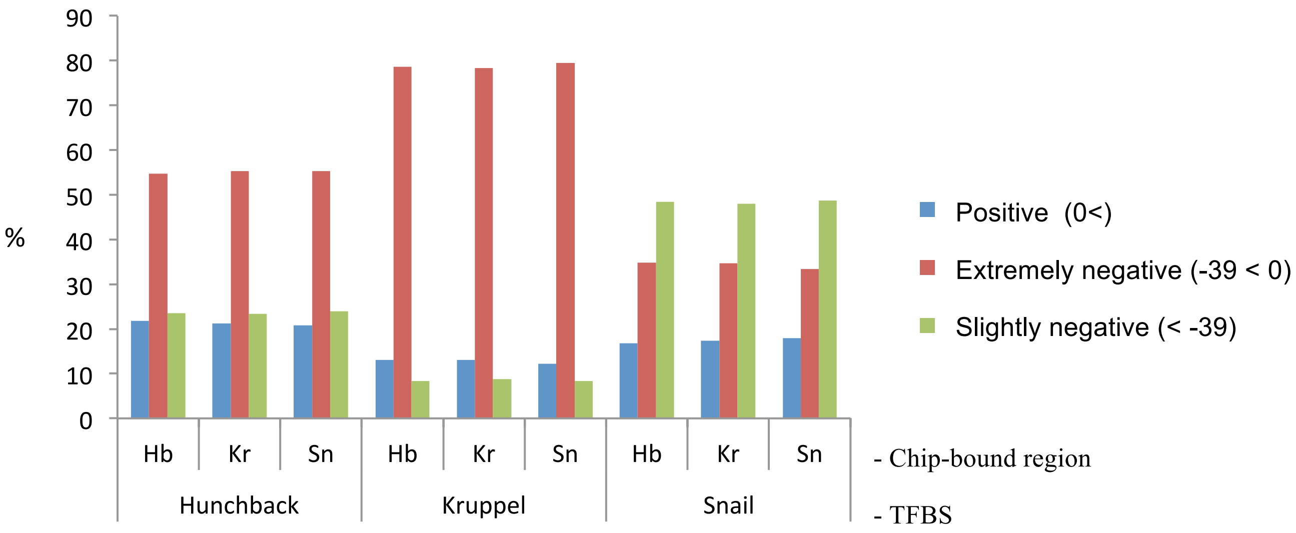


Figure 4.1: The effect on the PWM score by a mutation in a binding site. Blue bars represent increased in score of a binding site. Red bars represent slight negative effects and green extremely negative effects on the PWM score of the binding site. The bound regions look very similar depending on what factor is bound.

The analyses have, for the most part ignored the subtleties of binding site structure. I wanted to document where mutations were hitting within the 10 base pair binding sites. Is it random or are there some constrains on where and which SNPs are tolerated? The location of each SNP within each binding site was retrieved and I summed up all SNPs depending on position within the TFBS. The analysis was handled in R. The number of mutations is not uniform among positions within the binding sites (Figure 4.2, see appendix E Figure E.5 for statistical results). The SNPs within binding sites are not randomly distributed, for the three factors (HB, KR, Sn). The data is split up according to which factor they are bound by of the extracted areas. When the distribution of the SNPs within the binding site is compared the pattern looks very similar for each TF (Figure 4.2) and also for forward and reverse. That indicates that the SNPs are not randomly distributed within the binding site. There seem to be some constraints on where within the binding site mutations hit.

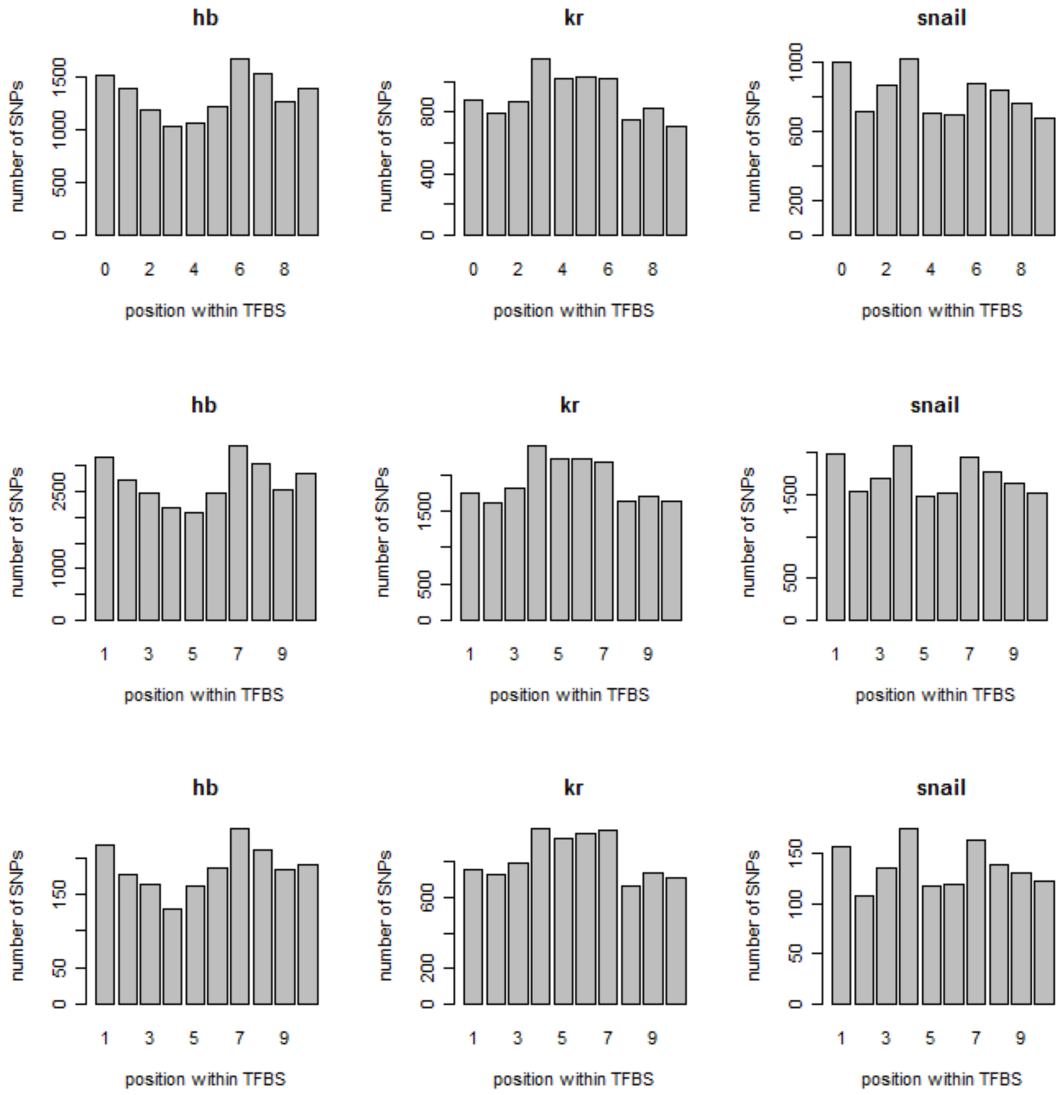


Figure 4.2: Distribution of SNPs in HB, KR and SNA binding sites TFBS in 1762 Hunchback bound regions (top), 3028 Krüppel bound regions (middle), 595 Snail bound regions (bottom). See appendix E Figure E.5 for statistical analysis.

One hypothesis was that more deletions of HB binding sites would be in higher frequency than for the reference factors, KR and SNA. We collaborated with Thomas Zichner at EMBL that had been looking at insertions and deletions in the DGRP data (ZICHNER *et al.* 2012). He looked for indels in the chip-chip areas and in TFBS within them, using the 200 sequenced DGRP lines.

For the large dataset, Zichner found numerous deletions. He both scanned all the chip-bound regions for all the factors and all the binding sites within them. Then he limited the run for the corresponding factor, or all binding sites for HB within HB-bound areas, KR binding sites within KR-bound areas and SNA binding sites within SNA-bound areas (Table 4.1). Chi-square tests show that a higher fraction of HB sites is damaged by deletions, compared to KR and SNA TFBSs (see Appendix E Tables E.1-E.4). A statistically significant difference between areas was not seen and neither the test of significance for difference in duplications between both areas and TFBS (Tables 4.1 - 4.2).

These results are the strongest indicator we have for the prediction that HB binding sites are subject to more deletions than the reference factors KR and SNA.

Table 4.1: Binding sites within the TF bound regions affected by deletion or duplication. Statistically significant that there is difference between factors.

TF	Number of TFBS	Affected by deletion	%	Affected by duplication	%
Hunchback	19421	1346	**6.9%**	661	3.4%
Krüppel	27284	1711	**6.2%**	988	6.3%
Snail	2494	100	**4.0%**	95	3.8%

Table 4.2: Regions affected by deletions or duplications within the TF bound areas.

TF	Number of Regions	Affected by deletion	%	Affected by duplication	%
Hunchback	1760	580	32.9%	130	7.3%
Krüppel	3028	1113	36.7%	277	9.1%
Snail	595	100	16.8%	31	5.2%

4.2 Polymorphisms in TFBS in developmental gene regions

The preceding data did not suggest a genome wide signal of more mutations in HB sites than the reference transcription factors. But it is still possible, that a more restricted tendency for loss of HB sites exists. Therefore, to test for a localized signal I randomly chose 21 developmental genes that are known to be regulated by transcription factors early in development (BRADLEY *et al.* 2010) at the same time/stage as *eve* is expressed and possibly under HB regulation. There is similar pattern of polymorphism in the developmental genes dataset as there is in the large dataset (Figure 4.3). Some constrains seem to be on where within the binding site mutations are likely to hit. The difference between this dataset and the genome wide datasets is that binding in the proximity of developmental genes is known to have real effects on gene regulation, whereas binding genome wide like the larger ChIP-chip dataset, includes many weakly bound less conserved and characterized regions which may reflect "background" or random binding (LI *et al.* 2008; MACARTHUR *et al.* 2009).

I performed the same analysis on the 21 developmental genes as the DGRP data and ChIP-chip data. I extracted the sequences for the 21 developmental genes, including the regulatory DNA around them. I then ran an algorithm to search for all possible binding sites that each factor can potentially bind to according to the PWM. I counted how many of the binding sites found had a SNP within the 10 base pair long sequence and how many did not contain SNPs. The counts were summed up in table 4.3 and did not show a distinct signal for HB compared to the other factors, however, it is KR predicted binding sites within the genes that have binding sites that are most often hit by SNPs.

Table 4.3: Count over what factor has the highest frequency of SNPs within binding site (see Table 4.3 and full table in Appendix E table E.7 for counts of each gene).

TF	BS
Hb	7
Krüppel	11
Snail	3
Total	21

Table 4.4: Ratio of total SNPs and ratio of all SNPs within BS for the three factors HB, KR and Sn. 4 of the 21 developmental genes (see full table in Appendix E table E.7).

Chr	Gene	TF	Within BS	SNPs in that gene	Total BS in gene	Ratio of total	Ratio of all BS
2L	***bowl***	**HB**	29	1061	63	**0.027**	**0.460**
		KR	24	1061	54	0.023	0.444
		SNA	27	1061	62	0.025	0.435
2L	***caud***	**HB**	10	475	58	0.021	0.172
		KR	16	475	47	**0.034**	**0.340**
		SNA	5	475	46	0.011	0.109
2L	***dpp***	**HB**	23	999	87	**0.023**	0.264
		KR	15	999	53	0.015	**0.283**
		SNA	28	999	103	0.028	0.272
3R	***hb***	**HB**	26	454	107	**0.057**	**0.243**
		KR	9	454	67	0.020	0.134
		SNA	9	454	60	0.020	0.150

The data from TFBS in the developmental gene regions confirmed what I had seen for the ChIP-chip areas. It is not random where the mutations hit within binding sites (Figure 4.3).

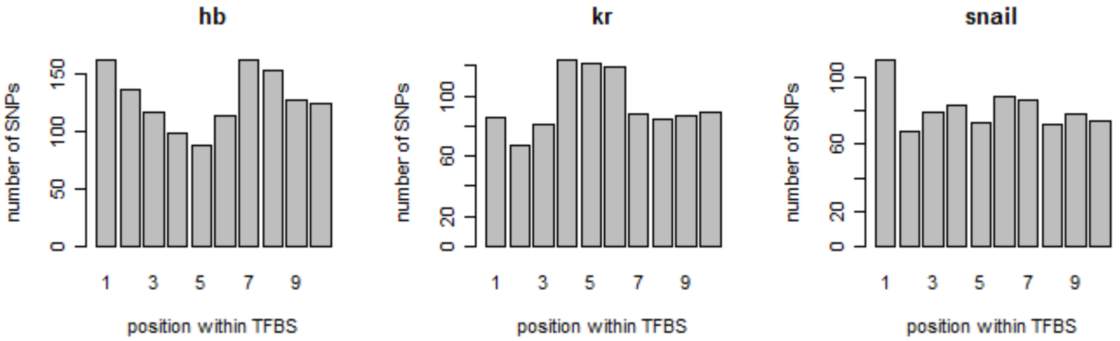


Figure 4.3: Distribution of SNPs within TFBS in the genomic regions of 21 developmental genes. (See statistical analysis in Appendix E Table E.6).

The developmental gene data, was smaller and contained much fewer TFBS and therefore it was possible to plot the data using different methods than for the larger dataset. Figure 4.4 does not indicate a distinct pattern of the frequency distribution of SNPs within the binding site. However an interesting result is in how many SNPs in high frequency are seen at position with high information content.

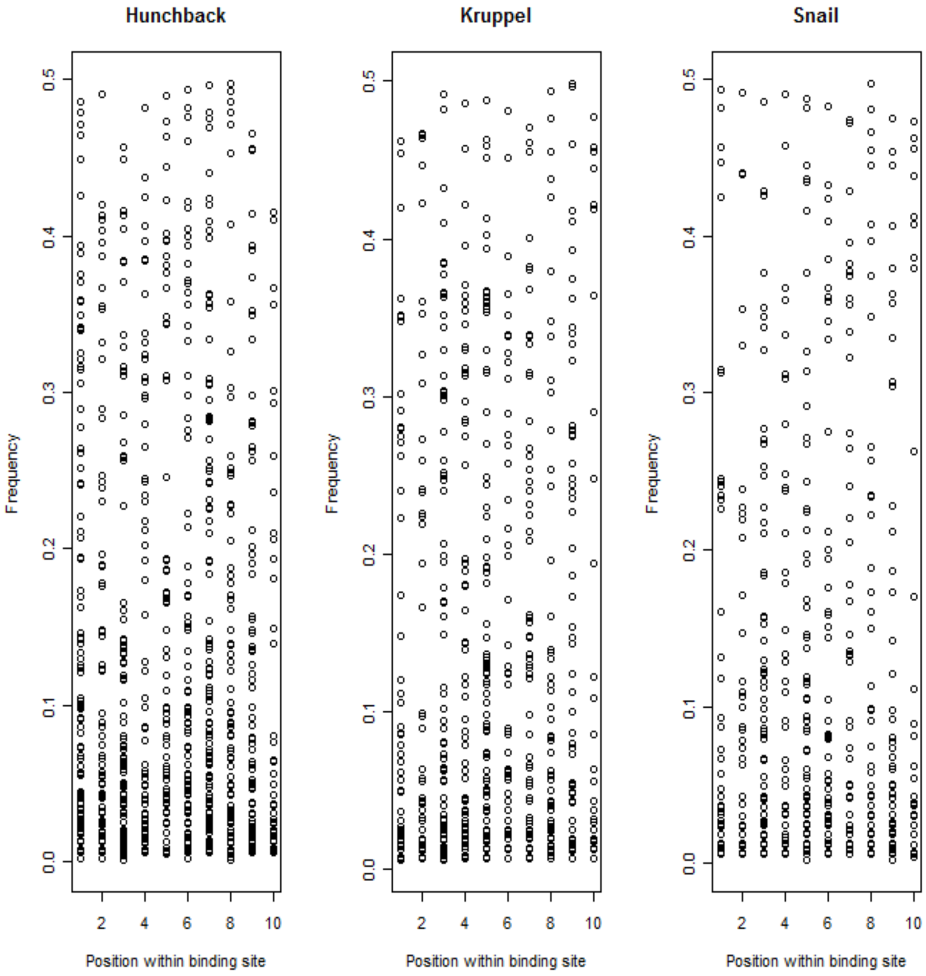


Figure 4.4: Frequency of SNPs at each position (1-10) within TFBS in the genomic regions of 21 developmental genes in binding sites for HB, KR and SNA.

I calculated the mean and median frequencies for HB binding sites of the developmental genes. The results are rather baffling, both the mean and the median are non-consistent with evolutionary predictions. The fact that bases number 4, 5, and 6 within the HB binding site, which have a PWM score of ~1 have the highest median frequency (Figure 4.5). That suggests that in those specific locations, there are quite a few SNPs at high frequency.

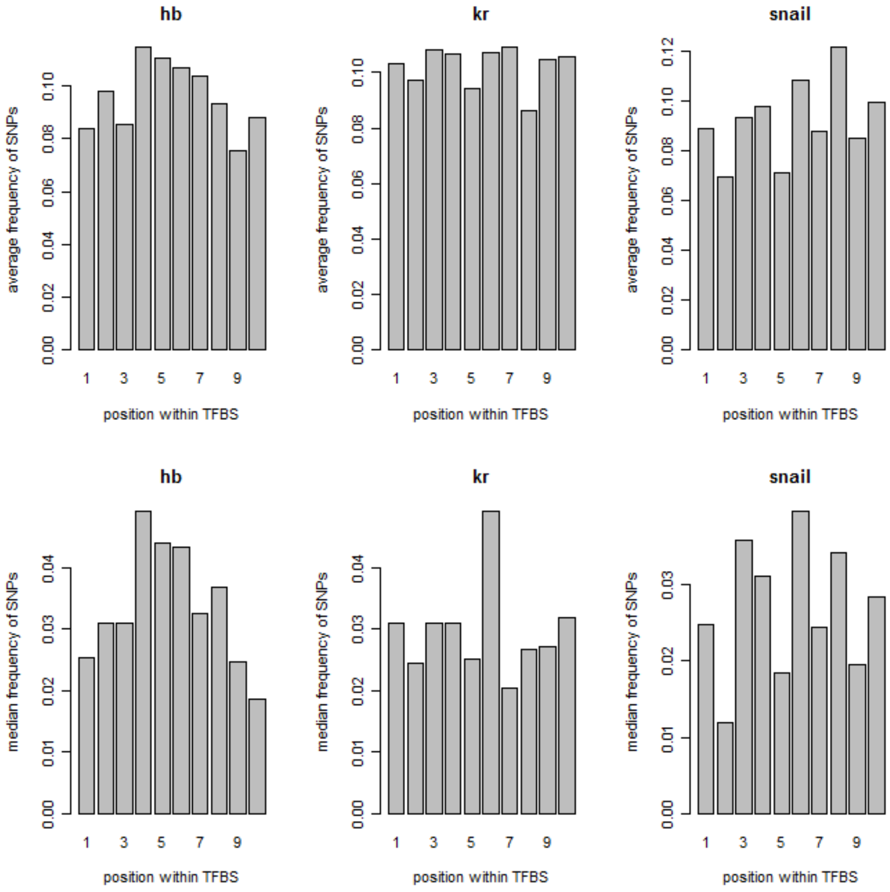


Figure 4.5: Mean (above) and median (below) frequency for SNPs in TFBS in Developmental gene regions.

4.3 *hunchback* protein

One of the predictions from our central model was that the *hb* protein may have changed. The possibilities include:

- Changes in DNA binding affinity and/or domain
- Changes in activation strength and/or domain
- Change in the stability of protein

To see if this was the case, I used comparative genomics. Alignment of the *hb* proteins from distinct species can reveal changes that might have occurred on the *Drosophila melanogaster* lineage. In the study I used two TFs for reference, chosen because they have the same DNA binding domain as HB (zinc-finger C2H2). Many things regarding the structure and function of those DNA-binding domains are unknown, for example i) if there is more than one functional domain in each protein, ii) are they all functioning at the same time, and iii) if there are multiple domains per TF, do they all have the same purpose or is one more commonly that is predominantly bound to the DNA? Perhaps the different domains are not all accessible and are more than one in a protein because of evolutionary history? Also, each protein functions differently and what seems to be a poor binding for one protein may be the optimal for another. It is beyond the scope of this study to answer those questions, but I wanted to see if there were any obvious signs of possible functional changes. I therefore compared the proteins using amino acid sequences of Hunchback, Snail and Krüppel and ran them through Pfam protein database. Pfam predicts, using HMM scores, what type of domains are found in the protein. For all the proteins, zinc-finger C2H2 DNA binding domains were found. Hunchback and Snail had three domains each, where Krüppel has four.

An interesting fact is that the scores for both Snail and Krüppel domains were higher than all the Hunchback domains, indicating that Hunchback might possibly be evolutionary less related or less constrained than the other two. The calculated HMM scores for Hunchback where 11-13, Krüppel domains scored 17-20 and Snail 21-29.

Figure 4.6 shows the alignment of HB from the 11 closest relatives of the *D. melanogaster* downloaded from UCSC genome browser in March 2010. The most diverged species *D. grimshawi*, is separated by 40-50 million years from *D. melanogaster* (RUSSO *et al.* 1995; CLARK *et al.* 2007). Comparison of the protein between the species shows that it is fairly similar structure in the species inspected. A closer comparison of only the DNA binding motifs (blue lines below the alignment in Figure 4.6) shows that they are identical in all species, except a single amino acid in the first motif shows some variation but not in the 4 closest relatives to *D. melanogaster*. Certainly there are some changes in the less characterized parts of the protein and functionally those regions cannot be ruled out to have significance for the protein function. However, the amino acid sequence is extremely similar between the 5 closes relatives, *D. melanogaster, D. simulans, D. sechelia, D. yakuba* and *D. erecta.* Furthermore, if the whole protein is studied there is not a single amino acid change on the *D. melanogaster* branch that is not seen in any of the other species. The amino acid sequence for *D. melanogaster* does not stand out as being different from its closest relatives. It is only the more distant ones that do show some diversity in the protein sequence.

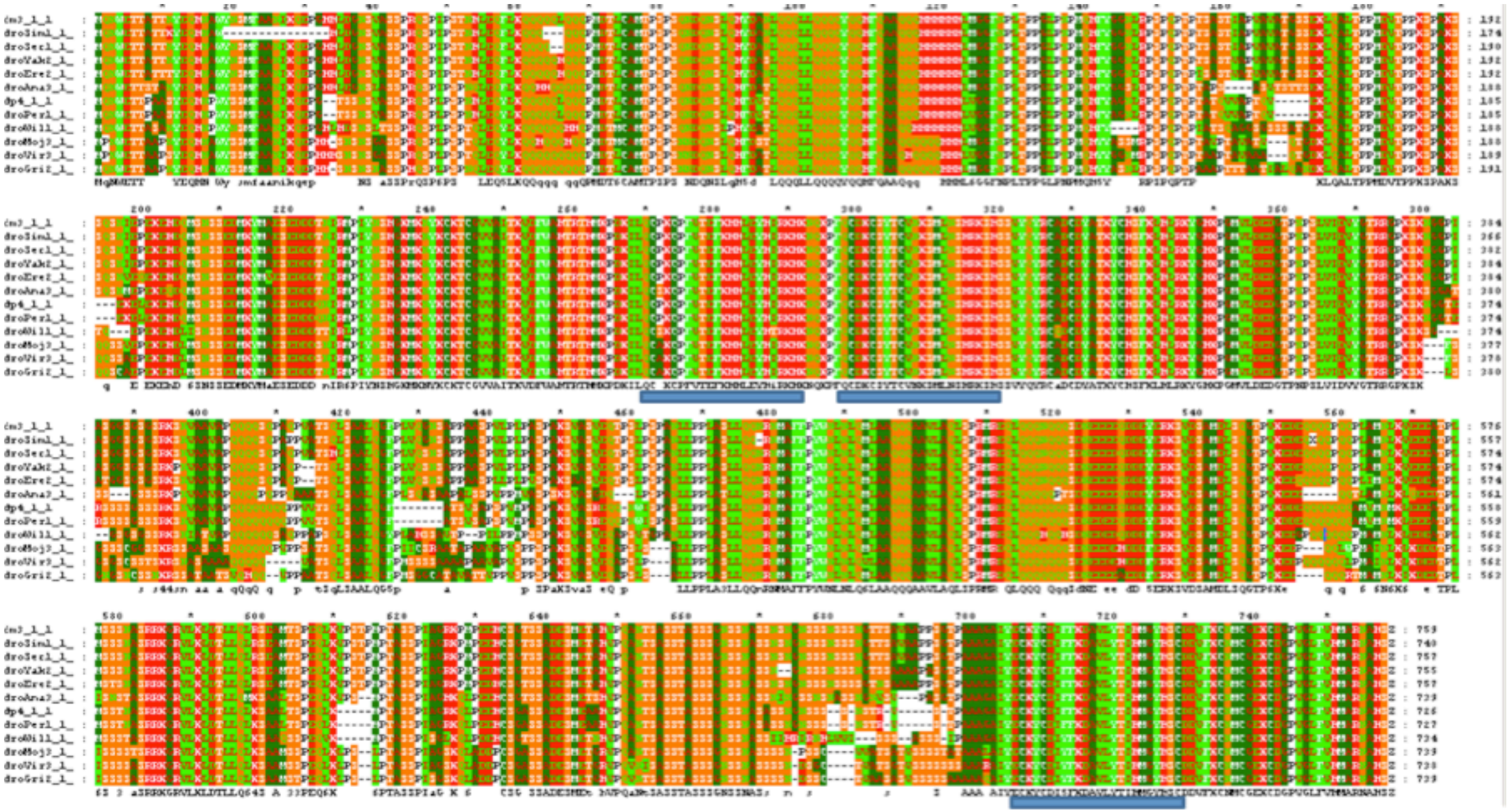


Figure 4.6: Alignment of closely related Drosophila *species of Hunchback. The blue lines give the location of DNA binding domains according to Pfam. Their exact locations are: 269- 291, 297-319 and 705-727. The species are from top* D. melanogaster, D. simulans, D. sechelia, D. yakuba, D. erecta, D.ananassae, D.pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis, *and* D. grimsawi.

4.4 Hunchback promoter between species

To look for other differences that might have lead to changes in the transcription of *hunchback* in the *D. melanogaster* linage we looked at the two characterized HB promoters and compared them between the close relatives of *D. melanogaster* or *D. erecta*, *D. yakuba*, *D. sechelia* and *D. simulans*. We checked both the zygotic and maternal promoters (WATSON 2008). When the sequences are aligned some differences are seen between the species (Figure 4.7 and 4.8). However, when the characterized binding sites for Bicoid or Hunchback in the promoter are examined, they are found to be conserved between the species. In the maternal promoter changes in two binding sites are found (Figure 4.7), for the zygotic promoter no SNPs are found between the species in the two characterized HB binding sites (Figure 4.8). However, those changes sit on the *D.sechelia* branch and indicate that divergence on that lineage, not the *D. melanogaster* lineage. I also checked patterns of polymorphism in the promoter within the *D. melanogaster* and used the SNPs from the 162 DGRP lines for that. The alignment showed that there are no SNPs in the characterized binding sites for either Bicoid or Hunchback in the HB promoters, which further supports the notion that these binding sites are under strong purifying selection to maintain their function.

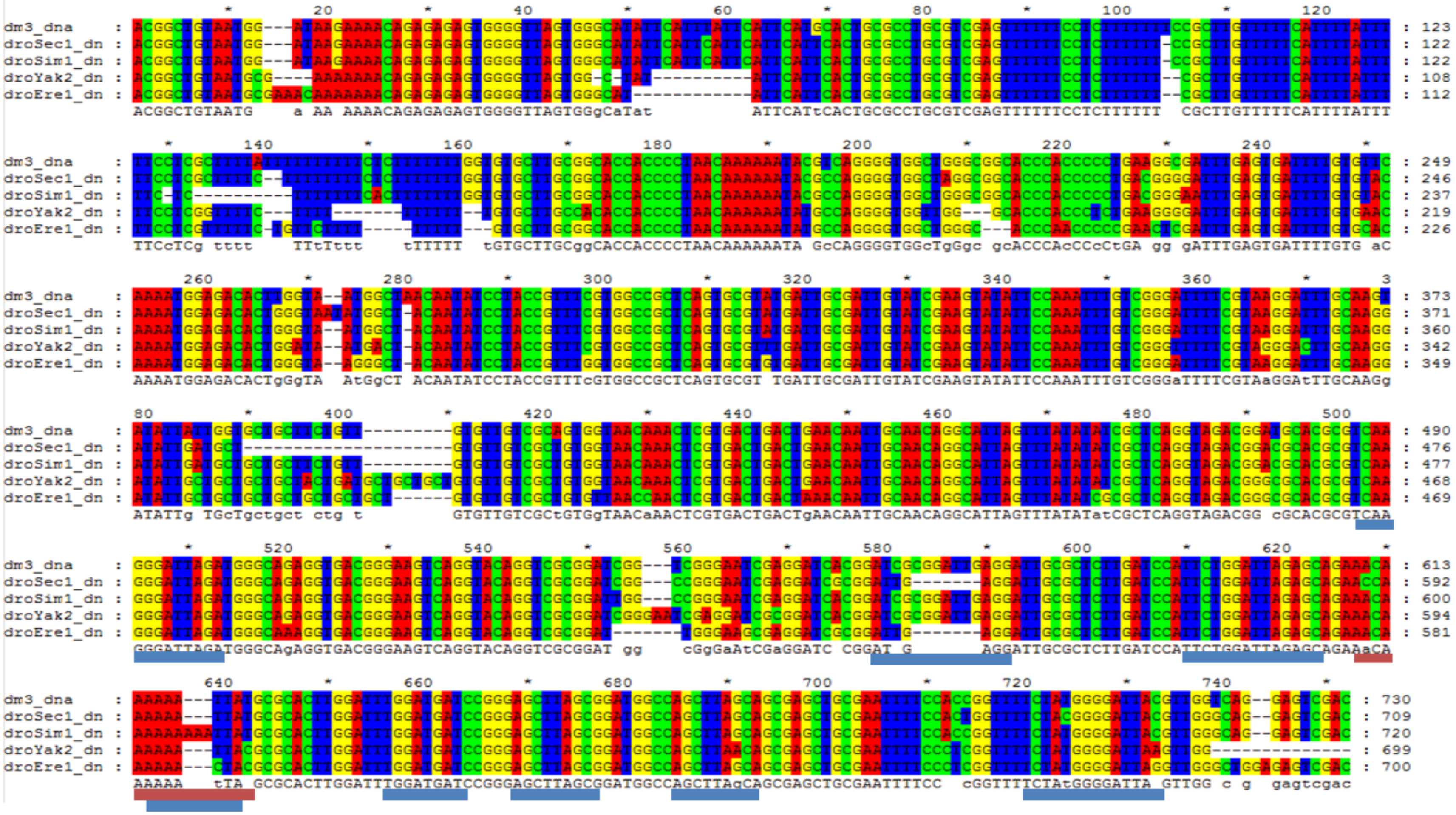


Figure 4.7: The alignment of the maternal hb promoter between D. melanogaster *and its closest relatives* D. sechelia, D. simulans, D. yakuba, *and* D. erecta*. All characterized (8) Bicoid binding sites in blue, one characterized HB binding site in red.*

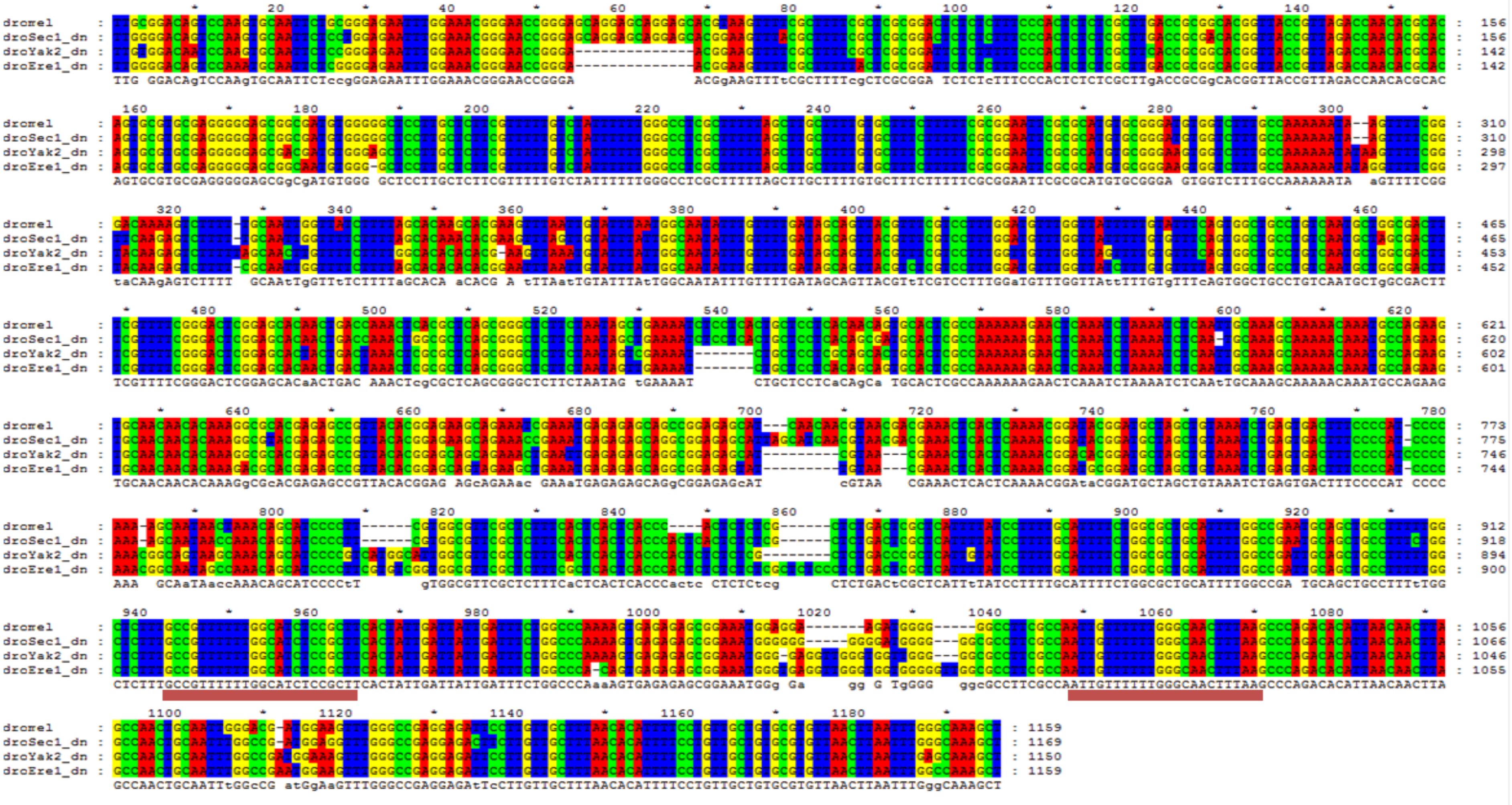


Figure 4.8: The alignment of the zygotic hb promoter between D. melanogaster *and its closest relatives* D. sechelia, D. simulans, D. yakuba*, and* D. erecta*. Characterized HB binding site in red.*

4.5 Changes in 3’ UTR

One possibility is that HB dose has changed, for instance through changes in mRNA stability or half-life. To evaluate this we looked for changes in functional elements in the 3’ UTR of *hunchback* mRNA. A few binding sites for proteins that are known to regulate HB expression in the embryo. Among them is a binding site for Pumilio, which is essential for translational regulation of *hunchback* (MURATA and WHARTON 1995). One of our hypotheses was that the expression of HB had shifted and the deletions in *eve* s3+7e were a response to that. Alignment of the 3’ UTR between the most related *Drosophila* species showed no change in the characterized Pumilio binding site. Another known developmentally important structure in the 3’UTR is a binding site for miRNA-8 (ROY *et al.* 2010). The multiple alignment revealed no change in that binding site on the branch leading to *D. melanogaster* (data not shown).

4.6 Polymorphism and tests for selection in the area surrounding *hb*

In order to look for signs of recent positive selection in the *hb* area we sequenced parts the gene and the surrounding regions. Recent positive selection leaves a mark of unusually long undisrupted haplotype(s), at high frequency or fixed, in a given region. In other words little polymorphism in a region, along with such haplotype structure can indicate positive selection (SABETI *et al.* 2007). I analyzed the polymorphism (measured by π) in 15 regions spanning 20 kb, 10 kb downstream and 7 kb upstream from *hb* (Figure 4.1). The results do not indicate positive selection in the area surrounding *hb*. In the proximity of the promoter and the protein itself there was a drop in polymorphism. However, that drop is normal and is usually seen in proximity to functionally important proteins due to purifying selection. At this point in the study the DGRP where released. For comparison I extracted the entire *hb* (23,746 bases) area from the DGRP data and also calculated the polymorphism (measured by π) in Tassel. A sliding window analysis of average pair wise polymorphism on the DRGP sequences was similar to my results (Figure 4.9). The promoter is located between 4519 kb and 4520 kb on chromosome 3R. The sequences showed no polymorphism in the promoter which is what is to be expected, however, directly upstream from it the value for π increased considerably. There is no evidence of long haplotypes, which would signal positive selection, in the genomic region around *hb* (data not show).

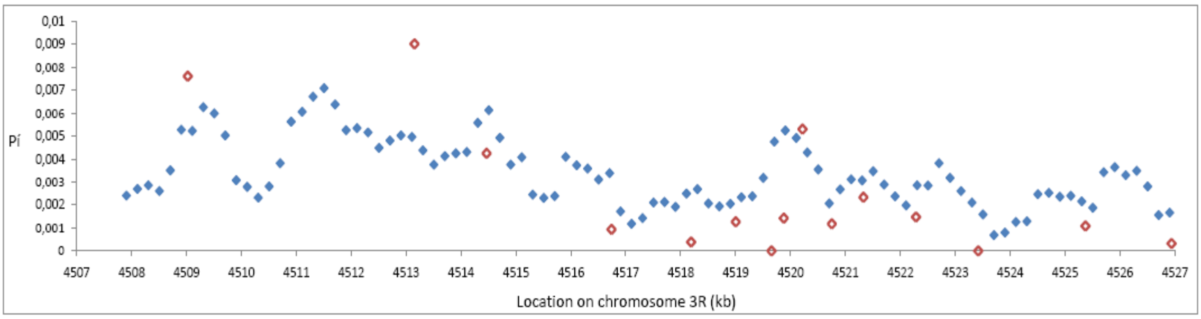


Figure 4.9: Sliding window of π in the HB region. For DGRP data of 162 individuals (blue) and the North Carolina data 15 regions (red, from 32 individuals). The window ran for the DGRP data is 1000 bases, sliding every 200 bases.

4.7 3+7 enhancer mutations, other binding sites within *eve*

As was described in the introduction, a sequencing survey of *eve* enhancers revealed two deletions taking out HB binding sites. In the study only a portion of the *eve* region was surveyed. I wanted to know if other mutations are affecting HB sites within the *eve* region. By aligning the regulatory region for *eve* from of the 162 DGRP strains and predicting Hunchback binding sites, I could see if any SNPs or other deletions are observed in TFBS. I found a HB binding site within the 3+7 enhancer that has the highest score possible for each base in the PWM. There was a SNP located at a site number 6 within the 10 base pair motif that has a PWM score of one. That means that in all binding sites examined, that base is always the same. I found 7 individuals that had that SNP which is about 4.3% frequency. The mutation sits on a different haplotype than the deletion Δhb8. In the dataset there were 22 individuals, or 13,6% of 162 that had the Δhb8. In the dataset I did not see the other mutation, ΔhbS1, (Palsson unpublished) because it is not in the *D. melanogaster* reference genome. The algorithms that are used to edit the sequences and build the DGRP SNP file throw out the section that spans the deletion. I wanted to see if there was a possibility if the deletion of the site was possibly creating a new binding site. When I scored the possible new site using PWMs the score was only around 4 that is not a high score and below my cutoff (data not shown).

Zichner scanned the *eve* region for deletions and insertions. For the *eve* region we only saw the previously characterized deletions that Palsson (unpublished) found, which confirms his findings (Table 4.5). The smaller deletion, 45 base pairs long, is documented as an insertion rather than deletion. The reason for that is it is not found in the *D. melanogaster* reference genome even though the 5 closest relatives of *D. melanogaster* have the deletion. Other deletion events in the *eve* regions did not hit within binding sites with the exception of two small insertions in Snail binding sites.

Table 4.5: Insertions and deletions in the eve region.

Location on chromosome 2R	Number of deleted bases	Number of inserted bases	Number of samples
5857929-5857931	1	0	20
5858473-5858484	10	14	20
5858668-5858670	1	0	66
5860495-5860497	1	0	56
5860574-5860583	8	0	28
5862205-5862208	2	0	51
5862449-5862451	1	0	50
5862449-5862457	7	6	42
5863472-5863535	**62**	**0**	**27**
5868513-5868522	8	0	67
5868944-5868949	4	0	47
5868961-5868964	2	0	24
5869107-5869115	7	23	55
5870850-5870852	1	7	33
5871409-5871417	7	0	110
5872003-5872005	1	0	31
5873269-5873277	7	9	94
5858471-5858472	0	4	152
5862672-5862673	0	1	23
5863526-5863527	0	10	53
5863532-5863533	0	10	48
5863775-5863776	**0**	**45**	**36**
5865068-5865069	0	14	32
5866508-5866509	0	2	154
5868593-5868594	0	1	29
5872003-5872004	0	1	74

4.8 Other results

4.8.1 LD between *eve* and Hb

One of our predictions (4) was co-evolution between HB and *eve*. If there is some functional interplay between any SNPs in *eve* and *hb* they would be linked together in a LD block. We tested that using Tassel (Bradbury *et al.* 2007) on the 162 DGRP lines. There were no interesting signals seen when we were testing for LD between them (data not shown).

4.8.2 Factors influencing frequency of SNPs in binding sites

We ran generalized linear model (*glm*) analysis to see what factors were affecting the frequency of SNPs within transcription factor binding sites. Each SNP was represented by the frequency of the major allele or minor allele. No information was provided on which allele was ancestral and which was derived. I built an increasingly more complex linear model, testing for the addition of each term. The most complex model we saw that explained the SNP frequency in HB TFBS in HB regions is composed of 5 factors (Tables 4.6 and 4.7). We saw that there was a difference between areas, in other words it matters in what area the binding site and SNP were found in. The change in score (as a consequence of the mutation) and the original score (of the TFBS) also contributed significantly. Beforehand we did imagine that the delta score would be a contributing factor but since the score is somewhat integrated into the delta score we were surprised to see that it also affected SNP frequency. Curiously, the primary sequence of the binding site also seemed to contribute. The one factor we were certain, before the analysis, that would be a contributing factor was the position of the SNP within the binding site. This was confirmed by the linear model. The best model that explains the frequency of SNPs in HB TFBS in HB regions were those 5 factors and in a biological context that makes considerable sense. Very comparable results were seen for KR sites in Kr-bound regions (data not shown).

Table 4.6: Statistical analysis of factors contributing to the frequency of SNPs in HB TFBS in Hb-bound regions. Here the addition of original score of the binding site as a contributing factor is tested.

Model	Equation	Resid - Df	Resid - Deviation	Df	Deviance	P-value
Reduced	Freq = area + delta + position	3437	81046			
Full	Freq = area + delta + position + score	3436	81015	1	30.407	3.502e-08

Table 4.7: Statistical analysis of factors contributing to the frequency of SNPs in HB TFBS in Hb-bound regions. Here the addition of original sequence of the binding site as a contributing factor is tested.

Model	Equation	Resid - Df	Resid - Deviation	Df	Deviance	P-value
Reduced	Freq = area + delta + score + position	3436	81015			
Full	Freq = area + delta + score + sequence + position	2497	52438	939	28577	2.2e-16

4.8.3 Fraction of binding sites on major or minor allele

During the processing of the SNP and binding site data we came across an interesting observation. To appreciate this result it is crucial to highlight the principles of the algorithm. First we edited in all the major alleles of all SNPs in a chip region, into the primary sequence for that region. Then we predicted binding sites, and classified them as TFBS on the major allele. Then we edited in the minor allele, and ran the prediction again, yielding TFBS that where essentially abolished by the major allele. When we did the bookkeeping we found that a significantly higher fraction of total predicted HB sites sit on minor allele (see Figure 4.10 and Appendix E for statistical analysis), compared to KR and SNA sites on the minor allele (Table 4.8). The same pattern was seen for the three types of chip regions, Hb-bound, Kr-bound and Sna-bound, only more pronounced in the HB region.

We have two possible explanations for this pattern, i) new HB binding sites are being created at a higher fraction then for the other TFs (perhaps via unknown primary sequence or mutational distance mechanism), or ii) binding sites that used to reside on the major allele are being selected against but are still seen on the minor allele. The latter explanation is consistent with our predictions. But it is a weak signal, and does not lend concrete support to our model.

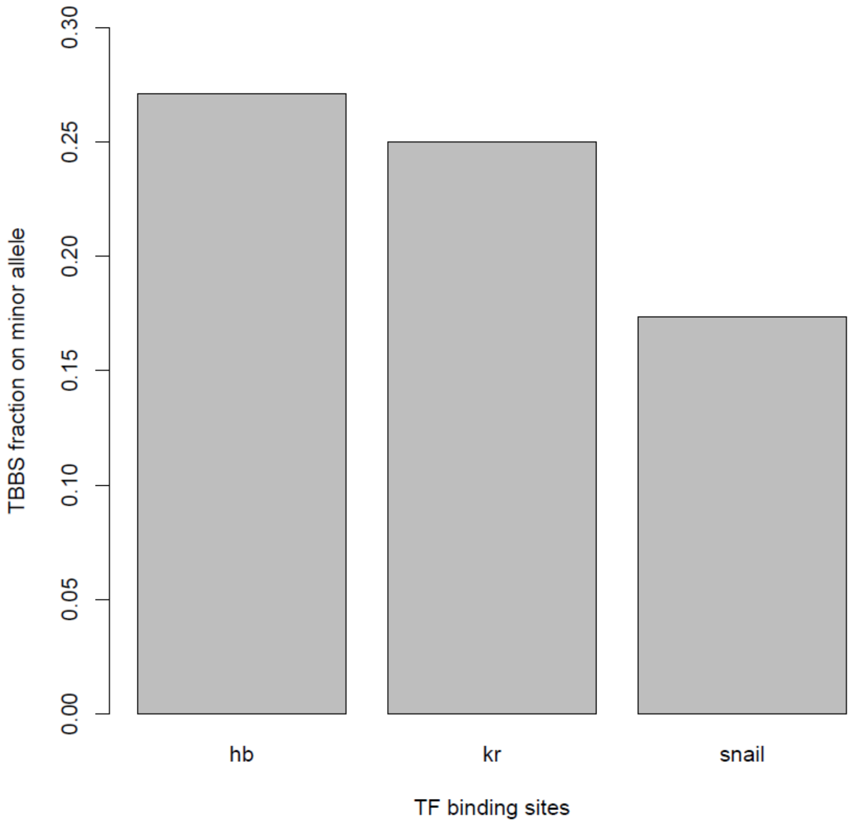


Figure 4.10: Fraction of binding sites that are found on minor allele in HB-bound areas.

Table 4.8: Chi-square tests for ratio of binding sites on minor allele.

Factors tested	Chi-square	Degrees of freedom	p-value
Hb-bound	249.5323	2	2.2e-16
Kr-bound	420.1387	2	2.2e-16
Sna-bound	48.6761	2	2.692e-11

4.8.4 Correlation between information content and frequency of SNPs

Based on evolutionary theory, a negative correlation between the PWM information content per base and the SNPs at each position within the TFBS is expected. Information content is another way to interpret PWMs for transcription factors. If a PWM score for a base at a certain position then the information content is 2, in other words, in all observed binding sites (for instance from functionally characterization using DNase I digestion). When we tested the correlation coefficient using t-test we saw a positive correlation signal for KR (t = 3.77, df = 8, p-value = 0.0054, corr =0.799). This is in the opposite direction to what we expected to see. We saw no signal for HB (t = -1.48, df = 8, p-value = 0.17) or SNA (t = 1.10, df = 8, p-value = 0.29) (Figure 4.11). The positive signal for KR is an interesting observation and indicates that there is something going on in KR binding sites. However, lack of a negative correlation in HB sites, is a weak indicator that we are seeing abnormally many SNPs in high frequency at well conserved positions within the binding site with high information content. The reason why no signal was found in SNA sites, might be that the predicted sites are fewer than for the other two factors as there are considerably fewer SNA-bound areas. If we had seen a negative correlation for KR and SNA we could put more faith in this result, however, seeing as we are unable to see this for the reference factors this result might simply be a fluke.

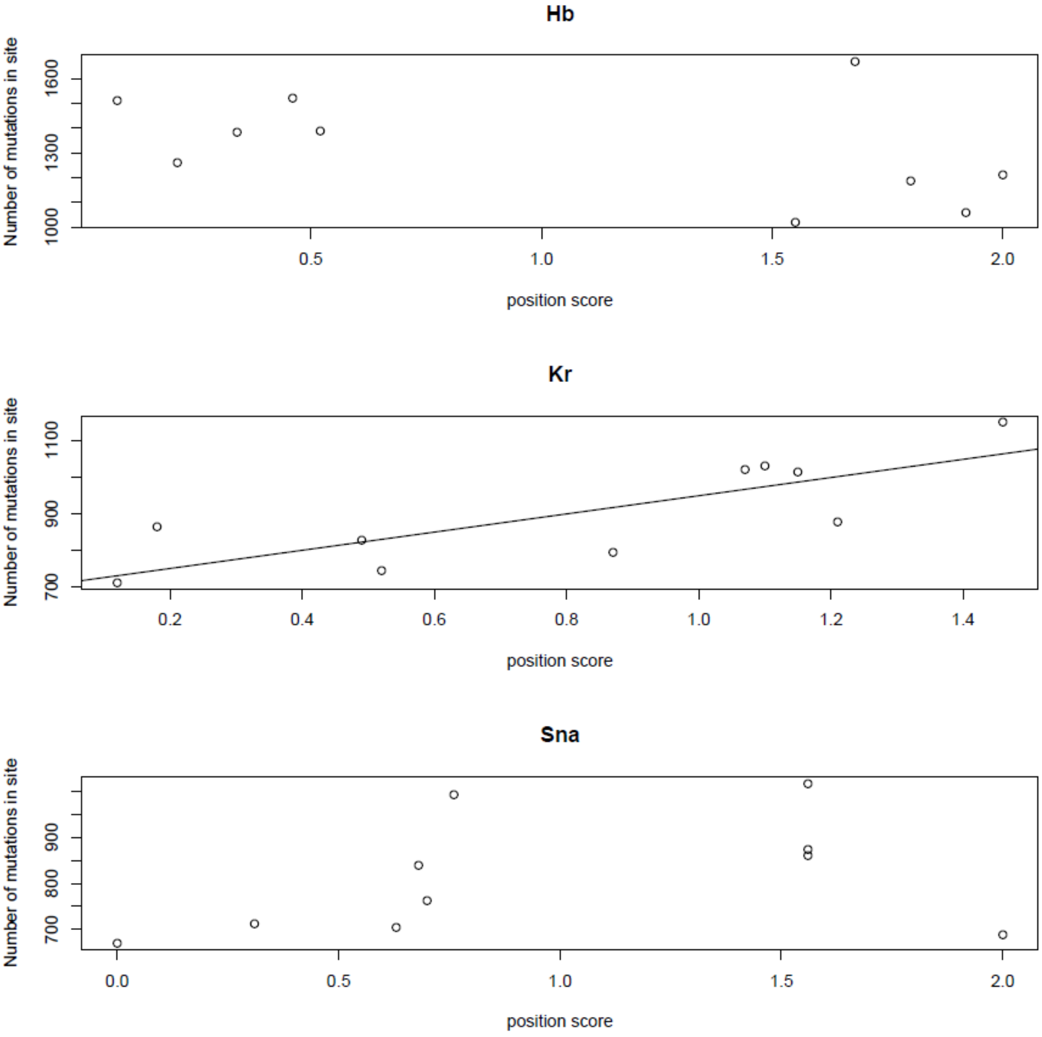


Figure 4.11: The correlation between information content (x-axis) and number of SNPs within position in binding sites. HB binding sites in Hb-bound areas (top), KR binding sites in Kr-bound areas (middle) and SNA binding sites in Sna-bound areas (bottom). (See statistical analysis in Appendix E Table E.9).

5. Discussion

5.1 Evaluating predictions about excess of SNPs and deletions in HB binding sites

The central hypothesis of this project is that there is/was co-evolution between the HB transcription factor and its target regulatory elements. This was based on preliminary evidence, outlined in the introduction. To reiterate, it was observed that two highly conserved HB binding sites in the s3+7e of eve are removed by two deletions, segregating at moderate frequency in *D. melanogaster* populations. Several more specific predictions follow from this general hypothesis. I wanted to know why this was happening; i) are the two deletions a local event, ii) is there a genome wide response – with a high fraction of HB binding sites in the genome being affected or iii) is there a chance that these deletions are simply an example of binding site turnover (LUDWIG *et al.* 1998; MOSES *et al.* 2006; LI *et al.* 2008; MACARTHUR *et al.* 2009). To evaluate these hypotheses, I used two genomic datasets. First, a polymorphism data from 162 fully sequenced *D. melanogaster* strains, DGRP (MASSOURAS *et al.* 2012) and ChIP-chip data for three zinc-finger TFs, surveyed between hours 2 and 3 of early *D. melanogaster* development (LI *et al.* 2008; MACARTHUR *et al.* 2009). I used bioinformatic and statistical methods to evaluate these specific predictions, both the entire set of Chipped regions as well as a focused set of developmental genes.

5.1.1 The number of, and the frequency of SNPs in three types of binding sites

We hypothesized that positive selection has or is favoring deletions of HB binding sites. The first prediction (1a) that there are more SNPs in HB binding sites than in binding sites for other factors. I chose two zinc finger TFs (Krüppel and Snail) as reference factors. The data did not give us any indications of such a signal. We performed the analysis on both TFBS sites predicted in chip-bound regions for the three separate factors (HB 1762, KR 3028 and SNA 595). The data do not indicate an excess of SNPs in HB sites compared to the two reference TFBS.

But the occurrence of a SNP in site is only a part the story. The frequency of mutations in populations can be a strong indicator of their fitness and molecular effects. Generally, common mutations are either neutral or even beneficial, and deleterious mutations tend to be rare (HARTL and CLARK 2007). Thus I also looked at the frequency of point mutations within HB, KR and SNA sites. Surprisingly, a comparison of the frequency of SNPs among positions within each TFBS, did not reveal a difference. This may be in part, because I only looked at frequency of all mutations, and did not take the impact of the SNP into question. This issue was addressed further via linear models (see below).

Thus the data did not suggest that SNPs are more commonly found in, or on average more frequent, in HB sites compared to the reference factors in this genome wide survey. But it is possible that a signal is found in more narrow set of candidate genes, for instance characterized developmental genes known to function around the same time in development as *eve* (Table 1.1). Thus we may be more likely to see a response as HB is known to affect the expression of some of those genes.

The data from the developmental genes did however echo the results from the genome wide survey. There was no evidence of more SNPs in HB sites overall or higher frequency of mutations in HB sites. There are several weaknesses to the approach as we implemented it. Perhaps the most crucial one is that we did not take evolutionary conservation or clustering of predicted TFBS into account in the study. Previous work has suggested that clustered TFBS are more likely to constitute functional enhancers (MURAKAMI *et al.* 2004), also in Drosophila species (BERMAN *et al.* 2004). One option to make this part of the study more focused would be to test for conservation. Also, we were using a dataset where each of these factors was bound genome wide. By doing so, we could possibly be drowning a localized signal, or the signal that is limited to a specific time in development or part of the embryo. Even though there was binding of the TF at that time-point in development the binding might be significant later in development or perhaps in adult life.

5.1.2 Do SNPs in binding sites affect function?

Transcription factor binding sites are generally 5-20 bp motifs, recognized by the TF factor. The central hypothesis can also be evaluated by looking for a relationship between SNPs and the information content in a binding site. Another of our predictions, is that HB sites may be more severely affected by SNPs with detrimental effects on the function of the binding site. I studied this by using PWMs to calculate the score for the binding site before and after it was affected by a mutation. The dataset was scanned for SNPs in binding sites and the impact of each exact substitution on PWM score was calculated. We predicted that the SNPs found in HB binding sites would have more negative effects on the binding site score than SNPs in TFBSs for the two reference factors. The results for this part of the study did not corroborate this prediction. What was noted was that a large fraction of SNPs in Krüppel binding sites did have more deleterious effects on the PWM score (~80%). In contrast, only ~55% of SNPs in HB sites decreased the PWM score this drastically. Oddly, more SNPs in HB binding sites did, increase the score of the site than for the reference factors. This result is somewhat puzzling. One possibility is that this just reflects noise in the data. Another is that the SNPs were only categorized as major vs. minor, not on whether they where ancestral or derived. Furthermore, those analyzes did not take frequency of the mutations into account (more on that later). Finally, the reason for this pattern might be in the PWM for the TFs used in the study. The prediction of TFBS depends on a good PWM for that site. Many studies have documented, and refined PWM for major TFS. A recent study examined the PWM of several TFs, and found that the HB PWM did give a fairly accurate prediction of HB affinity to the DNA (HE *et al.* 2011). The same study showed that the PWM for Knirps was not very reliable, and that further molecular and bioinformatic work greatly improved the PWM. It may be possible that the PWMs for the three factors used here were not all optimal, or at least equally good. He et al. (2011) also observed that using PWMs calculated from *D. mel* and *D. sim* did not matter, they showed the same result. This further supports the idea that transcription factors and binding sites are generally well conserved (HE *et al.* 2011).

5.1.3 Test of a relationship between information content and number of mutations at a specific position within a binding site

General evolutionary principles suggest that there should be a negative correlation between PWM information content per base and the number of segregating mutations in a population at each position within TFBS. It is somewhat puzzling that such a relationship was not seen for none of the factors (Figure 4.11). A couple of reasons could explain this lack of signal. One option is that the PWMs are not good enough. Also, it is possible that the data is insufficiently large to detect such pattern. Studies have confirmed that the HB PWM is accurate in predicting the effects of a mutation (HE *et al.* 2011), but similar studies are lacking for the other two factors. Furthermore, we had a limited number of SNA sites in the data. There were only about 600 bound SNA regions, compared to over 3000 for KR and 1700 for HB. The lack of relationship between SNP occurrence and information content at each position could possibly be explained by our model. However, such reasoning fails to account for the lack of correlation for the SNA sites and the opposite results that were seen for KR binding sites. Therefore, such reasoning from negative evidence is at best corroborative but at worst a fluke.

The analyses presented above only looked at one or two features at a time. I also did more systematic analyses, using a linear model in R with the frequency of major and minor alleles as a response variable. Thus I could ask which factors had significant impact on the frequency of the SNPs. By using a regression framework, one can simultaneously evaluate the impact of several factors, and using stepwise addition of terms, to test the influence impact of each factor. This was summarized above (Tables 4.6 and 4.7), it is clear that the factors that matter for the frequency of SNPs are the ChIP-chip area, the change in score, position of the SNP within binding site and the sequence of the binding site. To conclude from this model we have further indications that there seem to be some constraints on which positions tolerate more mutations which is consistent with some positions in TFBS being more important than others.

5.1.4 Deletions of/within HB binding sites

In the original outline of the project, we were most interested in studying the distribution of indels in chipped regions and TFBS. However, at the time only SNP predictions where available for the 162 DGRP lines. Predicting indels from scratch was beyond the scope of this study. However, through a series of fortunate events we got in contact with Thomas Zichner, a Ph.D. Student with Jan Korbel in EMBL Heidelberg. He has just recently published a paper documenting insertions and deletions, segregating in the DGRP lines (ZICHNER *et al.* 2012). He volunteered early access to his supplemental table and ran some bioinformatic analyses for us, on the HB, KR and SNA chipped regions and TFBS predictions generated in this project. Through this collaboration with Zichner we were able to scan our dataset for deletions affecting chipped regions, and binding sites. These results corroborate our central hypotheses. We did see that there are significantly more deletions in HB binding sites than for the reference factors. The data for insertions points in the same direction, though it must be stressed that a statistical support is lacking for a difference among TFBS classes. We also scanned the whole ChIP-chip areas to see if there was a difference between areas depending on which of the three factors (HB, KR and Sna) was bound. The difference between the areas also showed the same pattern even though it is not statistically significant. Note however, that a weaker pattern was found in chipped regions for KR and SNA, than in HB regions, was expected from our hypothesis; if the main force affecting the frequency of HB deletions was co-evolution of HB concentration and HB binding sites.

An interesting and general result of this work is that numerous deletions in the genome affect regions bound by TF and quite a large fraction of them (3-7%) takes out TFBS. The DGRP data opens up many possibilities in studying the distribution and impact deletions in the genome, on TFBS as was done here or other biological features. We can expect to see many investigations trying to understand why so many deletions are tolerated by the species.

The DGRP data along with Zichners method of scanning for deletions allows for studies of localized and specific deletion events. We wanted to further examine the *eve* area for deletions. A scan of the *eve* region 18379 bp did find the two deletions knocking out HB sites in the *eve* s3+7e described previously. No deletions hitting other HB binding sites were observed, suggesting very circumscribed effects. The only other indels affecting predicted sites, are two small insertions in two predicted Snail binding sites. We do not anticipate those to be of major relevance. An interesting follow up would be to run all the known developmental genes, active at this time-point in development and see if there are deletions of binding sites in their regulatory regions. Not only could we possibly find more HB binding site deletion, events but also add more pieces to what we know of gene regulatory networks. It is possible that we might see deletions of binding sites for known and characterized transcription factors, which would indicate that the HB deletions in *eve* s3+7e is simply allowed flux of binding sites and not a specific response to changes in expression or concentration of affected/affecting factors or genes.

5.2 Change in HB activity

We hypothesized that the *hb* protein or the gene might have been changed in some way, thus triggering a co-evolutionary response. We postulated that the protein structure or concentration had changed, or that the spatial expression of the gene had been altered. Among the features that make *Drosophila melanogaster* such a good model organism is the fact that it belongs to a rich phylogeny of close relative, and many of its closest relatives have had their genomes sequenced. This allows comparative genomic studies, alignments of species genomes and delineation of new changes on specific lineages. We made use of the available genomes for the close relatives of *D. melanogaster*. We aligned the protein sequence of 12 species, *D. melanogaster* and its 11 closest relatives with sequenced genomes. We had predicted the DNA binding motifs in Hunchback, Krüppel and Snail using HMM and wondered if there would be changes in the amino acid sequence affecting the function of the protein, the most serious ones would be seen in the DNA binding motif. Some changes are seen in the alignment, but not between the 4 closest relatives of *D. melanogaster*. The predicted DNA binding motifs are very well conserved between the species.

Other parts of the protein are however, different between D. *melanogaster* and its closest relatives, but as the function of the *hunchback* protein has not been studied in detail, it is very hard to predict their impact. Amino acids changes from a charged to a neutral, or bulky to a small, or switching of charges could alter the protein function quite profoundly. Such predictions, and experimental dissections of their effect, would be interesting but only, if sufficient data suggest that co-evolution did indeed occur. One could for instance replace the *D.melanogaster hb* with the *D. simulans hb* gene, and study the potential effects on development and biological function. This piece of data is far from sufficient to disprove the notion that the protein function or stability has changed.

If the HB concentration has changed, then the cause could possibly be found in the promoters for *hb*. One possibility in changed expression of HB could change in the Bicoid gradient. That is something that could be further pursued. Alignment of both the zygotic and maternal promoters did not reveal any clear functional candidate changes to explain changes in transcription of *hb*. However, when we were processing the 21 developmental genes we came across an interesting observation on *hb* genomic region. There are numerous HB binding sites in the regulatory region for *hunchback*. Self-regulation or auto-regulation, is a common theme in regulatory networks (DAVIDSON 2001). Most curiously, many of those predicted HB binding sites had SNPs in them. An interesting question is how many of those SNPs have detrimental effects on a binding site.

Testing for change in HB concentration in *D. melanogaster*, is somewhat complicated by the spatio-temporal dynamics of its expression and the fact that HB is transcriped both by the mother and the embryo. We would need information about the mRNA and protein concentration at specific time points, at specific locations in the embryo in a large number of individuals and the concentration is a localized signal. However, it is possible to perform a qPCR to measure the amount of mRNA and try to get some information on the difference in concentration both between *D. melanogaster* individuals and between *Drosophila* species.

Other known factors that could affect the function of HB, would be Pumilio and characterized miRNAs which impact translation of *HB* mRNA. The data did not give us any indication that there was something going on in HB translation, in the *D. melanogaster* lineage in comparison with the closest relatives.

5.3 Tests for selection in HB and LD of *eve* and Hb

5.3.1 Positive selection of Hb

Among the first ideas we proposed was that a gene in the proximity to *hb* had experienced (or was under) positive selection and that *hb* had been or was hitchhiking along. Thus, changes in HB function, admittedly undescribed would have been a side effect of some other selection effect. Strong positive selection can be detected if unusually long, undisrupted haplotypes are found in a genomic region, as time to fixation is so short that recombination is unable to break it up (SABETI *et al.* 2002; SABETI *et al.* 2007). We both sequenced areas surrounding *hb* and looked at the DGRP data for corresponding area. The data did not reveal long extended haplotypes in the HB region, nor was π unusually low in the proximity to *hb*. Thus there are no indications of a recent positive selection in either the *hb* gene itself or on a gene nearby. Positive selection is always interesting to study. Researchers have been trying to identify positive selection for quite some time and have developed programs that can run such tests on a small set of genomic regions but these programs were not used here.

5.3.2 LD between *eve* and Hb

Among our least likely hypothesis was that if there was co-evolution between *eve* and HB one could detect LD between the two genes. We tested this but we did not see a distinct signal linking *eve* and HB. What we concluded was that if the function of HB had changed then it was most likely evolutionarily older event.

5.4 Other possible scenarios

The data does not provide a clear confirmation of a genome wide selection against HB binding sites. What can explain the occurrence of the two deletions removing HB sites from *eve*? One possibility is simply, that those changes are there by chance. Maybe polymorphism in HB is acceptable. According to Fowlkes et al. (2008) differences in HB concentration did affect the expression and location of stripe 3 but not stripe 7 (based on imaging data to study patterns of expression in TFs) (FOWLKES *et al.* 2008). The result from the imaging data indicates that concentration of HB matters for stripe 3 to be expressed correctly. If we had seen one deletion of one binding site, that explanation might have been very plausible. However, we have observed two unrelated events, both taking out TFBS for HB. Another possibility is that they reflect a localized event. In other words we might be observing co-evolution within an enhancer, or with a *cis*-regulatory regions or the whole locus.

To elaborate other transcription factors that are known players in regulating *eve* stripe 3+7 that might be of relevance are DSTAT, Zelda and Knirps. The two former are activators and Knirps is a repressor. It would therefore be interesting to document any changes in the TFBS of those factors. More detailed investigation of *eve* s3+7e, like Ludwig et al. (2011) carried out for *eve* s2e where he used transgenes to study the enhancer, would be an interesting avenue of research (LUDWIG *et al.* 2011). Knowing more about the enhancer might provide us with answers to our questions. One might postulate that the effects of variations and the deletions in the s3+7 enhancer, might be favored in a response to a loss of activation capacity because of other TF, either abundance of TFs acting directly on the s3+7e or other functionally related sequences. Such spillover effects from other *cis*-elements were found for the s2e in detailed transgene experiments, whereby the deletion of s2e minimal enhancer, still retained a faint stripe 2 expression (LUDWIG *et al.* 2005). Furthermore, several shadow enhancers have been described in *Drosophila*. Those are elements, localized elsewhere in the *cis*-environment, which provide support for localized *cis*-regulatory modules (HONG *et al.* 2008; FRANKEL *et al.* 2010). Thus co-evolution within the entire regulatory region of a gene is quite possible, not only co-evolution within an enhancer (LUDWIG *et al.* 1998).

It is also possible that the enhancer is responding to changes in other regulatory mechanisms, like translation, localization or protein stability. We find this rather unlikely, except possibly for translational regulatory agents – which may have temporal and spatial capacity that corresponds to the function of the s3+7e. Recall, we found no other large deletions of HB sites in other *eve* enhancers. Other interesting screens could be for example only genes that HB is known to directly regulate. It would also be intriguing to scan the whole genome for deletions and then study which of the deletions do carry HB binding sites and cross reference those with regions that HB is known to regulate.

In my opinion the three most plausible explanations are i) co-evolution within the eve gene, ii) co-evolution between HB concentration and a subset of HB targeted *cis*-elements, or iii) the evolutionarily allowed flux of binding sites within species on a genome wide scale. On this last possibility, one could ask are SNPs in binding sites functionally or more severe as deletions of entire TFBS? One could also ask, how much variation can a binding site tolerate and what kind of mutations are acceptable? This varies between transcription factors and we need to characterize as many as we can to have an easier time predicting regulatory output. However, if that were done we could better predict what the effects of mutations will have. This is important work in the future especially for human transcription factors.

With the increasing technologies especially in genomic studies, more frequently large groups with the resources to do genome wide broad studies publish data that others can access. Their work paves the way for smaller groups to investigate more specific questions such as the one we were curious about. For the present study, we combined two large datasets and did a genome wide scan when trying to cast a light on a peculiar observation, namely the two HB deletions. Even though we did not see the pattern we expected we did see indications that deletions hit HB binding sites more commonly, then the TFBSs for Krüppel or Snail.

5.5 General results and future work

Even though we were unable to firmly confirm our predictions, the study does yield some general conclusions. There are plenty of SNPs in predicted binding sites segregating in wild populations of flies. This is even found for TFBS where the information content is very high. Why that is so is yet to be answered. Also, some of the numerous indels in the genome, affect chipped regions and TFBS, both in areas known to be functional as well as others not characterized as functional, at least not yet. For studies of *Drosophila* GRN (JAEGER and REINITZ 2006), like the interactions of the gap TFs, the fact that many enhancers have segregating variation must bring about a rethink. Jaeger and Reinitz (2006) showed that quantity of a TF plays a major role in the gap gene network and ultimately in deciding cell fates. In creating quantitative models of development the topology of the network is pivotal where changes in input quantity affect output.

The data we used for the study is an almost endless resource for questions like ours. Even though we were unable to get a distinct answer to our questions we did see other potential problems that the data could provide insight into. We only screened three transcription factors and their possible binding sites in a limited set of regions. It would be very interesting to both screen the whole genome as well as a more focused set. Future work could involve expanding the study and add more reference factors. Possible candidates are for example Huckebein and Schnurri that are also TFs with a C2H2 zinc finger motif. They might be a better reference than KR and SNA that possibly function very differently from HB, though the opposite could also be true, that they in fact have a similar function.

Bibliography

ARNOSTI, D. N., 2003 Analysis and function of transcriptional regulatory elements: insights from Drosophila. Annu Rev Entomol **48:** 579-602.

ASHBURNER, M., 2007 Drosophila Genomes by the Baker's Dozen. Preface. Genetics **177:** 1263-1268.

BAKKER, E. G., C. TOOMAJIAN, M. KREITMAN and J. BERGELSON, 2006 A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell **18:** 1803-1818.

BAMSHAD, M., and S. P. WOODING, 2003 Signatures of natural selection in the human genome. Nat Rev Genet **4:** 99-111.

BERMAN, B. P., Y. NIBU, B. D. PFEIFFER, P. TOMANCAK, S. E. CELNIKER *et al.*, 2002 Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A **99:** 757-762.

BERMAN, B. P., B. D. PFEIFFER, T. R. LAVERTY, S. L. SALZBERG, G. M. RUBIN *et al.*, 2004 Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol **5:** R61.

BIELER, J., C. POZZORINI and F. NAEF, 2011 Whole-embryo modeling of early segmentation in Drosophila identifies robust and fragile expression domains. Biophys J **101:** 287-296.

BIOPAUKER, J. D., 2008 Genkaskade kontrolliert die Musterbildung beim Drosophila-Embryo, pp.

BIRNEY, E., J. A. STAMATOYANNOPOULOS, A. DUTTA, R. GUIGO, T. R. GINGERAS *et al.*, 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447:** 799-816.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BRADLEY, R. K., X. Y. LI, C. TRAPNELL, S. DAVIDSON, L. PACHTER *et al.*, 2010 Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol **8:** e1000343.

CARROLL, S. B., J. K. GRENIER and S. D. WEATHERBEE, 2001 *From DNA to diversity : molecular genetics and the evolution of animal design*. Blackwell Science, Malden, Mass.

CELNIKER, S. E., L. A. DILLON, M. B. GERSTEIN, K. C. GUNSALUS, S. HENIKOFF *et al.*, 2009 Unlocking the secrets of the genome. Nature **459:** 927-930.

CHEUNG, V. G., and R. S. SPIELMAN, 2009 Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet **10:** 595-604.

CLARK, A. G., M. B. EISEN, D. R. SMITH, C. M. BERGMAN, B. OLIVER *et al.*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

CLYDE, D. E., M. S. CORADO, X. WU, A. PARE, D. PAPATSENKO *et al.*, 2003 A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature **426:** 849-853.

DAVIDSON, E. H., 2001 *Genomic regulatory systems : development and evolution*. Academic Press, San Diego.

DOMAZET-LOSO, T., and D. TAUTZ, 2010 A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature **468:** 815-818.

EWING, B., and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research **8:** 186-194.

FARNHAM, P. J., 2009 Insights from genomic profiling of transcription factors. Nat Rev Genet **10:** 605-616.

FINN, R. D., J. MISTRY, J. TATE, P. COGGILL, A. HEGER *et al.*, 2010 The Pfam protein families database. Nucleic Acids Res **38:** D211-222.

FOWLKES, C. C., K. B. ECKENRODE, M. D. BRAGDON, M. MEYER, Z. WUNDERLICH *et al.*, 2011 A conserved developmental patterning network produces quantitatively different output in multiple species of Drosophila. PLoS Genet **7:** e1002346.

FOWLKES, C. C., C. L. HENDRIKS, S. V. KERANEN, G. H. WEBER, O. RUBEL *et al.*, 2008 A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell **133:** 364-374.

FRANKEL, N., G. K. DAVIS, D. VARGAS, S. WANG, F. PAYRE *et al.*, 2010 Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature **466:** 490-493.

GIBSON, G., and S. V. MUSE, 2009 *A primer of genome science*. Sinauer Associates, Sunderland, Mass.

GILBERT, D. G., 2005 DroSpeGe, a public database of Drosophila species genomes, pp.

GILBERT, S. F., 2006 *Developmental biology*. Sinauer Associates, Inc. Publishers, Sunderland, Mass.

GOERING, L. M., P. K. HUNT, C. HEIGHINGTON, C. BUSICK, P. S. PENNINGS *et al.*, 2009 Association of orthodenticle with natural variation for early embryonic patterning in Drosophila melanogaster. J Exp Zool B Mol Dev Evol **312:** 841-854.

GORDON, D., 2003 Viewing and editing assembled sequences using Consed. Curr Protoc Bioinformatics **Chapter 11:** Unit11 12.

HARTL, D. L., and A. G. CLARK, 2007 *Principles of population genetics*. Sinauer Associates, Sunderland, Mass.

HE, B. Z., A. K. HOLLOWAY, S. J. MAERKL and M. KREITMAN, 2011 Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet **7:** e1002053.

HONG, J. W., D. A. HENDRIX and M. S. LEVINE, 2008 Shadow enhancers as a source of evolutionary novelty. Science **321:** 1314.

HOUCHMANDZADEH, B., E. WIESCHAUS and S. LEIBLER, 2002 Establishment of developmental precision and proportions in the early Drosophila embryo. Nature **415:** 798-802.

HULSKAMP, M., W. LUKOWITZ, A. BEERMANN, G. GLASER and D. TAUTZ, 1994 Differential regulation of target genes by different alleles of the segmentation gene hunchback in Drosophila. Genetics **138:** 125-134.

JAEGER, J., M. BLAGOV, D. KOSMAN, K. N. KOZLOV, MANU *et al.*, 2004 Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. Genetics **167:** 1721-1737.

JAEGER, J., and J. REINITZ, 2006 On the dynamic nature of positional information. Bioessays **28:** 1102-1111.

KALINKA, A. T., K. M. VARGA, D. T. GERRARD, S. PREIBISCH, D. L. CORCORAN *et al.*, 2010 Gene expression divergence recapitulates the developmental hourglass model. Nature **468:** 811-814.

KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. PRINGLE *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996-1006.

KIM, J., X. HE and S. SINHA, 2009 Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet **5:** e1000330.

KIMURA, M., 1979 The neutral theory of molecular evolution. Sci Am **241:** 98-100, 102, 108 passim.

KOSMAN, D., REINITZ,J., SHARP,D.H., 1999 Automated assay of gene expression at cellular resolution, pp. 6 - 17 in *Proceedings of the 1998 Pacific Symposium on Biocomputing*, edited by R. D. ALTMAN, K.; HUNTER, L.; AND KLEIN, T.

LANDRY, C. R., P. J. WITTKOPP, C. H. TAUBES, J. M. RANZ, A. G. CLARK *et al.*, 2005 Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics **171:** 1813-1822.

LARKIN, M. A., G. BLACKSHIELDS, N. P. BROWN, R. CHENNA, P. A. MCGETTIGAN *et al.*, 2007 Clustal W and Clustal X version 2.0. Bioinformatics **23:** 2947-2948.

LATCHMAN, D. S., 2010 *Gene control*. Garland Science, New York.

LI, X. Y., S. MACARTHUR, R. BOURGON, D. NIX, D. A. POLLARD *et al.*, 2008 Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol **6:** e27.

LUDWIG, M. Z., 2002 Functional evolution of noncoding DNA. Curr Opin Genet Dev **12:** 634-639.

LUDWIG, M. Z., MANU, R. KITTLER, K. P. WHITE and M. KREITMAN, 2011 Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet **7:** e1002364.

LUDWIG, M. Z., A. PALSSON, E. ALEKSEEVA, C. M. BERGMAN, J. NATHAN *et al.*, 2005 Functional evolution of a cis-regulatory module. PLoS Biol **3:** e93.

LUDWIG, M. Z., N. H. PATEL and M. KREITMAN, 1998 Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development **125:** 949-958.

LUKOWITZ, W., C. SCHRODER, G. GLASER, M. HULSKAMP and D. TAUTZ, 1994 Regulatory and coding regions of the segmentation gene hunchback are functionally conserved between Drosophila virilis and Drosophila melanogaster. Mech Dev **45:** 105-115.

MACARTHUR, S., X. Y. LI, J. LI, J. B. BROWN, H. C. CHU *et al.*, 2009 Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol **10:** R80.

MANU, S. SURKOVA, A. V. SPIROV, V. V. GURSKY, H. JANSSENS *et al.*, 2009 Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors. PLoS Comput Biol **5:** e1000303.

MASSOURAS, A., S. M. WASZAK, M. ALBARCA-AGUILERA, K. HENS, W. HOLCOMBE *et al.*, 2012 Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster. PLoS Genet **8:** e1003055.

MOSES, A. M., D. A. POLLARD, D. A. NIX, V. N. IYER, X. Y. LI *et al.*, 2006 Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol **2:** e130.

MURAKAMI, K., T. KOJIMA and Y. SAKAKI, 2004 Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. BMC Genomics **5:** 16.

MURATA, Y., and R. P. WHARTON, 1995 Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in Drosophila embryos. Cell **80:** 747-756.

NICHOLAS, K. B., N. H. B. JR. and D. D. W. II., 1997 GeneDoc: Analysis and Visualization of Genetic Variation, pp.

PAPATSENKO, D., and M. S. LEVINE, 2008 Dual regulation by the Hunchback gradient in the Drosophila embryo. Proc Natl Acad Sci U S A **105:** 2901-2906.

PATEL, N. H., and S. LALL, 2002 Precision patterning. Nature **415:** 748-749.

RAFF, R. A., 1996 *The shape of life : genes, development, and the evolution of animal form*. University of Chicago Press, Chicago.

RANEY, B. J., M. S. CLINE, K. R. ROSENBLOOM, T. R. DRESZER, K. LEARNED *et al.*, 2011 ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res **39:** D871-875.

ROCKMAN, M. V., and G. A. WRAY, 2002 Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol **19:** 1991-2004.

ROY, S., J. ERNST, P. V. KHARCHENKO, P. KHERADPOUR, N. NEGRE *et al.*, 2010 Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science **330:** 1787-1797.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386.

RUSSO, C. A., N. TAKEZAKI and M. NEI, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol **12:** 391-404.

RUVINSKY, I., and G. RUVKUN, 2003 Functional tests of enhancer conservation between distantly related species. Development **130:** 5133-5142.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-837.

SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449:** 913-918.

SAWYER, S. A., J. PARSCH, Z. ZHANG and D. L. HARTL, 2007 Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. Proc Natl Acad Sci U S A **104:** 6504-6510.

SELLA, G., D. A. PETROV, M. PRZEWORSKI and P. ANDOLFATTO, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet **5:** e1000495.

SMALL, S., A. BLAIR and M. LEVINE, 1996 Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. Dev Biol **175:** 314-324.

SOSINSKY, A., B. HONIG, R. S. MANN and A. CALIFANO, 2007 Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci U S A **104:** 6305-6310.

STANOJEVIC, D., T. HOEY and M. LEVINE, 1989 Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in Drosophila. Nature **341:** 331-335.

STERN, D. L., and V. ORGOGOZO, 2008 The loci of evolution: how predictable is genetic evolution? Evolution **62:** 2155-2177.

STORMO, G. D., 2000 DNA binding sites: representation and discovery. Bioinformatics **16:** 16-23.

STRUFFI, P., M. CORADO, L. KAPLAN, D. YU, C. RUSHLOW *et al.*, 2011 Combinatorial activation and concentration-dependent repression of the Drosophila even skipped stripe 3+7 enhancer. Development **138:** 4291-4299.

TIROSH, I., S. REIKHAV, A. A. LEVY and N. BARKAI, 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324:** 659-662.

TREIER, M., C. PFEIFLE and D. TAUTZ, 1989 Comparison of the gap segmentation gene hunchback between Drosophila melanogaster and Drosophila virilis reveals novel modes of evolutionary change. EMBO J **8:** 1517-1525.

VISEL, A., M. J. BLOW, Z. R. LI, T. ZHANG, J. A. AKIYAMA *et al.*, 2009 ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature **457:** 854-858.

WATSON, J. D., 2008 *Molecular biology of the gene*. Pearson/Benjamin Cummings; Cold Spring Harbor Laboratory Press, San Francisco Cold Spring Harbor, N.Y.

WILCZYNSKI, B., and E. E. FURLONG, 2010 Challenges for modeling global gene regulatory networks during development: insights from Drosophila. Dev Biol **340:** 161-169.

WITTKOPP, P. J., 2010 Variable transcription factor binding: a mechanism of evolutionary change. PLoS Biol **8:** e1000342.

WITTKOPP, P. J., B. K. HAERUM and A. G. CLARK, 2004 Evolutionary changes in cis and trans gene regulation. Nature **430:** 85-88.

WOLFF, C., R. SOMMER, R. SCHRODER, G. GLASER and D. TAUTZ, 1995 Conserved and divergent expression aspects of the Drosophila segmentation gene hunchback in the short germ band embryo of the flour beetle Tribolium. Development **121:** 4227-4236.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol **20:** 1377-1419.

WUNDERLICH, Z., and A. H. DEPACE, 2011 Modeling transcriptional networks in Drosophila development at multiple scales. Curr Opin Genet Dev **21:** 711-718.

ZICHNER, T., D. A. GARFIELD, T. RAUSCH, A. M. STUTZ, E. CANNAVO *et al.*, 2012 Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. Genome Res.

Appendix A

Receipts and Protocols

DNA Isolation

A GenElute Mammalian Genomic DNA Miniprep Kit from Sigma-Aldrich was used for the isolation. Whole flies were in freezer in ethanol. The ethanol was washed off with water and then an individual fly placed in a ependorf test tube.

1. A single fly is minced with a rod as much as possible.
2. 180 µl of Lysis solution followed by 20 µl of Proteinase K.
3. Sample then kept at room temperature and vortexed occationally for about 2 hours.
4. 200 µl of Lysis solution is then added and the mix is vortexed thoroughly.
5. Then incubation for 10 mins in 55°C waterbath.
6. Column preparation with column preparation solution and centrifuged at 12.000 x g for 1 min and liquid discarded.
7. 200 µl of ethanol (95-100%) to lysate. Vortex 5-10 sec.
8. The entire content is moved to a binding column and centrifuged for 1 min at 6500 x g. The fluid is discarded and binding column put in a new tube.
9. 500 µl of wash solution concentrate with ethanol added to the binding column then centrifuged at max speed 12.000 for 3 mins.
10. 100 µl of elution solution is pipette directly into the center of the binding column. Then centrifuged for 1 min at 6500 x g and incubated for 5 mins at room temperature.

PCR reaction

Table A.1: PCR recipe.

	Taq	**Teq**
	Amount (µL)	**Amount (µL)**
DNA	3	3
Buffer	2	2
dNTP	0,1	2
Primer F	0,2	0,4
Primer R	0,2	0,4
Polymerase	0,1	0,4
H_2O	12,5	11,8
Total	**20**	**20**

I used either Taq polymerase and the matching ThermoPol buffer from BioLabs or Teq polymerase from Prokarya and Reaction buffer provided. The units of Teq are lower than Taq so therefore there is a different receipt for that.

The same PCR program was used for every PCR run.

Table A.2: PCR program.

Step	Temperature (°C)	Time (mins)
1	94	3:00
2	94	1:00
3	52	0:30
4	72	1:00
	Steps 2-4 repeated 35 times	
5	72	7:00
6	4	Forever

Exo Sap

Table A.3: Exo sap cleaning step.

	Amount (µL)
H_2O	3,7
Phosphate buffer	1
Antartic phosphatase	0,2
Exonuclease 1	0,1
Total	**5**

Both the Exo1 and Ant. Phospatase are from BioLabs.

Sequencing reaction

Table A.4: Receipt for the sequencing reaction.

	Amount (µL)
dH_2O	5,25
Big dye buffer	2,75
BigDye	0,5
Primer (1pm/ µ)	1,5
Total	**10**

The BigDye that is used is 3.1 and comes from Applied Biosystem.

Ethanol precipitation

Table A.5: Solution 1 for Ethanol precipitation made for each tray of sequencing.

Solution 1 for 2 full trays	**Amount**
dH_2O	9 mL
Natriumacetat 3M (NaOAc)	1 mL
Glycogen	50 µL

Ethanol precipitation protocol:

1. 45 µl of solution 1 in each well.
2. 125 µl of 95% -20°C ethanol, mix carefully.
3. Centrifuge for 30 mins at 4000 RPM
4. Discard immediately.
5. Place tray upside down on 3 kim wipes and centrifuge on 300 RPM for 2 mins.
6. 250 µl 70% ethanol in each well, centrifuge for 5 mins on 4000 RPM.
7. Place tray upside down on 3 kim wipes and centrifuge on 300 RPM for 5 mins.
8. Place the trays in a dark place for drying for 15-20 mins.

The Glycogen used is from Fermentas.

HiDi dissolving

10 µL of HiDi are put in each well. The tray is then sealed extremely well and shaken for one minute. Then the tray is ready for the sequencer. The HiDi is from Applied Biosystems.

Appendix B

List of all primers

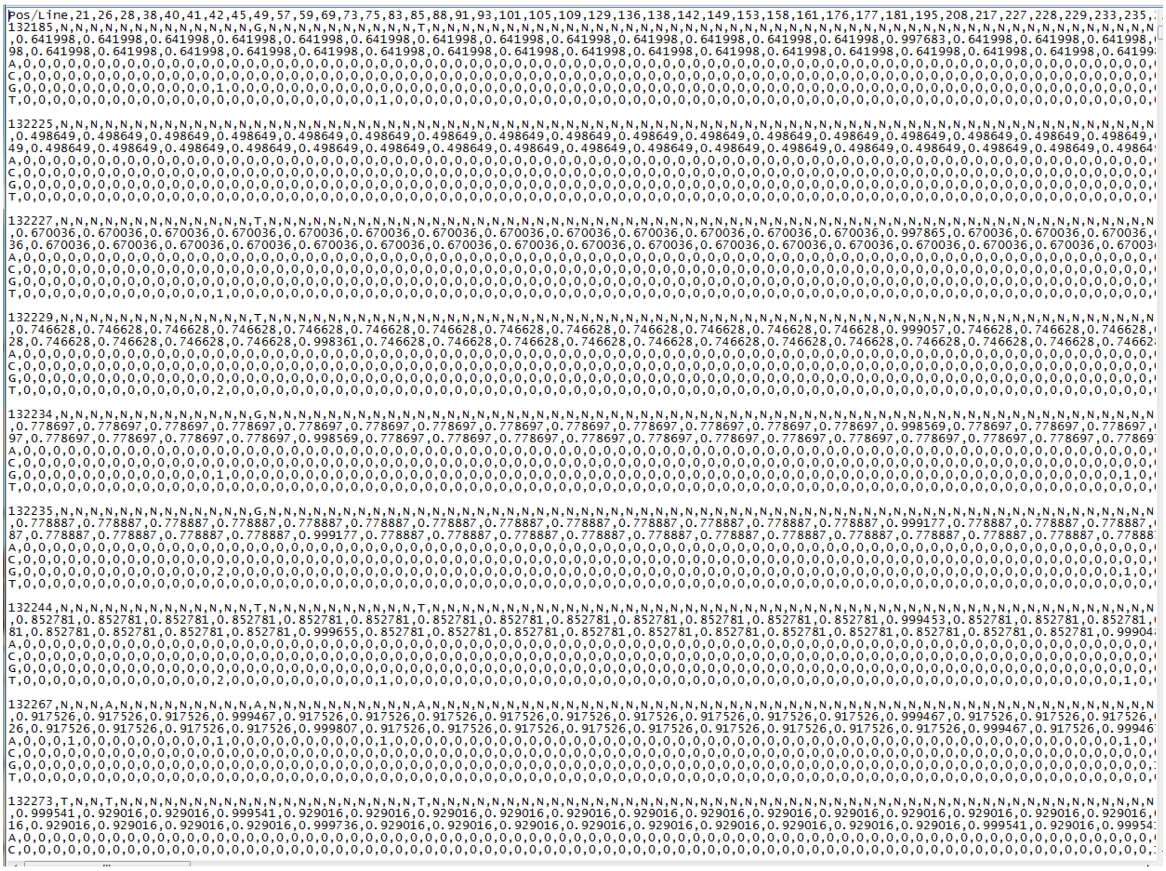
Table B.1: List of all primers.

Primer name	Sequence
CE1014a_f	ACGCCGTAAAATTGGCTATG
CE1014a_r	AGAAGATCGCGGCTTGTAAA
CE1014b_f	TTACAAGCCGCGATCTTCTT
CE1014b_r	ATTTCGCTGGCGACTTAGTG
Hb1_F	TCGAACTGGCACTGGTATTG
Hb1_R	CAATCTTCTGCCTCCTCTGG
Hb25k_F	GCAAACTGACCAAACGAGTC
Hb25k_R	AAACAAAATGCGTCCATCGT
mRpL19_F	GACTGGACATGCTGGATCG
mRpL19_R	AGCTGCGGTAGTCCTTCATC
Hb50k_F	TGGGTGGCAATAAAAATGTC
Hb50k_R	AGCCGAAATTAAAGCTCACG
hb56k_F	TGTTATCGCTGCACGTATCG
hb56k_R	TTAAGTACATGCGTTCACGTTT
Dhod_f	TGCAGCACTTGCTTCAAATC
Dhod_r	AGAGCTCCTGGAACAGGTCA
p_f	CAGCGCCTAACAATTTCCTC
p_r	GCCAGCATGTCCGATATTTT
CG8032_f	GTGTAGCAATGGGCGACAAG
CG8032_r	GGCAAATAGAGGGAACAGGA
CG9773_2_f	AGCAGACGTCGCAGGTATTT
CG9773_2_r	CGGATGAGTTCTCGATTTCC
cg9773_f	ATGGACATCGAAGGCATAGC
cg9773_r	CGGAAACAGACCAAGTCACA
Hb35_f	TGCGCTTTTCTCTGGATTCT
Hb35_r	TGGCACATTTAACACCTCCA
Hb30_f	TTGAGATCGCTGGCAATATG
Hb30_r1	CGAGGAGTGGGAGAAGTACG
Hb30_r2	CATTTAGGGACTTCGGCAAA
Hb15_f	CCCTGGCATTCTAGGCATAA
Hb15_r	ATGGTGGCCATTTTTAGCAG
Hb10_f	ATGCTGGGGTTCTGTTGAAG
Hb10_r	TCGAATCGAACGAAATCAAA
Hb2_f	GGGCTTGTGACCATACTTGC
Hb2_r	GAGCACGATCAGATGTCGAA
Hb3_f	AGATTGCCGCCATAGAAGTG
Hb3_r	GACCAACTACGAGCAGCACA
Hb4_f	GCAGGCTGTTTTGATCGTTT

Hb4_r	AGCCACCCCTGACGTATTTT
Hb5_f	TTTTCCGCTTGTTTTTCATTTT
Hb5_r	TCTGCCCATCTAATCCCTTG
Hb6_f	GGATGATCCGGGAGCTTAG
Hb6_r	AAAATGCAGCAACTGCACAA
Hb7_f	TGCACCACACAAAATGAAGC
Hb7_r	CTGCGTTTTCGAATTTTTCC
Hb8_f	GTGCCGAACTATTTGCCATT
Hb8_r	ATCGAGTGCTTCTTTTTCTGTC
Hb9_f	CCGCGGCACGGTTACCGTTAGAC
Hb9_r	TGTTGTGAGGAGCAGTGAGG
Hb11_f	TACGCAGTACGCAGGATCAC
Hb11_r	GCTTGGCTGCACATCTTACA
Hb12_f	GAAAATCGCGAGAAACTTCG
Hb12_r	AACGAGGCTCAAAGGACAAA
Cyp313b1_f	TATCGATGCCGATCTGAGTG
Cyp313b1_r	CCATCTTCCGACAGCTTCTC
Cyp313b1_f2	GTTCCTTGTGCAGCTTCTCC
Cyp313b1_r2	TGCCTTTAGCTTGACCGAAG
bel1_f	AGACCCACAGGATTGTCTGC
bel1_r	GCCGATTCTCAACCAGATGT
bel1_f2	ATGTGGCCGAGAACATAAGC
bel1_r2	TATCAATGGCCGGGATCTAA
CG8036_f	GGACAACCTTTGCGTGATCT
CG8036_r	GGTCGAGGTTCTTCAGCTTG
CG8036_f2	ATGTGGGCAAGAACTTCGAC
CG8036_r2	TCTCCGTCTAGGGCAACAAC
CG8043_f	ACGCCCATTTTCTCAACAAG
CG8043_r	GTTAGCTCCTGGCCCACATA
CG8043_f2	TCTTGTACCGCACCAACAAC
CG8043_r2	GTTACGTCCTGGCTGGAACC

Appendix C

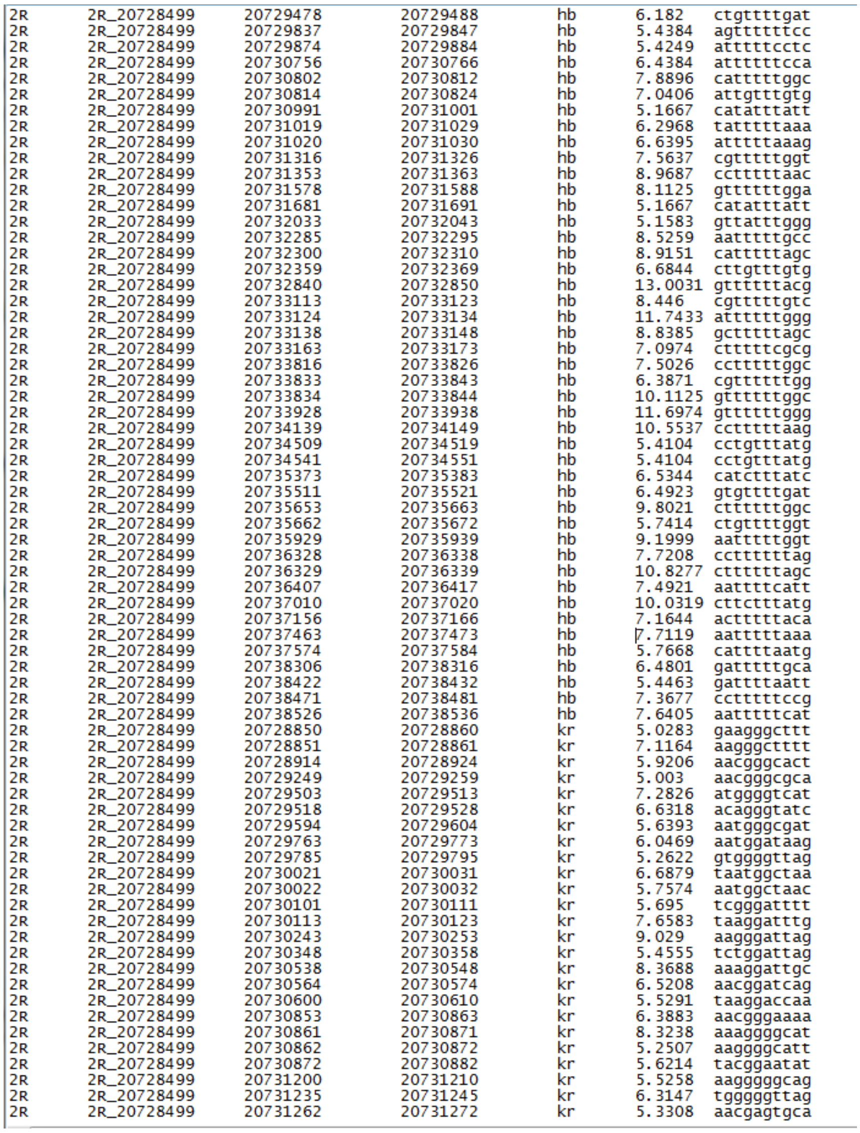
Snapshot of all input and output files for Python



```
Pos/Line,21,26,28,38,40,41,42,45,49,57,59,69,73,75,83,85,88,91,93,101,105,109,129,136,138,142,149,153,158,161,176,177,181,195,208,217,227,228,229,233,235,
132185,N,N,N,N,N,N,N,N,N,N,N,N,N,G,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.997683,0.641998,0.641998,0.641998,
98,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.641998,0.64199
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132225,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,
49,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.498649,0.49864
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132227,N,N,N,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.997865,0.670036,0.670036,0.670036,
36,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.670036,0.67003
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132229,N,N,N,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.999057,0.746628,0.746628,0.746628,
28,0.746628,0.746628,0.746628,0.746628,0.998361,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.746628,0.74662
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132234,N,N,N,N,N,N,N,N,N,N,N,N,N,G,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.998569,0.778697,0.778697,0.778697,
97,0.778697,0.778697,0.778697,0.778697,0.998569,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.778697,0.77869
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132235,N,N,N,N,N,N,N,N,N,N,N,N,N,G,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.999177,0.778887,0.778887,0.778887,
87,0.778887,0.778887,0.778887,0.778887,0.999177,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.778887,0.77888
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132244,N,N,N,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.999453,0.852781,0.852781,0.852781,
81,0.852781,0.852781,0.852781,0.852781,0.999655,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.852781,0.99904
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,

132267,N,N,N,A,N,N,N,N,N,N,N,N,N,A,N,N,N,N,N,N,N,N,N,N,A,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.917526,0.917526,0.917526,0.999467,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.999467,0.917526,0.917526,0.917526,
26,0.917526,0.917526,0.917526,0.917526,0.999807,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.917526,0.999467,0.917526,0.99946
A,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
G,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
T,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

132273,T,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,T,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N
,0.999541,0.929016,0.929016,0.999541,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,
16,0.929016,0.929016,0.929016,0.929016,0.999736,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.929016,0.999541,0.929016,0.99954
A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
C,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

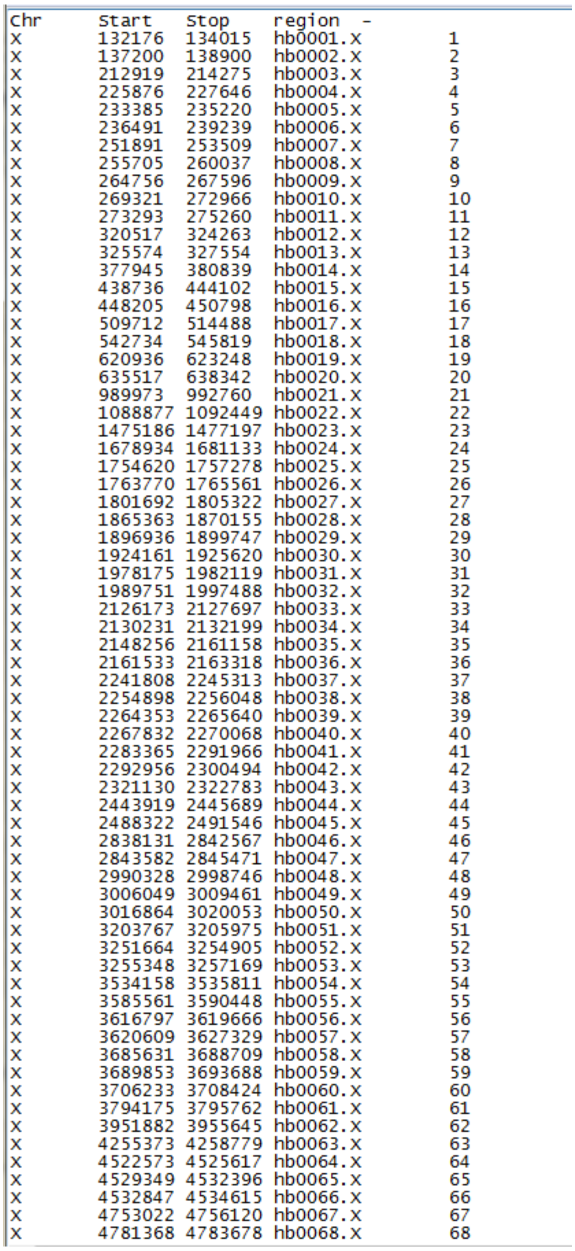
Figure C.1: Example of a SNP file published by DGRP. The first line lists all the inbred strains. Each 7 lines give information on the location of the SNP on a certain chromosome (this case X). Additional information is the coverage, the frequency (not seen in this figure) and from what strains the readings are coming from.

]



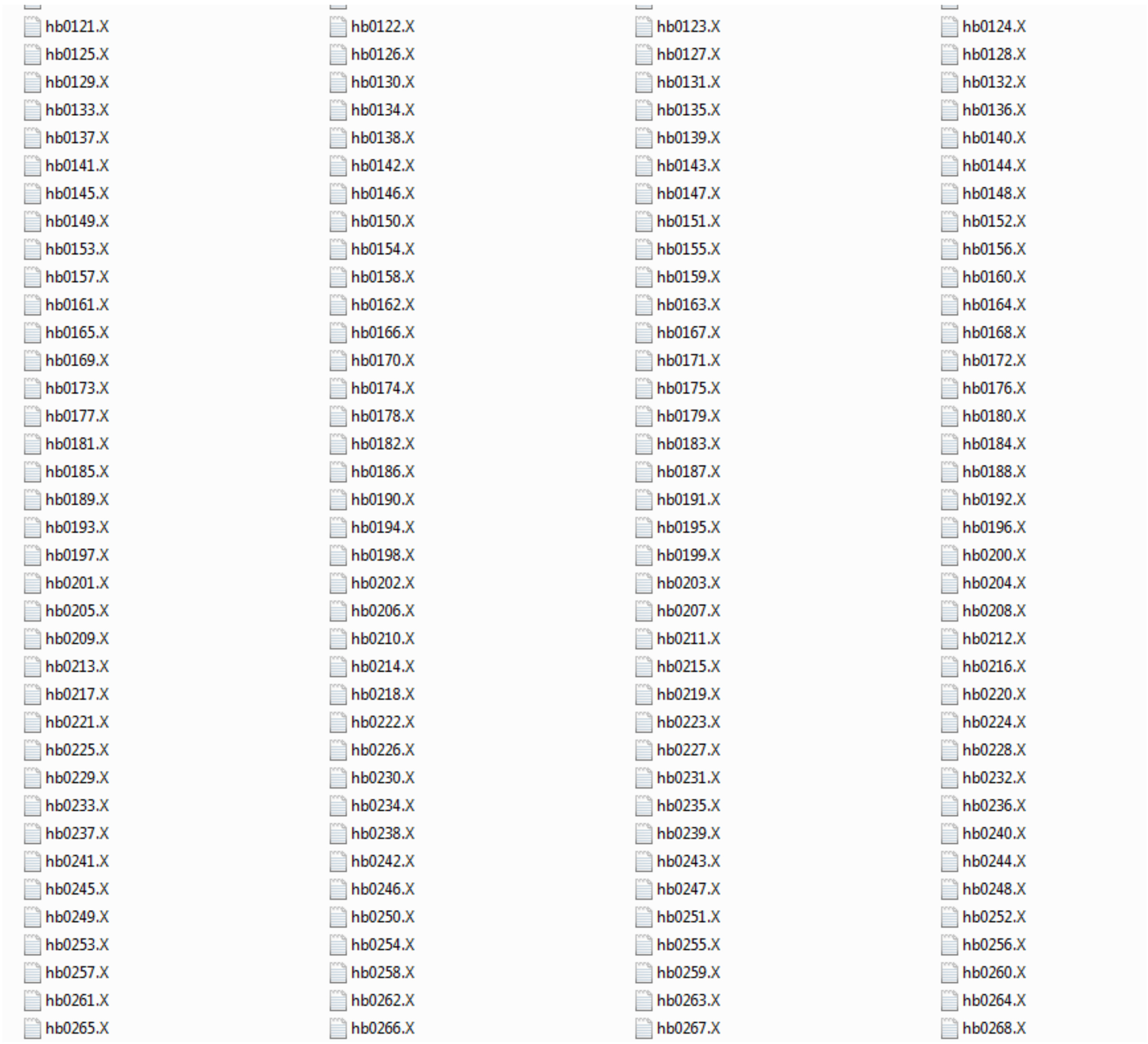
2R	2R_20728499	20729478	20729488	hb	6.182	ctgttttgat
2R	2R_20728499	20729837	20729847	hb	5.4384	agtttttcc
2R	2R_20728499	20729874	20729884	hb	5.4249	atttttcctc
2R	2R_20728499	20730756	20730766	hb	6.4384	attttttcca
2R	2R_20728499	20730802	20730812	hb	7.8896	catttttggc
2R	2R_20728499	20730814	20730824	hb	7.0406	attgtttgtg
2R	2R_20728499	20730991	20731001	hb	5.1667	catatttatt
2R	2R_20728499	20731019	20731029	hb	6.2968	tatttttaaa
2R	2R_20728499	20731020	20731030	hb	6.6395	atttttaaag
2R	2R_20728499	20731316	20731326	hb	7.5637	cgtttttggt
2R	2R_20728499	20731353	20731363	hb	8.9687	cctttttaac
2R	2R_20728499	20731578	20731588	hb	8.1125	gtttttttgga
2R	2R_20728499	20731681	20731691	hb	5.1667	catatttatt
2R	2R_20728499	20732033	20732043	hb	5.1583	gttatttggg
2R	2R_20728499	20732285	20732295	hb	8.5259	aatttttgcc
2R	2R_20728499	20732300	20732310	hb	8.9151	catttttagc
2R	2R_20728499	20732359	20732369	hb	6.6844	cttgtttgtg
2R	2R_20728499	20732840	20732850	hb	13.0031	gtttttttacg
2R	2R_20728499	20733113	20733123	hb	8.446	cgtttttgtc
2R	2R_20728499	20733124	20733134	hb	11.7433	attttttggg
2R	2R_20728499	20733138	20733148	hb	8.8385	gctttttagc
2R	2R_20728499	20733163	20733173	hb	7.0974	ctttttcgcg
2R	2R_20728499	20733816	20733826	hb	7.5026	ccttttttggc
2R	2R_20728499	20733833	20733843	hb	6.3871	cgttttttgg
2R	2R_20728499	20733834	20733844	hb	10.1125	gtttttttggc
2R	2R_20728499	20733928	20733938	hb	11.6974	gtttttttggg
2R	2R_20728499	20734139	20734149	hb	10.5537	cctttttaag
2R	2R_20728499	20734509	20734519	hb	5.4104	cctgtttatg
2R	2R_20728499	20734541	20734551	hb	5.4104	cctgtttatg
2R	2R_20728499	20735373	20735383	hb	6.5344	catctttatc
2R	2R_20728499	20735511	20735521	hb	6.4923	gtgttttgat
2R	2R_20728499	20735653	20735663	hb	9.8021	ctttttttggc
2R	2R_20728499	20735662	20735672	hb	5.7414	ctgttttggt
2R	2R_20728499	20735929	20735939	hb	9.1999	aatttttggt
2R	2R_20728499	20736328	20736338	hb	7.7208	cctttttttag
2R	2R_20728499	20736329	20736339	hb	10.8277	cttttttagc
2R	2R_20728499	20736407	20736417	hb	7.4921	aattttcatt
2R	2R_20728499	20737010	20737020	hb	10.0319	cttctttatg
2R	2R_20728499	20737156	20737166	hb	7.1644	actttttaca
2R	2R_20728499	20737463	20737473	hb	7.7119	aatttttaaa
2R	2R_20728499	20737574	20737584	hb	5.7668	cattttaatg
2R	2R_20728499	20738306	20738316	hb	6.4801	gatttttgca
2R	2R_20728499	20738422	20738432	hb	5.4463	gattttaatt
2R	2R_20728499	20738471	20738481	hb	7.3677	cctttttccg
2R	2R_20728499	20738526	20738536	hb	7.6405	aatttttcat
2R	2R_20728499	20728850	20728860	kr	5.0283	gaagggcttt
2R	2R_20728499	20728851	20728861	kr	7.1164	aagggctttt
2R	2R_20728499	20728914	20728924	kr	5.9206	aacgggcact
2R	2R_20728499	20729249	20729259	kr	5.003	aacgggcgca
2R	2R_20728499	20729503	20729513	kr	7.2826	atggggtcat
2R	2R_20728499	20729518	20729528	kr	6.6318	acagggtatc
2R	2R_20728499	20729594	20729604	kr	5.6393	aatgggcgat
2R	2R_20728499	20729763	20729773	kr	6.0469	aatggataag
2R	2R_20728499	20729785	20729795	kr	5.2622	gtggggttag
2R	2R_20728499	20730021	20730031	kr	6.6879	taatggctaa
2R	2R_20728499	20730022	20730032	kr	5.7574	aatggctaac
2R	2R_20728499	20730101	20730111	kr	5.695	tcgggatttt
2R	2R_20728499	20730113	20730123	kr	7.6583	taaggatttg
2R	2R_20728499	20730243	20730253	kr	9.029	aagggattag
2R	2R_20728499	20730348	20730358	kr	5.4555	tctggattag
2R	2R_20728499	20730538	20730548	kr	8.3688	aaaggattgc
2R	2R_20728499	20730564	20730574	kr	6.5208	aacggatcag
2R	2R_20728499	20730600	20730610	kr	5.5291	taaggaccaa
2R	2R_20728499	20730853	20730863	kr	6.3883	aacgggaaaa
2R	2R_20728499	20730861	20730871	kr	8.3238	aaaggggcat
2R	2R_20728499	20730862	20730872	kr	5.2507	aaggggcatt
2R	2R_20728499	20730872	20730882	kr	5.6214	tacggaatat
2R	2R_20728499	20731200	20731210	kr	5.5258	aaggggggcag
2R	2R_20728499	20731235	20731245	kr	6.3147	tgggggttag
2R	2R_20728499	20731262	20731272	kr	5.3308	aacgagtgca

Figure C.2: List of all possible binding sites for each ChIP-chip area (this case chromosome 2R of Hunchback bound regions). The first column gives the chromosome, the second is the start of the area and chromosome. Third and fourth column give the start and stop of a ten bp binding site. Fifth column is what transcription factor that binding site belongs to. Sixth column is the score calculated from PWM and last column gives the sequence for the binding site.



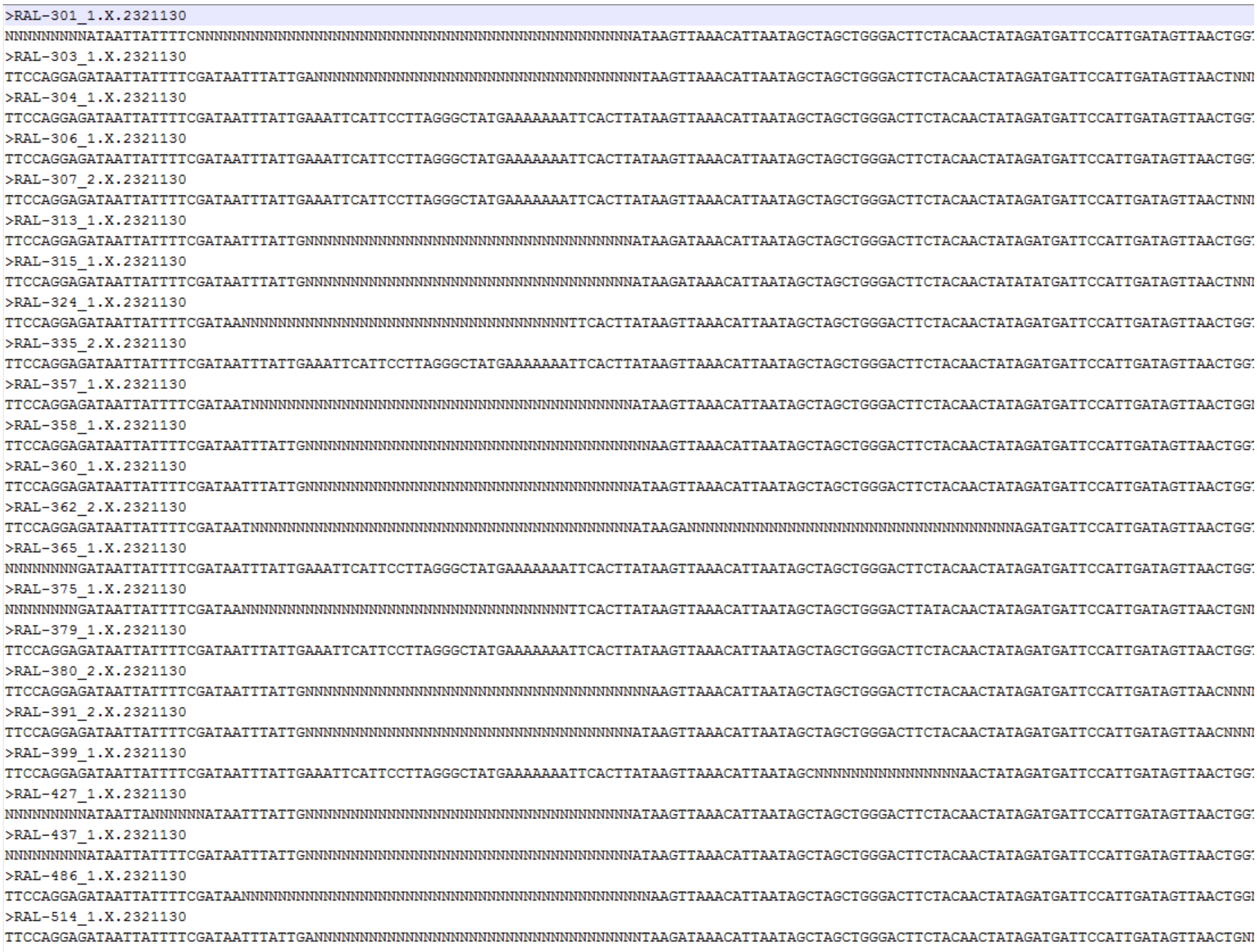
Chr	Start	Stop	region -	
X	132176	134015	hb0001.x	1
X	137200	138900	hb0002.x	2
X	212919	214275	hb0003.x	3
X	225876	227646	hb0004.x	4
X	233385	235220	hb0005.x	5
X	236491	239239	hb0006.x	6
X	251891	253509	hb0007.x	7
X	255705	260037	hb0008.x	8
X	264756	267596	hb0009.x	9
X	269321	272966	hb0010.x	10
X	273293	275260	hb0011.x	11
X	320517	324263	hb0012.x	12
X	325574	327554	hb0013.x	13
X	377945	380839	hb0014.x	14
X	438736	444102	hb0015.x	15
X	448205	450798	hb0016.x	16
X	509712	514488	hb0017.x	17
X	542734	545819	hb0018.x	18
X	620936	623248	hb0019.x	19
X	635517	638342	hb0020.x	20
X	989973	992760	hb0021.x	21
X	1088877	1092449	hb0022.x	22
X	1475186	1477197	hb0023.x	23
X	1678934	1681133	hb0024.x	24
X	1754620	1757278	hb0025.x	25
X	1763770	1765561	hb0026.x	26
X	1801692	1805322	hb0027.x	27
X	1865363	1870155	hb0028.x	28
X	1896936	1899747	hb0029.x	29
X	1924161	1925620	hb0030.x	30
X	1978175	1982119	hb0031.x	31
X	1989751	1997488	hb0032.x	32
X	2126173	2127697	hb0033.x	33
X	2130231	2132199	hb0034.x	34
X	2148256	2161158	hb0035.x	35
X	2161533	2163318	hb0036.x	36
X	2241808	2245313	hb0037.x	37
X	2254898	2256048	hb0038.x	38
X	2264353	2265640	hb0039.x	39
X	2267832	2270068	hb0040.x	40
X	2283365	2291966	hb0041.x	41
X	2292956	2300494	hb0042.x	42
X	2321130	2322783	hb0043.x	43
X	2443919	2445689	hb0044.x	44
X	2488322	2491546	hb0045.x	45
X	2838131	2842567	hb0046.x	46
X	2843582	2845471	hb0047.x	47
X	2990328	2998746	hb0048.x	48
X	3006049	3009461	hb0049.x	49
X	3016864	3020053	hb0050.x	50
X	3203767	3205975	hb0051.x	51
X	3251664	3254905	hb0052.x	52
X	3255348	3257169	hb0053.x	53
X	3534158	3535811	hb0054.x	54
X	3585561	3590448	hb0055.x	55
X	3616797	3619666	hb0056.x	56
X	3620609	3627329	hb0057.x	57
X	3685631	3688709	hb0058.x	58
X	3689853	3693688	hb0059.x	59
X	3706233	3708424	hb0060.x	60
X	3794175	3795762	hb0061.x	61
X	3951882	3955645	hb0062.x	62
X	4255373	4258779	hb0063.x	63
X	4522573	4525617	hb0064.x	64
X	4529349	4532396	hb0065.x	65
X	4532847	4534615	hb0066.x	66
X	4753022	4756120	hb0067.x	67
X	4781368	4783678	hb0068.x	68

Figure C.3: List of all areas bound by each transcription factor (this case Hb). First column gives the chromosome, second and third give the start and stop of the area. Fourth column is what is used for each alignment file of each area (total of 1762 for Hb). Fifth column is simply the count of each area.



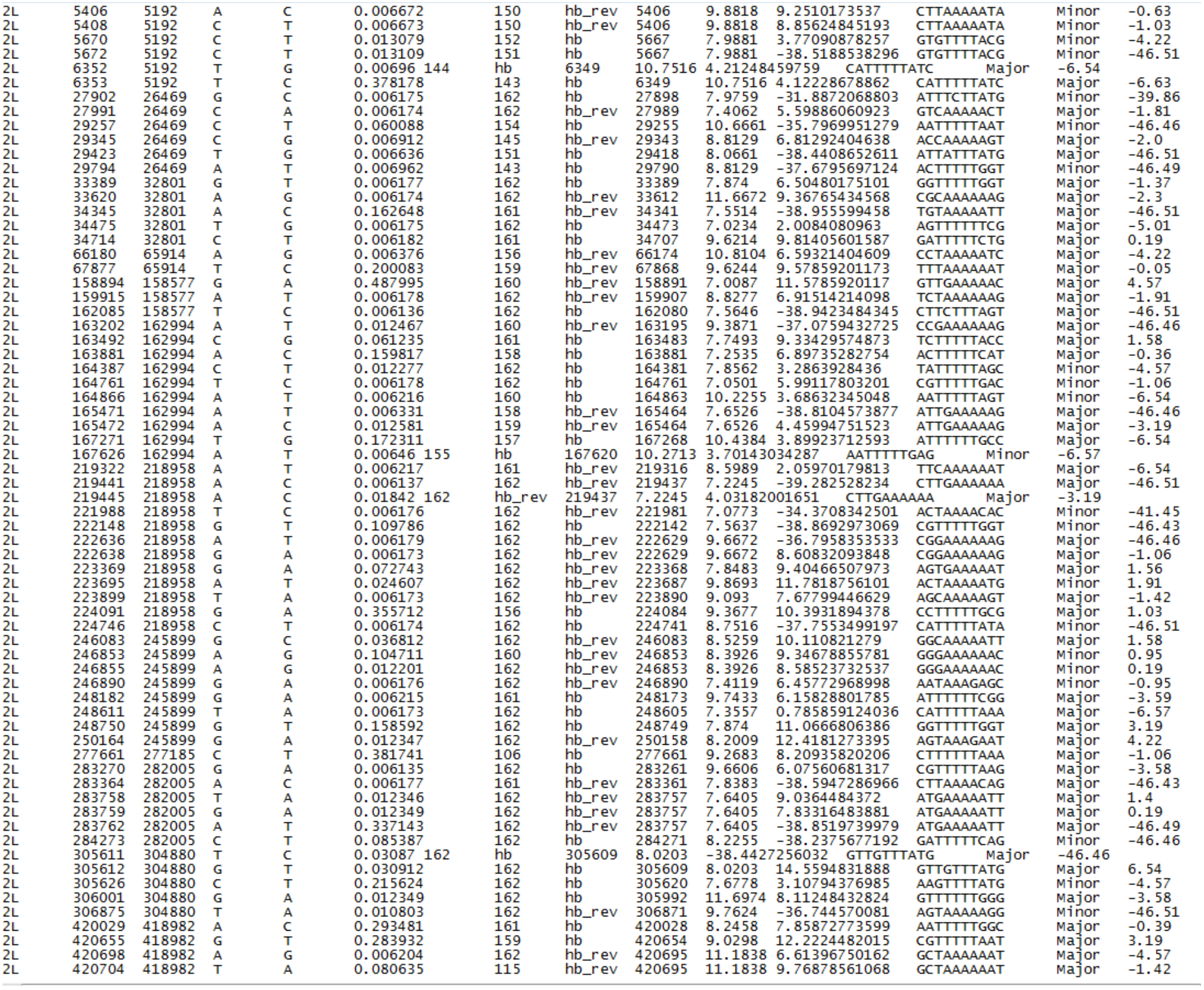
```
hb0121.X    hb0122.X    hb0123.X    hb0124.X
hb0125.X    hb0126.X    hb0127.X    hb0128.X
hb0129.X    hb0130.X    hb0131.X    hb0132.X
hb0133.X    hb0134.X    hb0135.X    hb0136.X
hb0137.X    hb0138.X    hb0139.X    hb0140.X
hb0141.X    hb0142.X    hb0143.X    hb0144.X
hb0145.X    hb0146.X    hb0147.X    hb0148.X
hb0149.X    hb0150.X    hb0151.X    hb0152.X
hb0153.X    hb0154.X    hb0155.X    hb0156.X
hb0157.X    hb0158.X    hb0159.X    hb0160.X
hb0161.X    hb0162.X    hb0163.X    hb0164.X
hb0165.X    hb0166.X    hb0167.X    hb0168.X
hb0169.X    hb0170.X    hb0171.X    hb0172.X
hb0173.X    hb0174.X    hb0175.X    hb0176.X
hb0177.X    hb0178.X    hb0179.X    hb0180.X
hb0181.X    hb0182.X    hb0183.X    hb0184.X
hb0185.X    hb0186.X    hb0187.X    hb0188.X
hb0189.X    hb0190.X    hb0191.X    hb0192.X
hb0193.X    hb0194.X    hb0195.X    hb0196.X
hb0197.X    hb0198.X    hb0199.X    hb0200.X
hb0201.X    hb0202.X    hb0203.X    hb0204.X
hb0205.X    hb0206.X    hb0207.X    hb0208.X
hb0209.X    hb0210.X    hb0211.X    hb0212.X
hb0213.X    hb0214.X    hb0215.X    hb0216.X
hb0217.X    hb0218.X    hb0219.X    hb0220.X
hb0221.X    hb0222.X    hb0223.X    hb0224.X
hb0225.X    hb0226.X    hb0227.X    hb0228.X
hb0229.X    hb0230.X    hb0231.X    hb0232.X
hb0233.X    hb0234.X    hb0235.X    hb0236.X
hb0237.X    hb0238.X    hb0239.X    hb0240.X
hb0241.X    hb0242.X    hb0243.X    hb0244.X
hb0245.X    hb0246.X    hb0247.X    hb0248.X
hb0249.X    hb0250.X    hb0251.X    hb0252.X
hb0253.X    hb0254.X    hb0255.X    hb0256.X
hb0257.X    hb0258.X    hb0259.X    hb0260.X
hb0261.X    hb0262.X    hb0263.X    hb0264.X
hb0265.X    hb0266.X    hb0267.X    hb0268.X
```

Figure C.4: An example of the output from the above input file. Each file contains the alignment of all 162 lines for a particular Hb-bound area. 1762 total for HB, 595 for Snail and 3028 for Krüppel. These files are used as input files for extended N and when they have gone through that an N is added at the end of the name.



```
>RAL-301_1.X.2321130
NNNNNNNNNATAATTATTTTCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-303_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTNN
>RAL-304_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-306_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-307_2.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTNN
>RAL-313_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGATAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-315_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGATAAACATTAATAGCTAGCTGGGACTTCTACAACTATATATGATTCCATTGATAGTTAACTNN
>RAL-324_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-335_2.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-357_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-358_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-360_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-362_2.X.2321130
TTCCAGGAGATAATTATTTTCGATAATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATGATTCCATTGATAGTTAACTGG
>RAL-365_1.X.2321130
NNNNNNNNGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-375_1.X.2321130
NNNNNNNNGATAATTATTTTCGATAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTATACAACTATAGATGATTCCATTGATAGTTAACTGN
>RAL-379_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-380_2.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACNNN
>RAL-391_2.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACNNN
>RAL-399_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGAAATTCATTCCTTAGGGCTATGAAAAAAATTCACTTATAAGTTAAACATTAATAGCNNNNNNNNNNNNNNNNAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-427_1.X.2321130
NNNNNNNNNATAATTANNNNNNATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-437_1.X.2321130
NNNNNNNNNATAATTATTTTCGATAATTTATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-486_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAGTTAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGG
>RAL-514_1.X.2321130
TTCCAGGAGATAATTATTTTCGATAATTTATTGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAAGATAAACATTAATAGCTAGCTGGGACTTCTACAACTATAGATGATTCCATTGATAGTTAACTGN
```

Figure C.5: An example of an alignment file from one Hb-bound area on chromosome X (Hb0043.X).



2L	5406	5192	A	C	0.006672	150	hb_rev	5406	9.8818	9.2510173537	CTTAAAAATA	Minor	-0.63
2L	5408	5192	C	T	0.006673	150	hb_rev	5406	9.8818	8.85624845193	CTTAAAAATA	Minor	-1.03
2L	5670	5192	C	T	0.013079	152	hb	5667	7.9881	3.77090878257	GTGTTTTACG	Minor	-4.22
2L	5672	5192	C	T	0.013109	151	hb	5667	7.9881	-38.5188538296	GTGTTTTACG	Minor	-46.51
2L	6352	5192	T	G	0.00696	144	hb	6349	10.7516	4.21248459759	CATTTTTATC	Major	-6.54
2L	6353	5192	T	C	0.378178	143	hb	6349	10.7516	4.12228678862	CATTTTTATC	Major	-6.63
2L	27902	26469	G	C	0.006175	162	hb	27898	7.9759	-31.8872068803	ATTTCTTATG	Minor	-39.86
2L	27991	26469	C	A	0.006174	162	hb_rev	27989	7.4062	5.59886060923	GTCAAAAACT	Major	-1.81
2L	29257	26469	C	T	0.060088	154	hb	29255	10.6661	-35.7969951279	AATTTTTAAT	Minor	-46.46
2L	29345	26469	C	G	0.006912	145	hb_rev	29343	8.8129	6.81292404638	ACCAAAAAGT	Major	-2.0
2L	29423	26469	T	G	0.006636	151	hb	29418	8.0661	-38.4408652611	ATTATTTATG	Major	-46.51
2L	29794	26469	A	T	0.006962	143	hb	29790	8.8129	-37.6795697124	ACTTTTTGGT	Minor	-46.49
2L	33389	32801	G	T	0.006177	162	hb	33389	7.874	6.50480175101	GGTTTTTGGT	Major	-1.37
2L	33620	32801	A	G	0.006174	162	hb_rev	33612	11.6672	9.36765434568	CGCAAAAAAG	Major	-2.3
2L	34345	32801	A	C	0.162648	161	hb_rev	34341	7.5514	-38.955599458	TGTAAAAATT	Major	-46.51
2L	34475	32801	T	G	0.006175	162	hb	34473	7.0234	2.0084080963	AGTTTTTTCG	Major	-5.01
2L	34714	32801	C	T	0.006182	161	hb	34707	9.6214	9.81405601587	GATTTTTCTG	Major	0.19
2L	66180	65914	A	G	0.006376	156	hb_rev	66174	10.8104	6.59321404609	CCTAAAAATC	Major	-4.22
2L	67877	65914	T	C	0.200083	159	hb_rev	67868	9.6244	9.57859201173	TTTAAAAAAT	Major	-0.05
2L	158894	158577	G	A	0.487995	160	hb_rev	158891	7.0087	11.5785920117	GTTGAAAAAC	Major	4.57
2L	159915	158577	A	T	0.006178	162	hb_rev	159907	8.8277	6.91514214098	TCTAAAAAAG	Major	-1.91
2L	162085	158577	T	C	0.006136	162	hb	162080	7.5646	-38.9423484345	CTTCTTTAGT	Major	-46.51
2L	163202	162994	A	T	0.012467	160	hb_rev	163195	9.3871	-37.0759432725	CCGAAAAAAG	Major	-46.46
2L	163492	162994	C	G	0.061235	161	hb	163483	7.7493	9.33429574873	TCTTTTTACC	Major	1.58
2L	163881	162994	A	C	0.159817	158	hb	163881	7.2535	6.89735282754	ACTTTTTCAT	Major	-0.36
2L	164387	162994	C	T	0.012277	162	hb	164381	7.8562	3.2863928436	TATTTTTAGC	Minor	-4.57
2L	164761	162994	T	C	0.006178	162	hb	164761	7.0501	5.99117803201	CGTTTTTGAC	Minor	-1.06
2L	164866	162994	A	T	0.006216	160	hb	164863	10.2255	3.68632345048	AATTTTTAGT	Minor	-6.54
2L	165471	162994	A	T	0.006331	158	hb_rev	165464	7.6526	-38.8104573877	ATTGAAAAAG	Major	-46.46
2L	165472	162994	A	C	0.012581	159	hb_rev	165464	7.6526	4.45994751523	ATTGAAAAAG	Major	-3.19
2L	167271	162994	T	G	0.172311	157	hb	167268	10.4384	3.89923712593	ATTTTTTGCC	Major	-6.54
2L	167626	162994	A	T	0.00646	155	hb	167620	10.2713	3.70143034287	AATTTTTGAG	Minor	-6.57
2L	219322	218958	A	T	0.006217	161	hb_rev	219316	8.5989	2.05970179813	TTCAAAAAAT	Major	-6.54
2L	219441	218958	A	C	0.006137	162	hb_rev	219437	7.2245	-39.282528234	CTTGAAAAAA	Major	-46.51
2L	219445	218958	A	C	0.01842	162	hb_rev	219437	7.2245	4.03182001651	CTTGAAAAAA	Major	-3.19
2L	221988	218958	T	C	0.006176	162	hb_rev	221981	7.0773	-34.3708342501	ACTAAAACAC	Minor	-41.45
2L	222148	218958	G	T	0.109786	162	hb	222142	7.5637	-38.8692973069	CGTTTTTGGT	Minor	-46.43
2L	222636	218958	A	T	0.006179	162	hb_rev	222629	9.6672	-36.7958353533	CGGAAAAAAG	Major	-46.46
2L	222638	218958	G	A	0.006173	162	hb_rev	222629	9.6672	8.60832093848	CGGAAAAAAG	Major	-1.06
2L	223369	218958	G	A	0.072743	162	hb_rev	223368	7.8483	9.40466507973	AGTGAAAAAT	Major	1.56
2L	223695	218958	A	T	0.024607	162	hb_rev	223687	9.8693	11.7818756101	ACTAAAAATG	Minor	1.91
2L	223899	218958	T	A	0.006173	162	hb_rev	223890	9.093	7.67799446629	AGCAAAAAGT	Major	-1.42
2L	224091	218958	G	A	0.355712	156	hb	224084	9.3677	10.3931894378	CCTTTTTGCG	Major	1.03
2L	224746	218958	C	T	0.006174	162	hb	224741	8.7516	-37.7553499197	CATTTTTATA	Minor	-46.51
2L	246083	245899	G	C	0.036812	162	hb_rev	246083	8.5259	10.110821279	GGCAAAAATT	Major	1.58
2L	246853	245899	A	G	0.104711	160	hb_rev	246853	8.3926	9.34678855781	GGGAAAAAAC	Minor	0.95
2L	246855	245899	A	G	0.012201	162	hb_rev	246853	8.3926	8.58523732537	GGGAAAAAAC	Minor	0.19
2L	246890	245899	G	A	0.006176	162	hb_rev	246890	7.4119	6.45772968998	AATAAAGAGC	Minor	-0.95
2L	248182	245899	G	A	0.006215	161	hb	248173	9.7433	6.15828801785	ATTTTTTCGG	Major	-3.59
2L	248611	245899	T	A	0.006173	162	hb	248605	7.3557	0.785859124036	CATTTTTAAA	Major	-6.57
2L	248750	245899	G	T	0.158592	162	hb	248749	7.874	11.0666806386	GGTTTTTGGT	Major	3.19
2L	250164	245899	G	A	0.012347	162	hb_rev	250158	8.2009	12.4181273395	AGTAAAGAAT	Major	4.22
2L	277661	277185	C	T	0.381741	106	hb	277661	9.2683	8.20935820206	CTTTTTTAAA	Major	-1.06
2L	283270	282005	G	A	0.006135	162	hb	283261	9.6606	6.07560681317	CGTTTTTAAG	Major	-3.58
2L	283364	282005	A	C	0.006177	161	hb_rev	283361	7.8383	-38.5947286966	CTTAAAACAG	Major	-46.43
2L	283758	282005	T	A	0.012346	162	hb_rev	283757	7.6405	9.0364484372	ATGAAAAATT	Major	1.4
2L	283759	282005	G	A	0.012349	162	hb_rev	283757	7.6405	7.83316483881	ATGAAAAATT	Major	0.19
2L	283762	282005	A	T	0.337143	162	hb_rev	283757	7.6405	-38.8519739979	ATGAAAAATT	Major	-46.49
2L	284273	282005	C	T	0.085387	162	hb	284271	8.2255	-38.2375677192	GATTTTTCAG	Minor	-46.46
2L	305611	304880	T	C	0.03087	162	hb	305609	8.0203	-38.4427256032	GTTGTTTATG	Major	-46.46
2L	305612	304880	G	T	0.030912	162	hb	305609	8.0203	14.5594831888	GTTGTTTATG	Major	6.54
2L	305626	304880	C	T	0.215624	162	hb	305620	7.6778	3.10794376985	AAGTTTTATG	Minor	-4.57
2L	306001	304880	G	A	0.012349	162	hb	305992	11.6974	8.11248432824	GTTTTTTGGG	Major	-3.58
2L	306875	304880	T	A	0.010803	162	hb_rev	306871	9.7624	-36.744570081	AGTAAAAAGG	Minor	-46.51
2L	420029	418982	A	C	0.293481	161	hb	420028	8.2458	7.85872773599	AATTTTTGGC	Major	-0.39
2L	420655	418982	G	T	0.283932	159	hb	420654	9.0298	12.2224482015	CGTTTTTAAT	Major	3.19
2L	420698	418982	A	G	0.006204	162	hb_rev	420695	11.1838	6.61396750162	GCTAAAAAAT	Major	-4.57
2L	420704	418982	T	A	0.080635	115	hb_rev	420695	11.1838	9.76878561068	GCTAAAAAAT	Major	-1.42

Figure C.6: All SNPs found within binding sites for each transcription factor and the change in score (this case HB regions chromosome 2L). Second column gives the location of the SNP. Third is the start of the chip-chip area that was bound. Fourth and fifth show the major and minor allele and sixth column is the frequency of minor allele. Seventh column is coverage (how many individuals have reads at that location). Eigth column gives the name of the transcription factor of the SNP. Ninth column is the start of the binding site. Tenth is the original score. Eleventh cloumn is the sequence.Tvelfth is the allele the binding site is located on. Last column is the delta score of the binding site.

Appendix D

All Python algorithms

Algorithm D.1: Algorithm that extracts all areas from ChIP-chip data (Hb_areas_alignment).

```
################################
# Create new folders from a table folder and sequence folder
########################################
# - this code does not clean up. it appends to each file.
# - Before it is ran again old files have to be removed.
###############################
import os
fasta_dir =os.getcwd()+ "\\"+"170_fa"
fasta_list = []
#chr_dir = =os.getcwd()+ "\\"+"CHR_areas"
#########################
#-Read fasta to array
########################
raw_fasta_dir = os.listdir(fasta_dir)
for item in  raw_fasta_dir:
  out =item.split('.')
  if (len(out) >1):
    if (out[1] == "fa"):
      fasta_list.append(item)
#------------------------
#########################
#-Read Chr's to seperate arrays
########################
f = open('hX.chr', 'r')
chr_hX = f.readlines()
chr_hX = [item.rstrip() for item in chr_hX]
f.closed
f = open('h2L.chr', 'r')
chr_h2L  = f.readlines()
chr_h2L = [item.rstrip() for item in chr_h2L]
f.closed
f = open('h2R.chr', 'r')
chr_h2R  = f.readlines()
chr_h2R = [item.rstrip() for item in chr_h2R]
f.closed
f = open('h3L.chr', 'r')
chr_h3L  = f.readlines()
chr_h3L = [item.rstrip() for item in chr_h3L]
f.closed
f = open('h3R.chr', 'r')
chr_h3R  = f.readlines()
chr_h3R = [item.rstrip() for item in chr_h3R]
f.closed
#f = open('eve_area.chr', 'r')
#chr_eve_area  = f.readlines()
#chr_eve_area = [item.rstrip() for item in chr_eve_area]
#f.closed
#f = open('Hb_area.chr', 'r')
#chr_Hb_area  = f.readlines()
#chr_Hb_area = [item.rstrip() for item in chr_Hb_area]
#f.closed
ChrX  = chr_hX[1:]                                    #removes first line in tables
Chr2L = chr_h2L[1:]
Chr2R = chr_h2R[1:]
Chr3L = chr_h3L[1:]
Chr3R = chr_h3R[1:]
```

```
#chr_eve_area = chr_eve_area[1:]
#chr_Hb_area = chr_Hb_area[1:]
#####################################
#Creating new folders for each are with all individuals together in same file
#####################################
os.chdir(os.getcwd()+ "\\"+"Hb_areas_alignment_170")             #Destination file
for fileX in fasta_list:
  f = open(fasta_dir+"\\"+fileX, 'r')
  active_file = f.readlines()
  active_file = [item.rstrip() for item in active_file]
  active_file = active_file[1:]
  chr_total =""
  for item in active_file:
    chr_total = chr_total+item
  fname_front,trash=  fileX.split(".")
  SP_line,chrome = fname_front.split("_")
  for item in eval(chrome):                                          #what table is being used to extract areas
     chrs,start,stop,region,counter = item.split()           #what is in each line
     start = int(start)
     stop = int(stop)
     sequence = chr_total[start:stop]
     outfile = open(region, 'a')
     entry_to_write=">"+SP_line+"."+str(chrs)+"."+str(start)+"\n"                    #the file is made in fasta format. each
sequence with same name as the folder it was taken from.
     outfile.write(entry_to_write+sequence+"\n")             #how the new file is made
     outfile.close()
  f.closed
```

Algorithm D.2: Algorithm that finds all possible binding sites, scores them and writes them in a table (Binding_sites).

```
from math import log
#####################################
class Matrix: # Encapsulates one sequence matrix
####################################
  def __init__(self,line):
    self.name = line[0].lstrip(">")
    self.A = [float(item) for item in line[1].split()[1:] ]
    self.C = [float(item) for item in line[2].split()[1:] ]
    self.G = [float(item) for item in line[3].split()[1:] ]
    self.T = [float(item) for item in line[4].split()[1:] ]
    self.motif_length=len(self.A)
    if abs((self.A[0]+self.C[0]+self.G[0]+self.T[0])-100) <10:
      #we have percentages,convert
      # and change 0 to 0.0000000001 so that we do not get crass for log2(0)
      self.A = [(item/100) for item in self.A]
      self.C = [(item/100) for item in self.C]
      self.G = [(item/100) for item in self.G]
      self.T = [(item/100) for item in self.T]
    # and change 0 to 0.0000000001 so that we do not get crass for log2(0)
    self.A = [max((item),0.00000000000001) for item in self.A]
    self.C = [max((item),0.00000000000001) for item in self.C]
    self.G = [max((item),0.00000000000001) for item in self.G]
    self.T = [max((item),0.00000000000001) for item in self.T]
    #lets find first perfectly conserved base.
    self.A_score = [log(item/0.25,2) for item in self.A]
    self.C_score = [log(item/0.25,2) for item in self.C]
    self.G_score = [log(item/0.25,2) for item in self.G]
    self.T_score = [log(item/0.25,2) for item in self.T]
    self.col_score = []
    self.tot_score =0
    counter = 0
    while counter <len(self.A_score):
      #print [self.A_score[counter]*self.A[counter],self.C_score[counter]*self.C[counter],self.G_score[counter]*self.G[counter],
self.T_score[counter]*self.T[counter]]
      #print self.C[counter]
      #print self.C_score[counter]
      self.col_score.append(self.A_score[counter]*self.A[counter])
      self.col_score[counter]+= self.C_score[counter]*self.C[counter]
      self.col_score[counter]+= self.G_score[counter]*self.G[counter]
      self.col_score[counter]+= self.T_score[counter]*self.T[counter]
      self.tot_score = self.tot_score+self.col_score[counter]
    counter+=1
#def Report(self):
  #print self.name
  #print "model length = " + str(len(self.A)) + " Cols"
  #print "total information content = " +str(self.tot_score) + " bits"
  #print [str(a) for a in self.A_score]
  #print [str(a) for a in self.A]
  #print [str(a) for a in self.C]
  #print [str(a) for a in self.G]
  #print [str(a) for a in self.T]
  #print [str(a) for a in self.col_score]
def score(self,sequence,chrome,start,stop): #Optional index
  #sequence=sequence[500:]
  sequence= sequence.upper()
  offset = 0
  score = 0
  len_seq = len(sequence)
  mc = 0 #motif counter
  while offset + self.motif_length <  len_seq:
    mc = 0
    while mc < self.motif_length:
      if sequence[offset+mc] == 'A':
        score += self.A_score[mc]
      elif sequence[offset+mc] == 'C':
        score += self.C_score[mc]
      elif sequence[offset+mc] == 'G':
        score += self.G_score[mc]
      elif sequence[offset+mc] == 'T':
        score += self.T_score[mc]
      mc+=1
      #sort out long sequence of same bases
      a_counter = c_counter = g_counter = t_counter = 0
      match = sequence[offset:offset+self.motif_length]
      for letter in match:
        if letter == "A":
          a_counter += 1
        elif letter == "C":
          c_counter += 1
        elif letter =="G":
          g_counter += 1
        elif letter == "T":
          t_counter += 1
      non_major = 100
      if a_counter >= len(match)-3:
        non_major =  len(match)  -a_counter
      if c_counter >= len(match)-3:
        non_major =  len(match) -c_counter
      if g_counter >= len(match)-3:
        non_major =  len(match)  -g_counter
      if t_counter >= len(match)-3:
        non_major =  len(match)  -t_counter
      if non_major < 2:
        score = 0
      if non_major == 2:
         score = score * .75
      if non_major == 3:
         score = score * 0.8
     #End of penalizing long same base sequences
    if score>5: #Choose threshold of reporting here
      rep_line = chrome+ "\t"
      rep_line = rep_line+chrome+"_"+str(start)+"\t"
      rep_line = rep_line+str(offset+start)+"\t"
      rep_line = rep_line+str(offset+start+10)+"\t"
      rep_line = rep_line+self.name+ "\t"
      rep_line = rep_line+ str(round(score, 4))+ "\t"
      rep_line = rep_line+sequence[offset:offset+self.motif_length].lower()
      #print self.name
      #print float(score)
      #print offset+start
      #print sequence[offset:offset+self.motif_length].lower()
      print rep_line
    score = 0
    offset+=1
    #print score
```

```
f = open('motifs_all.txt', 'r')
matrix_data = f.readlines()
matrix_data = [item.rstrip() for item in matrix_data]
f.closed
matrix_list = []
active =[]
while matrix_data:
  current = matrix_data.pop(0)
  if len(current) == 0:
    continue
  if current[0] in [" ","#"] :
    continue
  if current[0] == ">":
    current == current
    #print "processing "+current.lstrip(">")
  active.append(current)
  active.append(matrix_data.pop(0))
  active.append(matrix_data.pop(0))
  active.append(matrix_data.pop(0))
  active.append(matrix_data.pop(0))
  temp = Matrix(active)
  matrix_list.append(temp)
  active=[]
#for matrix in matrix_list:
  #matrix.Report()
#print "going into score"
#print "forward"
#matrix_list[0].score(AB004572)
#print "reverse"
#matrix_list[1].score(AB004572)
import re
f = open('AA_freq_table.txt', 'r')
read_data = f.readlines()
f.closed
f = open('fasta.txt', 'r')
read_fasta = f.readlines()
f.closed
string = ""
#Parsing header##
#header = read_fasta[0]
#(h1,h2) = header.split("species")
#(chrome,loc)=  h1.split(":")
#chrome = chrome[1:]
#(start,stop)= loc.split("..")
#read_fasta[0] = ""
string =""
for line in read_fasta:
  if line[0] == ">":
    (h1,h2) = line.split("species")
    (chrome,loc)=  h1.split(":")
    chrome = chrome[1:3]
    (start,stop)= loc.split("..")
    start =int(start)
    stop = int(stop)
    if string:
      for mat in matrix_list:
        mat.score(string,chrome,start,stop)
    string = ""
  else:
    line = line.rstrip()
    string = string +line.upper()
  #DNA = string
#complement_map = {'C': 'G', 'G': 'C', 'A': 'T', 'T': 'A'}      # map for use in finding reseverse sequence
#def reverse_complement(DNA):                                   # defining the action of finding reverse sequence
    #complist = map(complement_map.get, DNA)
    #complist.reverse()
    #return ''.join(complist)
#DNArev = reverse_complement(DNA)                                 # repeate commands for reverse sequence
#DNArev = DNArev[0:]
#for mat in matrix_list:
#  mat.score(string)
#for mat in matrix_list:
  #mat.score(DNArev)
ammino = {}
```

```
for line in read_data:
  line = line.rstrip()
  Char,Name,Codon,Protein_Freq,Codon_usage_Freq = line.split()
  sub_dict = {}
  sub_dict['AA'] = Name
  sub_dict['AA_Freq'] = Protein_Freq
  sub_dict['Codon_Rel_Freq'] = Codon_usage_Freq
  sub_dict['Char'] = Char
  ammino[Codon] = sub_dict
```

Algorithm D.3: Algorithm that counts all SNPs within a binding site of total SNPs in area(SNP_call_extract).

```
import os,fileinput
f = open('FV2L.txt', 'r')
SNP_FV2L = f.readlines()
f.closed
#f = open('Final_2L_HB.txt', 'r')
#F_2L_HB = f.readlines()
#f.closed
#f = open('Final_2L_Snail.txt', 'r')
#F_2L_Sn = f.readlines()
#f.closed
#f = open('Final_2L_Krüppel.txt', 'r')
#F_2L_KR = f.readlines()
#f.closed
#joke = SNP_FV2L[1].split(',')
#print joke[0]
f = open('dummypos.txt', 'r')
dummy = f.readlines()
dummy = [item.rstrip() for item in dummy]
f.closed


#print dummy
#f = open('Hunchback_new_7.txt', 'r')
#chr_hb = f.readlines()
#f.closed
#f = open('Snail_new_7.txt', 'r')
#chr_sn = f.readlines()
#f.closed
#f = open('Krüppel_new_7.txt', 'r')
#chr_kr = f.readlines()
#f.closed
#for item in chr_hb:                                          #what table is being used to extract areas
#  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()   #tf-transcription factor, bs_seq-binding site sequence
#  start = int(start)
#  stop = int(stop)
#for item in chr_sn:                                          #what table is being used to extract areas
#  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()   #tf-transcription factor, bs_seq-binding site sequence
#  start = int(start)
#  stop = int(stop)
#for item in chr_kr:                                          #what table is being used to extract areas
#  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()   #tf-transcription factor, bs_seq-binding site sequence
#  start = int(start)
#  stop = int(stop)
counter_hb = 0
line_counter = 0
SNP_FV2L = SNP_FV2L[1:]
#chr_hb = chr_hb[1:]
#chr_kr = chr_kr[1:]
#chr_sn = chr_sn[1:]
#print(dummy)
nr2 = []
for line in SNP_FV2L:
  items  = line.split(',')
  if len(items) >0 and len(items[0])>0 :
    if items[0][0] in  ['0','1','2','3','4','5','6','7','8','9']:
      items[0] =int(items[0])
      nr2.append(line)

for line in nr2:
  items  = line.split(',')
  line_counter = line_counter + 1
  for line in dummy: #dummy is array of binding sites location
    chrs,chr_start,start,stop,tf,score,bs_seq = line.split()
    start = int
    stop = int
    if str(chrs) == "2L" and (str(tf) == "hb" or str(tf) == "hb_rev"):
      if ((int(items[0]) >= start) and (int(items[0]) <= stop)):
          counter_hb = counter_hb+1
print counter_hb
print line_counter
```

Algorithm D.4: Algorithm that extracts all SNPs found within bound regions from original released SNP file (SNPs_reduced).

```
import os,fileinput
TRUE =1
FALSE = 0
#f = open('FV2L.txt', 'r')#SNP_FV2L = f.readlines(6)
#print SNP_FV2L
#for line in fileinput.input(['Final_Variants_X.txt']):
#  print (line)
f = open('h3L.chr', 'r')
chr_h3L = f.readlines()
chr_h3L = [item.rstrip() for item in chr_h3L]
chr_h3L = chr_h3L[1:]
f.closed
loc_list = []
for step in chr_h3L:
  chrs,start,stop,region,counter = step.split()
  start = int(start)
  stop = int(stop)
  loc_list.append([start,stop])
def check_area(loc):
  loc =int(loc)
  for item in loc_list:
    if (loc >= item[0] and loc <= item[1]):
      return(TRUE)
  return(FALSE)
line = "XXX"
line_nr =1
f = open('Final_Variants_3L.txt', 'r')
f_out = open('Final_3L_HB.txt', 'w')
line =f.readline()
f_out.write(line)
while line :
  line = f.readline()
  if line_nr==1:
    fields=line.split(",")
    loc = fields[0]
    if check_area(loc): #THIS IS WHERE YOU TEST FOR AREA
      TO_USE = TRUE
    else:
      TO_USE = FALSE
  if TO_USE == TRUE:
    f_out.write(line)
  line_nr = line_nr+1
  if line == "\n" :
     line_nr = 1
f.closed
f = open('Final_Variants_X.txt', 'r')
SNP_FVX= f.readlines(8)
#print SNP_FVX
f.closed
f = open('Hunchback_new_7.txt', 'r')
chr_hb = f.readlines()
f.closed
f = open('Snail_new_7.txt', 'r')
chr_sn = f.readlines()
f.closed
f = open('Krüppel_new_7.txt', 'r')
chr_kr = f.readlines()
```

```
f.closed
for item in chr_hb:                                          #what table is being used to extract areas
  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()          #tf-transcription factor, bs_seq-binding site sequence
  start = int(start)
  stop = int(stop)
for item in chr_sn:                                          #what table is being used to extract areas
  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()          #tf-transcription factor, bs_seq-binding site sequence
  start = int(start)
  stop = int(stop)
for item in chr_kr:                                          #what table is being used to extract areas
  chrs,chr_start,start,stop,tf,score,bs_seq = item.split()          #tf-transcription factor, bs_seq-binding site sequence
  start = int(start)
  stop = int(stop)
counter_hb = 0
line_counter = 0
#print line_counter/7
```

Algorithm D.5: Algorithm that extends N of 7 more N‘s from the sequence alignment files (extended_N).

```
##################################
# What files are being used
################################
import os
chrom_dir =os.getcwd()+ "\\"+"Hb_EVE"       #folder containing alignment files
chrom_list = []
##############################
# Loop for extending N areas
#########################
TRUE = 1
FALSE =0
def extend_N(nr,string):
# nr = number how many N to add to each side
# string = the string to pad the N in.
 array = list(string)
 length = len(string)
 counter =0
 in_N = FALSE
 while counter < length:
   if (array[counter] == "N" and (in_N == FALSE)):                    #finds N and extends front
     #Extend front
     in_N =TRUE
     for i in range(nr+1):
       to_change = counter-i
       to_change = max(to_change,0)
       array[to_change] = "N"
   elif (array[counter] <>"N" and (in_N == TRUE)):                    #finds not N and extends back
      #Extend back
      in_N =FALSE
      for i in range(nr):
        to_change = counter+i
        to_change = min(to_change,length-1)
        array[to_change] = "N"
      counter =counter + nr                                  #jumps forward as many N's as were extended
   #print ''.join(array)
   counter = counter+1
 #print array
 return(''.join(array))
##############################
#-Read area files to array
##############################
raw_chrom_dir = os.listdir(chrom_dir)                         #list of files read in
for item in  raw_chrom_dir:
  out =item.split('.')                                       #split on a .
  if (len(out) >1):
    #if (out[1] == ".X" or ".2L" or ".2R" or ".3L" or ".3R"):            #what files are used (the ending)
      chrom_list.append(item)
chrom_list = ['primer_area.e']
#-------------------------------------
os.chdir(os.getcwd()+ "\\" + "Hb_EVE")                         #destination file
for fileX in chrom_list:
  f = open(chrom_dir+"\\"+fileX, 'r')                            #File X is the active file
  active_file = f.readlines()
  #active_file = [item.rstrip() for item in active_file]
  line_counter = 0
  while line_counter < len(active_file):
    #print active_file[line_counter]
    #print "X------------------------X"
    if active_file[line_counter][0] <> ">":                              #skips lines beginnin with >
      active_file[line_counter] = extend_N(7,active_file[line_counter])        #how many N's are changed
    #print active_file[line_counter]
    line_counter=line_counter+1
  outfile = open(chrom_dir+ "\\" + "N_"+fileX, 'w')                       #what is in outfile
  #print active_file
  outfile.writelines(active_file)
  outfile.close()
  f.closed
```

Algorithm D.6: Algorithm that writes out SNPs within binding sites and lists minor allele frequencies and the genotypes for major and minor alleles (SNP_merge).

```
import os,fileinput,sys
FALSE = ''

#loc_file = sys.argv[1]         #file containing binding sites
#snp_file = sys.argv[2]         #file containing SNP locations

f = open("Hunchback_new_5.txt", 'r')        #open binding sites file
LOC_raw = f.readlines()
f.closed

f = open("Final_2L_HB.txt", 'r')        #open SNP file
SNPs = f.readlines()
f.closed
SNPs = SNPs[1:]                 #removing 1st line

#f = open('Final_2L_HB.txt', 'r')
#F_2L_HB = f.readlines()
#f.closed
#F_2L_HB = F_2L_Hb[1:]

#f = open('Final_2L_Snail.txt', 'r')
#F_2L_Sn = f.readlines()
#f.closed
#F_2L_Sn = F_2L_Sn[1:]

#f = open('Final_2L_Kruppel.txt', 'r')
#F_2L_KR = f.readlines()
#f.closed
#F_2L_KR = F_2L_Kr[1:]


#f = open('Hunchback_new_5.txt', 'r')
#chr_hb = f.readlines()
#f.closed

#f = open('Snail_new_5.txt', 'r')
#chr_sn = f.readlines()
#f.closed

#f = open('Kruppel_new_5.txt', 'r')
#chr_kr = f.readlines()
#f.closed

loc_list = []                  #modify binding site file
for item in LOC_raw:
  item = item.rstrip()
  line = item.split()
  loc_list.append(line)


#Now lets loop
reduced = []
while len(SNPs) >5:            #as long as there are more than 5 lines
  data1 = SNPs[0].split(",")   #what is in each line and split on comma
  pos   = data1[0]              #this is the loc and what each individual's genotype
  coverage = (len(data1)- 2 - data1.count('N') )
  data4 = SNPs[2].split(",")    #this is A
  data5 = SNPs[3].split(",")    #this is C
  data6 = SNPs[4].split(",")    #this is G
  data7 = SNPs[5].split(",")    #this is T
  count_A = float(data4[-2])                #use the second last number of this line which is the frequency for A
  count_C = float(data5[-2])                #use the second last number of this line which is the frequency for C
  count_G = float(data6[-2])                #use the second last number of this line which is the frequency for G
  count_T = float(data7[-2])                #use the second last number of this line which is the frequency for T
  if (count_A > 0) and (count_A <= 0.5):    #defining of a nucleotide is minor og major allele
    minor = "A"
    minor_freq = count_A
  if (count_A >= 0.5):
    major = "A"


  if (count_C > 0) and (count_C <= 0.5):    #defining if a nucleotide is minor og major allele
    minor = "C"
    minor_freq = count_C
  if (count_C >= 0.5):
    major = "C"

  if (count_G > 0) and (count_G <= 0.5):    #defining if a nucleotide is minor og major allele
    minor = "G"
    minor_freq = count_G
  if (count_G >= 0.5):
    major = "G"

  if (count_T > 0) and (count_T <= 0.5):    #defining if a nucleotide is minor og major allele
    minor = "T"
    minor_freq = count_T
  if (count_T >= 0.5):
    major = "T"

### Missing is coverage - how many are genotyped at each position
  #coverage=sum(count((data4[4:10])))
  #(data4[4:15])

  string = str(pos)+" " + major+ " " +minor+ " " + str(minor_freq) +" " +str(coverage)
  line = [pos,major,minor,minor_freq,coverage]
  reduced.append(line)
  SNPs = SNPs[7:]
del SNPs

def check_area(SNPs):
  loc = int(SNPs[0])
  for item in loc_list:
    if item[0] == "2L":
      if (loc >= int(item[2]) and loc <= int(item[3])):
        SNP_item = item[0]+ "\t"                          #chromosome
        SNP_item = SNP_item + str(SNPs[0])+ "\t"          #SNP location
        SNP_item = SNP_item + SNPs[1]+ "\t"               #Major allele
        SNP_item = SNP_item + SNPs[2]+ "\t"               #Minor allele
        SNP_item = SNP_item + str(SNPs[3])+ "\t"          #minor allele frequency
        SNP_item = SNP_item + str(SNPs[4])+ "\t"          #Coverage
        SNP_item = SNP_item + item[4]+ "\t"               #name of TF
        SNP_item = SNP_item + item[2]+ "\t"               #location of start of binding site
        SNP_item = SNP_item + item[5]+ "\t"               #score of binding site
        SNP_item = SNP_item + item[6]+ "\n"                #sequence of binding site
        return(SNP_item)
  return(FALSE)

os.chdir(os.getcwd()+ "\\"+"Large_tables")              #Destination file

outfile = open("test.txt",'a')                         #opening the outfile that has to be previously made

for item in reduced:
  ret = check_area(item)                               #ret gives all the SNPs found within a site
  outfile.write(ret)                            #in outfile is written ret
#  if ret:
#    print ret
outfile.close()
```

Appendix E

Additional tables and figures

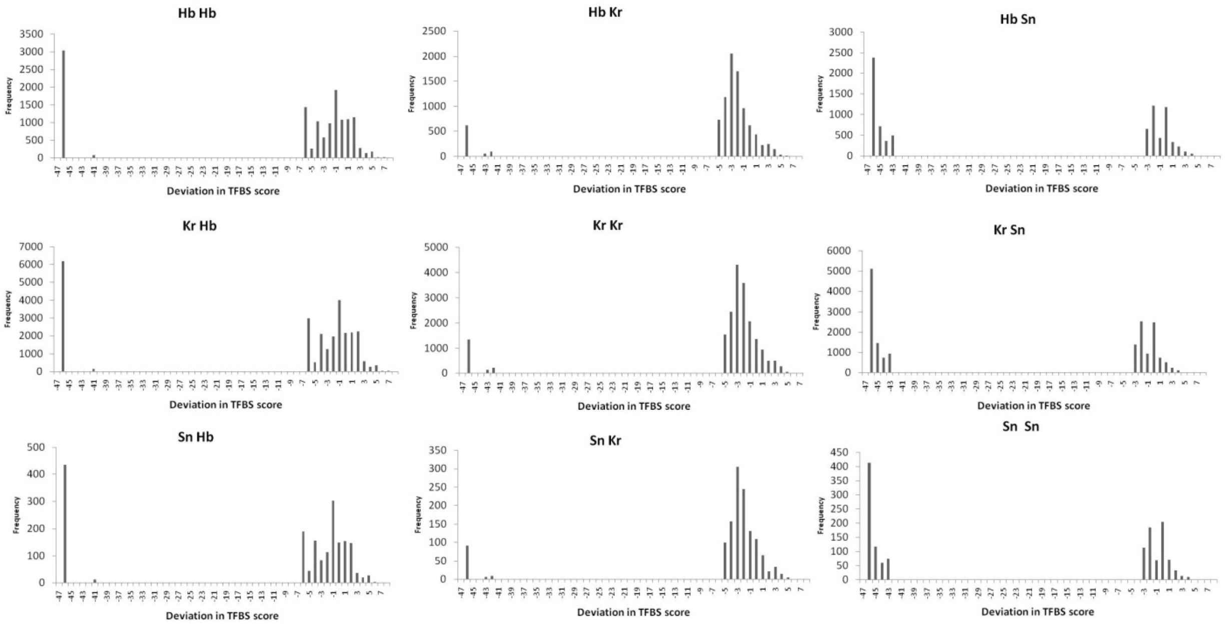


Figure E.1: Change in score when a SNP affects a binding site.

Table E.1: Pearsons Chi-sqare test on deletions within binding sites.

Factors tested	Chi-square	Degrees of freedom	p-value
All	33.5696	2	5.134e-08
HB vs. KR	7.9612	1	0.004779
KR vs. SNA	20.0662	1	7.48e-06
HB vs. SNA	33.5696	2	5.134e-08

Table E.2: Pearsons Chi-sqare test on insertions within binding sites.

Factors tested	Chi-square	Degrees of freedom	p-value
All	2.1087	2	0.3484
HB vs. KR	1.5145	1	0.2185
KR vs. SNA	0.1799	1	0.6715
HB vs. SNA	2.1087	2	0.3484

Table E.3: Pearsons Chi-sqare test on deletions within binding sites.

Factors tested	Chi-square	Degrees of freedom	p-value
All	89.2467	2	2.2e-16
HB vs. KR	6.8754	1	0.00874
KR vs. SNA	87.9771	1	2.2e-16
HB vs. SNA	89.2467	2	2.2e-16

Table E.4: Pearsons Chi-sqare test on insertions within binding sites.

Factors tested	Chi-square	Degrees of freedom	p-value
All	12.2864	2	0.002148
HB vs. KR	4.2172	1	0.04002
KR vs. SNA	9.4138	1	0.002154
HB vs. SNA	12.2864	2	0.002148

Table E.5: Statistical analysis for distribution of SNPs within binding sites.

Region	TFBS	Chi-square	Degrees of freedom	p-value
HB-bound	Hb	300.9493	9	2.2e-16
	Kr	204.3143	9	2.2e-16
	Sn	174.372	9	2.2e-16
KR-bound	Hb	607.5041	9	2.2e-16
	Kr	425.3733	9	2.2e-16
	Sn	258.9042	9	2.2e-16
SNA-bound	Hb	48.2955	9	2.249e-07
	Kr	178.3244	9	2.2e-16
	Sn	32.2284	9	0.0001818

Table E.6: Statistics for distribution of SNPs within binding sites.

Factors tested	Chi-square	Degrees of freedom	p-value
HB	46.4883	9	4.887e-07
KR	36.1776	9	3.687e-05
SNA	15.9667	9	0.06758

Table E.7: Counts for all SNPs in binding sites in 21 developmental gene regions.

Chr	Gene	TF	Within BS	SNPs in that gene	Total BS in gene	Ratio of total	Ratio of all BS
2L	***bowl***	**HB**	29	1061	63	**0.027**	**0.460**
		KR	24	1061	54	0.023	0.444
		SNA	27	1061	62	0.025	0.435
2L	***caud***	**HB**	10	475	58	0.021	0.172
		KR	16	475	47	**0.033**	**0.340**
		SNA	5	475	46	0.010	0.109
2L	***dpp***	**HB**	23	999	87	**0.023**	0.264
		KR	15	999	53	0.015	**0.283**
		SNA	28	999	103	0.028	0.272

Table E.7: Continued.

Chr	Gene	TF	Within BS	SNPs in that gene	Total BS in gene	Ratio of total	Ratio of all BS
2L	***slp1+2***	**HB**	64	2521	134	**0.025**	0.478
		KR	58	2521	107	0.023	**0.542**
		SNA	63	2521	123	0.025	0.512
2L	***sna***	**HB**	26	522	43	**0.050**	**0.605**
		KR	17	522	32	0.033	0.531
		SNA	8	522	37	0.015	0.216
2R	***eve***	**HB**	12	402	55	**0.030**	0.218
		KR	5	402	50	0.012	0.100
		SNA	12	402	38	**0.030**	**0.316**
2R	***kr***	**HB**	42	736	130	**0.057**	**0.323**
		KR	27	736	92	0.037	0.293
		SNA	15	736	59	0.020	0.254
2R	***twi***	**HB**	18	553	35	**0.033**	0.514
		KR	10	553	23	0.018	0.434
		SNA	17	553	23	0.031	**0.739**
3L	***D***	**HB**	31	699	74	**0.044**	**0.419**
		KR	15	699	52	0.021	0.288
		SNA	17	699	42	0.024	0.405
3L	***kni***	**HB**	7	267	70	0.026	0.100
		KR	8	267	55	**0.030**	**0.145**
		SNA	5	267	36	0.019	0.139
3R	***hb***	**HB**	26	454	107	**0.057**	**0.243**
		KR	9	454	67	0.020	0.134
		SNA	9	454	60	0.020	0.150
3R	***hkb***	**HB**	1	83	52	0.012	0.019
		KR	4	83	31	**0.048**	**0.129**
		SNA	3	83	34	0.036	0.088
3R	***ftz***	**HB**	23	951	158	0.024	0.146
		KR	23	951	119	**0.024**	**0.193**
		SNA	20	951	111	0.021	0.180
3R	***opa***	**HB**	13	380	110	0.034	0.118
		KR	15	380	78	**0.039**	**0.192**
		SNA	3	380	86	0.008	0.035
3R	***tll***	**HB**	17	417	35	**0.040**	**0.486**
		KR	9	417	30	0.022	0.300
		SNA	13	417	31	0.031	0.419
X	***brk***	**HB**	18	472	83	**0.038**	0.217
		KR	14	472	54	0.030	**0.259**
		SNA	7	472	57	0.015	0.123
X	***btd***	**HB**	8	371	55	**0.022**	0.145
		KR	7	371	34	0.019	**0.206**
		SNA	6	371	46	0.016	0.130
X	***gt***	**HB**	32	620	187	**0.052**	0.171
		KR	24	620	100	0.039	0.240
		SNA	15	620	55	0.024	**0.273**
X	***run***	**HB**	25	751	113	0.033	0.221
		KR	26	751	84	**0.034**	**0.310**
		SNA	9	751	56	0.012	0.161
X	***sog***	**HB**	35	1191	137	**0.029**	0.255
		KR	21	1191	80	0.018	**0.262**
		SNA	19	1191	108	0.016	0.176
X	***vnd***	**HB**	7	356	67	**0.020**	**0.104**
		KR	5	356	76	0.014	0.066
		SNA	4	356	90	0.011	0.044