A systems-level study of *giant* regulation in *Drosophila melanogaster*

Astrid Hoermann

TESI DOCTORAL UPF / 2014

DIRECTOR DE LA TESI:

Dr. Johannes Jaeger

Centro de Regulación Genómica (CRG), EMBL / CRG - Systems Biology Unit

dedicated to
my parents

“Two years’ work wasted,
I have been breeding those flies for all that time
and I've got nothing out of it.”

Thomas Hunt Morgan

Acknowledgements

- Reinitz lab: John Reinitz, Ah-Ram Kim and Carlos Martinez for the xtransc code, the evoplot tool and for answering all my questions about them
- Damjan Cicin-Sain for modifying the matlab scripts in order to quantify lacZ, for making the code run on our machines and on the cluster, for contributing several scripts to handle different kind of files in batch, for adapting the FlyGUI to my needs, for his patience, for making back-ups, and for solving each and every computational problem immediately
- Johannes Jaeger: for counterbalancing my pessimism with his optimism
- For their contribution to the *Clogmia* project:
  Ultrasequencing Unit (Heinz, Maik, Rebecca, Sarah)
  Bioinformatics Core (Ernesto, Luca) for the assembly
  Saurabh Sinha, Yinan Zhang, Abdol Majid Kazemian for identifying the CREs
- Eva Jiménez: for feedback on the entire thesis manuscript, for keeping my flies alive while I was on holidays, for second senior opinions, for her help while handling the library membrane, for discussing other work-related issues and for acting in my video about *in-situ* hybridizations
- Bàrbara Negre: for double checking crossing schemes, for setting up crosses while I was away, for looking at cuticle phenotypes and for feedback on parts of the thesis manuscript
- Anton Crombach: for ideas and advice on modeling and plotting and for feedback on the modeling parts of the thesis manuscript
- Karl Wotton: for explaining me the microscope, for the hb and cad probes, the hb RNAi fragment, for showing me how to do RNAi and hand-devitellinisation and for useful feedback on my application for a scientific workshop
- Anna Alcaine: for keeping my flies alive while I was on holidays; for help with flipping stocks, collecting virgins, fixing embryos, isolating gDNA; for doing some control-PCRs
- Brenda Gavilán: for help with flipping stocks, screening flies and fixing embryos
- Núria Bosch, Alejandro Thérèse Navarro: for help with flipping stocks
- Heleia Roca: for feedback on the Spanish version of the abstract
- CRG kitchen and ALMU (Timo, Raquel, Arrate, Javi)
- Kenneth Barr: for sharing his list of candidates for cooperativity
- Marc Battler for help with the FUJI-Imager
- la caixa: for funding
- The generous fly community for plasmids, flies, advice or sharing protocols, data, codes:
  Steve Small, Miki Fujioka, Jack Bateman, Robert Zinzen, Guillaume Junion, Rupinder Sayal, Susan Celniker, Ulrike Gaul, Mark Biggin, Svetlana Surkova, Konstantin Kozlov, Maria Samsonova, Pavel Tomancak, Radoslaw Kamil Ejsmont, Jean Paul Vincent, Alex Stark, Alistair Boettiger, Thomas Gregor, Michael Perry, Max Staller, Angela DePace…

Abstract

Deutsche Kurzfassung:

Studie auf Systemebene über die Regulation von *giant* in *Drosophila melanogaster*

Diese Dissertation entschlüsselt die Regulierung der Transkription des Gap-Gens *giant (gt)* im *Drosophila* Blastoderm Embryo mit einem Reverse-Engineering Ansatz: ein mathematisches Modell extrahiert die zugrunde liegenden Mechanismen aus quantitativen Expressionsdaten vom Wildtyp. Das Modell wird an Reporter-mRNA angepasst, die von den *cis*-regulatorischen Elementen (CRE) von *gt* gesteuert wird. Es ist ein leistungsfähiges Werkzeug um zu erforschen, wie das Expressionsmuster auf der molekularen Ebene von den verschiedenen Bindungsstellen für die Transkriptionsfaktoren gebildet wird, und es gibt uns die Möglichkeit, die Musterung in Mutanten vorauszusagen. Diese Studie verdeutlicht, dass zwei aneiandergrenzende *gt* CREs unterschiedlich reguliert werden, und erbringt den ersten experimentellen Beweis für Gt Auto-Aktivierung durch Mutagenese seiner Regulations-Elemente. Nach der Optimierung der Parameter in einem Wildtyp-Hintergrund, kann das Modell die beobachteten Veränderungen in den *Krüppel* und *tailless* Mutanten richtig vorausberechnen. Beiträge von andere Transkriptionsfaktoren, die das Modell vorgeschlagen hat, werden durch systematische Auswertung der CREs in den entsprechenden Mutanten bestätigt.

Resumen en español:

Estudio a nivel sistémico de la regulación de *giant* en *Drosophila melanogaster*

Esta tesis revela la regulación transcripcional del gen gap *giant (gt)* en el embrión blastodermal de *Drosophila* por ingeniería inversa: un modelo matemático infiere los mecanismos subyacentes de datos cuantitativos de expresión recopilados en un fondo genético silvestre. El modelo se amolda a mRNA reportero controlado por elementos reguladores en *cis* (CRE) de *gt*. Es una herramienta potente para investigar cómo se forma el patrón a nivel molecular por los sitios de unión de factores de transcripción y permite predecir la expresión en cepas mutantes. La presente tesis esclarece la regulación diferencial de dos CRE adyacentes de *gt* y presenta la primera evidencia experimental de auto-activación de *gt* mediante mutagénesis de sus elementos reguladores. Tras la optimización de los parámetros en un fondo de tipo silvestre, el modelo predice correctamente los cambios observados en mutantes de *Krüppel* y *tailless.* Otras contribuciones reglamentarias sugeridas por el modelo son confirmadas por la evaluación sistemática de los CREs en mutantes.

English abstract:

A systems-level study of *giant* regulation in *Drosophila melanogaster*

This thesis unravels the transcriptional regulation of the gap gene *giant* (*gt*) in the *Drosophila* blastoderm embryo via a reverse-engineering approach: a mathematical model infers the underlying mechanisms from quantitative expression data collected in the wild-type background. The model is fit to reporter mRNA driven by *cis*-regulatory elements (CRE) of *gt*. It is a powerful tool to investigate how the pattern is formed at the molecular level from transcription factor binding sites and it gives us the ability to predict the expression in mutants. This thesis elucidates the differential regulation of two adjacent *gt* CREs and presents the first experimental evidence for Gt auto-activation via site-directed mutagenesis of its enhancers. After optimizing the parameters in the wild-type background, the model correctly predicts the observed changes in *Krüppel* and *tailless* mutants. Other regulatory contributions suggested by the model are confirmed by systematic evaluation of the CREs in mutants.

Preface

One of the big challenges of today is to quantitatively understand transcriptional regulation in eukaryotes due to its crucial role in development and pattern formation. In the last decade, modelling attempts brought new insights but also raised many new questions. In contrast to most of these approaches in the field of transcriptional regulation, this study does not aim for genome-wide predictions, but rather for an in-depth analysis of the regulatory mechanisms of an endogenous gene with high resolution in space and time.

Segment determination in the *Drosophila* embryo is guided by a limited set of genes, which are subdivided into maternal, gap, pair-rule and segment polarity genes. The gap gene *giant* (*gt*) is expressed in a broad anterior and a posterior domain during the early blastoderm stage of development. Before gastrulation the anterior domain refines into two stripes and also expression at the anterior tip becomes visible. The regulatory region of *gt* contains binding sites for several transcription factors (TF). Among them are the activators Bicoid and Caudal, as well as the repressors Hunchback, Tailless, Knirps and Krüppel.

The aim of this PhD thesis is to understand by which mechanisms this network acts on the different *cis*-regulatory elements (CRE) of *gt* and how that leads to the observed expression domains. I want to capture the dynamics of the system and investigate how different transcription factor binding sites (TFBS) form a CRE and how these elements together establish the entire expression pattern of the gene.

These questions are addressed by quantitative analysis and mathematical modelling of spatio-temporal patterns. In contrast to the traditional genetic approach, which first perturbs the system to draw conclusions, this reverse-engineering approach extracts information about regulatory interactions from quantitative expression data collected in embryos carrying reporter constructs in a wild-type background. The model of transcriptional control is fit to the expression of a reporter mRNA driven by different *gt* CREs. The output is the combinations of activator and repressor sites on the DNA sequence required for correct expression over time. Hence, it is a powerful tool to investigate how the *gt* pattern is formed at the molecular level.

This thesis elucidates the differential regulation of two adjacent CREs driving the posterior domain of *gt*. Expression driven by the element gt-3 arises earlier and is activated by Caudal, whereas gt-1, which additionally drives the anterior domain of *gt*, comes up slightly later and depends on auto-activation. I provide the first experimental evidence for Giant auto-activation via meticulous site-directed mutagenesis of its enhancers. Other regulatory contributions predicted by the model are also confirmed by systematic evaluation of the CREs in mutants. After optimizing the parameters in a wild-type background, quantitative datasets of the *tailless* and the *Krüppel* mutant are plugged into the model, which accurately predicts the pattern of the CREs in these mutants. Hence, this is the first report of a validated quantitative model derived from wild-type, able to predict the expression of enhancers in mutants from their sequence.

Contents

Abbreviations

a.u.	arbitrary units
aa	amino acids
AP	alkaline phosphatase
A-P	anterior-posterior
approx.	approximately
BF	bright field
bp	base pairs
BTM	basal transcriptional machinery
b-ZIP	basic leucine zipper
C	cleavage cycle
cDNA	complementary DNA
CDS	coding sequence
conc.	concentration
CRE	*cis*-regulatory element
DEPC	diethylpyrocarbonate
DIC	differential interference contrast
Dmel	*Drosophila melanogaster*
dNTP	deoxyribonucleotide
DPE	downstream core promoter element
Dpse	*Drosophila pseudoobscura*
D-V	dorsal-ventral
AP	alkaline phosphatase
EtOH abs.	absolute ethanol
FISH	fluorscent *in-situ* hybridization
gDNA	genomic DNA
GLC	germ line clones
h	hours
hs	heat shock
kb	kilobases
max	maximum
MCS	multiple cloning sites
min	minutes
mRNA	messenger RNA
MSE	minimal stripe element
neg. ctrl.	negative control
nt	nucleotides
OLS	ordinary least squares
max	maximum
ORF	open reading frame
PCR	polymerase chain reaction
pos.	positive
PWM	positional weight matrix
RMS	root mean square
RNAi	RNA interference
sh	short hairpin
smFISH	single molecule FISH
TF	transcription factor
TFBS	transcription factor binding sites
UAS	upstream activating sequence
UTR	untranslated region
WLS	weighted least squares
WT	wild-type

Drosophila **genes**

bcd	*bicoid*
bks	*brakeless*
btd	*buttonhead*
cad	*caudal*
cic	*capicua*
ems	*empty spiracles*
en	*engrailed*
eve	*even-skipped*
exu	*exuperantia*
ftz	*fushi tarazu*
gsb	*gooseberry*
gt	*giant*
h	*hairy*
hb	*hunchback*
hh	*hedgehog*
hkb	*huckebein*
kni	*knirps*
Kr	*Krüppel*
l(2)gl	*lethal (2) giant larvae*
nos	*nanos*
odd	*odd-skipped*
odt	*orthodenticle*
osk	*oskar*
prd	*paired*
run	*runt*
slp	*sloppy paired*
tll	*tailless*
tor	*torso*
tsl	*torso-like*
vas	*vasa*
wg	*wingless*
wol	*wollknäuel*
zld	*zelda (vielfältig)*

List of figures

List of tables

Appendix

1 Introduction

It is fascinating how a simple fertilized egg can develop into a complex organism and that all the required information is encoded in the genome. The core task of this process is to switch on genes, fine-tune their expression patterns, ensure the maintenance of certain levels and also to shut them down when necessary. This is achieved by sets of transcription factors (TF) bound in concert to short regions on the DNA called *cis*-regulatory elements (CRE). Researchers set out to decipher the regulatory code and discovered that certain rules exist, but many exceptions and unsolved puzzles remain. A quantitative understanding of eukaryotic transcription is lacking partly due to the absence of an *in vitro* reconstitution assay. We want to achieve mechanistic insights into enhancer function via quantitative modelling of spatio-temporal expression patterns of the gap gene *giant* (*gt*). It is a TF involved in the regulation of other *Drosophila* segmentation genes. Although it has a relatively simple expression pattern, it tends to be troublesome when modeling protein interaction networks. Previous modeling attempts in wild-type, as well as in the *tailless* mutant, only worked if the two domains were treated separately (Jaeger et al. 2007). Modeling of the gap gene network of a distantly related fly species, *Clogmia albipunctata*, infers unexpected influences from Caudal on *gt* (Crombach et al. 2014). Also a model of transcriptional control fit to *even-skipped* (*eve*) stripe enhancers was not able to reproduce the expression patterns driven by the CREs of *gt* (Kim et al. 2013).

1.1 Principles of eukaryotic transcriptional regulation

1.1.1 Transcription factor binding sites and *cis*-regulatory elements

Transcription factor binding sites (TFBS) are short stretches on the DNA sequence that can be recognized by specific DNA-binding proteins such as activators or repressors (Latchman 1997). These motifs can occur randomly and binding of one transcription factor (TF) alone to an isolated site only, might not provoke a response. But if several TFBS cluster to form an element, a specific effect can be achieved. Such *cis*-regulatory elements or modules (CRE or CRM) can be further subdivided into enhancer, silencer or insulator, according to their function (Blackwood and Kadonaga 1998). Most research has focused on enhancers, since they define when and where a eukaryotic gene is expressed. Hence, in the literature the term CRE usually refers to the enhancers.

From an experimental point of view, enhancers were defined as DNA fragments of about 1kb in length that can be placed in front of a promoter to drive expression of a reporter gene independent of the orientation and location of the sequence. This was taken even further by introducing a DNA fragment from another species, which occasionally resembled partially or entirely the endogenous transcriptional output, if the TFs were sufficiently conserved (Tautz 2000).

Enhancers were further categorized based on their mode of action as enhanceosomes or billboards (Arnosti and Kulkarni 2005). Enhanceosomes need to be loaded with certain proteins that form a complex and act highly cooperatively and coordinated. Disruption of the physical interaction of one of the members of the complex usually renders them inactive. The billboard model proposes that TFBS can be distributed flexible, since their individual inputs somehow sum up to an overall output. The arrangement of the binding sites can be changed as long as certain subunits are maintained. Recently, a third model of enhancer action was discovered and termed “TF collective” (Junion et al. 2012, Spitz and Furlong 2012) or “HOT (highly occupied target) regions” (Kvon et al. 2012). Such enhancers are depleted in motifs for the bound TFs and rely on a higher degree of cooperative binding resulting in even more flexible motif composition.

The regulatory region of an insect gene tends to be quite compact via lining up a few CREs within less than 10 kb of an intergenic stretch. Nevertheless, they do not have to be intergenic, but can even reside in introns or exons. It is not entirely clear how CREs find the promoter of their target gene to activate the basal transcriptional machinery (BTM), when they can be located several kilobases up- or downstream from the transcription start site. It was shown via chromosome conformation capture sequencing (4C-seq) that enhancers form loops with their promoters in advance and transcription initiates through release of paused polymerase (Ghavi-Helm et al. 2014).

1.1.2 Activating and repressing mechanisms

A CRE is usually bound by multiple activators, which can interact with each other or with the BTM via different mechanisms. This results in non-linear transcriptional responses to changing activator concentrations (Carey 1998). The term "synergy'' in this context refers to a greater than multiplicative effect of several activator molecules simultaneously stimulating the BTM (Han et al. 1989, Green 2005). Another way of achieving such an effect solely on the level of DNA binding is cooperative binding, where an activator bound to the DNA facilitates the recruitment of others (Ma et al. 1996, Lebrecht et al. 2005).
Activators can interact with the core promoter over large distances and long-range repression can cause the silencing of an entire locus via the assembly of repressosomes (Courey and Jia 2001). In contrast, short-range repression or quenching is caused by repressors situated within 100-200 bp from an activator site and they function in a concentration-dependent manner (Gray et al. 1994, Gray and Levine 1996). Long- as well as short-range repressors interact with co-repressors. In the *Drosophila* embryo, Groucho and C-terminal binding protein (CtBP) serve as co-repressors for both repressing mechanisms (Parkhurst 1998, Mannervik et al. 1999, Payankaulam and Arnosti 2009). Chromatin immunoprecipitation and micrococcal nuclease mapping showed that the short-range repressor Knirps induces local changes of histone density and acetylation, whereas the long-range repressor Hairy causes widespread histone de-acetylation and subsequent inhibition of the BTM (Li and Arnosti 2011).
The functional independence of CREs is attributed to long-range activation combined with short-range repression. The facts that TFBS clustering is actually used to identify CREs and that such DNA fragments drive reporter expression independently from their original surrounding, are strong evidences already for the existence of short-range repression.

1.1.3 Complementarity versus redundancy

A gene usually owns several CREs and one might imagine that more complex expression patterns require more enhancers in order to account for all details of the endogenous gene in space and time. Interestingly, the seven-stripes of the pair-rule genes in the *Drosophila* embryo are triggered by several stripe-specific CREs, as well as a 7-stripe element able to establish the entire pattern (Schroeder et al. 2011). Only one early and one late element were identified in the regulatory region of *broad* which is expressed in a complex pattern in the *Drosophila* follicular epithelium and predetermines the formation of eggshell appendages (Fuchs et al. 2012). The early element responsible for a part of the pattern arising early and fading out soon is repressed by Mirror, whereas the late element is activated by Mirror and persists longer. These patterns seem to add up to give the endogenous expression of *broad*.
In recent years, the term shadow enhancer was invented for redundant CREs located more distal than the other proximal CREs (Hong et al. 2008). They are not essential for viability under normal conditions, but confer robustness at unusual temperatures or in certain genetic backgrounds. Later it turned out that the distance is not relevant, since a *snail* CRE right next to the promoter was shown to be the shadow enhancer, whereas the distal enhancer is required in any circumstance (Perry et al. 2010, Dunipace et al. 2011).

1.1.4 Auto-regulation

Precision and robustness in pattern formation can also be achieved by positive or negative auto-regulatory loops. The *eve* locus of *D.melanogaster* contains a 900 bp auto-activatory region triggering expression during gastrulation and germband elongation, which is abolished in *eve* mutants (Goto et al. 1989). A 200 bp minimal element from this auto-activatory region drives reduced levels (Harding et al. 1989), but multimerization of four copies reestablishes the optimal expression (Jiang et al. 1991). Mutagenesis of the *eve* binding sites showed that they are functionally relevant. Auto-activation has also been suggested for Hunchback and Krüppel based on their expression in mutant backgrounds, but mutagenesis of their binding sites was not performed (see Jaeger 2011 for review).
In zebrafish hindbrain segmentation, the FGF gradient activates Sprouty4 (Spry4), which controls the relative size of the rhombomeres in the r3-r5 region. Krox20 expression in the rhombomers 3 and 5 is driven by two initiator and one auto-regulatory element, responsible for the later amplification and maintenance of expression (Labalette et al. 2011). Spry4 fine-tunes FGF-activation of the initiator elements to ensure the appropriate timing of Krox20. In contrast, neither FGF signaling nor Spry4 influence the auto-regulatory enhancer.

1.2 Enhancer studies: state of the art

In recent years, many different types of enhancer studies, accompanied by the development of novel techniques, brought new insights into transcriptional regulation (for reviews see Lagha et al. 2012, Yáñez-Cuna et al. 2013). The availability of larger datasets for enhancers made it possible to infer context-dependent rules governing in certain subclasses. BAC recombineering allows analyzing the role of seemingly redundant CREs for robustness. Synthetic CREs were engineered to assess the relationship between variation in the content and spacing of motifs and their activity during *Drosophila* development. The literature is vast, and I can only focus on a few representative examples here.
A quantitative analysis of the short-range transcriptional repressors Giant, Knirps, Krüppel, and Snail was conducted to test activator-repressor spacing and stoichiometry, arrangement and promoter proximity, as well as activator number and affinity (Fakhouri et al. 2010). The lacZ read-out triggered by Twist and Dorsal activator sites was used to infer distance-dependent rules by a fractional occupancy-based model that uncovered a non-linear behaviour for the quencher Giant. In a similar study for mesoderm development, 63 synthetic CREs were designed to contain different kinds of TF motifs and numbers of sites with distinct relative spacing and orientation between them (Erceg et al. 2014). The motif configuration of an enhancer able to drive robust activity in the visceral mesoderm can be very flexible. In contrast, heart expression was very sensitive to subtle sequence changes.
A dataset of Bicoid (Bcd)-dependent enhancers showed that this morphogen gradient requires additional input from repressors to spatially organize target gene expression in the *Drosophila* blastoderm (Chen et al. 2012). 34 previously validated and 32 newly identified Bcd-dependent CREs were classified into three categories based on the position of their posterior boundaries. The search for a sequence-motif overrepresented in the second class and underrepresented in the third class discovered the pair-rule protein Runt (Run), which is expressed in an opposing gradient at early stages. It turned out to be necessary and sufficient to repress the second enhancer class.
The individual contributions of the three *hunchback* (*hb*) enhancers to the endogenous expression pattern were investigated with the help of BAC transgenesis and quantitative imaging methods (Perry et al. 2012). Hb is regulated by an enhancer immediately upstream of the promoter, a distal shadow enhancer, as well as a stripe enhancer, which drives a central and the posterior stripe in the *Drosophila* blastoderm. The stripe enhancer is repressed by other gap

genes and Hb itself to ensure a precise border position for the anterior Hb domain and its removal leads to cuticular defects.
Spatio-temporal *cis*-regulatory activity can be predicted without prior knowledge of TF motifs, function or expression. *In vivo* binding of five TFs at five time points was used to predict mesodermal CREs and subsequently, their expression was characterized and categorized into five tissue-classes (Zinzen et al. 2009). The binding profiles of the CREs were used to train a support vector machine (SVM) for each tissue-class to decide if an enhancer is a member of the class or not. The trained SVM was capable of correctly predicting CREs not included in the training set and unexpected plasticity in TF occupancy for CREs from the same tissue was observed.
The Vienna Tiles (VT) library comprises over 7000 transgenic flies with reporter constructs containing *D.melanogaster* enhancer candidates. Their expression patterns were characterized at six time intervals of embryogenesis and 46% of the candidates are active *in vivo* (Kvon et al. 2014). They were assigned to target genes via manual comparison with the mRNA expression of the target. 36% of all enhancers are intragenic and 88% are located in the vicinity of their targets. 79% of the intragenic enhancers regulate their host genes. A SVM was trained to identify *cis*-regulatory motifs that are predictive and required for class-specific enhancer activity. The results were validated experimentally via mutation of the GAGA-, GATA- and Ttk-like motifs in several enhancers.
STARR-seq (self-transcribing-active-regulatory-region-sequencing) allows the identification of CREs in a direct and quantitative manner in entire genomes (Arnold et al. 2013). It consists of cloning randomly sheared genomic DNA downstream of a minimal promoter, followed by a polyadenylation site, such that active CREs transcribe themselves, and the abundance of their own RNA is a measure for their activity. After transfection of this genome-wide reporter library into *Drosophila* S2 cells and paired-end sequencing of the selected RNAs, the fragments are mapped to the genome and their enrichment is quantified.

1.3 The model system: the gap gene network acting in the *Drosophila* blastoderm

The segmentation gene network of *Drosophila melanogaster* is definitely a convenient model system to study eukaryotic transcriptional regulation for the following reasons: (1) The blastoderm is the simplest biological entity without any growth or tissue movements. (2) Thanks to numerous mutant experiments, most of the interactions in the network are well-understood. (3) Resources for spatio-temporal gene expression data are publicly available (Poustelnikova et al. 2004). (4) The regulation is highly combinatorial and almost without post-transcriptional contributions.

1.3.1 *Drosophila* embryogenesis

The *Drosophila* life cycle is temperature-dependent and comprises the stages embryo, 1st, 2nd and 3rd instar larvae, pupae and the adult fly (Lawrence 1992). The process from egg deposition until hatching of the larvae is called embryogenesis and lasts approximately 24h (Campos-Ortega and Hartenstein 1985). The embryo starts as a syncytium without cell membranes, which undergoes 13 mitotic divisions within 3h without tissue growth. The blastoderm is the cortical layer formed by the nuclei between cleavage cycle 10 (C10) and C14A (Figure 1).

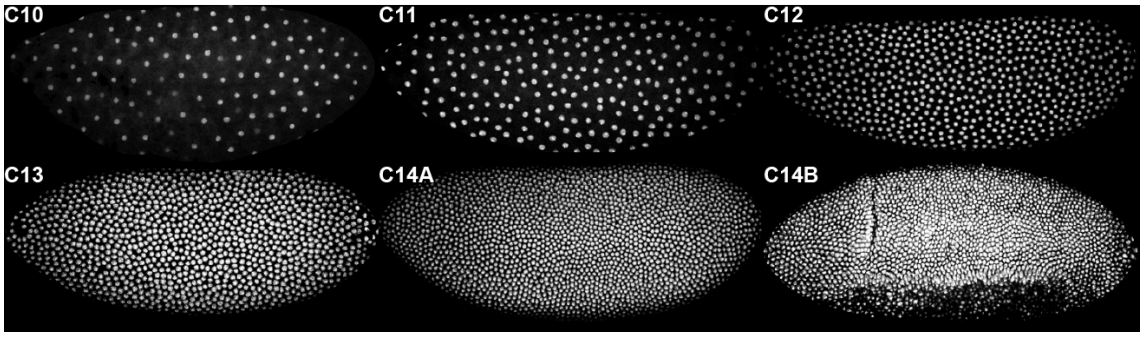


Figure 1: The blastoderm stage.
Distribution of the nuclei during the cleavage cycles 10 to 14B in the *Drosophila* embryo. The blastoderm stage comprises C10 to C14A. In C14B gastrulation starts and the head and ventral furrows form. The nuclei were stained with Hoechst34580 and imaged on a confocal microscope. All embryo images in this thesis are oriented with anterior to the left and dorsal up, unless otherwise stated.

The first nine cycles are short, approx. 10 min, but their length increases to 15-20 min in C10 until C13 and finally, C14A lasts 50 min (Foe and Alberts 1983, Foe 1989). At the beginning, the nuclei are localized in the middle of the embryo and then migrate to the surface between cycle 7 and 10. We further subdivide C14A into eight time classes (T1-T8) based on the Even-skipped (Eve) protein expression pattern and the membrane morphology (Surkova, Kosman, et al. 2008). The cell membranes invaginate in-between the nuclei during cellularization and in C14B gastrulation rearranges the embryo into three layers called ectoderm, endoderm and mesoderm. Subsequently, other tissue movements such as germ-band extension, germ-band retraction, head involution and dorsal closure restructure the embryo further.

1.3.2 Segmentation genes

The body plan of insects is pre-patterned in the early embryo through segment determination, which delimits para-segment boundaries. While most insects add segments sequentially via growth (short germ-band development), *Drosophila* undergoes long germ-band development with all segments being determined simultaneously in the blastoderm (Sanders 1976). Most of the genes involved in this process encode TFs and were identified by saturating mutagenesis screens (Nüsslein-Volhard and Wieschaus 1980, Jürgens et al. 1984, Nüsslein-Volhard et al. 1984, 1987, Wieschaus et al. 1984, Schüpbach and Wieschaus 1986). They were categorized into maternal, gap, pair-rule and segment polarity genes based on the mutant phenotype of the larvae. The mother deposits the mRNA of the maternal coordinate genes into the egg. They form protein gradients and hence establish asymmetry and polarity. The gap genes are expressed as one or two domains and their mutation abolishes several adjacent segments in the larvae. Pair-rule genes are expressed in seven stripes and their phenotypes manifest in the deletion of alternating segments. Segment-polarity genes form 14 thin stripes after gastrulation. Their mutants maintain the same number of segments, but a part of each segment is deleted and replaced by a mirror-image duplication of the remaining part and hence, shows reversed polarity (Johnston and Nüsslein-Volhard 1992). At late blastoderm stage, homeobox (Hox) gene expression starts, which determines the segment identity of the adult fly.
The segmentation gene network (Figure 2) is arranged in a hierarchical way with higher layers regulating lower ones, but not *vice versa* (Akam 1987). *Bicoid* (*bcd*) mRNA is placed at the anterior pole and its protein diffuses towards the posterior (Berleth et al. 1988, Little et al. 2011), thereby translationally repressing the initially ubiquitous *caudal* (*cad*) mRNA (Mlodzik and Gehring 1987). *Hunchback* (*hb*) is also contributed maternally and repressed by Nanos (Nos) in the posterior (Tautz 1988). Additionally it is transcribed from a zygotic promoter and belongs to the gap genes (Tautz et al. 1987).

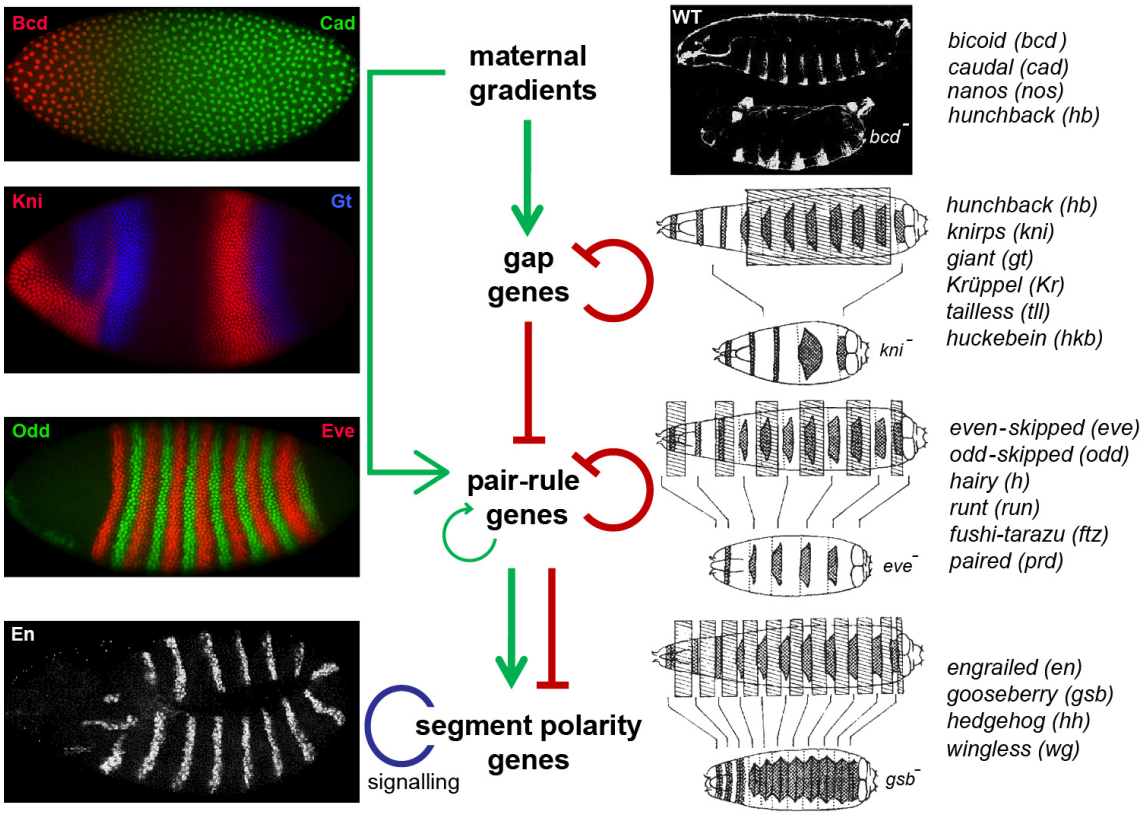


Figure 2: The segmentation gene hierarchy.
Each horizontal panel represents one gene category, showing an embryonic protein expression pattern of one or two representatives and a mutant larval phenotype. Green arrows represent activation and red T-bar connectors indicate repression. Signaling pathways influence the segment polarity genes (blue circle). Embryo images were taken from FlyEx[1] (Pisarev et al. 2009), except of the picture of En, which was taken by Carlos Vanario-Alonso. Larvae cuticle images taken from (Gilbert 2000) and larvae drawings, which indicate the deleted segments, from Nüsslein-Volhard and Wieschaus 1980.

The maternal gradients activate the gap and the pair-rule genes, whereas the gap genes repress the pair-rule and the Hox genes. The gap genes are further subdivided into trunk gap genes (*Krüppel*, *knirps*, *giant*), terminal gap genes (*tailless*, *huckebein*) and head-gap genes (*orthodenticle*, *empty spiracles, buttonhead, sloppy paired*). The terminal gap genes are not regulated by other zygotic segmentation genes (Broenner and Jaeckle 1991). The pair-rule genes can activate or repress the segment-polarity and the Hox genes. The cross-regulation within the same class of segmentation genes leads to further refinement of their initial patterns. In the case of the pair-rule genes, most of these interactions are repressing, apart from the activation of *odd-skipped* (*odd*) by *fushi tarazu* (*ftz*) (Schroeder et al. 2011).

1.3.3 The gap gene network

The trunk gap genes emerge as broad domains in C10-C12 and intensify and refine over time (Figure 3). Hunchback (Hb) is expressed in the anterior half of the embryo and in a posterior stripe close to the pole. Krüppel (Kr) forms a stripe in the middle of the embryo and a weaker one in the anterior at late stages. Knirps (Kni) has a prominent stripe in the posterior that overlaps with Kr and a very thin one in the anterior. Additionally, it is expressed at the anterior tip reaching ventrally until it fuses with the thin stripe. Giant (Gt) protein expression starts as two broad domains in the anterior and in the posterior. The anterior domain refines into two stripes and at late stages also expression at the anterior tip becomes visible.

[1] http://urchin.spbcas.ru/flyex

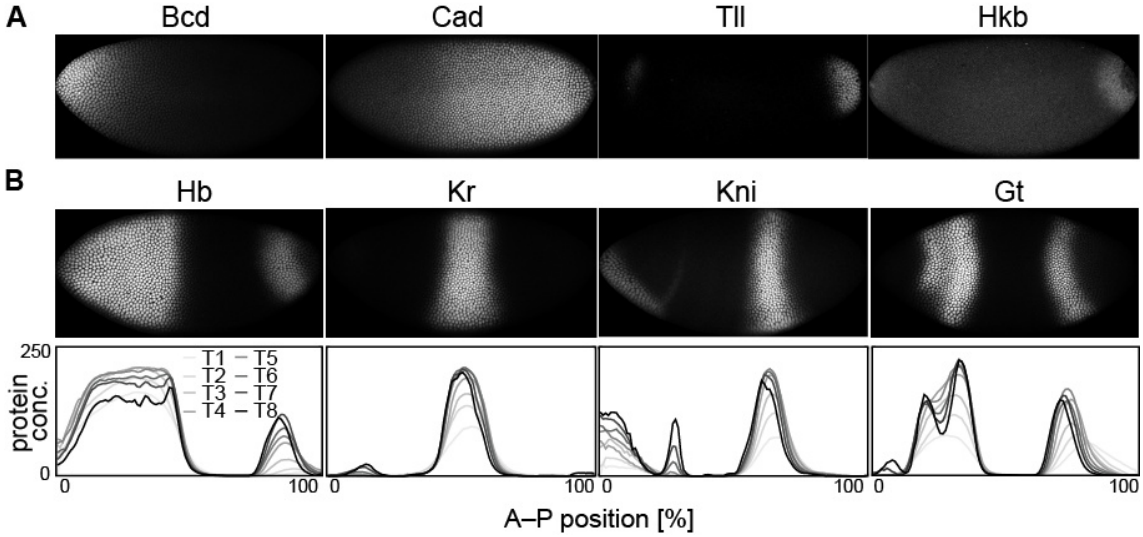


Figure 3: Expression patterns of genes involved in the gap gene network.
(A) Protein expression of the maternal and terminal genes at time class T5 in C14A (images from the FlyEx database). **(B)** Protein expression of each of the four trunk gap genes at time class T5 in C14A and quantified protein concentration (in arbitrary units) over all eight time classes of C14A (starting with light grey and ending with black). The anterior pole is defined as 0% anterior-posterior (A-P) position. Adapted from Jaeger 2011.

The gap gene network receives its initial conditions from the maternal protein gradients of Bcd and Cad (Figure 4). Cross-repression between the gap genes leads to the sharpening of the initial broad domains. In particular, we observe strong repression between the mutually exclusive domains of Hb and Kni, as well as Kr and Gt. These interactions are sometimes referred to as "alternating cushions" mechanism of gap gene regulation (Jaeger 2011). Additionally, weaker repression between overlapping domains operates from the more posterior towards the more anterior domain. This appears like a directional harmonica starting from the posterior Hb domain towards the posterior Gt stripe and going further via Kni and Kr in order to end at the anterior Hb. As a net result of this asymmetric behavior, all posterior gap domains shift towards the anterior over time (Jaeger et al. 2004).

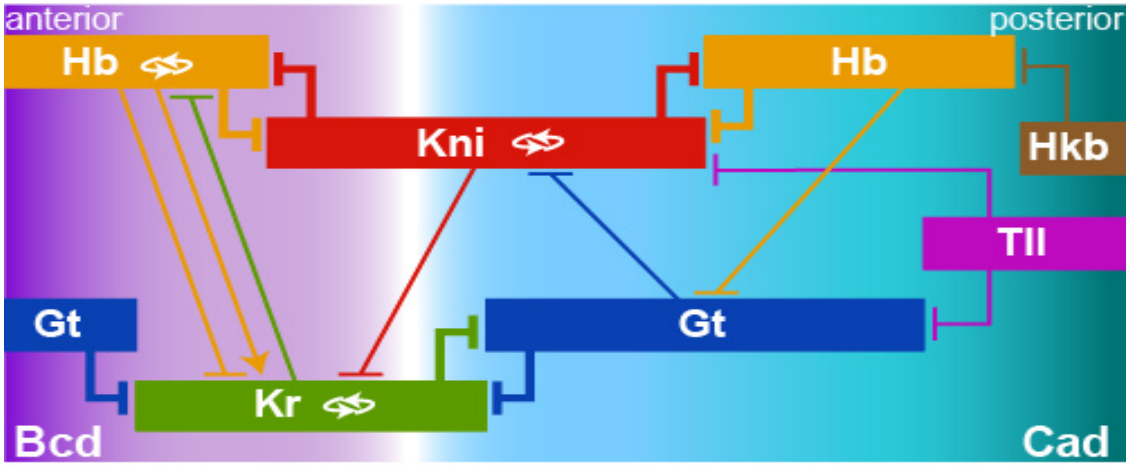


Figure 4: The gap gene network.
Schematic overview of the activating and repressing interactions within the network. The position of the gap domains is shown along the A–P axis (colored boxes) with a focus on the posterior half of the embryo. The background gradients indicate the predominant maternal activators Bicoid and Caudal. Circular arrows indicate auto-activation. Thick T-bar connectors represent major repression between mutual exclusive domains. Thinner T-bar connectors show weaker repressive interactions. Modified from (Ashyraliyev et al. 2009).

The terminal system acts via Tll onto posterior Gt and Kni, and Hkb represses the posterior Hb domain. Finally, there is some evidence for auto-activation based on observations in mutants or over-expression experiments, but this is probably indirect (Jaeger 2011).

1.4 The gap gene *giant*

1.4.1 Discovery of *giant* and its mutant phenotype

gt was discovered in 1925 (Bridges and Gabritschevsky 1928) and named based on its phenotype like most *Drosophila* genes. Viable *gt* mutants manifest during 3rd instar in an extended growth period leading to huge larvae and a delay of metamorphosis[2]. This is probably the consequence of lower ecdysone levels and hence, defects in the regulation of DNA synthesis yielding cells with twice as much DNA than normal (Kaufman et al. 1973, Schwartz et al. 1984, Narachi and Boyd 1985).
In the first mutagenesis screen (Nüsslein-Volhard and Wieschaus 1980), only *Kr*, *kni* and *hb* were categorized as gap genes, whereas *gt* had to wait to be discussed in the follow-up paper for defects in the head and in the abdominal segments A5-A7 (Wieschaus et al. 1984), but was still denied the title "gap gene". Reason therefor was its slightly more complex pattern and hence unconventional phenotype with two small gaps (for detailed descriptions see (Gergen and Wieschaus 1986, Petschek et al. 1987, Mohler et al. 1989)). Not until ten years after the mutagenesis screen, it was finally nominated a *bona fide* gap gene, when more details about its role in the segmentation gene network were discovered (Reinitz and Levine 1990, Eldon and Pirrotta 1991, Kraut and Levine 1991a, 1991b).

1.4.2 The transcription factor Giant and its protein domains

gt[3] is located on the X chromosome at cytological position 3A3 (Mohler et al. 1989). The ORF is 1780nt long and contains a 75nt intron at sequence position 164 (Capovilla et al. 1992). It encodes a transcription factor of 448 amino acids[4] with a basic leucine zipper (b-ZIP) (Capovilla et al. 1992). The b-ZIP proteins (Vinson et al. 1989) contain a sequence specific DNA-binding region constituted of basic residues such as arginine (R) and lysine (K). Nearby this domain, several leucines (L) are positioned with a spacing of exactly 6 aa in-between them. This heptad repeat leads to the formation of an alpha helix and provokes dimerization via hydrophobic interactions between the leucines of the two monomers. Subsequently, a coiled coil configuration arises and the amino acids of the two basic stretches form hydrogen bonds with the DNA bases of the major groove. In the case of *gt*, the first of the five leucines was substituted with isoleucine (I) and the last one with phenylalanine (F), which are hydrophobic residues as well. Gt protein becomes phosphorylated and posttranslational modification is required for repression (Capovilla et al. 1992). It is a short-range repressor containing the evolutionarily conserved motif VLDLSRR at residues 98-104, which partially corresponds to the dCtBP consensus motif PxDLSxR/K/H (Nibu and Levine 2001, Strunk et al. 2001). According to Nibu *et al.*, the minimal repression domain encompasses the residues 60-133, whereas Strunk *et al.* defined it as aa 89-205. Nevertheless, the experiments of both labs show that the CtBP co-factor is required in some but not all cases, depending on the enhancer-context.

[2] *Giant* (*gt*) should not be confused with the gene *lethal (2) giant larvae* (*l(2)gl*), which exhibits a similar larvae phenotype.
[3] CG7952, FBgn0001150
[4] GenBank X61148.1 and AAF45780

1.4.3 Expression of *giant*

gt was first cloned by Mohler et al. (1989) and its mRNA was visualized by *in-situ* hybridization. The mRNA is expressed in two broad domains (Figure 5A), reaching from 60-82% egg length (EL) and 0-33% EL in cycle 12 (Mohler et al. 1989). Both domains intensify and retract from the poles and their boundaries become steeper during the first half of C14A (Figure 5B). The anterior domain then splits into two stripes at 75-83 and 62-70% EL and finally also expression at the anterior tip arises, giving four lateral stripes in total (Figure 5C). The first and the second stripe are not covering the ventral regions and stripe 4 shifts to the anterior and disappears at gastrulation (Figure 5F). An additional small ventral stripe becomes visible at later stages (Figure 5D and E).

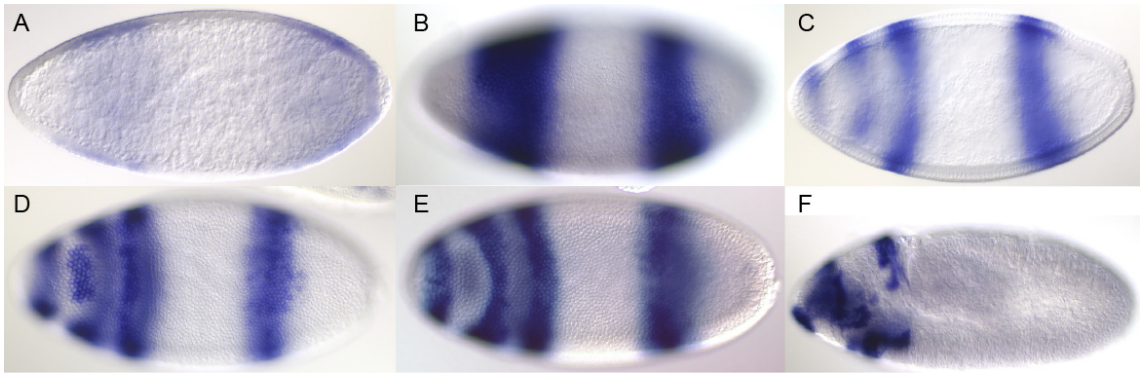


Figure 5: *giant* mRNA expression.
Enzymatic in-situ hybridization to *gt* mRNA from the Berkeley Drosophila Genome Project[5] (Tomancak et al. 2002, 2007), classified as developmental stages 1-3 (**A**), 4-6 (**B-E**) and 11-12 (**F**). Anterior is to the left in all images. (**B, C and F**) show a lateral, (**D**) a ventral and (**E**) a dorsal view.

Gt is expressed exclusively in the embryo without any traces, neither in the larvae nor in the adult (FlyAtlas Anatomical Expression Data[6]). The protein is localized in the nuclei (Kraut and Levine 1991a) and appears at C12 (Eldon and Pirrotta 1991). Protein and mRNA expression are very similar, but no detailed comparison has been reported yet. A protein dataset of high temporal and spatial resolution is available on FlyEx (Pisarev et al. 2009), but quantification of gap gene mRNA has only been published for the very early stages C10–C13 (Jaeger et al. 2007). In this thesis, a detailed comparison of protein with mRNA expression over C14A was conducted (see Results and discussion 4.3.1) and used to fit a model of post-transcriptional regulation (Becker et al. 2013).

1.4.4 Regulation of *giant*

Extensive genetic investigations have revealed how the gap genes are regulated (for review see Jaeger 2011). The interpretation of the results from the mutants is not always straightforward, especially in gene networks with extensive cross-regulation. In particular, de-repression of a repressor can cause indirect effects that might be misinterpreted as activation.

Regulation of the anterior

The two *gt* domains start to form at opposite ends in the embryo and hence are situated in different contexts receiving distinct regulatory inputs. The anterior *gt* domain gets abolished in *bcd* mutants (Eldon and Pirrotta 1991, Kraut and Levine 1991a), but is still present in maternal and zygotic *cad* backgrounds. A *bcd* and *cad* double mutant does not show any *gt* expression at all (Rivera-Pomar et al. 1995). The anterior *gt* stripe 3 is slightly shifted toward the anterior in zygotic *hb* mutants (Eldon and Pirrotta 1991, Kraut and Levine 1991a, Yu and Small 2008), which is probably caused by Kr expanding to the anterior (Jaeckle et al. 1986). Anterior *gt*

[5] http://insitu.fruitfly.org
[6] www.flyatlas.org

expression persists (Kraut and Levine 1991b) in embryos overexpressing Hb protein ubiquitously at high levels via a heat shock-construct (hs-*hb*).
Kr and Gt are always complementary, even in mutant combinations such as *vasa torso exuperantia* (*vas tor exu*) (Struhl et al. 1992), but the two Gt domains do not meet in the middle in *Kr* mutants. The direct evidence for repression of anterior Gt by Kr is ambiguous. Kraut *et al.* claimed that anterior Gt expands towards the center in *Kr* mutants, but the depicted embryo is not lateral (Kraut and Levine 1991a, 1991b), and in fact two other papers show that anterior Gt is normal (Mohler et al. 1989, Surkova et al. 2013).
Anterior Gt is not affected in *kni* (Mohler et al. 1989) and *tll* mutants (Eldon and Pirrotta 1991). In *hkb tll* double mutants, stripe 1 is missing and stripes 2 and 3 are shifted to the anterior (Eldon and Pirrotta 1991). Entire *gt* gets removed from embryos expressing Tll via a heat shock promoter (hs-*tll*) (Kraut and Levine 1991b). In *tor* or *torso-like* (tsl) mutants, the anterior tip is missing and stripe 2 is shifted anteriorly (Eldon and Pirrotta 1991, Kraut and Levine 1991a).
Anterior *gt* is affected by several head gap genes, but not by *buttonhead (btd)*. In *orthodenticle (otd)* mutants, stripe 1 expands towards the posterior and stripe 2 is shifted posteriorly (Eldon and Pirrotta 1991). Stripe 2 and 3 fail to split and refine in *empty spiracles (ems)* mutants (Eldon and Pirrotta 1991) and in germ line clones of the putative co-repressor *brakeless* (*bks*), the separation of the anterior *gt* domain into two stripes is delayed (Haecker et al. 2007). Capicua (Cic) participates in establishing the posterior boundary of the Bcd-responsive anterior *gt* tip, whereas the anterior *gt* domain is not affected in *cic* mutants (Löhr et al. 2009).
In summary, the anterior Gt domain is independent of Hb, Kr and Kni. It is activated by Bcd and the posterior boundary might be set by a Bcd threshold. Tll, Hkb, Cic, Ems, as well as the dorso-ventral (DV) system, are involved in refining the anterior into a pattern of three stripes.

Regulation of the posterior
Posterior *gt* is shifted towards the anterior in *bcd* mutants (Eldon and Pirrotta 1991, Kraut and Levine 1991a) and absent in maternal and zygotic *cad* mutants (Rivera-Pomar et al. 1995), as well as in hs-*hb* embryos (Kraut and Levine 1991b). It is also missing in *nos Kr* double mutants (Kraut and Levine 1991a) and in *nos* or *oskar* (*osk)* mutants, in the latter cases probably indirectly via Hb (Eldon and Pirrotta 1991). The posterior *gt* domain is derepressed in zygotic as well as maternal & zygotic *hb* mutants (Eldon and Pirrotta 1991, Struhl et al. 1992).
In *Kr* mutants, posterior *gt* expands towards the center, but the expression level is reduced (Kraut and Levine 1991a, 1991b). Premature reduction of posterior *gt* was also observed in *kni* mutants (Eldon and Pirrotta 1991, Kraut and Levine 1991a). In *tll* mutants, posterior *gt* fails to retract from the pole (Eldon and Pirrotta 1991, Kraut and Levine 1991a) and there is no *gt* expression at all in hs-*tll* embryos (Kraut and Levine 1991b). Posterior *gt* expands until the pole in *tor* or *tll hkb* double mutants (Broenner and Jaeckle 1991, Eldon and Pirrotta 1991) and becomes abolished in hs-*hkb* embryos (Broenner et al. 1994).
Reduction of posterior *gt* was observed in *wollknäuel* (*wol*) and in 25% of the embryos in *bks* germ line clones (Haecker et al. 2007). Gt recovers at later stages in the *wol* mutants and it was suggested that this maternal gene affects either Cad protein stability or the efficiency of Cad translation (Haecker et al. 2008). Finally, Gt auto-activation was suggested, because the stripes do not intensify during cycle 14 in a *gt* null background (Eldon and Pirrotta 1991).
In summary, the posterior domain is activated by Cad and repressed by Hb, Kr, Tll and maybe Hkb. The other observed effects are most likely indirect and the DV-system is not involved in the regulation of posterior Gt.

Early versus late regulation and the role of Hb
Based on modelling results, Jaeger et al. (2007) claimed that Hb sets the anterior boundary of the posterior *gt* domain at C11 and C12 (Figure 6). In C13, Kr protein becomes detectable in the central region of the embryo and takes over this role. This leads to the question how strong early Hb repression is overcome in the anterior domain at C11 and C12. Their models required equilibrium between Bcd activation and Hb repression to correctly position the gap domains,

but that was not achieved for early *gt* expression. Local neutralization of Hb repression was needed in the anterior, which was modeled by two different approaches: spatially specific co-regulators were taken into account, meaning that Hb is considered as an activator in the presence of Bcd in the anterior and as a repressor in the presence of Cad in the posterior. In the second approach, both domains were treated independently, as if controlled by CREs implementing distinct regulatory mechanisms. However, it remained unclear how the balance between Bcd activation and Hb repression could be achieved at the molecular level and how the information from separate CREs leads to the expression pattern of the entire, endogenous *gt* gene.

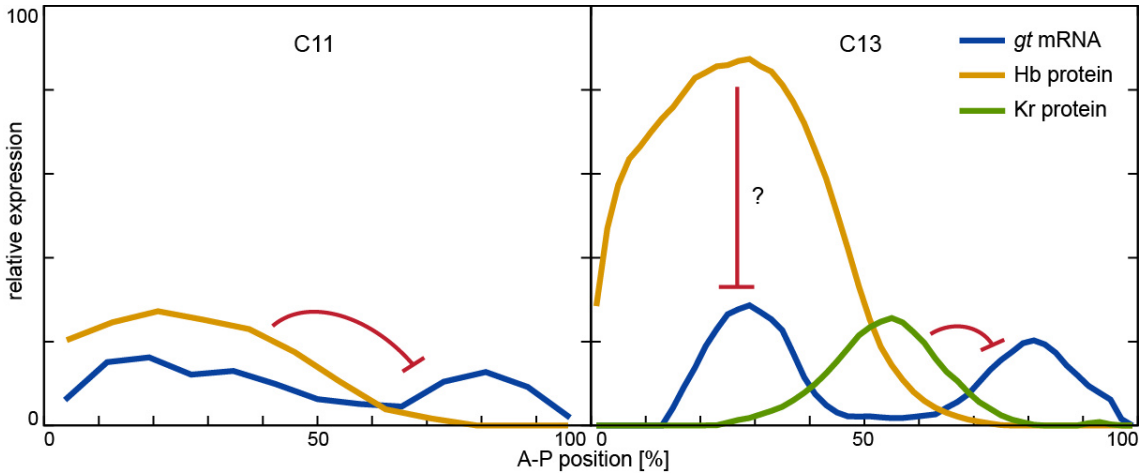


Figure 6: Regulation of *giant* mRNA by Hunchback protein at early stages.
Hb sets the anterior boundary of the posterior *gt* domain at C11. At C13, Kr takes over this role, but it is not clear how the strong repression from Hb is neutralized in the anterior. *gt* mRNA data were taken from Jaeger 2007 and Hb and Kr protein from the FlyEx database. Kr protein was not detected at C11. Red T-bar connectors represent repression.

1.5 Prediction and evaluation of *giant* CREs

30 years ago, CREs were identified via labor-intensive "promoter bashing" which consists of cutting the region upstream of the transcription start site of the gene of interest into fragments and testing these *in vivo*. Nowadays, *in silico* approaches allow for genome-wide predictions. They primarily exploit the clustering of transcription factor binding sites (TFBS) and some of them secondarily take advantage of evolutionary conservation to refine the CRE boundaries. For a detailed description of the algorithms used to identify the *gt* CREs, see section 1.5.3.

1.5.1 The CREs of *giant*

In total, eight *gt* CREs were annotated by different groups and named either based on their position relative to the transcription start site, with the number of stripes they drive or arbitrary (Figure 7, Table 1). They are about 1-2 kb long and lie within 10 kb upstream from the *gt* gene, except of the distal downstream enhancer gt+36. Some partially overlap and trigger the same domain. For example, the smallest element gt1 constitutes only the middle part of the largest enhancer gt-6.

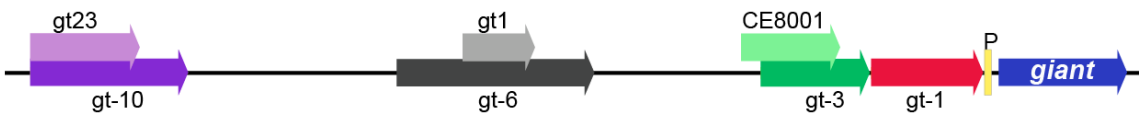


Figure 7: Genomic locus of *giant* with its CREs.
Indicated are the previously identified CREs, except gt+36, as well as the promoter (P) and the *gt* gene (blue).

Two overlapping CREs were identified for each subpart of the *gt* expression pattern, namely the anterior tip (gt1, gt-6), the anterior domain (gt23, gt-10) and the posterior domain (CE8001, gt-3). Two further elements are capable of driving the anterior and the posterior domain: one of them (gt-1) is located right in front of the promoter, whereas the other one lies downstream (gt+36).

CRE	pattern	length [bp]	reference	method
gt-1	both domains	1240	Schroeder 2004	Ahab
gt-3	posterior	1210	Schroeder 2004	Ahab
gt-6	head tip	2183	Schroeder 2004	Ahab
gt-10	anterior	1746	Schroeder 2004	Ahab
CE8001	posterior	1099	Berman 2002	cis-analyst
gt1	head tip	805	Ochoa-Espinosa 2005	bcd cluster
gt23	anterior	1214	Ochoa-Espinosa 2005	bcd cluster
gt+36	both domains, blurry	1756	Perry 2011	not mentioned

Table 1: Summary of the eight *giant* CREs previously identified.
See section 1.5.3 for a detailed description of the methods used for their identification and the Appendix for their sequences.

Some general considerations about these CREs have to be kept in mind. They were only tested qualitatively and no spatio-temporal expression data is available (Figure 8A). The computationally identified DNA fragments might not always drive exactly the endogenous pattern. It is not trivial to define their boundaries and hence, they might not represent the minimal stripe elements or even worse, they could be missing important TFBS. An evolutionary filter can help to make the decisions where they actually start and end. Interestingly, gt-1 is separated by only 6 bp from the promoter and from gt-3.

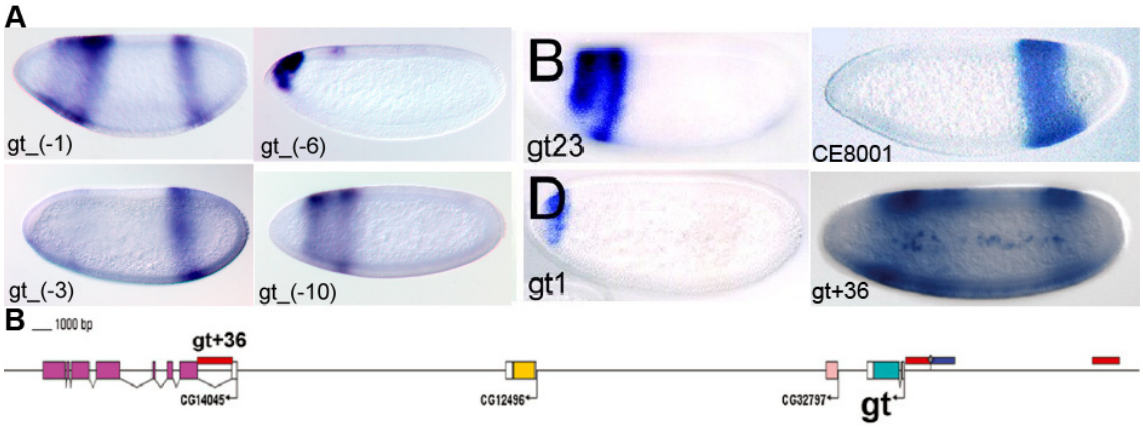


Figure 8: Expression driven by the *giant* CREs.
(A) Expression of lacZ mRNA driven by the indicated CREs. Perry 2011 stated that gt+36 shows patchy, stochastic expression. Images were taken from Schroeder 2004, Ochoa-Espinosa 2005, Berman 2002 and Perry 2011. **(B)** The distal downstream element gt+36 is located in an intron of a CG gene. Taken from Perry 2011.

1.5.2 Identification of transcription factor binding sites

Most approaches for CRE detection rely on searching the DNA for TFBS with positional weight matrices (PWM). It is important to be aware that these matrices can derive from completely different methods, such as footprints, Selex and bacterial-one-hybrid (B1H) screens.
Initially, TFBS were identified via DNA footprinting (Galas and Schmitz 1978). A DNA fragment with potential binding sites is radioactively labeled at one end and the purified protein of interest is added. Treatment with DNase and subsequent polyacrylamide gel electrophoresis will yield a fragmentation pattern, which is visualized by radiography. If a site was bound by the

protein, the DNA is protected against DNase cleavage at this position, which will be reflected by a break in the pattern compared to the control DNA without protein.
Another *in vitro* method is Selex (**S**ystematic **E**volution of **L**igands by **EX**ponential enrichment) (Oliphant et al. 1989, Tuerk and Gold 1990). An oligonucleotide library synthesized from the genome is applied onto a column containing the protein of interest. The bound DNA fragments are eluted and amplified by PCR to yield the pool of DNA for the next selection cycle with higher stringency.
A more recent *in vivo* technique is the B1H screen (Meng et al. 2005, Noyes et al. 2008). It requires a bait plasmid containing the TF fused to the ω-subunit of the bacterial RNA polymerase and a library of prey plasmids carrying randomized 28 bp nucleotides in front of a weak promoter and the HIS3 and URA3 genes. After transformation into bacteria lacking these genes as well as the ω-subunit, only those clones which received a DNA fragment able to bind the TF, will survive on minimal media without histidine. The ω-knockout strains are viable and the polymerase can be actively recruited by the subunit fused to the DNA-bound TF. The URA3 serves as a negative marker in this case to eliminate false-positives via growing on a media supplemented with 5-FOA, which is converted into a toxic compound by the uracil biosynthesis pathway.
The sequences derived from the above mentioned methods need to be aligned and trimmed with a pattern discovery tool in order to identify the core motif. The non-coding DNA of the yeast genome could simply be scanned for overrepresented words (Bussemaker et al. 2000), but higher eukaryotic transcription is more complex. Instead of the consensus sequences, a PWM is constructed and used to search for binding sites (Stormo et al. 1982). The observed frequency of each of the four bases at each position is divided by the background frequency of the genome and the log likelihood ratio is calculated. This is based on the assumption of independence between nucleotides, which in general is sufficient, although might not always be the case.

1.5.3 Prediction of *cis*-regulatory elements

The following four studies discovered the *gt* enhancers and numeral other CREs active in the *Drosophila* embryo.
Berman *et al.* used heterotypic cluster analysis of the factors Bcd, Cad, Hb, Kr and Kni with PWMs constructed from DNase footprints (Berman et al. 2002). Binding site search in the genome was performed with PATSER (Hertz and Stormo 1999) using a certain cut-off value for each factor. They developed the program CIS-ANALYST to scan non-coding DNA for clusters of TFBS with the requirement to contain a minimum amount of sites in a sliding window of 700 bp. They took this approach even further with eCIS-ANALYST by including the comparison with the *D. pseudoobscura* genome (Berman et al. 2004) and predicted several hundred new CREs. Considering conservation of clusters in two fly species allowed discarding false-positive hits found in the *D.melanogaster* genome and hence increased specificity.
The three combinable algorithms Argos, Gibbs Sampler and Ahab, were developed by Rajewsky *et al.* for detecting different kinds of regulatory information (Rajewsky et al. 2002). Argos requires only the genome as input and without any training data generates a score for clusters of overrepresented motifs within a certain window size. It yielded about one putative element every 5 kb on average. With the Gibbs sampler, repeated motifs can be inferred from known CREs. The *D.melanogaster* genome was scanned with Ahab using PWMs and a background model to account for local variations in sequence composition and degeneracy of motifs. In the follow-up article (Schroeder et al. 2004), 16 out of the 32 novel elements predicted by Ahab were evaluated *in vivo* and 13 gave faithful expression patterns. In contrast to other methods, no predefined factor-dependent cut-offs were applied, permitting for the detection of clusters of weak sites, which might be relevant *in vivo*.

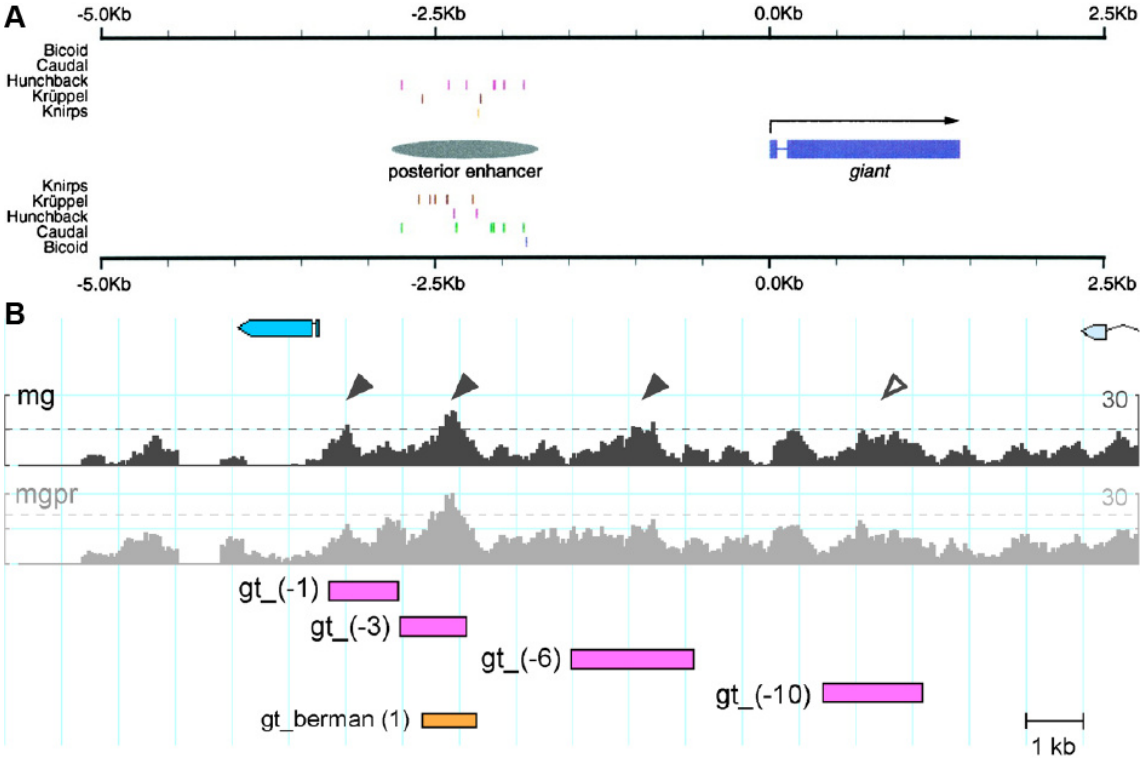


Figure 9: Prediction of CREs based on TFBS clustering.
(A) Taken from Berman 2002: Cluster of binding sites found between 2.9 kb and 1.8 kb upstream of *gt*. **(B)** Taken from Schroeder 2004: Free energy profiles for two Ahab runs (mg and mgpr). The free energy cut-offs are marked by dotted lines; mg run predictions with scores greater than 15 are marked by black arrowheads, tested sub-threshold peaks with scores below 15 by open arrowheads. The transcribed region of the locus is marked in blue, the experimentally tested genomic regions are marked by pink bars and named according to distance from transcription start site to the middle of the enhancer, and the previously known module is marked by an orange bar.

Lifanov *et al.* analysed homotypic clustering in over 60 known CREs and claimed that this is a wide-spread phenomenon in developmental genes (Lifanov et al. 2003). They examined the *gt* locus for Bcd binding and predicted a region later shown to drive the anterior tip (Schroeder et al. 2004, Ochoa-Espinosa et al. 2005). Additionally, they discovered a cluster immediately down-stream of the coding sequence but with a PWM score right at the cut-off value (Figure 10). This fragment was never picked up in any other study and was never tested *in vivo*.

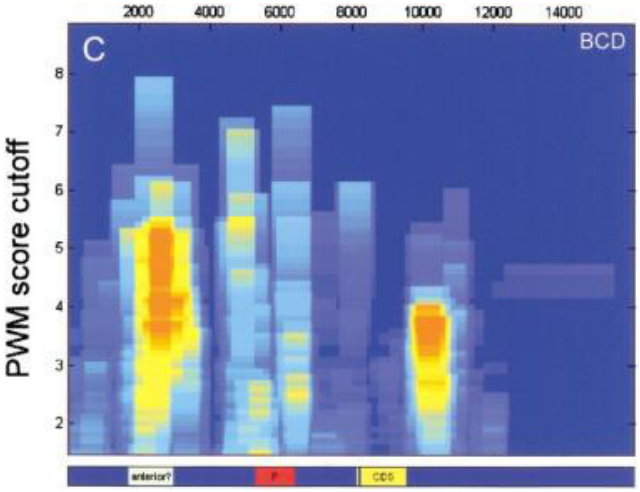


Figure 10: Putative downstream CRE of *gt*. Taken from Lifanov *et al.* 2003. The color intensity scale represents statistical significance of constitutive clusters. The Bicoid cluster located downstream of the *gt* CDS (at position ~10000) contains low-affinity sites. Its PWM score is below the established optimal cut-off of 4.2. An anterior and a posterior CRE upstream of the *gt* CDS were confirmed by other groups.

Ochoa-Espinosa *et al.* also relied on homotypic clustering of Bcd (Ochoa-Espinosa et al. 2005). They scanned a training-set of known Bcd-dependent enhancers and identified the following common features: the CREs harbour at least 6 Bcd sites above the threshold in a 550 bp stretch with at least one of them being a high-affinity site and at least 2 well scoring sites within 200bp. These parameters were used to search 20 kb up- and downstream from the CDS of 10 target genes, including *gt*. In order to determine the fragment boundary of the newly identified enhancers, an alignment with the *D. pseudoobscura* genome using the VISTA browser (Couronne et al. 2003) was performed. The start and end of the element were set where the conservation identity dropped below 50%. With this combination of Bcd-cluster analysis and evolutionary conservation, seven previously unknown CREs were discovered and validated. Additionally, they wanted to investigate the mechanisms of A-P patterning by the Bcd morphogen gradient. However, no correlation between the estimated binding strength of the Bcd cluster and the posterior border positions of the CRE expression patterns was observed.

1.6 Modeling transcriptional regulation

1.6.1 Reverse-engineering

Traditional genetics relies on examining expression patterns in mutants, which definitely contributed to our knowledge about developmental systems (Figure 11). This bottom-up approach also has its limitations. Such mutants represent a disturbed system with indirect effects, which can be difficult to interpret. For example, in *cad* mutants the posterior *gt* domain is absent (Rivera-Pomar et al. 1995), which in this case easily leads to the conclusion that Cad activates this domain. In *bcd* mutants (Eldon and Pirrotta 1991, Kraut and Levine 1991a), the anterior *gt* domain is absent and additionally, the posterior is shifted towards the anterior, which is indirectly caused by Kr de-repression.

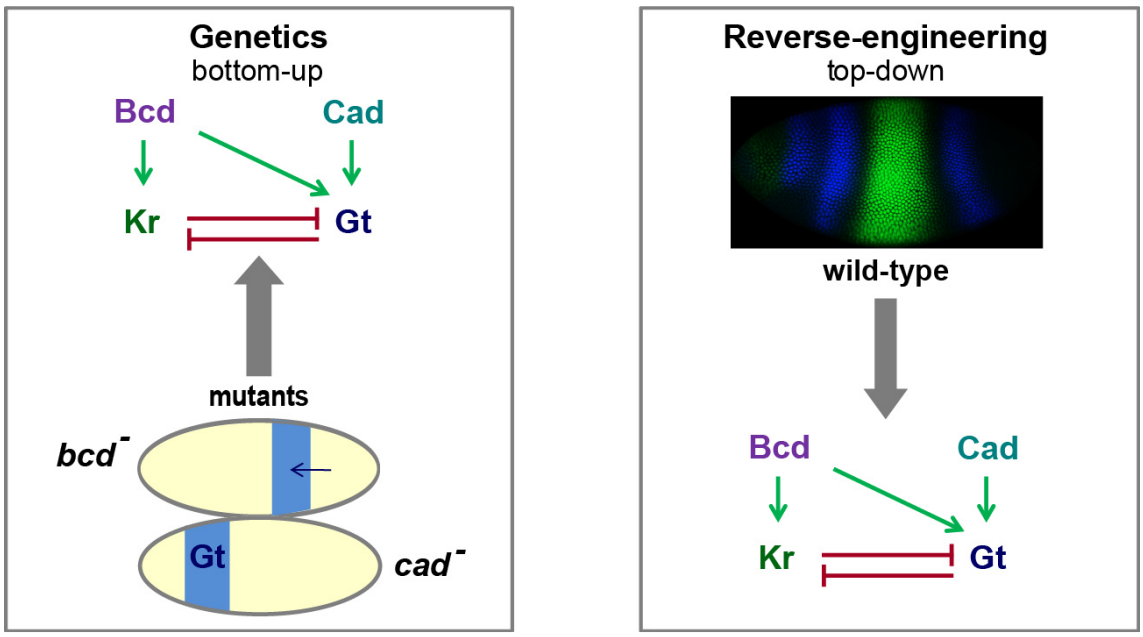


Figure 11: The concept of reverse-engineering.
The genetics approach tries to infer information about regulatory contributions from mutant experiments. Shown in this example is the expression of Gt in *bcd* and *cad* mutant embryos. The reverse-engineering approach extracts regulatory contributions after collecting expression data in wild-type embryos. The image of Gt and Kr protein expression was taken from FlyEx.

In contrast to these mutant studies, the reverse engineering or top-down approach extracts information about regulatory interactions from quantitative expression data collected in wild-type embryos with the help of a mathematical model (Jaeger and Monk 2010, Ashworth et al. 2012). This is especially useful for non-model organisms which cannot be genetically modified. In *D.melanogaster*, we can take this approach even further, since the molecular and genetic tools are available to inject reporter constructs with CREs. Expression data can then be collected from individual CREs instead of the entire endogenous protein pattern, in order to obtain a deeper insight into transcriptional regulation.

1.6.2 Modeling anterior-posterior patterning in the *Drosophila* embryo

The logical next step after the identification of the huge amount of novel CREs was to predict their expression patterns from their regulatory sequence. The following three modelling studies of the *D.melanogaster* A-P system, including *gt* CREs, are on a grand scale. These models were trained with expression patterns of about 40 CREs at mid-blastoderm stage and implemented distinct mechanisms via different mathematical approaches.

Segal *et al.* presented a thermodynamic model for the segmentation gene network that requires the DNA sequence, the concentrations of the TFs and their PWMs as input (Segal et al. 2008). The framework does not assume *a priori* whether a factor is an activator or repressor and it considers competition between TFs at overlapping sites. However, the concepts of short-range repression and activator synergy were not taken into account. 44 gap and pair-rule gene CREs served as training set and the derived model was used to predict their expression patterns, as well as those of the 11 CREs driving expression in the anterior (predicted by Ochoa-Espinosa et al. 2005) and 15 *D.pseudoobscura* modules. The model first computes the occupancy distribution of the TFs along the A-P axis at mid-blastoderm stage without applying pre-determined PWM thresholds. Then the probabilities of all possible binding configurations are calculated and converted into expression levels via a logistic function. These expression contributions are weighted by multiplying with their probability and finally, the sum of these over all possibilities is calculated. The PWMs were fitted as well, which changed the binding preferences in some cases. It is not entirely clear whether this procedure actually improved the fits substantially. Additionally, due to this increased amount of free parameters, the fitting procedure becomes even more computationally exhausting. They used a sampling based algorithm, which could have been inadequate to explore sufficient solutions. They claim that the avoidance of predetermined thresholds allows for the contribution of weak sites and that the short-range homotypic clustering of such weak sites facilitates cooperative binding, which increases predictive power.

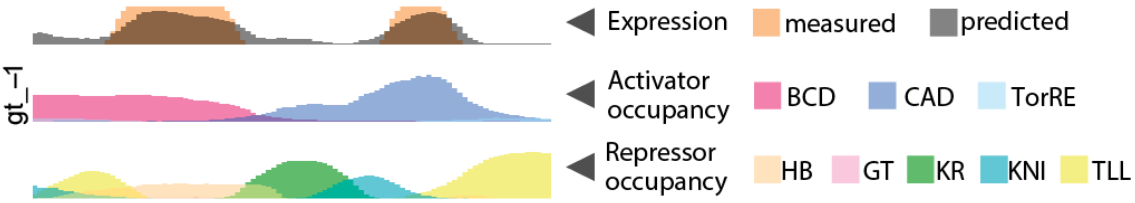


Figure 12: Expression and regulation predicted for gt-1 by Segal *et al.*
Taken from Segal *et al.* (2008). The predicted expression level (gray) is superimposed on the measured expression (orange). Occupancies of participating factors are superimposed and separated by activators (middle) and repressors (bottom). Note that the Torso response element (TorRE) is not a TF, but the binding site for the repressor Capicua.

Unfortunately, the biological context of one of the input factors was mistaken, which casts doubts on the reliability of certain model outputs. They considered the Torso response element (TorRE) as an activator and claimed that TorRE is its corresponding binding motif (Figure 12). In fact, the TorRE is recognized by the repressor Capicua (Cic), which is expressed in the trunk region of the embryo only, because it is post-transcriptionally degraded in response to Torso-signalling at the poles via the MAPK pathway (Jiménez et al. 2000). As input concentration

they used the expression profile of Torso, which is not a TF, and therefore no PWM is available. Additionally, it might have been too ambitious to model the entire embryo from 0 – 100% A-P position without regarding the possible influence of the head gap genes. Since the model needs to compensate for the missing head factors with the available TFs, the resulting regulatory contributions might reflect artefacts rather than the real underlying mechanisms.

He *et al.* presented ***G**ene **E**xpression **M**odeling based on **Sta**tistical **T**hermodynamics* (GEMSTAT) (He et al. 2010). This approach is similar to the model from Segal *et al.*, but with different assumptions about underlying mechanisms. They distinguish *a priori* between activator and repressor and they explicitly consider short-range repression as well as synergistic activation of the basal transcriptional machinery (BTM). They use the data set from Segal *et al.*, excluding the TF Tll, the TorRE and a couple of CREs bringing down their number to 37. They only model from 20 – 80% of the A-P-axis and use dynamic programming methods for training. Their model needs to calculate two main terms: one is the fractional occupancy of the TFs at the DNA and the second one describes the interactions of the bound TFs with the BTM. Based on this, it is then possible to implement different modes of activation and repression. One can choose between two quite different variants of repressions. Their direct interaction model assumes that the repressors act directly on the BTM, independent of their distance, whereas in the short-range repression (SRR) model, a bound repressor renders an activator site inaccessible if it lies within a certain distance. On the other hand, the model does not account for overlapping sites, unless when treating them as a special case of the SSR model with a distance limit of 10 bp. Under this condition, the mechanism of short-range repression as such is not considered anymore. In order to account for distinct mechanisms of multiple bound activators, an additive and a multiplicative effect model were proposed. The first only allows for one bound TF to contact the BTM, for the case that factors would need to interact with the same subunit of the complex. In contrast, the latter model explicitly implements synergy and assumes that the activators might affect different steps of transcription or distinct parts of the BTM. Additionally, it is possible to allow for homo- or heterotypic cooperative binding of predefined pairs of TFs, independent of their mode of action. The strength of this contribution is distance-dependent and the corresponding sites need to be adjacent without other occupied sites in between.

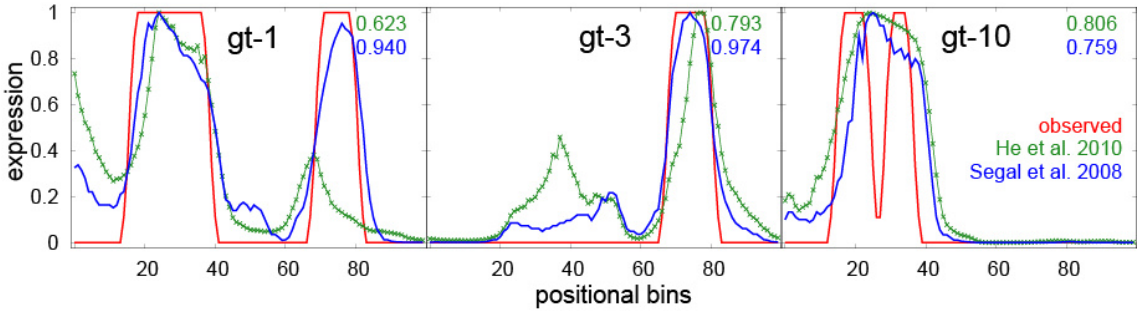


Figure 13: Comparison of predicted expression patterns from two different thermodynamic models. Taken from He et al. 2010. Observed expression patterns (red), Segal 2008 (blue) and He 2010 (green, DirectInt-Coop model with homotypic cooperative interactions of Bcd and Kni). The correlation coefficient between a model's prediction and the known readout is indicated in the top right corner of the panel.

Although the conceptual framework of such a flexible model sounded very promising, a lot of baseline and ectopic expression were observed (Figure 13). Derepression of the *runt* stripe CREs and some others at 80% A-P-position might by due to the exclusion of Tll as input factor. Since experimental evidence supports short-range function for the four inhibitors included in this study, the authors themselves were surprised that the SSR model did not achieve any improvements over the direct interaction model. One has to be aware that the performances of different models were compared via the Pearson correlation coefficient, which was averaged

over all CREs. A complementing task, when judging model outputs, is inspecting the contributions from each factor at each A-P-position as well, but unfortunately such a representation was not shown in the article. The different model combinations usually had either one or the other approach for activation and repression, but did not test the possibility of including several mechanisms at once. It is a pity that overlapping sites are not considered explicitly, since steric hindrance due to a bound protein is physically the most logical explanation for reducing fractional occupancy of another potential binding-protein. It would be interesting to know the performance of a model including overlapping binding sites, short-range repression, synergy and cooperative binding. On the other hand, there is a risk of over-fitting when including more mechanisms and therefore more free parameters.
In the follow-up article (Samee and Sinha 2014), the predictions from GEMSTAT for several individual enhancers from the intergenic locus of the target gene were linearly combined to fit the entire endogenous gene expression pattern.
Kazemian *et al.* described a different approach for CRE identification combined with pattern prediction (Kazemian et al. 2010). First, a logistic regression model was learned from a dataset of 46 CREs, considering the same TFs as above, but including Cic, Tll, Hkb and Forkhead (Fkh). The STUBB program searches for TFBS with the PWMs and determines a factor motif score for each TF across the genome (Sinha et al. 2006). This score is multiplied with the concentration of the TF and its weight, which will be positive for activators and negative for repressors. The weights, as well as the role of the TFs, need to be estimated by fitting the model. The product of these three values corresponds to a weighted fractional occupancy, which is then converted into an expression profile along the A-P-axis via logistic regression.

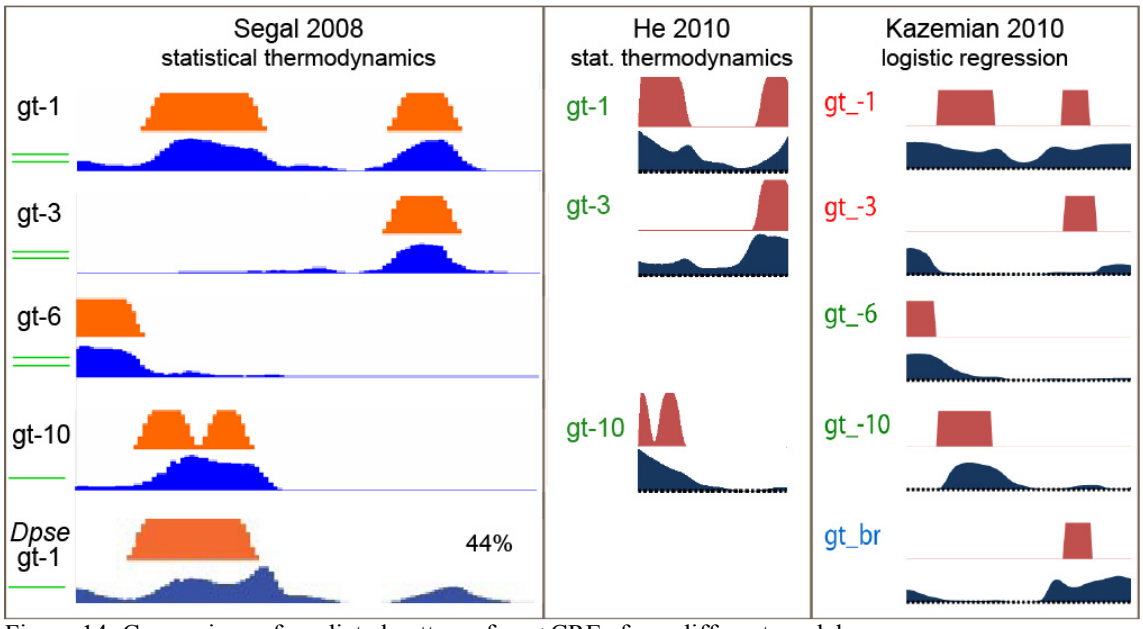


Figure 14: Comparison of predicted patterns for *gt* CREs from different models.
Graphs were taken from Segal *et al.* 2008, He *et al.* 2010 and Kazemian *et al.* 2010. Observed expression patterns are shown in red (above) and predicted ones in blue (below). Note that Segal 2008 and Kazemian 2010 model from 0-100% and He 2010 from 20-80% of the A-P-axis. Segal 2008 classified the modules subjectively as good (underlined twice in green), fair (underlined once in green) and poor (n.a.). The last prediction shows the *Dpse* gt-1 ortholog, which has a sequence identity of 44% and does not drive the posterior *gt* domain. He 2010 shows expression profiles from the DirectInt model (with Bcd and Kni self-cooperativity) on a scale from 0 to 1. Labels in green indicate CREs where the correlation coefficient is greater than 0.65. Predictions in Kazemian 2010 were categorized as good (green label), fair (blue label) or bad (red label). *gt_br* corresponds to CE8001 from Berman *et al.*

This approach is much simpler compared to the thermodynamic models of Segal 2008 and He 2010 (Figure 14), but the fits are slightly better than in the latter one, since there is less baseline and ectopic expression. There are several possible explanations for this paradox: the inclusion of Cic and Tll, overfitting of the other factors or the abstraction of details could also lead to an improvement. The authors claimed that the fits improved dramatically via usage of multi-species motif profiles including 10 other *Drosophila* species. The *Dpse* gt-1 ortholog was an exception, because the single-species model showed better performance due to the functional divergence of this CRE. When using the model with *in vivo* binding data from ChIP-peaks instead of *in silico* motif scores, the performance was worse, including assignation of Gt as an activator and less statistical significance for other repressors.
Based on the logistic regression, they developed a new measure for comparing predicted patterns with endogenous gene expression. This so-called Pattern Generating Potential (PGP) has several advantages over the RMSE or correlation coefficient. It allows for sub-domains, is sensitive to shape and magnitude and avoids biases concerning too broad or narrow domains. A window of 1 kb slides from 10 kb up- until 10 kb-downstream of the target gene and predicts expression patterns. Then the prediction is averaged separately in expressing and non-expressing bins, giving a reward and a penalty term, respectively. The penalty, weighted thrice, is subtracted from the reward term and finally, linear scaling results in a PGP score between -1 and 1.

2 A mathematical model of transcriptional control

2.1 The scope: from genome-wide to high spatio-temporal resolution

The thermodynamic models described above (Segal et al. 2008, He et al. 2010, Kazemian et al. 2010) were very ambitious since they aimed to fit and predict a huge set of CREs involved in A-P patterning. We know that the considered input factors and mechanisms are probably not sufficient to explain the functioning of all the CREs of the training set and therefore compensatory artefacts might be misinterpreted as an underlying regulatory interaction. In particular, cross-regulation among the pair-rule and segment-polarity genes was neglected as well as short-range repression. Regardless if one categorizes a model and its fits as good or bad, it can still give us new insights. The results obtained reflect, that we are further from understanding the regulatory mechanisms than we thought. The reason therefore could be that although there might be something like a "grammar" for *cis*-regulation, there are a lot of exceptions to it as well. It is probably necessary to step back from genome-wide and high-throughput approaches and to concentrate on subsets of genes or even single CREs. For this aim it is recommendable to increase temporal and spatial resolution, in order to capture details of differential expression based on dynamic changes. Furthermore, there are probably more cases of context-dependent switches between activator and repressor than those known so far from specific experiments (Small et al. 1991, 1996, Rembold et al. 2014). So far, it has proven difficult to design functional synthetic enhancers from scratch.
I use the model of transcriptional regulation from our collaborator John Reinitz (Reinitz et al. 2003), which has been used to undertake detailed studies of the regulation of the *eve* gene (Janssens et al. 2006, Kim et al. 2013). For these analyses, the model was fit to regulatory fragments of the gene of interest only, instead of training it with an entire data set of CREs from different gene families. In particular, it was fit to expression data from a lacZ reporter construct, driven by the 1.7 kb sequence upstream of *eve*, which includes the Minimal Stripe Element (MSE) 2 that triggers expression of stripe 2 and also partially stripe 7. The model was able to correctly predict the spatio-temporal expression of lacZ mRNA over eight time classes (Janssens et al. 2006). It suggested previously unknown regulatory input from Kni and Cad, showing that individual binding sites outside the so-called minimal element can also be important. To study the effects of rearrangements of non-coding DNA, four different fusion constructs of the MSE2 and MSE3 with and without spacers of different length in between them were designed (Kim et al. 2013). Their expression was quantified and used to train the model, which then predicted the expression pattern of *eve* enhancers for stripes 5 and 4/6 of *D.melanogaster*, as well as stripes 2 and 3/7 from various drosophilid and sepsid species. It was also capable of predicting other *Drosophila* genes, such as *runt*, but it was not able to reproduce the expression of *gt* CREs. This is due to the distinct underlying regulatory influences of different TFs. For example, repression by Gt is required to set the anterior boundary of *eve* stripe 2, while on the other hand it might not be auto-regulating at all, even if binding sites are present. In a different scenario, Gt might be auto-repressing or auto-activating in one, but not another of its CREs. Such a bimodal behavior is usually context-dependent, similar to the case of Hb, which was shown to repress MSE3, but activates MSE2 if Bcd is present (see section 2.2.2 for details).

2.2 Structure, assumptions and mechanisms of the model

The model requires as input the concentration, the PWM and the role of the TFs, as well as the DNA sequence of the CRE (Figure 15). The aim is to optimize certain parameters of the mathematical framework, which formulates TF binding to the DNA and protein-protein interactions, thereby considering different activating and repressing mechanisms. The parameters are inferred by starting with random values and fitting the model to quantitative expression data of one or more CREs over different stages. The code searches for binding sites and optimizes towards their number and identity. The output is the expression pattern of the CRE from the DNA sequence across different time classes and the combinations of sites required for correct expression.

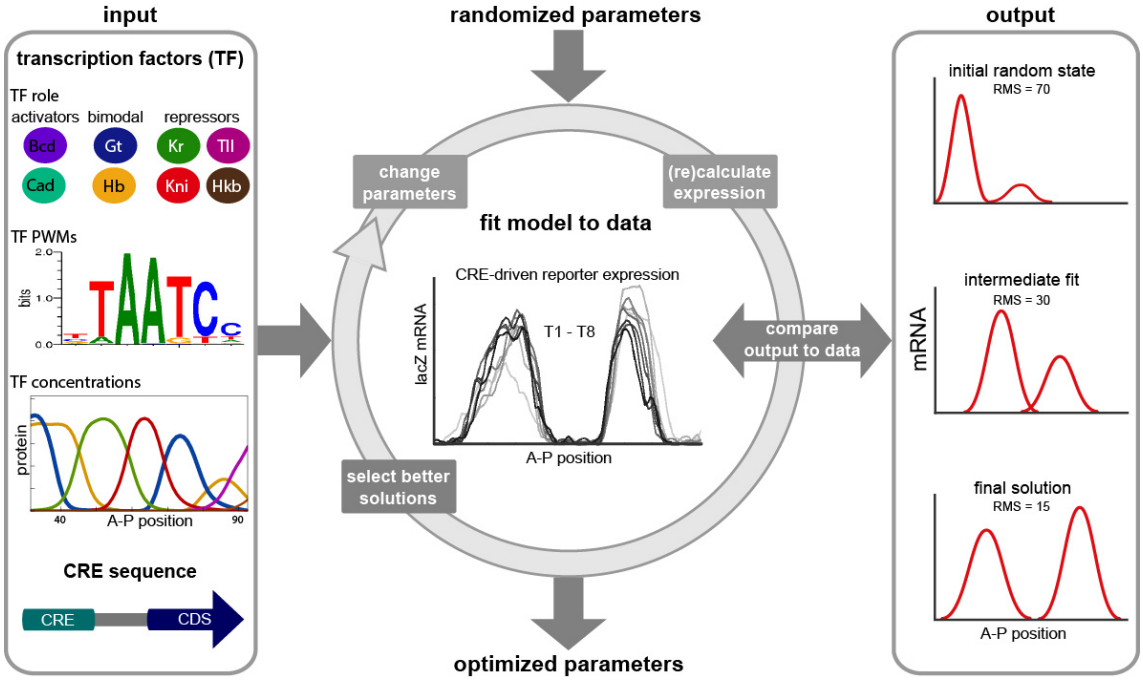


Figure 15: Modelling transcriptional control.
See text for explanation.

The optimization is performed with a method called Lam-simulated annealing (Lam and Delosme 1988a, 1988b). One can imagine this algorithm walking around on a landscape, but a higher-dimensional one, since it represents the entire parameter space. Starting on a random spot of it, the aim is to find the global minima. It is allowed to climb uphill again, in order to avoid to get trapped in local minima. The algorithm tries to improve the solution over many iterations by minimizing the difference between the model output and the expression data, represented by a number called the root mean square (RMS) error. The optimized parameters can then be utilized to predict the expression of the CREs in mutant backgrounds or of CREs with mutated TFBS. Additionally, also CREs from other species or different types of DNA fragments, such as intergenic regions, can be tested. In order to predict mutants lacking a certain TF, the model requires an entire dataset with the concentrations of all the TFs in this particular genetic background.

2.2.1 Assumptions of the model

The model is phenomenological in the sense that it mathematically describes observed phenomena without knowing all the details about how CREs and TFs communicate with the BTM. The main difference in its mechanistic basics compared to other thermodynamic models, is the concept that so-called *adaptor factors* or *mediators* are recruited by the activators as a

functional bridge towards the BTM (Berger et al. 1990, 1992, Lemon and Tijian 2000, Näär et al. 2001). Some of these proteins could be identified and they show ubiquitous expression from maternal mRNA in the *Drosophila* blastoderm (Tamkun et al. 1992, Park et al. 2001, Saurin et al. 2001). One can imagine that transcription initiation is the rate-limiting step, which needs to be calculated. The model is based on the assumption that the adaptor factors lower the activation energy in order to initiate the transcriptional process. This is an analogy to reaction kinetics, where enzymes can catalyze by reducing the energy barrier ΔA (Figure 16), which is described by the Arrhenius law: $k \propto \exp(-\Delta A / RT)$, with k the reaction rate, ΔA the activation energy in Joule/mole, R the gas-constant in Joule/K mole and T the temperature in Kelvin.

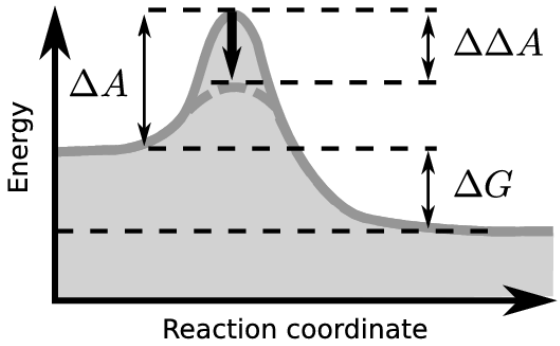


Figure 16: Reaction energies.
The activation energy barrier ΔA needs to be overcome and the Gibbs free energy ΔG is released during a reaction. Catalysis by enzymes or activators decreases the energy barrier by $\Delta\Delta A$. Taken from Kim 2013.

2.2.2 Regulatory mechanisms considered in the model

Repressive mechanisms
The model considers competitive binding at overlapping sites, disregarding whether the TF was set as an activator or repressor (Figure 17A). The steric hindrance is assumed to cover a range of 14 bp based on the size of the footprinted site for Bcd. This value tends to be greater than a usual binding motif and can be changed by the user. Short-range repression or quenching (Figure 17B) refers to the effect of bound repressors on nearby activators (Gray et al. 1994, Gray and Levine 1996), which is assumed to prevent the binding of the adaptor factors. It is implemented via a distance-dependent function, restricting the range of this mechanism to 150bp. Direct repression (Figure 17C) from repressors bound near the transcription start site onto the BTM is formulated in a similar distance-dependent manner as quenching and can be turned on or off by the user.

Activating mechanisms
Independent *in vitro* experiments and *in vivo* assays in yeast have demonstrated pairwise cooperative binding (Figure 17D) of Bcd to adjacent sites (Ma et al. 1996, Burz et al. 1998). This ability could be attributed to single amino acids in the homeodomain and subsequently fly lines carrying mutations for these cooperativity residues were generated (Burz and Hanes 2001, Lebrecht et al. 2005). It is not entirely clear over which distances this mechanism works, but it turned out to be still functional over 41 bp in the *hb* promoter (Ma et al. 1996) and therefore the upper limit in the model was set to 60 bp. Another special mechanism is co-activation (Figure 17E), because only isolated cases were studied in sufficient detail. Hb was shown to function as a repressor in the MSE3 of *eve* (Small et al. 1996). Transient co-transfection assays in *Drosophila* Schneider cells demonstrated that Hb is activating the *eve* MSE2 in a multiplicative manner with Bcd (Small et al. 1991). This synergy was confirmed by an independent study using artificial enhancers, indicating that the effect was maintained even when separated by 100bp (Simpson-Brose et al. 1994). There is no experimental evidence for co-activation of Hb by Cad, nevertheless it was also included in the model by Kim *et al.* (2013) in a distance-dependent fashion. Finally, the model assumes that the recruitment of the adaptor factors by the activators (Figure 17F) stimulates the BTM and initiates transcription.

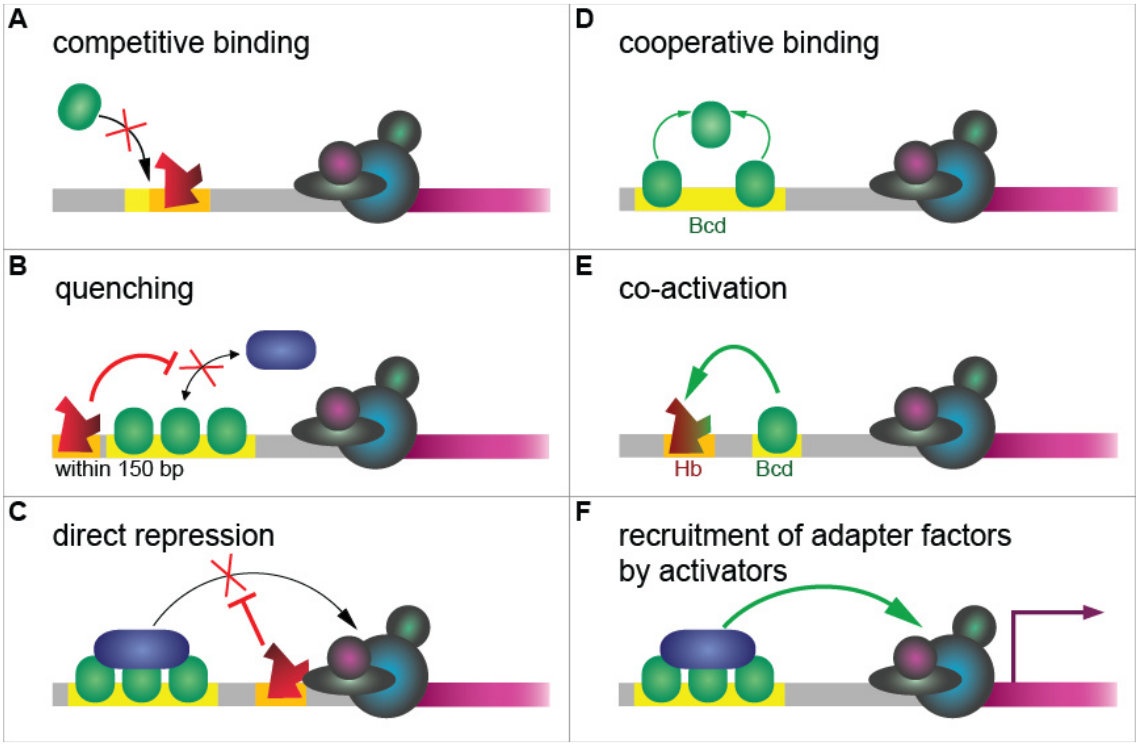


Figure 17: Mechanisms considered in the model.
See text for explanation.

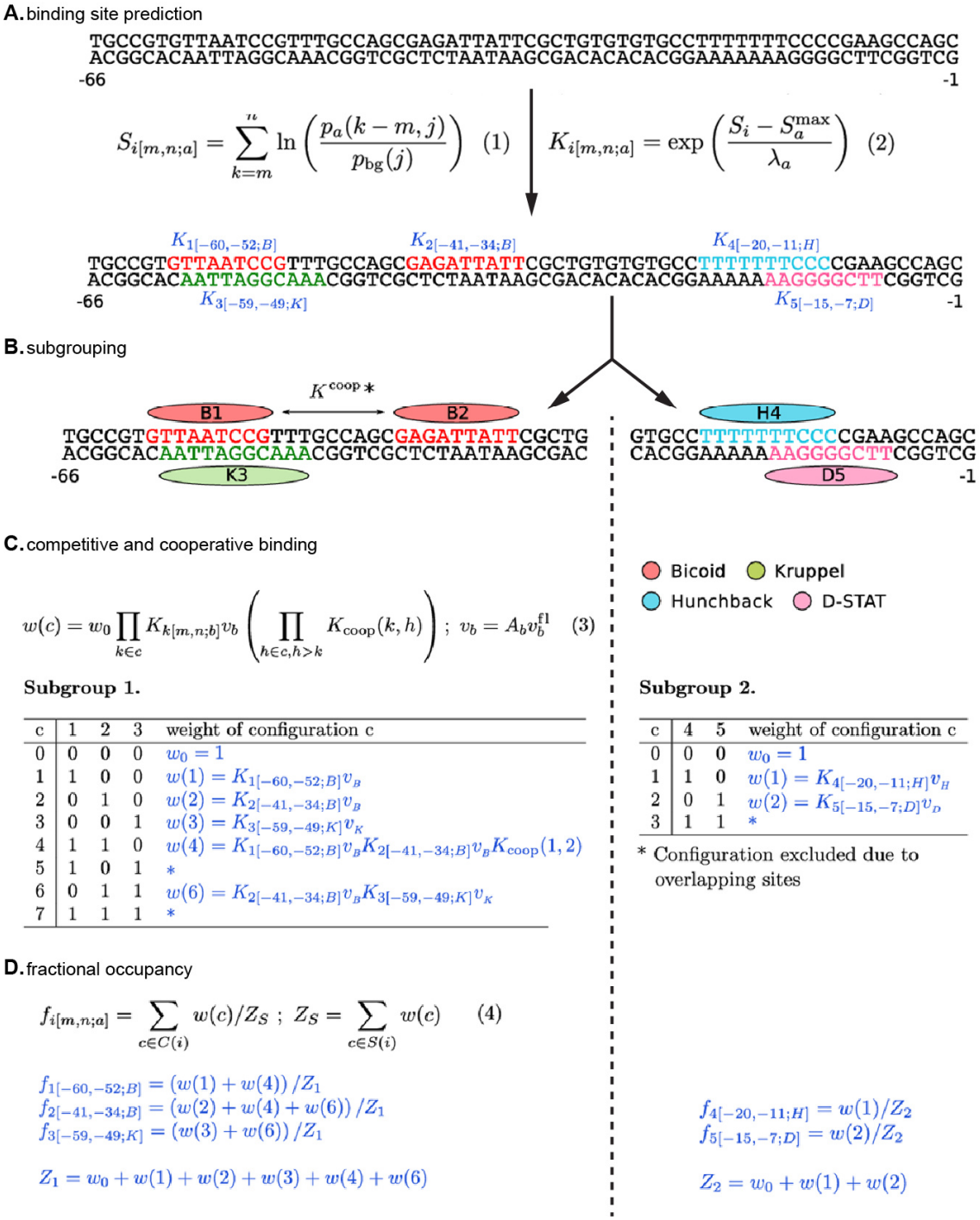
2.3 Model equations

The model is constructed in three layers, starting with TF binding to the DNA (Figure 18), then incorporation of protein-protein interactions (Figure 19) and finally integration of the inputs and formulation inspired by the Arrhenius law (Figure 20).

2.3.1 Transcription factor binding to DNA

First, the PWM is used to search for binding sites and for each found site i a score S is assigned for binding TF a to the DNA sequence from bp position m to n (Equation 1 in Figure 18A). This score is the sum of the probabilities to find the base j (A, C, G or T) at each position k within the site over the background frequency for this base in the genome of *D. melanogaster* (Table 2). A binding site is considered if its score is above a certain threshold, which can be fixed or adjusted between certain limits. The score is then converted into an affinity K by taking into account the maximum possible score S^{max}, representing a perfect fit to the consensus motif, and a proportionality constant λ for the TF a (Equation 2).
Based on statistical thermodynamics, all possible states of TF binding to DNA need to be explored (Figure 18C, D). In this step, competitive binding and cooperative binding at nearby sites are implemented (Equation 3). In order to avoid combinatorial explosion and thus a potentially enormous computational effort, it is necessary to define subgroups based on interacting sites (Figure 18B). The weights w of all possible configurations c are calculated by multiplying the affinities K with the concentrations v of the TFs (Equation 3). This is accomplished at each A-P position and for each time point, because the TF concentrations vary over space and time. Since we can only measure relative fluorescent intensities v^{fl}, we adjust to "absolute" concentrations by multiplying with a scaling parameter A. Cooperative binding is incorporated via one free parameter K_{coop} per interacting TF pair. Competition is included by prohibiting configurations with simultaneous binding of TFs at overlapping sites (asterisk for such configurations in Figure 18C). Next, the fractional occupancy of each site i with TF a is

calculated by summing up the weights of the configurations that included this site and dividing it with the sum of the weights of all possible configurations within the corresponding subgroup *S* (Equation 4).



c	1	2	3	weight of configuration c
0	0	0	0	$w_0 = 1$
1	1	0	0	$w(1) = K_{1[-60,-52;B]} v_B$
2	0	1	0	$w(2) = K_{2[-41,-34;B]} v_B$
3	0	0	1	$w(3) = K_{3[-59,-49;K]} v_K$
4	1	1	0	$w(4) = K_{1[-60,-52;B]} v_B K_{2[-41,-34;B]} v_B K_{\mathrm{coop}}(1,2)$
5	1	0	1	*
6	0	1	1	$w(6) = K_{2[-41,-34;B]} v_B K_{3[-59,-49;K]} v_K$
7	1	1	1	*

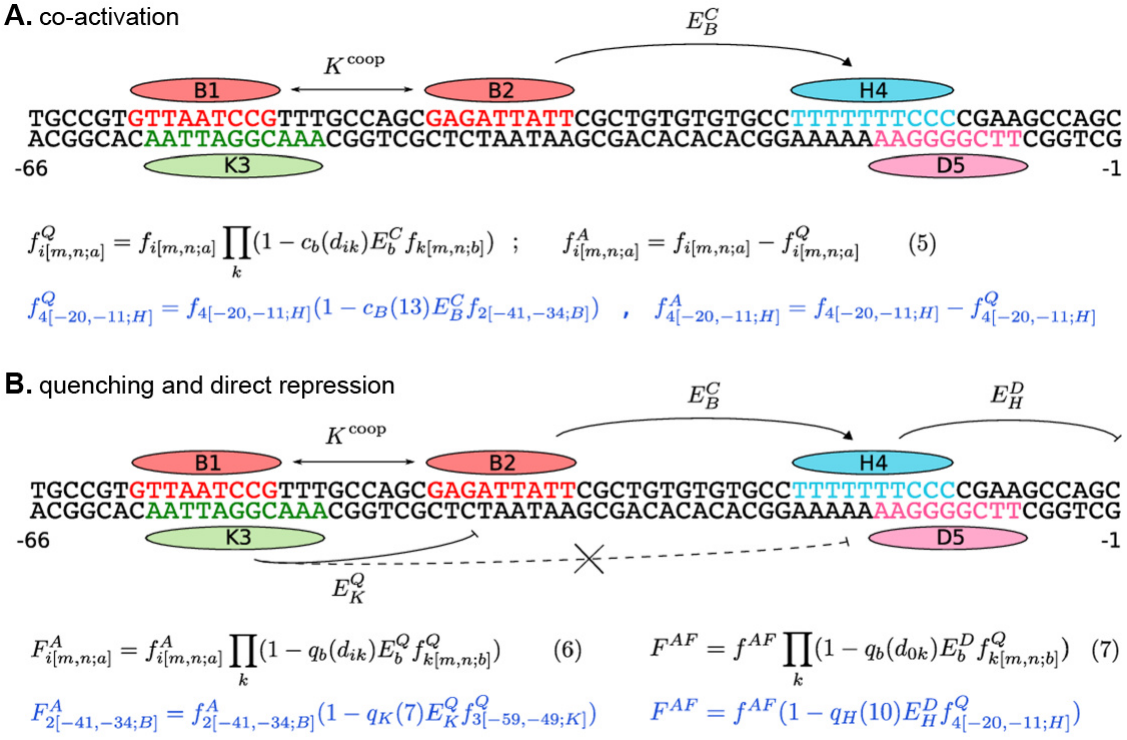
c	4	5	weight of configuration c
0	0	0	$w_0 = 1$
1	1	0	$w(1) = K_{4[-20,-11;H]} v_H$
2	0	1	$w(2) = K_{5[-15,-7;D]} v_D$
3	1	1	*

* Configuration excluded due to overlapping sites

Figure 18: Equations for transcription factor binding to DNA.
Adapted from Kim 2013. Shown are the equations 1 until 4 and examples are in blue. See text for explanation of variables and indexes.

2.3.2 Protein-protein interactions

In the second layer of the model, repression and co-activation are incorporated into the fractional occupancies depending on the distance between the interacting proteins (Figure 19). The distance function for the different cases will be explained in more detail (section 2.3.4). Each mechanism is assigned a strength E for each TF, which is allowed to vary between 0 and 1. From now on, the TFs are distinguished based on their role and their physical fractional occupancies obtain the superscript A for activators and Q for quenchers. Co-activation (Figure 19A) is not implemented as a complete switch of a repressor to an activator site, but allows for the co-existence of both roles (Equation 5). It is mathematically constrained by assuming that the fractional occupancies as repressor and activator sum up to the total physical fractional occupancy. The f^Q term of this special case incorporates the co-activation strengths E^C of all nearby co-activators. Short-range repression reduces the fractional occupancy of an activator by considering the repressive strengths E^Q of all nearby quenchers, resulting in F^A (Equation 6). In a similar way, direct repression acts on the fractional occupancy of the adaptor factors f^{AF}, which is set to 1. This value is reduced, depending on the distance of the quenchers to the BTM and on their direct repression strengths E^D (Equation 7).



$$f^Q_{i[m,n;a]} = f_{i[m,n;a]} \prod_k (1 - c_b(d_{ik}) E^C_b f_{k[m,n;b]}) \quad ; \quad f^A_{i[m,n;a]} = f_{i[m,n;a]} - f^Q_{i[m,n;a]} \qquad (5)$$

$$f^Q_{4[-20,-11;H]} = f_{4[-20,-11;H]}(1 - c_B(13) E^C_B f_{2[-41,-34;B]}) \quad , \quad f^A_{4[-20,-11;H]} = f_{4[-20,-11;H]} - f^Q_{4[-20,-11;H]}$$

$$F^A_{i[m,n;a]} = f^A_{i[m,n;a]} \prod_k (1 - q_b(d_{ik}) E^Q_b f^Q_{k[m,n;b]}) \qquad (6)$$

$$F^{AF} = f^{AF} \prod_k (1 - q_b(d_{0k}) E^D_b f^Q_{k[m,n;b]}) \qquad (7)$$

$$F^A_{2[-41,-34;B]} = f^A_{2[-41,-34;B]}(1 - q_K(7) E^Q_K f^Q_{3[-59,-49;K]})$$

$$F^{AF} = f^{AF}(1 - q_H(10) E^D_H f^Q_{4[-20,-11;H]})$$

Figure 19: Equations for protein-protein interactions.
Taken from Kim 2013. Shown are the equations 5 until 7 and examples are in blue. Note that f^{AF} is set to 1. See text for explanation.

2.3.3 Integration of activating inputs to obtain the mRNA output

Finally, the recruitment of the adaptor factors is calculated via multiplying the corrected fractional activator occupancy of a site i with the activator strength E^A of the corresponding activating TF a and a summation of all these inputs (Figure 20, Equation 8). This value is reduced in case of direct repression (Equation 9). Transcriptional activation is then simulated as a simple enzymatic process lowering the energy barrier of transcription initiation (Equation 10). There is experimental evidence that TFs can activate in a greater than multiplicative manner (Han et al. 1989). This synergy can be modeled by using a diffusion-limited Arrhenius rate law. In this equation, θ corresponds to the activation energy barrier ΔA, which is reduced by a

decrement *M*. The mRNA output follows a logistic function with sigmoid shape (Figure 20), allowing to emphasize two important features of transcription: the above-mentioned synergy, since the expression is exponential within a certain range of *M*, and the saturation represented by R_{max}. At this point, the polymerase already started to function and diffusion of new molecules would become the rate-limiting step. R_{max} is usually fixed to 255, which is the maximum value for a greyscale pixel (in 8 bit color space), extracted from the fluorescence intensities.

$$N = \sum_a \left(E_a^A \sum_i F_{i[m,n;a]}^A \right) \quad (8)$$

$$M = N.F^{AF} \quad (9)$$

$$[mRNA] \propto \frac{d[mRNA]}{dt} = R_{max} \left(\frac{\exp(M-\theta)}{1+exp(M-\theta)} \right) \quad (10)$$

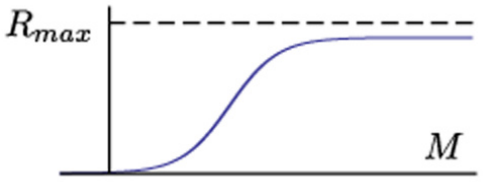


Figure 20: Integration of activating inputs and logistic function with sigmoid shape.
Adapted from Kim 2013. Equations 8 until 10. See text for explanation.

2.3.4 The distance function

The edge-to-edge distance between the TFs is taken into account when calculating short-range and direct repression, as well as co-activation. The strength of the repression or activation equals 1 if the distance is smaller than a threshold D1 and 0 if it is bigger than the limit D2 (Figure 21). Between these two values, the strength is calculated by linear interpolation. In the case of short-range and direct repression, the threshold D1 was fixed to 100 bp and the limit D2 was set to 150 bp (Figure 21A). In contrast, when the co-activation distance is adjusted, D2 is set to 1.1x D1 in order to add only one additional free parameter (Figure 21B). Based on detailed studies of the junctions between the *eve* enhancers MSE2 and MSE3 with different spacing, the possible distance for co-activation of Hb by Bcd appears to be tightly constrained (Kim et al. 2013). Hence, the threshold D1 was allowed to vary between 150 and 200 bp in the case of Bcd, whereas for Cad no such observations were made and it was adjusted within 10 and 200 bp.

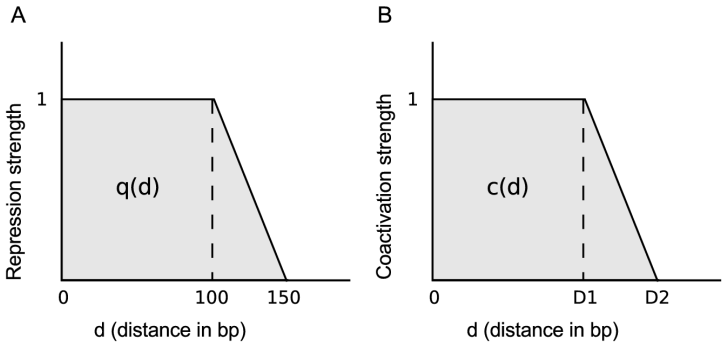


Figure 21: Distance-function for repression and co-activation. **(A)** The distance limits are fixed for repression. **(B)** The limits are adjusted in the case of co-activation. Taken from Kim 2013.

2.3.5 Parameter estimation

Many of the variables used in the aforementioned equations need to be optimized during the fitting procedure. For each TF *a*, the protein scaling factor A_a, the PWM scaling factor λ_a, the PWM threshold T_a and either the activation or the quenching energy E^A or E^Q are adjusted within predefined limits (Table 2). Furthermore, for certain mechanisms we need additional parameters, which are the co-activation energies E^C, the affinity for cooperative binding K^{coop}, the delta for co-activation and the direct repression energies E^D for each quencher. The number of free parameters will vary depending on the combinations of TFs and mechanism considered

during an optimization run. In general, with 8 TFs and excluding cooperativity, co-activation and direct repression, it will yield 32 free parameters, plus one more for θ.

		parameter	value
general	f^{AF}	fractional occupancy of adaptor factors	1
	R^{max}	maximum rate	120 – 255
	θ	activation threshold	5 – 20
	Q	constant	1
PWMs	λ_a	PWM scaling factor	0.5 – 5
	T_a	PWM threshold	0 – 5
		pseudocount	1
		size of binding site	14 or 24
activators or repressors	A_a	protein scaling factor	0. 000001 – 4
	E^A	activation coefficient	0. 0001 – 20
	E^Q	quenching coefficient	0. 0001 – 1
quenching and direct repression	d_{ik}	quenching distance	100
		delta of quenching distance	50
	E^D	direct repression coefficient	0. 0001 – 1
	d_{0k}	direct repression distance	100
		delta of direct repression distance	50
co-activation and cooperativity	E^C	co-activation coefficient	0 – 1
	d_{ik}	co-activation distance for Bcd	150 – 200
	d_{ik}	co-activation distance for Cad	10 – 200
		delta of co-activation distance	10
	K^{coop}	cooperativity	1 – 500
		cooperativity distance	60
construct		position effect scale factor	1
		AT background frequency	0. 297
		CG background frequency	0.203

Table 2: Parameters of the model.
Parameters can either be fixed to a certain value or adjusted within a low and high limit. Note, that in several optimizations runs the threshold for Bcd and Hb were fixed to 1.71 and 0.63, respectively. Q is the amount that each bound adaptor factor lowers θ. Because Q is a constant that multiplies all the E^A coefficients, it can be set to 1 without any loss of generality (Martinez et al. 2014).

3 Objectives

The specific objectives of this PhD thesis are:

1. Collect qualitative data of the available reporter fly lines harboring the previously identified *giant* CREs in order to decide, which ones are most suitable for further quantitative analyses.
2. Create new fly lines containing reporter constructs with different *gt* CREs by a method for site-specific integration into the same locus in order to better compare them.
   The available transgenic flies were generated by random integration of the reporter construct into the genome and hence, the insertion locus might have position effects on the expression.
3. Quantify the expression patterns of the CREs with high spatio-temporal resolution.
   Data acquisition comprises fluorescent *in-situ* hybridization with simultaneous immuno-staining, confocal microscopy, image processing and time classification of the embryos. The generation of such precise datasets is labor-intensive.
4. Fit the model of transcriptional control to these datasets, in order to infer the regulatory contributions of potential activators and repressors.
   Thereby I can test different input combinations, potential scenarios and distinct mechanisms, such as cooperativity and co-activation.
5. Utilize the optimized model parameters to predict the expression of CREs in mutants.
   This is only possible if a mutant dataset with the expression profiles of all TFs is available.
6. Evaluate *in vivo* the regulatory mechanisms suggested *in silico* and the model predictions for the mutants.
   The CREs can be introduced into mutants lacking a particular TF via different techniques, such as germ line clones, transgenic RNAi or recombination. Hypotheses about regulatory influences can also be verified via modified enhancers with mutated TFBS. Subsequently, the mutated cassette needs to be integrated into the same target fly lines.

Specific scientific questions

The two elements driving the expression of the posterior *giant* domain are of particular interest. Gt-3 is exclusively responsible for the posterior domain, whereas gt-1 also drives the anterior domain. Several questions concerning their expression dynamics and regulation were raised:

- Do the expression boundaries of both CREs coincide with the pattern of the endogenous mRNA over time or is one CRE responsible for setting the anterior and the other one the posterior boundary of the posterior domain?
- Are there differences in the timing of the distinct CREs, which could reflect varying underlying regulation for early or late expression?
- What are the functions and synergies of these two seemingly redundant elements and are both of them really required for viability? It could be that gt-1 acts as a booster for later stages or that it confers robustness under difficult environmental conditions or if perturbations of the system occur.

Gt-1 contains no binding sites with sufficient affinity for the obvious anterior and posterior activators, Bicoid and Caudal. Nevertheless, it is able to drive both domains. Therefore, the puzzling question is: how can it be activated at all? Segal *et al.* (2008) claimed that Bcd cooperativity is the driving force based on their modeling results. Other hypotheses are co-activation by the bimodal factor Hunchback or a previously unknown maternal gradient. I contemplate these as well as alternative theories. Connected with this issue is the paradox of the repressing inputs on the early expression in the anterior versus the posterior region. Previous modeling for the endogenous *gt* pattern only worked if the two domains were treated separately (Jaeger et al. 2007). In particular, Hb repression has to be neutralized in the anterior region, but it remained unclear how that could be achieved at the molecular level.

4 Results and discussion

4.1 Expression dynamics and TFBS content of *giant* CREs

4.1.1 Expression dynamics of *giant* CREs

The first step of this PhD thesis was to collect qualitative data of the available reporter fly lines harboring the previously identified *gt* CREs (Table 1, Figure 7) in order to decide which ones were the most suitable for further quantitative analyses. In the original publications (Berman et al. 2002, Schroeder et al. 2004, Ochoa-Espinosa et al. 2005) only one image at mid-blastoderm stage is shown (Figure 8), but we wanted to capture the dynamics of the expression patterns. For this purpose, I performed enzymatic *in situ* hybridization against lacZ mRNA and staged the embryos by visual inspection of the membrane morphology (see Figure 22 and Figure 23 for a summary, and Figure 72 and Figure 73 in the Appendix for versions with more stages). In general, the observed spatio-temporal patterns driven by the CREs appear to coincide with the endogenous *gt* mRNA expression (for quantified expression profiles over eight time classes see Figure 31), but slight deviations were observed in several cases.

Gt1 and gt-6 are overlapping CREs of different lengths that drive expression at the anterior tip of the embryo (Figure 22A). In this particular case, the two patterns differ quite substantially from each other in terms of timing, strength and ectopic patches. Gt1 is much stronger compared to gt-6 and it appears sooner. The anterior tip of the endogenous mRNA arises at T5, whereas for gt1 it is already clearly visible earlier (Figure 72). Furthermore, gt1 shows an additional slight anterior stripe and gt-6 drives ectopic dorsal expression at late stages.

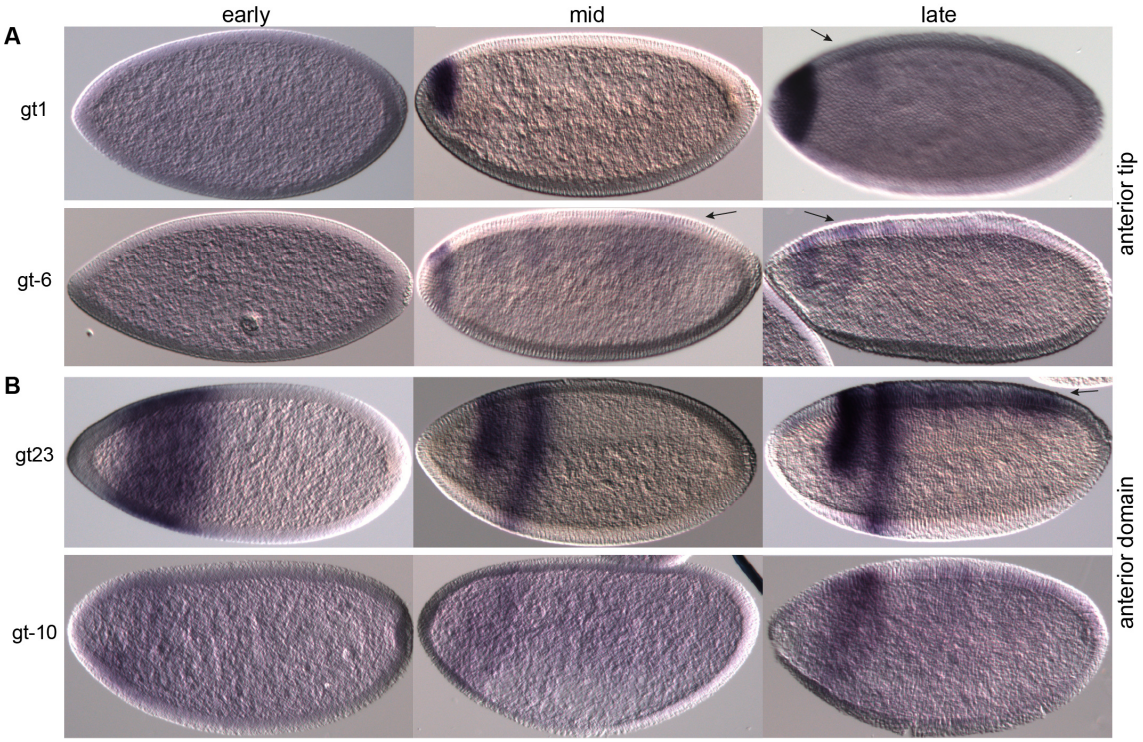


Figure 22: Previously available reporter-fly lines with *gt* CREs driving expression in the anterior.
The CREs drive expression of the anterior tip **(A)** or the anterior domain **(B)**. Shown are DIC images of embryos at early, mid and late cleavage cycle 14, stained by enzymatic *in situ* hybridization with a DIG-labelled lacZ probe. Ectopic expression is indicated with arrows. For more images at different stages see Appendix (Figure 72).

Gt23 and gt-10 (Figure 22B) are supposed to drive the anterior domain, which splits into two stripes over time. The pattern of gt23 resembles the endogenous anterior expression quite well, including the clear refining of one broad domain into two separate stripes, but it also triggers ectopic dorsal expression at late stages. None of the available fly lines carrying gt-10 was able to achieve a well-defined anterior domain. The slight expression at later stages rather seems to be the ectopic expression from the vector backbone (see section 4.2 for detailed explanation of this artefact).
The CREs **CE8001 and gt-3** activate the posterior domain (Figure 23A). CE8001 seems to be slightly broader than the endogenous pattern and it triggers an additional faint anterior domain at late stages. Gt-3 does not show any obvious deviations from the endogenous expression.
Gt-1 drives both domains (Figure 23B), but their boundaries appear less sharp, resulting in a rather blurry aspect in general, and the anterior domain does not refine properly into two separate stripes. The posterior expression of gt-1 is hardly visible during the earlier stages, which is not the case in the endogenous pattern.

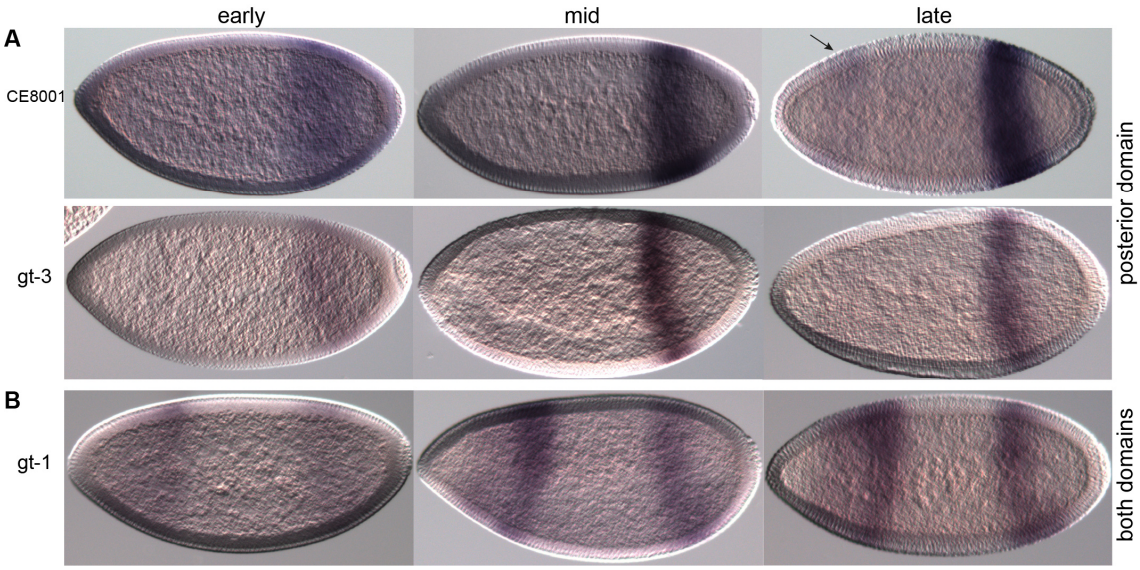


Figure 23: Previously available reporter-fly lines with *gt* CREs driving expression in the posteriror.
The CREs drive expression of the posterior domain only **(A)** or of both *gt* domains **(B)**. Shown are DIC images of embryos at early, mid and late cleavage cycle 14, stained by enzymatic *in situ* hybridization with a DIG-labelled lacZ probe. Ectopic expression is indicated with arrows. For more images at different stages see Appendix (Figure 73).

The observed expression patterns driven by these CREs are related to their TFBS content. In order to get an idea how these CREs might be regulated, we wanted to know which activators and repressor sites are present on their DNA sequence.

4.1.2 TFBS content of *giant* CREs

The original publications considered different TFs and used distinct algorithms to detect them. I wanted to double-check the TFBS content with the same tool for all CREs. For this purpose, I analyzed them with the program Windowfit (Sinha et al. 2006), which predicts the position and strength of TFBS (Figure 24). I decided to include the activators Bcd, Cad and Dstat (*signal-transducer and activator of transcription protein at 92E*), as well as the repressors Hb, Kr, Gt, Tll, and Cic, which binds to the Torso Response Element (TorRE) (Jiménez et al. 2000). I used the default settings of the program, which included the PWMs as in Schroeder et al. (2004). Additionally, I searched for binding sites with the Selex PWMs and the trimmed version of the Cad matrix from Berman et al. (2002). Note that the program applies the same threshold for all TFs and that PWMs from different sources can give distinct outputs.

Figure 24: TFBS found in the *gt* CREs.
TFBS were predicted by the program Windowfit (Sinha et al. 2006). The numbers on top indicate the nucleotide position on the DNA sequence. Whenever TFBS overlap, they are drawn on a new line and the height indicates their strength. The numbers next to the TFs represent their integrated contribution in arbitrary units. The PWMs as in Schroeder et al. (2004) were used for the predictions. For gt-1 also the output with the Selex PWMs and the Cad matrix from Berman et al. (2002) is shown (with different color coding).

The tip-CRE gt1 is much shorter and entirely included in gt-6. Both fragments contain Bcd activator sites, Kni repressor sites and the TorRE. Cic was shown to establish the posterior boundary of the anterior tip (Löhr et al. 2009). The anterior CRE gt23 contains mainly Bcd activator sites and Kr repressor sites. There are also sites for Kni, which has an expression pattern complementary to *gt* in the anterior, including the DV-asymmetry. The posterior CRE gt-3 carries activator sites for Cad and repressors sites for Kr, Hb and Tll. A Bcd-cluster at the beginning is probably compensated by the Hb sites in the same region. Analysis of the CRE gt-1 revealed only minor activating input from Cad, expected to trigger expression of the posterior domain. Most interestingly, the program did not find any site for Bcd, which was supposed to activate the anterior, and with the Selex matrix only one very weak site was identified (last panel in Figure 24). The element contains several strong Kr repressor sites as well as Tll, Hb and Gt sites.

4.1.3 Early vs. late regulation

If we would only consider the complete endogenous protein pattern, then we might assume that the anterior boundary of the posterior *gt* domain is set by Hb via strong repression at early stages (section 1.4.4). This would create a problem for the anterior domain, which overlaps with anterior Hb. At later stages, *Kr* is expressed and takes over the repressing role of Hb. It was not known how Hb could act differentially on the two *gt* domains, but if we think in terms of separate CREs, timing and TFBS, it can easily be achieved. For these reasons gt-1, which drives both domains, was of special interest for me, because it would be a paradox if it could drive expression for both domains at early stages (Figure 25). I checked the early expression of all elements by *in-situ* hybridization, which showed that gt-1 only arises at cleavage cycle 14 and therefore cannot trigger expression at early stages, whereas the other elements show expression at C12 already. Hence, it seems that at early stages expression is driven by separate CREs for each domain.

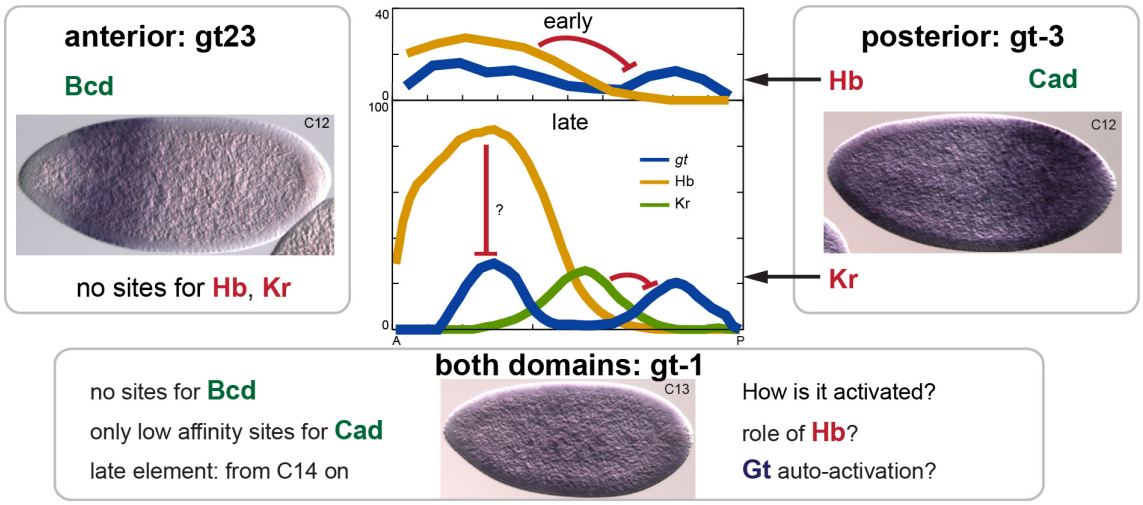


Figure 25: The early anterior and posterior *gt* domains are driven by different CREs.
The anterior domain is initially triggered by Bcd-activation via the CRE gt23. No sites of sufficient strength for Hb or Kr were found, suggesting that the posterior boundary of the anterior domain might be set by a Bcd threshold. The posterior is initially activated by Cad via gt-3. Hb sets the anterior boundary of the posterior domain at early stages and Kr takes over this role at later stages. The late element gt-1, which drives expression of both domains from C14 on, has no Bcd and only low affinity Cad sites.

The anterior element gt23 harbors only one weak site for Hb and Kr and hence it does not seem to be repressed by these factors. The posterior boundary might be set by a Bcd threshold in this case. In contrast, gt-3 carries several binding sites for Hb, which is the reason that this CRE does not drive expression in the anterior. Hb can set its anterior boundary at early stages and it also contains sites for Kr, which takes over the role of Hb at later stages.

4.1.4 Choice of CREs for further quantitative analyses

Due to their different functions, timing and regulation, the CREs gt-1 and gt-3 were chosen for further analyses with a focus on the posterior domain. For the modelling, we need the concentrations of all TFs that might be involved in their regulation. The head-gap genes *otd*, *ems*, *btd* and *slp* might play a role for the refining of the anterior. Unfortunately, no expression profiles are available for these TFs and hence, the anterior CREs cannot be considered for *in silico* analyses. Nevertheless, the posterior boundary of the anterior *gt* domain is still included in the model.
Gt-1, the element driving both domains, was of particular interest, because it lacks TFBS for the two main activators, Bcd and Cad. Therefore it is not clear how it becomes activated at all and several questions were raised: Did sequence motifs, probably coding for weaker binding sites, remain undetected? Are unknown additional factors involved in its activation? How important is

Gt auto-activation for later stages? What is the role of Hb, as it is a bimodal factor, which can be considered as an activator if Bcd is present? Segal et al. (2008) claimed that gt-1 relies on cooperativity between several weak Bcd sites. Since the model of transcriptional control used in this thesis incorporates distinct regulatory mechanisms such as cooperative binding and co-activation, these hypotheses can be tested *in silico*.
Another reason for the choice of these CREs was their interesting location in the genomic locus. They are juxtaposed and gt-1 is situated right in front of the promoter. Since several mechanisms are distance dependent, I was wondering if cross-talk between TFBS of the two CREs is operating. In particular, short-range repression can act over approximately 150 bp, and hence the spacing between CREs is an important regulatory feature to ensure their independence in the endogenous locus. I decided to create an additional reporter construct with the sequence spanning both CREs as in the genome, including the 6 bp spacer. It will be interesting to evaluate if the patterns of the separate CREs add up to the expression of the combined fragment.

4.2 Creation of transgenic fly lines via site-specific integration

The available transgenic flies (Berman et al. 2002, Schroeder et al. 2004, Ochoa-Espinosa et al. 2005) were generated by random integration of the reporter construct into the genome and hence the integration locus might have effects on the expression pattern (Markstein et al. 2008). In order to avoid position effects and integration of the vector backbone and to place all constructs at the same genomic locus for quantification, I applied site-specific integration via recombinase-mediated cassette exchange (RMCE) with ΦC31 integrase (Bateman et al. 2006) into predefined attP target lines. Thereby the cassettes can be inserted either in 3´or 5´orientation.
Initially, the idea was to use the *giant* core promoter in order to resemble the endogenous situation as closely as possible. Due to unexpected expression features and low levels (explained further below), additional controls with the *eve* basal and the *hsp70* promoter were carried out.
Furthermore, I performed negative controls in order to evaluate if sequences other than the CRE are able to activate expression. For this purpose, I injected the empty vectors without any CRE, but with the corresponding promoter and the lacZ reporter gene.

reporter construct	target line	orientations	expression pattern
gt-3 – Pgt - lacZ	37B	3´	very weak P
gt-3 – Pgt - lacZ	89B	5´	P + ectopic A
Pgt - lacZ (neg. ctrl)	37B	both	ectopic A if 5´
Pgt - lacZ (neg. ctrl)	89B	both	ectopic A if 5´
gt-3 - Peve - lacZ	37B	both	P + ectopic A if 5´
Peve - lacZ (neg. ctrl)	37B	5´	ectopic A
gt-3 - Phsp - lacZ	37B	both	P + ectopic A if 5´
Phsp - lacZ (neg. ctrl)	37B	both	ectopic A if 5´
gt-1 - Pgt - lacZ	89B	both	A + P (similar levels)
combined - Pgt - lacZ	89B	5´	A + stronger P

Table 3: Reporter-constructs injected into site-specific target lines.
The endogenous *giant* (Pgt), the *even-skipped* basal (Peve) and the heat-shock protein 70 (Phsp) promoters were tested. Plasmids containing the CREs gt-3, gt-1 and the combined fragment were injected, as well as the empty vectors without any CRE (negative control). The fly lines from Bateman et al. (2006) with integration sites at 37B (II. chromosome) and 89B (III. chromosome) were used. The integration of the cassettes can happen either in 3´or 5´orientaion. The pattern and expression levels depend on the orientation but not on the target line. The CREs drive expression in the anterior (A), posterior (P) or in both regions (A+P). In the case of gt-1, the expression levels in the anterior were similar to the levels in its posterior, whereas for the combined fragment the posterior was stronger than its anterior domain. The observed ectopic expression in the anterior (ectopic A) derives from the P-element fragments left-over in the target lines.

Target line independence, orientation dependence and ectopic expression

First, I tested which of the existing target lines are appropriate for my aims. Flies with integration sites at chromosomal positions 37B and 89B (II. and III. chromosome, respectively) were injected with a plasmid carrying gt-3 (Figure 26A) under the control of the *gt* promoter. Additionally, a negative control without CRE (Figure 27A) was injected. Unfortunately, a non-specific anterior stripe is present in the negative controls and in the lines harboring gt-3, if the orientation of the cassette is 5´ (arrows in Figure 26 and Figure 27). In contrast, if the orientation is 3´, this stripe is not visible, but the expression level from gt-3 is too low to be quantified. Therefore, I was only interested in 5´oriented insertions. The injection of the line 37B with the plasmid carrying gt-3 did not yield any 5' orientation. Hence, the remaining constructs gt-1 and gt-1-gt-3 combined were injected into 89B only.

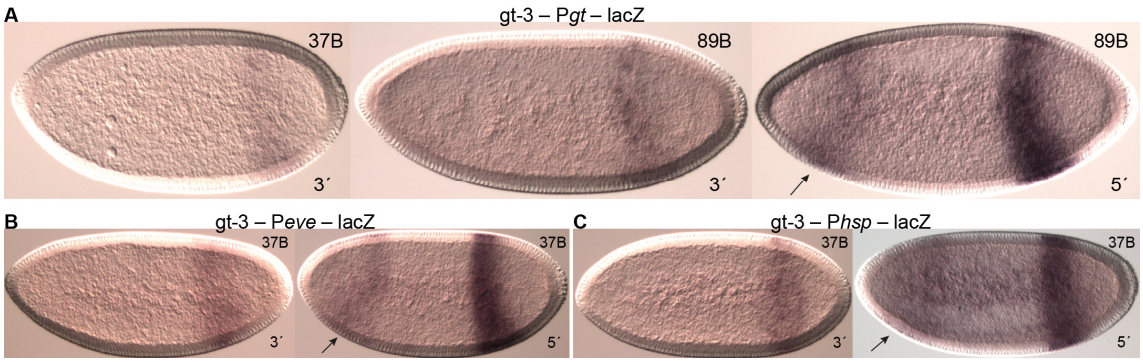


Figure 26: Expression driven by gt-3 under the control of different promoters in different target lines.
Enzymatic *in-situ* hybridizations against lacZ mRNA under the control of the endogenous *gt* core promoter **(A)**, the *eve* basal **(B)** or the *hsp70* promoter **(C)**. The reporter cassettes were integrated into the target lines 37B or 89B in 3' or 5' orientation, as indicated. Unfortunately, no 5´oriented fly line emerged from the injections of Pgt-gt-3-lacZ into 37B. Shown are DIC images of embryos at mid cycle 14 in lateral view with anterior to the left and dorsal up. Arrows indicate the ectopic anterior stripe.

In order to exclude that the low expression level in the lines with 3´orientation is due to an intrinsic property of the *gt* promoter, I injected the landing line 37B with plasmids harboring either the *eve* basal or the *hsp70* promoter (Figure 26 B, C). It turned out, that all three promoters show similar strength of expression. Nevertheless, the ectopic anterior expression could not be abolished by the usage of another promoter (Figure 27 B, C).

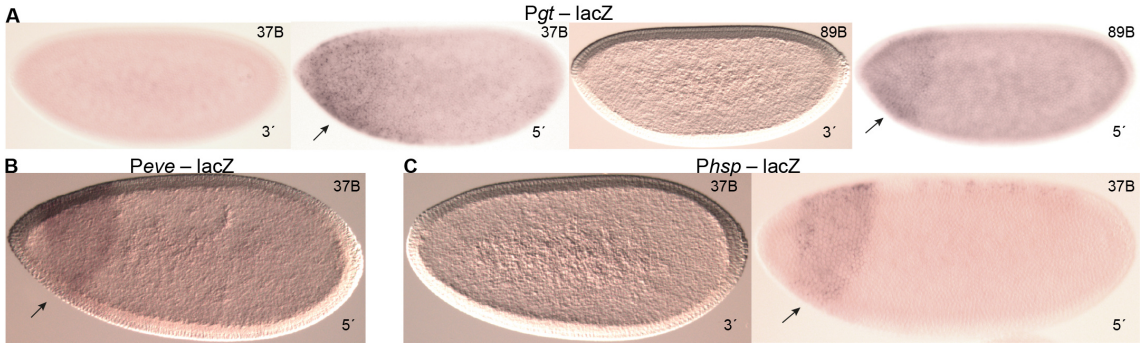


Figure 27: Negative controls carrying lacZ reporter cassettes with different promoters but without CREs, integrated into different target lines.
Enzymatic *in-situs* against lacZ mRNA under the control of the endogenous *gt* core promoter **(A)**, the *eve* basal **(B)** or the *hsp70* promoter **(C)**. The cassettes were integrated into the lines 37B or 89B in 3' or 5' orientation, as indicated. No 3´oriented fly line emerged from the injections of Peve -lacZ into 37B. The constructs drive ectopic anterior expression (arrows) depending on the orientation. Shown are BF or DIC images of embryos at mid cycle 14 in lateral view with anterior to the left and dorsal up.

Bateman et al. (2006) did not report about any differences regarding cassette orientation and other site-specific methods can integrate only in one orientation (Groth et al. 2004, Oberstein et al. 2005, Bischof et al. 2007, Pfeiffer et al. 2008). The ectopic anterior expression was mentioned previously (Small et al. 1992) and vector sequences in the P-transposon were thought to trigger it. The constructs with the *gt* CREs from the labs of Gaul and Small are based on the pCaSpeR vector for P-element transformation. The target lines from Bateman themselves were also generated by P-element transformation using the same vector backbone, which results in the integration of these left-over fragments of only 326 bp at the 3' and 372 bp at the 5' end. Interestingly, in an article from Zinzen *et al.* (Zinzen et al. 2009), this ectopic anterior expression appears in almost all of the pre-gastrulating embryos, although they used another site-specific integration method (Bischof et al. 2007). These target lines were generated with the mariner element, which shows some sequence similarity with the P-element.
In summary, the expression levels and the non-specific anterior stripe are independent of the target line and the promoter, but orientation-dependent. Based on my experimental evidences and information from the literature, I conclude that the ectopic patch is triggered by left-over sequences from the P-transposon or the mariner element used to generate the target lines.
Since gt-1 and gt-3 border each other in the genome, it was a concern to elucidate whether their expression patterns simply add up or if they interact via distance-dependent mechanism. For that purpose, a fly line was generated carrying the entire combined fragment of gt-1 and gt-3, exactly like in the *D.melanogaster* genome. The combined fragment appears to coincide with gt-1 in terms of boundary positions (Figure 28). Gt-1 drives both domains with similar expression levels, while the combined sequence triggers a posterior that is stronger compared to its anterior. In contrast to the endogenous pattern, gt-1 does not show proper refinement into two separate stripes in the anterior. This continuing leakage between the two stripes gets also reflected in the combined fragment. All the observations taken together suggest an additive behavior of the two adjacent CREs, although we cannot conclude whether their expression levels are strictly summing up.

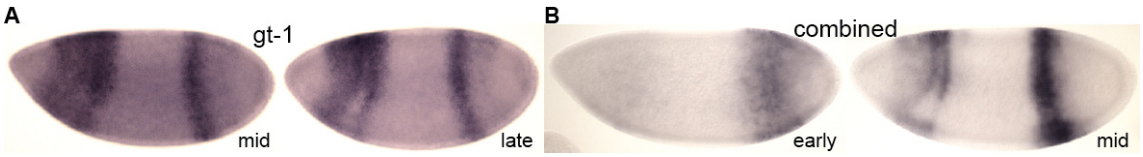


Figure 28: Expression driven by gt-1 and the combined CRE.
Enzymatic *in-situ* hybridizations against lacZ mRNA driven by gt-1 **(A)** or the combined CRE **(B)**. Both cassettes contain the endogenous *gt* core promoter and were integrated into the target line 89B in 5' orientation. Shown are BF images of embryos at early, mid or late cycle 14 (as indicated) in lateral view.

4.3 Quantitative fluorescent datasets

The generation of quantitative fluorescent datasets with high resolution in space and time was required for the model fitting. The fluorescent method allows to achieve more precise expression dynamics of the CREs, because the enzymatic *in-situ* hybridizations give a non-linear signal due to the accumulation of converted color substrate. The embryos were stained for the lacZ mRNA and additionally for *gt* mRNA, Eve protein and the nuclei. The datasets are based on at least 10 laterally oriented embryos per time point. For details about stainings and imaging, see Materials and Methods. The quantitative datasets include the cleavage cycles C12, C13 and C14, and the latter one is further subdivided into eight time classes (T1-T8). The embryos were staged by visual inspection of the Eve pattern and the membrane morphology. Figure 29 shows how highly dynamic certain expression profiles can be within C14A and hence, emphasizes the importance of such high temporal resolution.

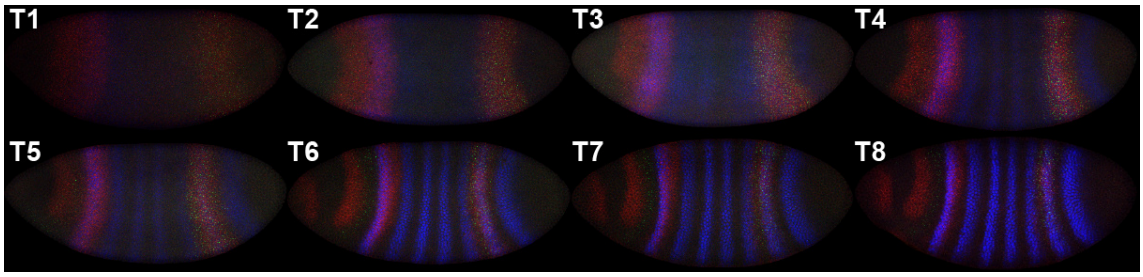


Figure 29: Expression dynamics during cleavage cycle 14A.
Merged channels of fluorescent stainings against gt-3 driven lacZ mRNA (green), endogenous *gt* mRNA (red) and Eve protein (blue) over the eight time classes (T1-T8) of C14A.

As explained before, I focused on the posterior and quantified expression driven by the CREs gt-3 and gt-1. Additionally, since it is particularly interesting that these two CREs are right next to each other, I also quantified the reporter-expression from this combined fragment as it is in the genome (Table 4).

	gt-1			**gt-3**			**gt-1 – gt-3 combined**		
time point	lacZ	*gt*	Eve	lacZ	*gt*	Eve	lacZ	*gt*	Eve
C12	10	10	10	10	8	10	15	19	19
C13	14	11	15	84	74	89	23	40	40
T1	43	41	43	87	90	90	12	15	15
T2	20	20	20	43	43	43	15	15	15
T3	23	23	23	68	68	68	17	17	17
T4	26	26	26	59	60	60	13	13	13
T5	45	43	45	91	91	91	19	19	19
T6	18	18	18	60	60	60	16	16	16
T7	23	23	23	69	69	69	14	14	14
T8	12	12	12	27	26	27	13	13	13

Table 4: Embryo counts of the quantitative fluorescent datasets.
The quantitative datasets include the cleavage cycles 12, 13 and 14, and the latter one is further subdivided into eight time classes (T1-T8). Numbers refer to integrated embryos per time point for lacZ mRNA, endogenous *gt* mRNA and Eve protein from the datasets for gt-1, gt-3 and the combined CRE.

Data processing comprises several steps, which are explained in detail in Materials and Methods. Briefly, the confocal images were cropped and aligned and an embryo mask was created. Subsequently, they were subjected to image segmentation to measure the average RNA or protein concentrations in each nucleus. Non-specific background from different antisera was removed. For the last steps, the middle 10% strip of the embryo between 45 and 55% of the DV-axis were taken into account (Figure 30).

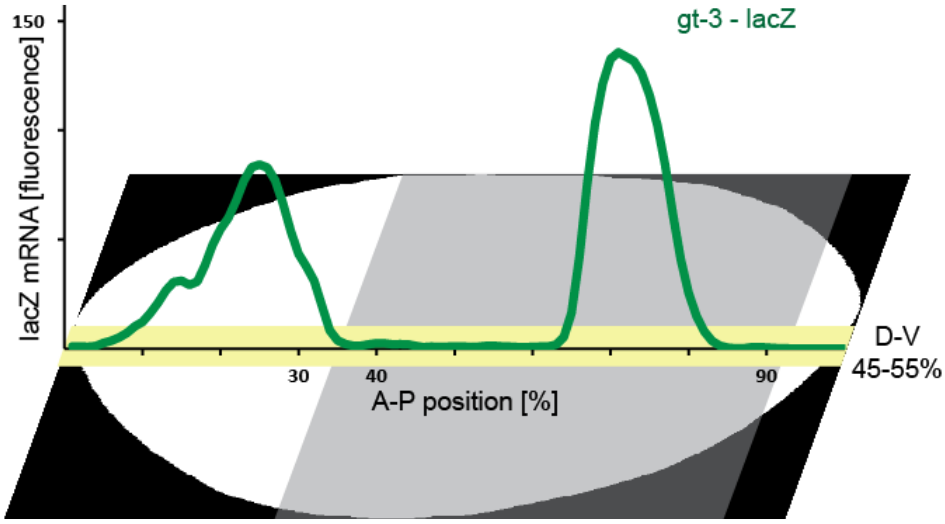


Figure 30: Extraction of 1D expression profiles from fluorescent stainings.
Embryo mask overlaid with the gt-3 expression profile at T5 (green) considering the middle 10% strip (yellow) along the dorso-ventral (D-V) axis. The fluorescence intensity is a relative measure for the lacZ mRNA. The trunk region of the embryo is indicated in grey from 35 – 92% of the A-P axis.

During data registration, individual expression features are aligned based on the Eve pattern in order to remove embryo-to-embryo variability. Finally, the data were integrated, meaning grouping the nuclei from different embryos into positional bins and calculating the average concentration.

4.3.1 Quantified endogenous *giant* mRNA

The translation of the messenger RNA into the protein takes a certain amount of time and post-transcriptional modifications might occur. This is reflected in the comparison between *gt* mRNA and its protein product (Figure 31). In the anterior, their boundary positions almost coincide, apart from the protein domain being slightly broader than the mRNA. The refining into two stripes of the protein starts approximately two time classes later and also the emergence of the anterior tip has a delay of at least one time class. The peak expression of the posterior domain shifts to the anterior in both cases, but while their anterior boundaries precisely coincide, the posterior protein boundary retracts later from the pole and subsequently trails behind the mRNA. In contrast to the posterior, their anterior expression profiles do not shift over time. We can compare the levels between time classes, but not directly between mRNA and protein levels. Nevertheless, we can clearly see that the difference between their expression levels increases over time and that at T8, the mRNA has almost disappeared, whereas the protein expression is maintained at high concentrations.

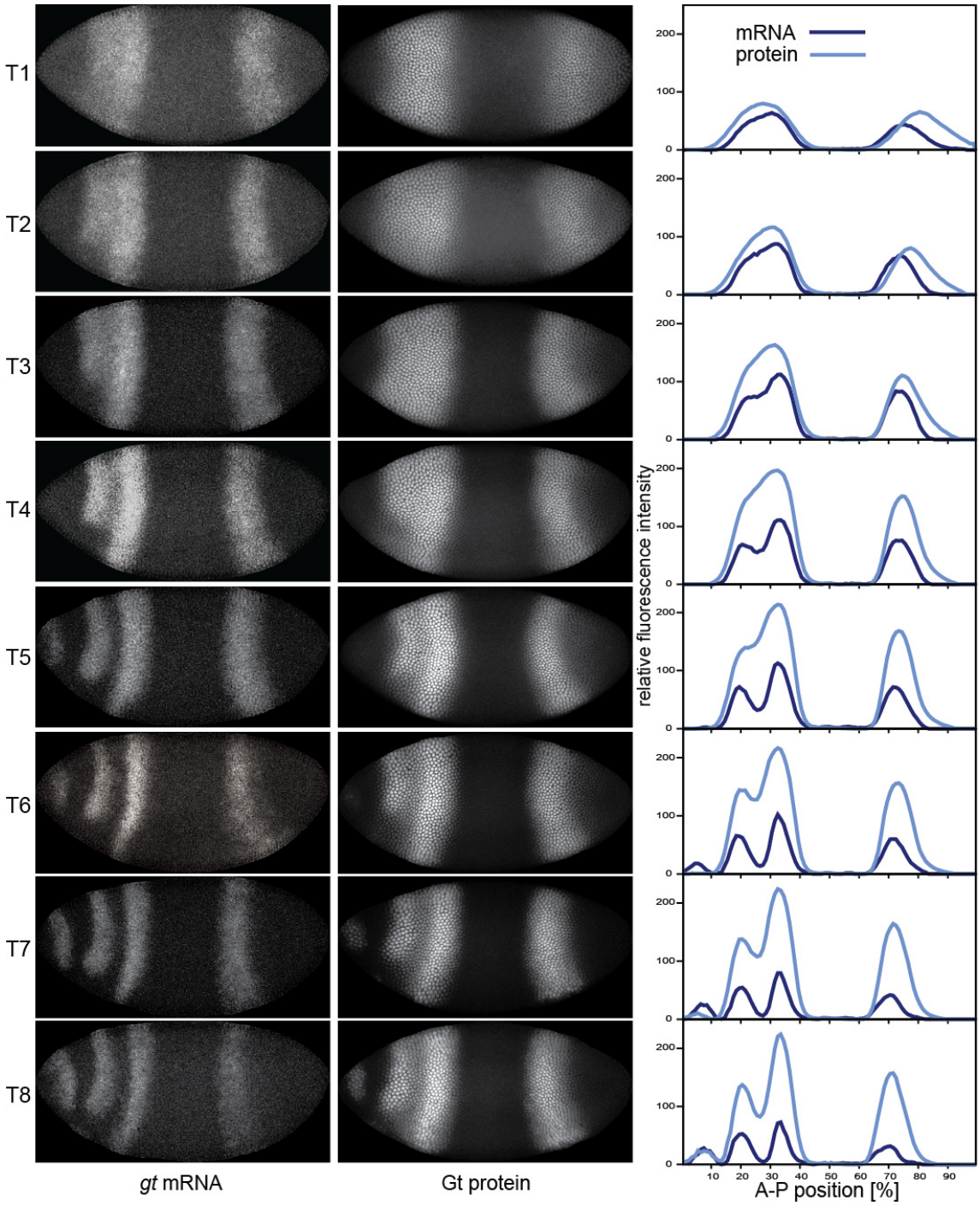


Figure 31: Expression of *gt* mRNA vs. Gt protein.
Left panel: *in-situ* hybridizations against endogenous *gt* mRNA from the gt-3 dataset and Gt protein staining from the FlyEx database over the eight time classes (T1-T8) in cleavage cycle 14A. Right panel: integrated data from at least ten lateral embryos per time class along the anterior-posterior embryo axis.

4.3.2 Modelling post-transcriptional regulation

My quantitative dataset of *gt* mRNA expression was used to fit and solve a model of post-transcriptional regulation (Becker et al. 2013) in order to estimate the protein production delay τ, which includes splicing, nuclear export and translation. τ was calculated to be approximately 2 ½ minutes, with a production rate of 0.11, negligible diffusion rates (0.01) and a decay rate of 0.11. Gt has a short open reading frame of 1780 bp and contains only one short intron of 75 bp.

That helps to keep the protein production delay short, in order to cope with fast nuclear divisions during blastoderm segmentation. This reverse engineering approach indicated that post-transcriptional regulation is not required for pattern formation in the case of Gt, Kr and Kni. It might only be required for the proper tuning of the protein levels, since translation rates should be adapted for quick accumulation at early stages and higher protein stability ensures the maintenance of high concentrations at later stages.

4.3.3 Quantified expression of gt-1, gt-3 and the combined CRE

The fluorescent lacZ mRNA expression driven by gt-1, gt-3 and the combined CRE is shown in Figure 32 and the quantified expression profiles compared to the endogenous *gt* mRNA is shown in Figure 33. The boundaries of the posterior domain of the separate CREs gt-1 and gt-3 coincide with the endogenous *gt* mRNA, whereas the combined fragment drives a slightly broader posterior stripe. It has to be kept in mind that the datasets were generated during separate time periods on different confocal microscopes with distinct lasers. Nevertheless, the posterior boundary of the anterior domain of the combined CRE overlaps with gt-1 and with the endogenous *gt* mRNA. The posterior boundary of the posterior domain formed by gt-1 is less steep than the ones from gt-3 and the combined CRE at several time classes. All three reporter-constructs, as well as the endogenous *gt* mRNA form a tiny additional peak in T7 and T8 at 80% A-P position. This subtle feature could be observed by eye in some but not all raw images and proves that the quantification procedure is actually capable of picking up such details.

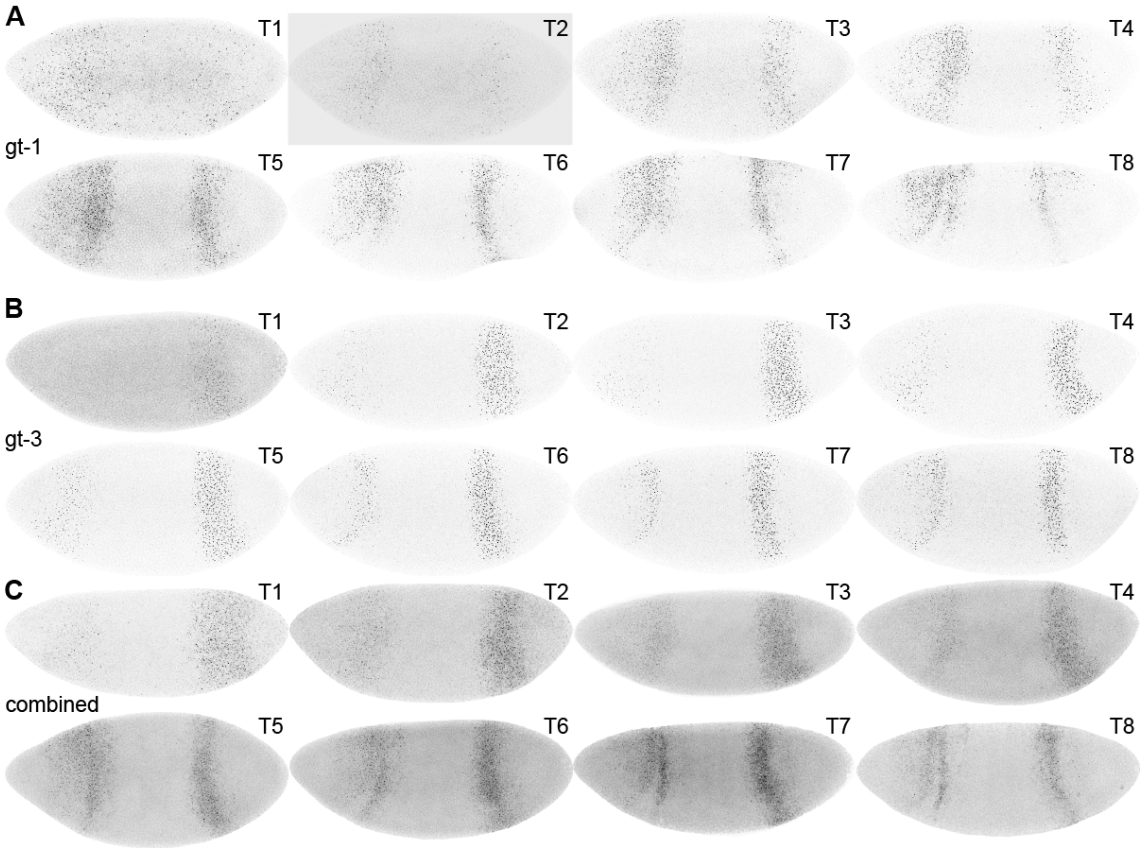


Figure 32: mRNA expression driven by gt-1, gt-3 and the combined CRE.
Fluorescent *in-situ* hybridizations against lacZ mRNA over the eight time classes of C14A. The reporter gene is driven by gt-1 (A), gt-3 (B) and the combined CRE gt-1-gt-3 (C). The confocal images were inverted.

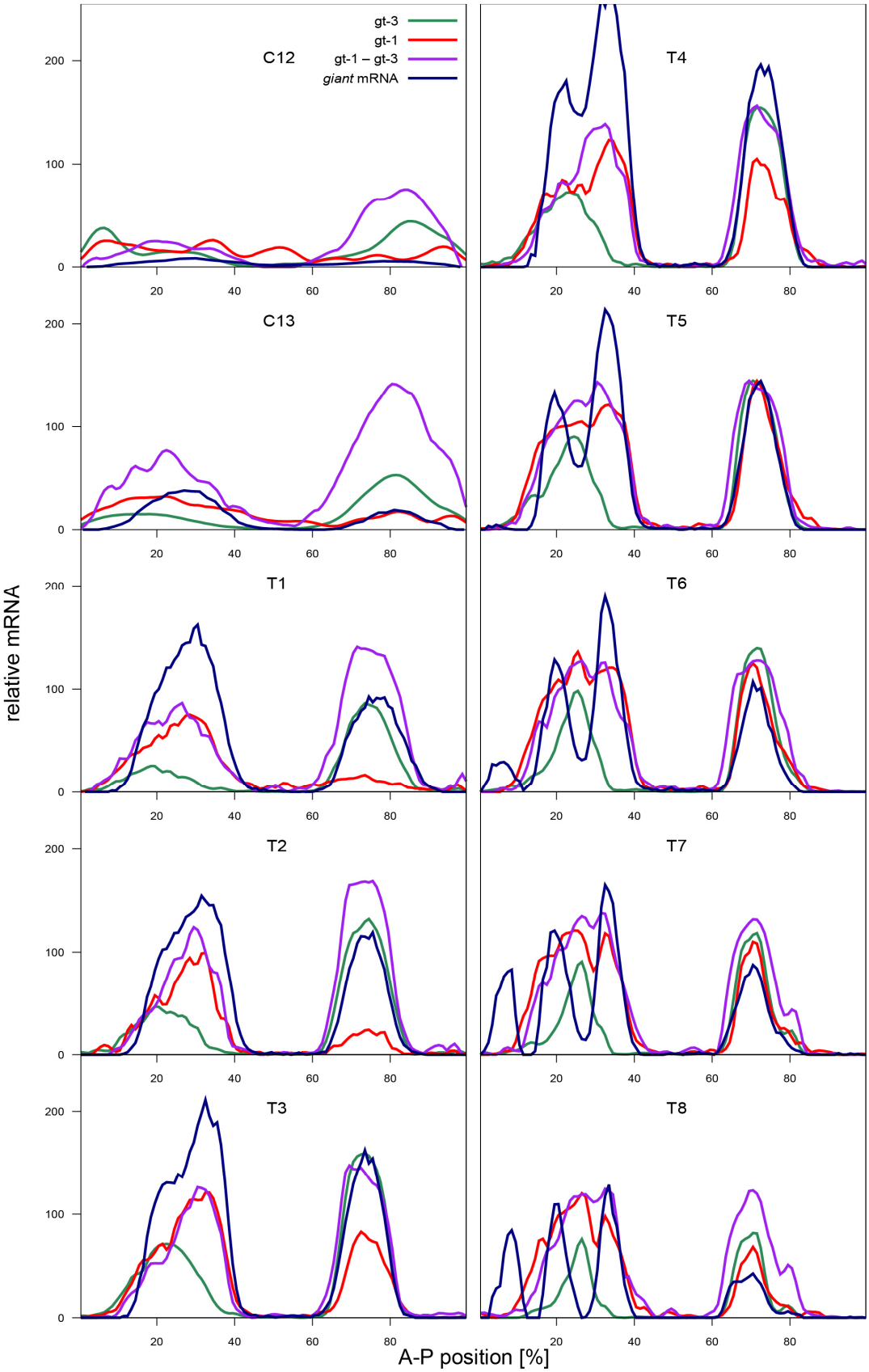


Figure 33: Expression profiles of gt-1, gt-3 and the combined CRE compared to *gt* mRNA.
Integrated and smoothed datasets from fluorescent *in-situ* hybridizations against lacZ or endogenous *gt* mRNA at cleavage cycles 12 and 13, as well as over the eight time classes (T1-T8) of C14A.

Neither gt-1 nor the combined fragment achieves the separation in-between the two anterior stripes and their anterior boundary (of the anterior domain) seems to be less sharp than the one of endogenous *gt*. Both observations could be an artefact from the ectopic vector expression.
Gt-3 and the combined fragment are expressed at C12 already, whereas gt-1 only gives a few isolated dots without forming any defined domains and hence, it was set to zero for the modelling. Gt-1 emerges at C13 and only slowly increases until finally reaching its peak expression at T5. It consistently stays at much lower levels than gt-3 or the combined fragment.

Limitations of the dataset
Since it is not possible to measure absolute concentrations via FISH and subsequent image processing, we can only compare boundary positions, but not the levels of mRNA or protein expression. In the case of lacZ mRNA this drawback is a bit more pronounced due to its punctate accumulation (Figure 32). The group of David Arnosti (Sayal et al. 2010) claimed that the appearance of lacZ depends on the 5´-UTR after testing several constructs with different promoters, 5´ and 3´-UTRs. In particular, they observed that with a construct containing the *twist* core promoter fused to the *eve* 5´-UTR, a diffuse lacZ pattern could be achieved. On the other hand, I observed a diffuse distribution of the endogenous *gt* mRNA, which was easily quantified. In fact, I tried to reconstruct this endogenous situation as far as possible in my reporters, by using the *gt* core promoter and its 5´-UTR as it is in the genome. Hence, if the distribution depended exclusively on the 5´-UTR as claimed by Sayal et al. (2010), then my approach should have worked. Nevertheless, I was forced to include a thresholding step in the quantification procedure in order to yield higher signals in the processed data. Only the pixel values above the threshold were considered for the calculation of the average intensity in the nucleus and its surrounding cytoplasm (see Materials and methods section 6.3 for details). Thereby the levels at the very early stages (C12 and C13) might have been slightly overestimated and the ectopic anterior expression was artificially emphasized in all time classes (Figure 33). Unfortunately, the information that the posterior of the combined fragment is much stronger compared to its anterior over all time classes, as seen in the enzymatic *in-situ* hybridizations (Figure 28), was lost during the quantification of the fluorescent *in-situs* due to the thresholding step.

For the modelling, we additionally need the concentrations of the regulating TFs, which are only available for a certain subset of them (Pisarev et al. 2009) and the head gap genes are not included, for example. Based on these restrictions, we can only model the trunk region of the embryo from 35 – 92% A-P position (Figure 30). This still includes the posterior boundary of the anterior *gt* domain. The ectopic anterior expression is only fading out towards 35% and hence negligible and was set to zero in the case of gt-3. In general, the fluorescent datasets were cleaned up by setting the values of the non-expressing regions in the middle of the embryo to zero, as well as the expression of gt-1 in C12. This post-processing of the data is recommended to avoid that the model is fit to noise. Nevertheless, no artefacts were observed in the model outputs even when this step was omitted.

4.4 Fitting the model to expression data from CREs in a wild-type background

The quantitative fluorescent datasets with high spatio-temporal resolution of the lacZ mRNA driven by the different CREs were used to fit the model of transcriptional regulation. The model serves as a tool to explore different regulatory mechanisms and possible scenarios. It keeps track of the fractional occupancy of each activator and quencher at each position along the A-P axis and on the DNA sequence. In particular, I tested several hypotheses concerning the activation of gt-1 by including the mechanisms of Bcd cooperativity and Hb co-activation. Additionally, I tried to elucidate the role of the TF Giant on its own regulation. Afterwards, the optimized parameters of the model were used for predicting the expression of the CREs in mutants.

4.4.1 Settings in the model

As explained before, the trunk region of the embryo from 32% or 35% until 92% of the A-P position was considered. Normally, the model was fit to all 10 time points, except for some controls, where either only the very early stages or the eight time classes of C14 were regarded. Since fitting with weighted least squares (WLS) based on the standard deviations of the data did not improve the output, ordinary least squares (OLS) were used for most of the runs. In general, all the relevant transcription factors (Bicoid, Caudal, Hunchback, Tailless, Knirps, Huckebein, Giant, Krüppel) were included apart from some control runs, where certain regulators were omitted. The same combination of PWMs as in Kim *et al.* (2013) were used, except for some runs, which were carried out exclusively with B1H matrices[7] for all TFs (Noyes et al. 2008). The PWM thresholds were either adjusted within the same limits for all TFs or fixed for Bcd (1.71) and Hb (0.63) to the same values as in Kim *et al.* 2013. Bcd and Cad were defined *a priori* as activators, and Kr, Kni, Tll and Hkb as repressors. Since Hb has a bimodal role in the regulation of *eve*, the possibility of a similar differential behavior on *gt* was tested with the model by considering it either as a repressor or by allowing for co-activation by Bcd and Cad. In a similar manner, the hypothesis of cooperativity between nearby Bcd sites and direct repression of the BTM were evaluated (see Chapter 2 for details and values). The mode of action of Gt on itself was previously unknown and hence, was explored with the help of the model via repeating each run with Gt either turned off, set as a repressor or as an activator.

The model was fit either to one CRE dataset per time or to two or three CREs at once. Ideally, one would fit to as much data as possible. The reason for fitting to one or multiple datasets was that I suspected that gt-1 and gt-3 are regulated differently based on the preliminary results. In particular, due to the discrepancy between early and late regulation and the role of Hb and Gt itself, it might not be possible to achieve a satisfactory fit of the model to all datasets simultaneously.

In total, 164 different input combinations were tested and each of them was repeated with random initial conditions at least thrice and up to 10 times, depending on their importance. See Table 5 for the corresponding input combinations of the runs presented in this section, and the Appendix for a table with all optimization runs (Table A. 13).

The model outputs the relative mRNA concentration driven by the corresponding input CRE sequence, as well as the final activation energy decrease after considering direct repression and the fractional occupancy of the quenchers. The quality of the output was judged by visual observation of the expression pattern and the contribution of the TFs, since the RMS scores were not sufficiently distinct and hence not informative.

[7] http://pgfe.umassmed.edu/ffs

combi	run	CRE	Giant	coop&coact	TFs	threshold	AP %
25	1	gt-3	A	no	only Gt	B, H fixed	35-92
26	1	gt-1	A	no	only Gt	B, H fixed	35-92
27	1	gt-3	A	no	no B, C	B, H fixed	35-92
28	1	gt-1	A	no	no B, C	B, H fixed	35-92
52	2	gt-3, gt-1	off	no	all	all fitted	35-92
53	3	gt-3, gt-1	A	no	all	all fitted	35-92
54	1	gt-3, gt-1	R	no	all	all fitted	35-92
90	1	combined	off	no	all	all fitted	31-92
91	1	combined	A	no	all	all fitted	31-92
92	2	combined	R	no	all	all fitted	31-92
93	3	combined	off	yes	all	all fitted	31-92
114	1	all 3	off	no	all	all fitted	31-92
115	1	all 3	A	no	all	all fitted	31-92
132	1	gt-3	off	no	all	all fitted	31-92
133	2	gt-3	A	no	all	all fitted	31-92
134	1	gt-3	R	no	all	all fitted	31-92
135	2	gt-3	off	yes	all	all fitted	31-92
138	2	gt-1	off	no	all	all fitted	31-92
139	1, 2	gt-1	A	no	all	all fitted	31-92
140	2	gt-1	R	no	all	all fitted	31-92
141	1	gt-1	off	yes	all	all fitted	31-92

Table 5: Summary of selected models.
The first column shows the number of the input combination (combi) and the second column the number of the corresponding run shown in the figures. The models were fitted either to one CRE per time or two or three simultaneously. Gt protein was either set as an activator (A) or repressor (R) or turned off (off). The mechanisms of Bcd cooperativity and Hb co-activation (coop&coact) can be considered or not. For the control runs (25-28), certain transcription factors were excluded. The same PWMs as in Kim et al. (2013) were used. The threshold for each PWM can either be fixed or adjusted within certain limits during the fitting procedure. The region from 32 or 35 until 92 % of the A-P position was modeled.

4.4.2 Fitting to one CRE per time

Fitting to gt-3
A minimal model fit to gt-3 without considering Gt auto-regulation nor cooperativity, co-activation or direct repression, is able to resemble the observed posterior domain including the shift towards the anterior over time (Figure 34). The output reflects the boundary positions quite precisely, but there is an issue with the levels, which are too high in C12 and C13 and too low at later time classes. As expected, the model predicts activation by Cad and repression by Kr in the middle region, as well as by Tll and Hkb from the posterior pole. The model also suggests that, Hb coming up in the posterior, provokes the shift of the *gt* domain towards the anterior. The role of anterior Hb is clearly to avoid leakage in the entire region that it covers. Maternal Hb is probably also involved in setting the anterior boundary of the posterior *gt* domain at very early stages (C11 or earlier), before the emergence of zygotic Kr. Since it was very difficult to quantify such low expression levels of the dotty lacZ mRNA, these time points were not included in the model, although expression was monitored at C11 (data not shown).

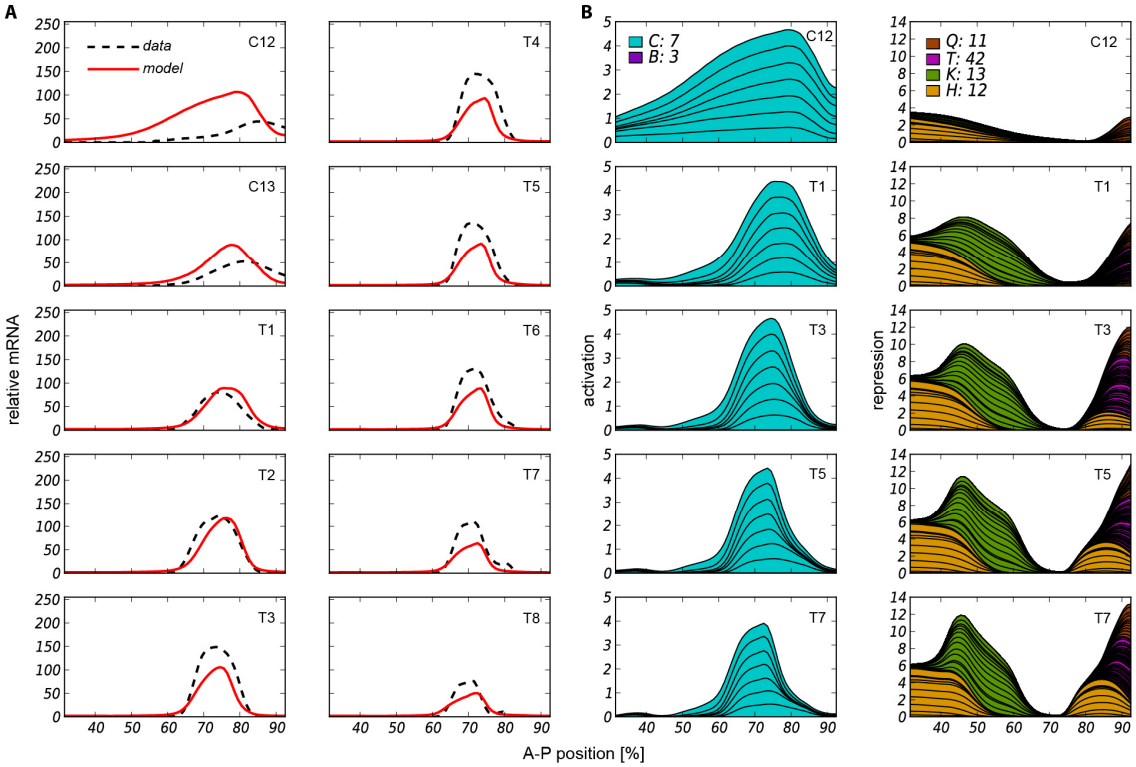


Figure 34: Minimal model fit to gt-3.
Output from run 132_1 without considering Gt auto-regulation nor cooperativity, co-activation or direct repression. **(A)** Model fits showing the relative mRNA concentration (in a.u.) of the model (red) versus the observed expression data (dashed black line) over all ten time points. **(B)** Regulatory contributions of the TFs in cleavage cycle C12 and in time classes T1, T3, T5, and T7 of C14. Activation is the final activation energy decrease after considering all possible mechanisms of the model (in a.u.). Repression shows the fractional occupancy of the quenchers (in a.u.). Each additional regulator is plotted starting from the last TF and not from the x-axis. Each black line within the region of the same color stands for one binding site and the total number of sites for each TF is indicated in the legend. Note that the scale of the y-axis of these plots varies between different model outputs and that these values are relative numbers.

Models including Gt as a repressor (Figure 35A), give almost exactly the same fits and TF contributions as without Gt. The reason is that it is very unlikely to achieve a good fit having Gt repressing itself. Hence, the model falls into solutions without this mechanism.
When fitted excluding Gt or considering it as a repressor (runs 132 and 134), the model tends to overshoot in C12 and C13 and to underestimate the levels of later time points. Gt auto-activation equilibrates this discrepancy and Cad remains as a prominent activator (run 133, Figure 35B). The model suggests that Hb represses much stronger than Kr in the middle of the embryo. Nevertheless, Kr is still responsible for setting the anterior boundary of the gt-3 domain. Additionally, the model predicts repression by Kni (all runs of combination 133).
I tested the influence of Bcd-cooperativity and co-activation of Hb by Bcd and Cad with the model. When including these mechanisms for gt-3, the model predicts only minor or no activating contributions from Bcd or Hb binding sites (Figure 35C).

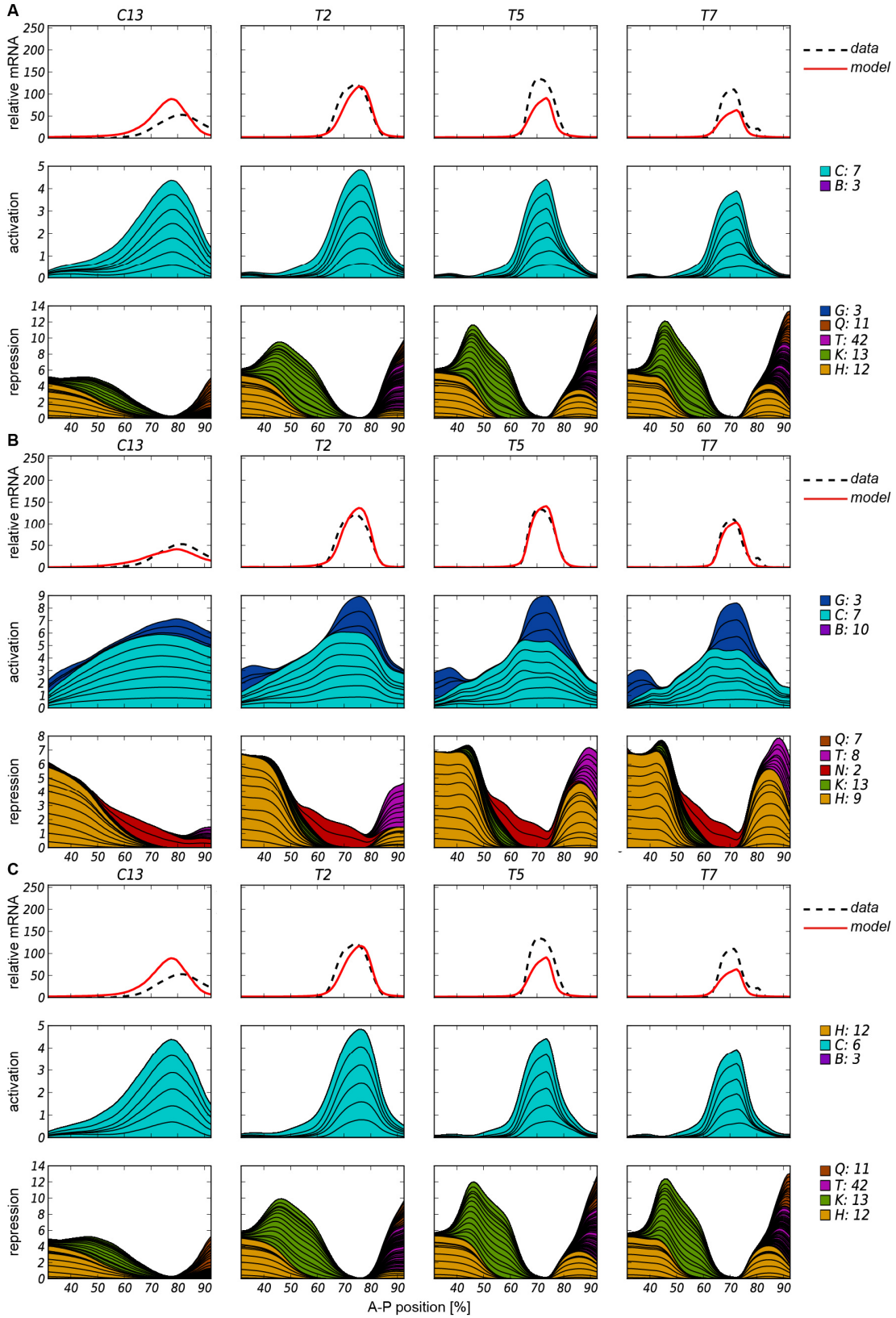


Figure 35: Alternative models for gt-3.
(A) Output from run 134_1 considering Gt as repressor, **(B)** run 133_2 considering Gt as activator or **(C)** run 135_2, without Gt auto-regulation but including co-activation and cooperativity. Shown are the relative mRNA conc. of the model (red) vs. the observed data (dashed black line) and the regulatory contributions at four time points (C13, T2, T5, T7). For detailed explanation of the graphs see Figure 34.

Fitting to gt-1

A model fit to gt-1 without considering Gt auto-regulation, achieves both domains over all time classes (run 138, Figure 36). It has to be kept in mind that only the trunk region of the embryo and therefore only the posterior boundary of the anterior domain can be modeled. The boundary positions are not very precise and the levels are too low from T3 until T7. On the other hand, the model gives quite some ubiquitous expression at C12 (not shown), even though this data point was set to zero.

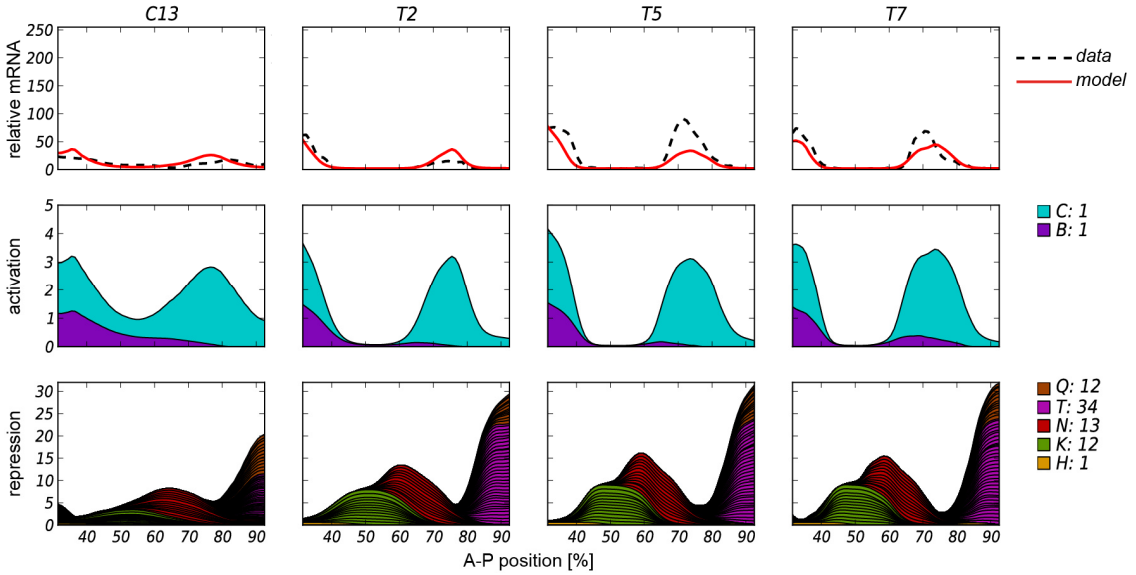


Figure 36: Model fit to gt-1, without considering Gt auto-regulation.
Output from a model excluding input from Gt (run 138_2). Relative mRNA concentration of the model (red) versus the observed expression data (dashed black line) and regulatory contributions at four time points (C13, T2, T5, T7). For detailed explanation of the graphs see Figure 34.

Although the model predicts activation by Bcd and Cad, it only finds one binding site for each of them. As expected, repression by Hb is insignificant, Kr represses in the middle region, and Tll and Hkb from the posterior pole. The model also suggests input from Kni, which is less likely. When Gt is included as a repressor, the model finds almost the same solution as without Gt, with the only difference that now the model predicts a lot of Gt sites, which are explicitly allowed to auto-repress (data not shown).
Like in gt-3, the issue with the expression levels becomes solved by considering Gt auto-activation in gt-1 (run 139, Figure 37). Nevertheless, the big difference between the two enhancers is that in gt-3 the model still predicts Cad to be the main activating factor and only minor input from Gt, whereas in gt-1 its contribution is at least 50%. Usually, the model falls into very similar solutions in consecutive runs from the same combination of inputs. Interestingly, when considering auto-activation in gt-1, the outputs show the same quality of fit, but more diverse possibilities in terms of regulatory contributions are found (Figure 37). In particular, the ratio of Cad versus Gt activation and the identity of the repressors differ. Run 139_1 gives almost the same emphasis on the contribution from a single Cad site like on the 3 Gt sites (Figure 37A). The posterior domain receives very different repressive strengths from the two sites by Kni and Tll. The posterior boundary of the anterior domain seems to be set by Gt itself. In contrast, run 139_2 suggests relatively much more auto-activation and obtains repression from Kr and Tll with similar strengths, as well as minor input from Hb (Figure 37B). Based on these observations run 139_2 appears to be the more plausible output.

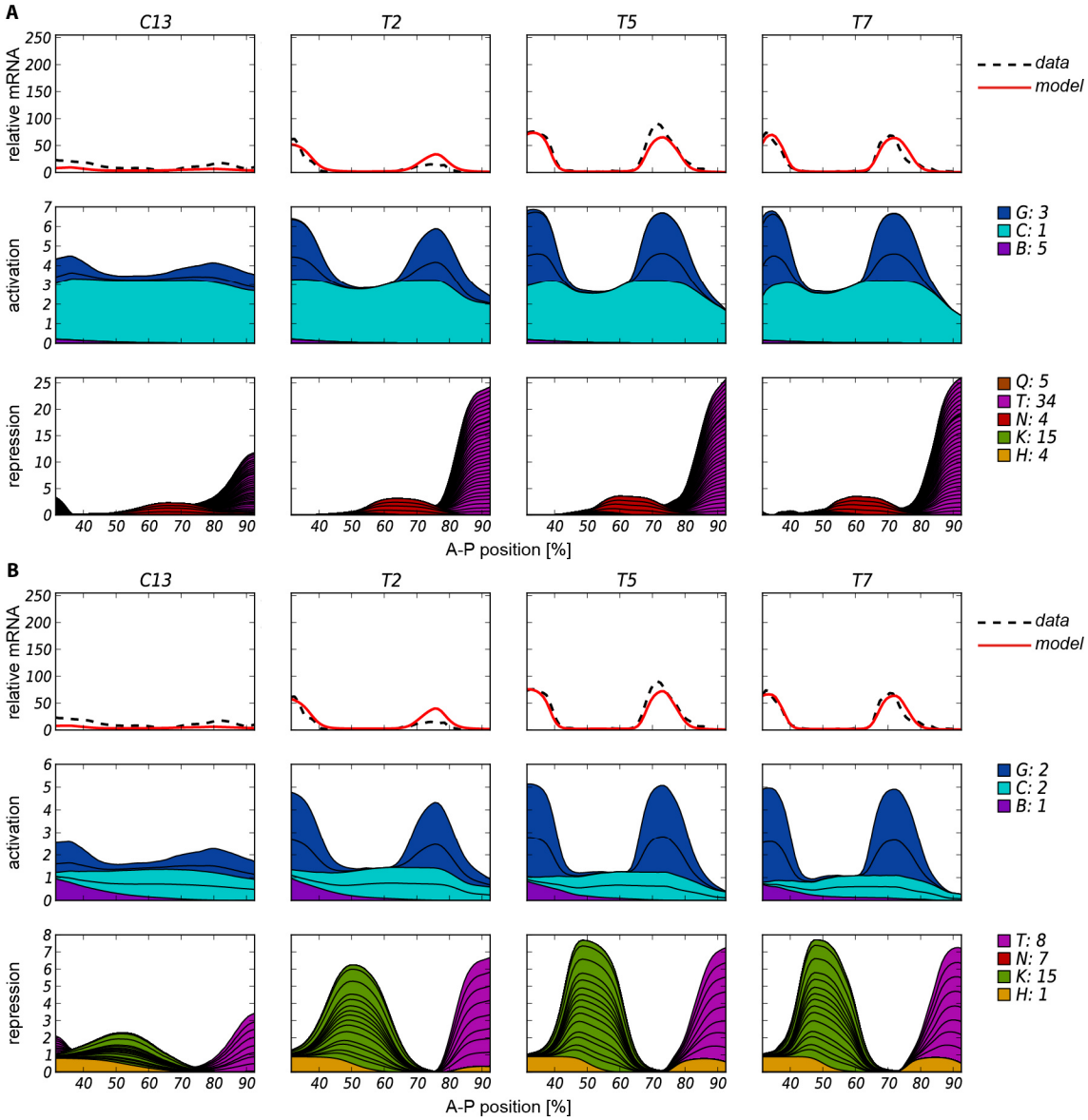


Figure 37: Model fit to gt-1, considering Gt as an activator.
Model fits from run 139_1 **(A)** and 139_2 **(B)**, showing the relative mRNA concentration of the model (red) versus the observed expression data (dashed black line) and the regulatory contributions at four time points (C13, T2, T5, T7). For detailed explanation of the graphs see Figure 34.

Bcd-cooperativity or co-activation of Hb are not capable of achieving the same quality of fit as auto-activation for gt-1 (Figure 38, run 141). Although somewhat higher influence from Bcd was observed in other runs including cooperative binding (but excluding co-activation, data not shown), there will always be a conflict with the fact that gt-1 is a late element. When fitting gt-1 with both mechanisms at once, the model still does not show activation from Bcd but allows for a bit of co-activation. The main activation comes from only one Cad site, which is rather unlikely.

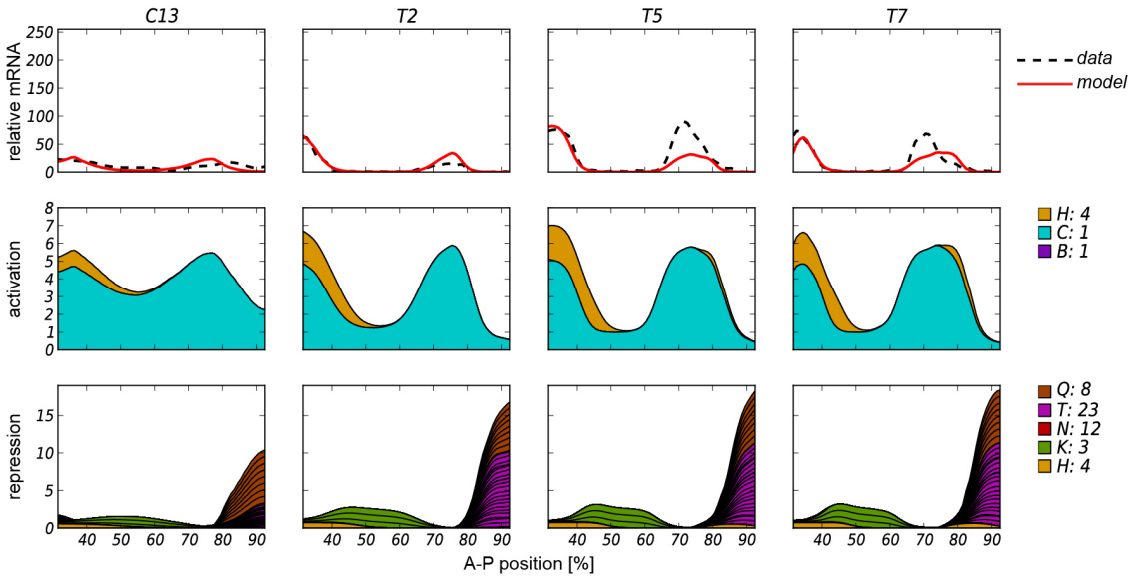


Figure 38: Model fit to gt-1, without Gt auto-regulation but including co-activation and cooperativity.
Model fits from run 141_2 showing the relative mRNA concentration of the model (red) versus the observed expression data (dashed black line) and the regulatory contributions at four time points (C13, T2, T5, T7). For detailed explanation of the graphs see Figure 34.

Fitting to the combined CRE

For the combined fragment gt-1-gt-3, the model gives almost the same output when Gt is considered a repressor (Figure 39B, run 92_2) as when it is turned off (Figure 39A, run 90_1), because it does not fall into a solution where auto-repression is permitted. In both cases, no anterior expression is predicted by the model. This suggests that the anterior, presumable driven by the gt-1 fragment of the combined sequence, cannot be activated by Bcd alone, or that it is overwhelmed by Hb repression from TFBS in gt-3.

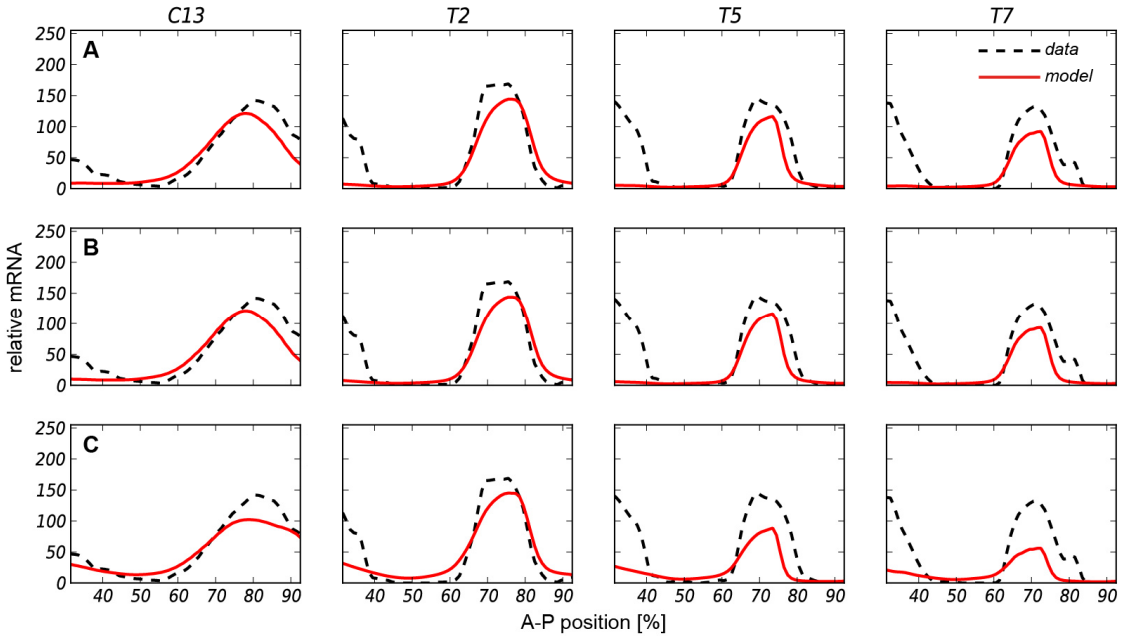


Figure 39: Models fit to the combined CRE without considering Gt auto-activation.
(A) Output, without Gt auto-regulation (run 90_1), **(B)** considering Gt as a repressor (run 92_2), or **(C)** without auto-regulation, but including co-activation and cooperativity (run 93_3). Shown is the relative mRNA conc. (in a.u.) of the model (red) vs. the observed data (dashed black line) at C13, T2, T5 and T7.

Considering the same input-combinations, but including co-activation and cooperativity, the model tries to compensate by predicting activation by Hb in the anterior, and therefore a little bit of expression appears, but also derepression in the middle of the embryo (Figure 39C, run 93_3). Additionally, there are some defects regarding the boundary positions and levels of the posterior domain. The problem of the missing anterior domain gets solved via including Gt auto-activation in the model (Figure 40). The output fits all boundaries with the desired precision, but the levels are slightly lower for the anterior and slightly higher for the posterior, resulting in a clear difference between the two domains. The model suggests minor contribution from Bcd, some Cad and quite some auto-activation. Repressing inputs come from Kr, Hb, Tll and insignificant influence from Kni in the region of the posterior *gt* domain.

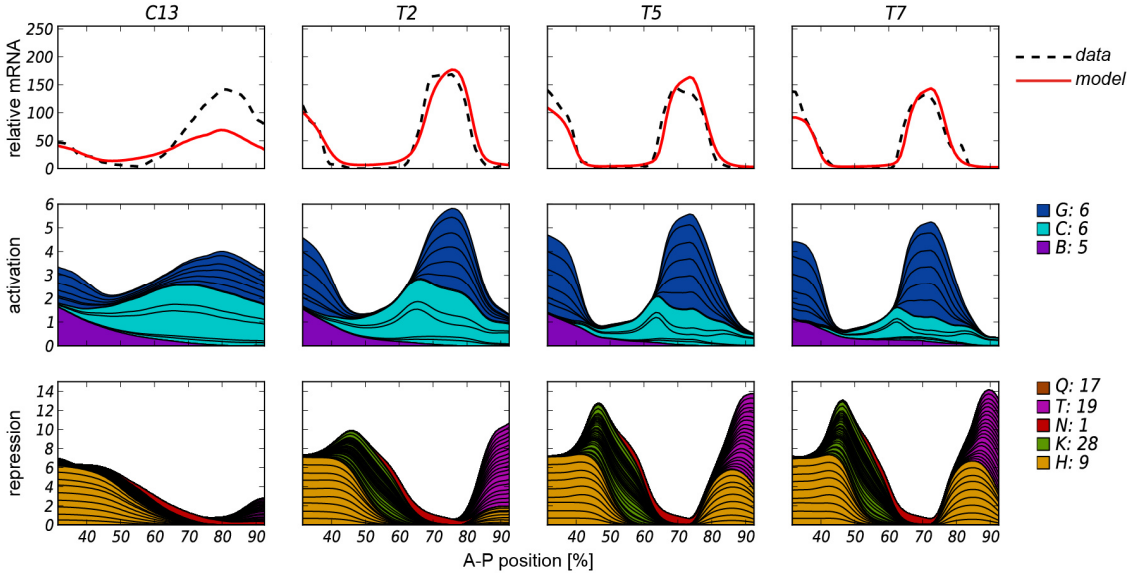


Figure 40: Model fit to the combined CRE, considering Gt as an activator.
Model fits from run 91_1 showing the relative mRNA conc. of the model (red) vs. the observed data (dashed black line) and the regulatory contributions at four time points (C13, T2, T5, T7).

4.4.3 Fitting to multiple datasets simultaneously

I also performed optimization runs with two or three datasets at once. Fitting to gt-1 and gt-3 simultaneously achieves similar quality of fit and TF contributions as with the separate optimization, depending on the role of Gt (Figure 41).

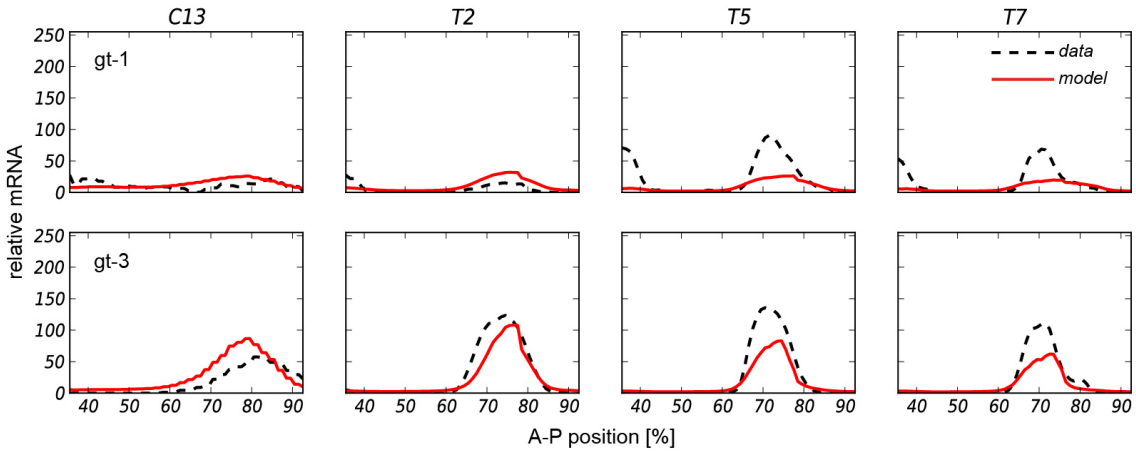


Figure 41: Model fit to gt-1 and gt-3 simultaneously without Gt auto-regulation.
Model fits from run 52_2 showing the relative mRNA concentration (in a.u.) of the model (red) versus the observed expression data (dashed black line) at four time points (C13, T2, T5, T7).

The main difference is that the model predicts much more Gt auto-activation for gt-3 (Figure 42). In this particular run, the activation is triggered by three Gt sites only but overwhelms the 30 predicted Cad sites. Such a scenario is unrealistic and rather an artefact of the model trying to find a suitable solution for both CREs.

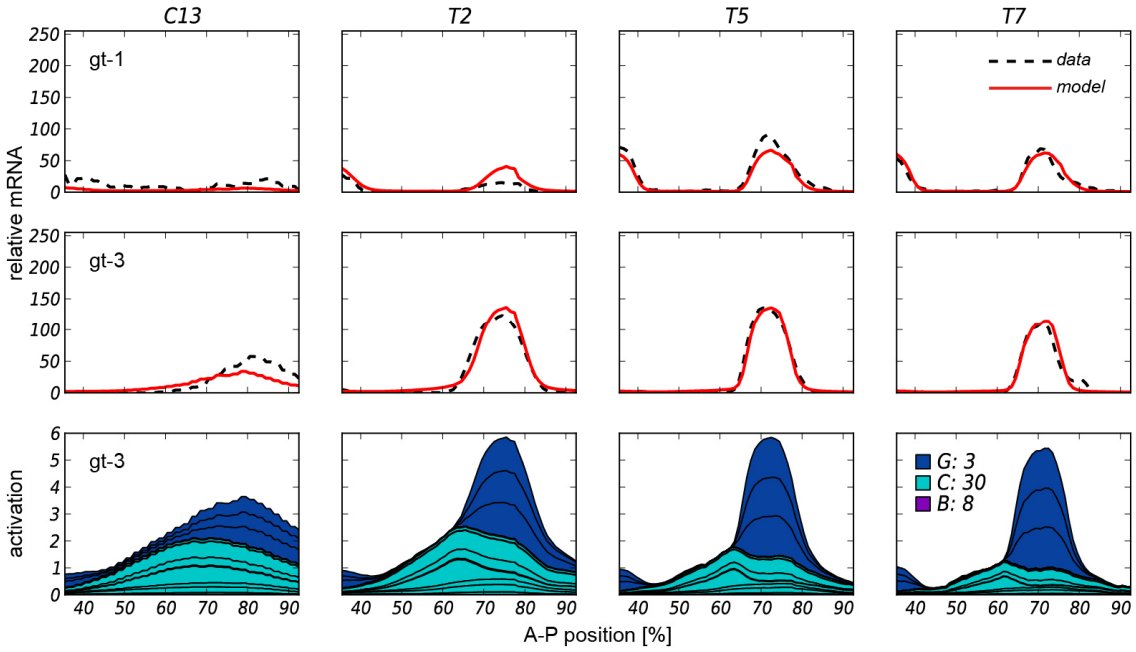


Figure 42: Model fit to gt-1 and gt-3 simultaneously, considering Gt as an activator.
Model fits from run 53_3 showing the relative mRNA concentration of the model (red) versus the observed expression data (dashed black line) for gt-1 and gt-3 at four time points (C13, T2, T5, T7). The regulatory contributions are shown for gt-3. For detailed explanation of the graphs see Figure 34.

Models fitted to all three datasets at once without considering auto-regulation were not able to find an acceptable compromise for all three patterns. They result in no expression at all in any time class for gt-1 and also the anterior domain of the combined CRE is missing, whereas gt-3 looks reasonable (Figure 43A, run 114). Setting Gt as a repressor gives basically the same output, since the prediction does not allow for auto-repression (data not shown). After including auto-activation, the model fits for the two individual CREs are almost perfect (Figure 43B, run 115). For the combined CRE, the model comes to the same conclusion as when fit individually: it is impossible that it drives both domains with the same intensity, instead the posterior must be stronger compared to the anterior. The posterior now almost reaches the maximum rate and the difference in expression levels between the two domains is huge, while it was much less pronounced in the individual optimization. On the other hand, the model suggests a huge influence from Gt, only slight activation by Cad and repression by Kni (data not shown), which is a biologically unlikely combination of contributions.

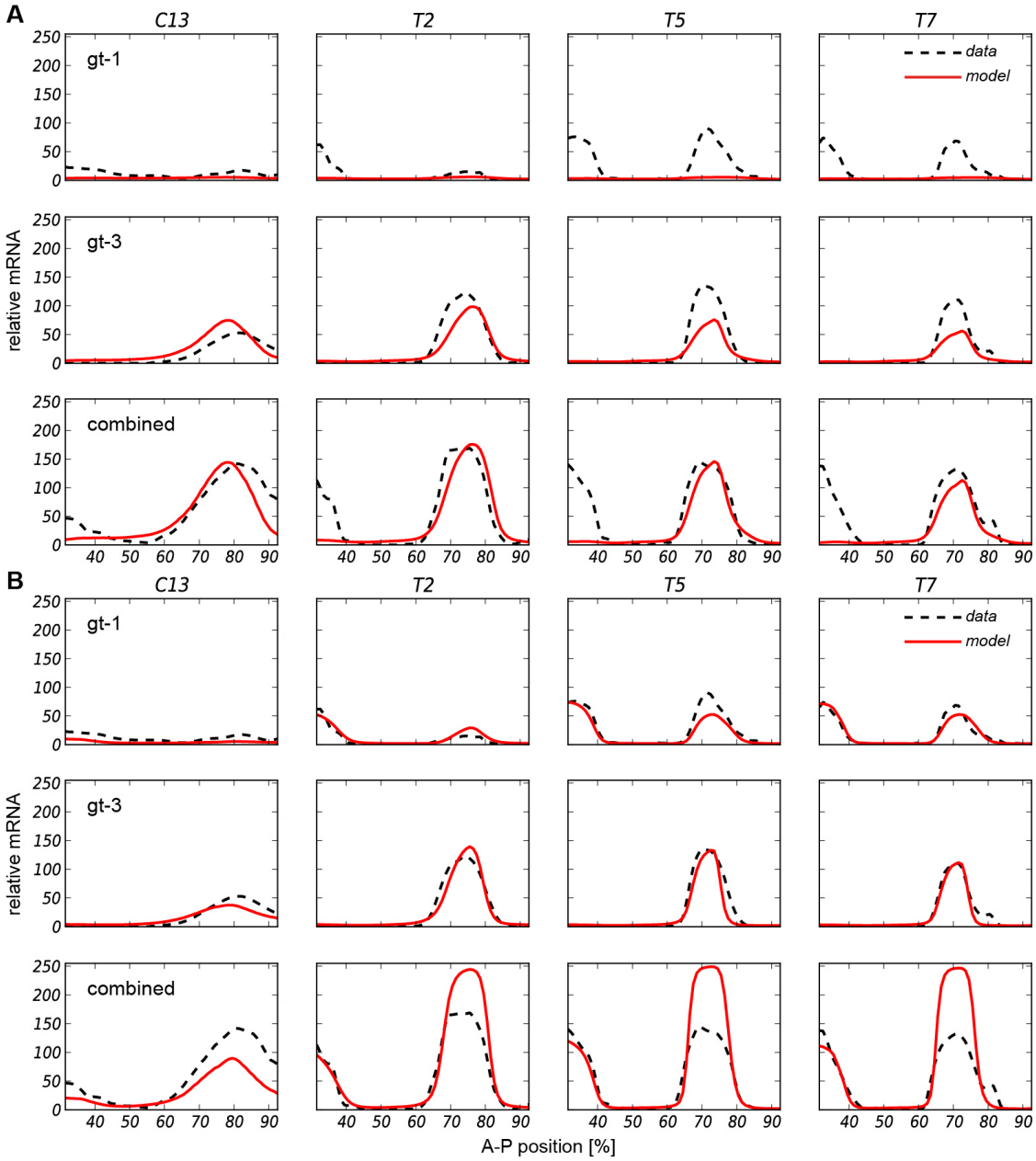


Figure 43: Model fit to all three datasets simultaneously.
(A) Model without Gt auto-regulation (run 114_1), and **(B)** model with auto-activation (run 115_1). Shown is the relative mRNA concentration (in a.u.) of the model (red) versus the observed expression data (dashed black line) for gt-1, gt-3 and the combined CRE at four time points (C13, T2, T5, T7).

4.4.4 Insights from the fitting procedure

The trend that the model falls into similar solutions starting from randomized initial parameters is a sign that the simulated annealing algorithm is exhaustive and capable to find the global minimum. As shown above, this model does not simply fit everything and the rejected, as well as the proposed mechanisms, seem reasonable. In general, the models set the boundary positions correctly and no ectopic expression was observed, although in some cases entire domains were missed. Figure 44 shows all the parameter values of the selected solutions.

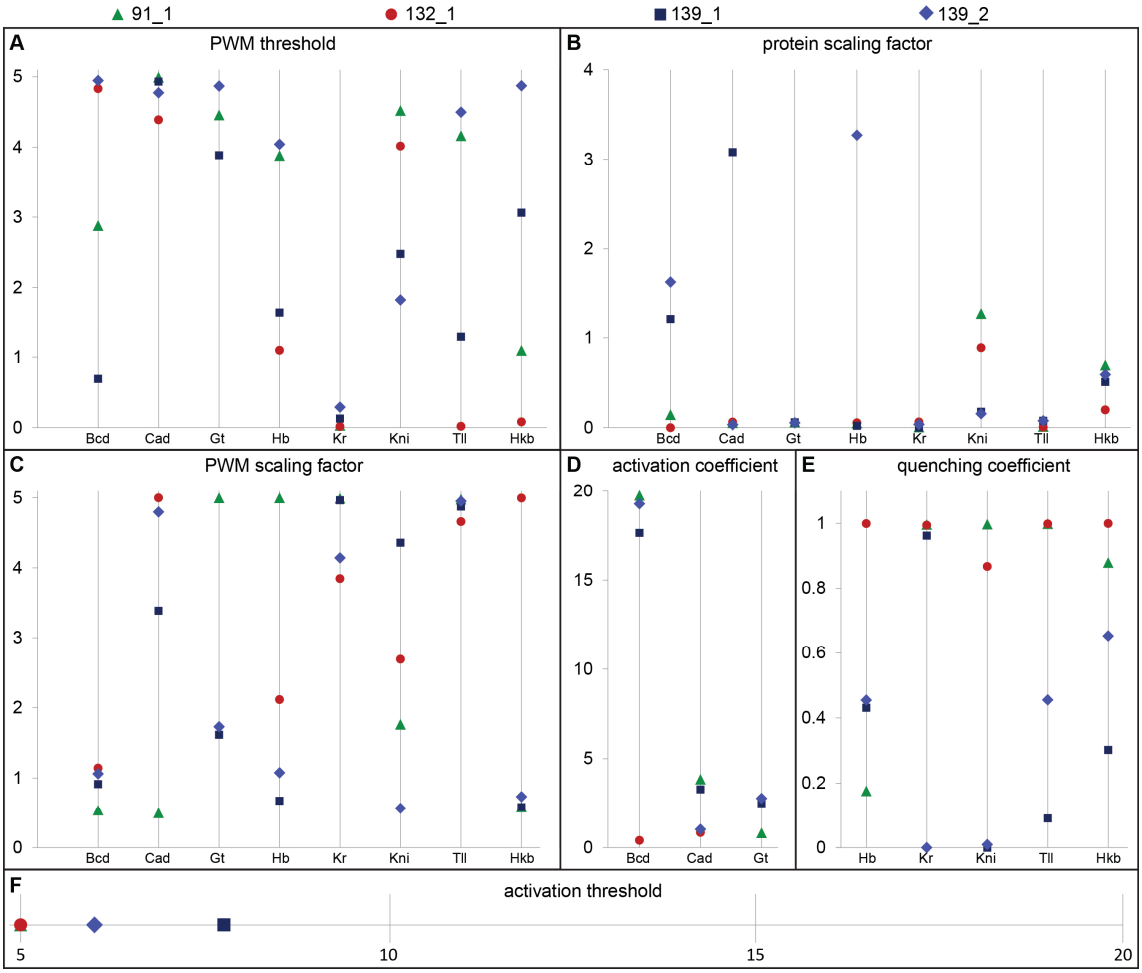


Figure 44: Parameter values of selected models.
The selected models were fit to the combined CRE with auto-activation (91_1, green triangle), to gt-3 without considering Gt (132_1, red circle) and to gt-1 with auto-activation (139_1, dark blue square and 139_2, light blue rhombus, as indicated at the top). Shown are the PWM threshold T_a **(A)**, the protein scaling factor A_a **(B)** and the PWM scaling factor λ_a **(C)** for each TF, as well as the activation coefficient E^A for Bcd, Cad and Gt **(D)** and the quenching coefficient E^Q for Hb, Kr, Kni, Tll and Hkb **(E)**. The activation threshold θ of each run was adjusted within 5 and 20 **(F)**. All values are in arbitrary units.

Contribution from binding sites

Models using either the same combination of PWMs as Kim *et al.* (2013) or exclusively B1H matrices tend to fall into solutions with the same quality of model fits and similar regulatory contributions, apart from the emergence of slight repression by Kni with the B1H matrix, which might be a threshold issue (data not shown). No obvious differences were observed between keeping the thresholds fixed for Bcd and Hb or adjusting all PWM thresholds. Nevertheless, the model occasionally includes much more TFBS than it is probably the case in reality, depending on the input combinations. For example, in several runs, a huge amount of binding sites are predicted for Tll. Such issues could have been tackled by fine-tuning the limits of the thresholds or testing PWMs from different sources, but it would not drastically change the predicted regulation by a certain factor in general. On the other hand, the model suggests a relatively high contribution from just one Cad site in models fit to gt-1 without considering auto-activation (Figure 36 and Figure 38). It is biologically unrealistic that a single binding site can trigger transcription. The modelling approach has certain freedom for adjusting parameters in order to fit the data as closely as possible, which might result in some artefacts. Nevertheless, the model is not able to achieve the observed levels in these cases. This shows that the model still has constraints and cannot entirely compensate for missing input.

Expression levels
The main difference between model solutions from distinct input combinations lies in the expression levels. In most cases this has a biological meaning and suggests regulatory mechanisms (see 4.4.5 for detailed discussion). Moreover, it turned out that the model has predictive power concerning levels when optimized to semi-quantitative data. In particular, when fitting to the combined CRE (Figure 40) considering Gt auto-activation, it looks like the model does not match the quantified dataset with the desired precision. It suggests levels slightly lower for the anterior and slightly higher for the posterior, resulting in a clear difference between the two domains. In fact, the enzymatic *in-situ* hybridizations showed that the posterior is much stronger compared to the anterior (Figure 28). This information was lost during the quantification of the fluorescent *in-situs* due to the thresholding step. Hence, the model qualitatively predicts the levels correctly, even though it was fit to slightly deformed input datasets. This should be interpreted as strength of the modelling approach.

Temporal resolution
The model of transcriptional regulation for *eve* in Kim et al. (2013) was fit to T6 only. In contrast, this model for *gt* is capable of dealing with such a high temporal resolution of 10 time points. Nevertheless, when fitting to all time classes without auto-activation, the model tends to overshoot in C12 and C13 and to underestimate the levels of later time points. An improvement is achieved for gt-3, but not for gt-1, via fitting to the 8 time classes of C14 only (runs 162 and 163). Not surprisingly, the model also resembles the observed expression quite well for both CREs when fit only to the very early stages C12, C13 and T1 (data not shown), since it does not have to find a compromise with the remaining seven time points anymore. Hence, apart from auto-activation also reduction of the temporal resolution can result in improved outputs, because both ways somehow represent compensatory mechanisms, although the first increases and the latter reduces complexity in the model.

4.4.5 Regulatory mechanisms concluded from the model

Repressive and activating contributions
Direct repression refers to the effect from repressor sites nearby the BTM. It is not applicable to gt-3 in the endogenous locus since this CRE is separated from the promoter by more than 1 kb. No major differences concerning quality of fits and contributing TFs were observed depending on the inclusion or exclusion of this mechanism. Subsequently, direct repression was turned off for most of the runs.
The distance-dependent mechanism of short-range repression implemented in the model is capable of suggesting regulatory input in accordance with the literature. Contribution by Tll from the pole is found in all model outputs and has been observed for endogenous *gt* in *tll* mutants (Kraut and Levine 1991b). The model tends to identify Kr as the major repressor in the middle region of the embryo, as previously suggested (Kraut and Levine 1991a). Hb turned out to be an important repressive input for gt-3, even when co-activation was included (Figure 35). It has the potential to set the anterior boundary of the posterior domain at C11 (not modeled). At C12, repression from Kr is already appearing in the model (data not shown). Hb also seems to be required in this CRE to shut-down expression in the anterior region. This has not been previously reported, because such information cannot simply be deducted from changes of the endogenous *gt* expression in *hb* mutants. The contribution of Hb in a model fit to gt-1 considering auto-activation, is negligible (Figure 37). It would have the potential to repress the anterior domain of this CRE, which needs to be avoided in the model. The boundary positions of the endogenous posterior *gt* domain are not altered in *kni* mutants, but the expression levels are reduced (Kraut and Levine 1991b, Surkova et al. 2013). Usually, the model does not suggest contribution by Kni, but compensatory mechanisms between TFs were observed in some cases. For example, a model fit to gt-3, considering auto-activation, allows for repression by Kni

(Figure 35). This is probably due to increased activation by Gt, which can now overcome the negative input from Kni.
The models tend to find six or seven Cad sites in gt-3 (Figure 34), which should be sufficient to trigger transcription in the posterior *in vivo*. In contrast, only one or two Cad and Bcd sites are predicted from the gt-1 sequence (Figure 36), which is probably not enough to achieve a response. For this reason and in order to assess the issue with the levels in gt-3, other activating possibilities were explored *in silico*.

Bcd-cooperativity and co-activation
The model was used to test for Bcd-cooperativity and co-activation of Hb by Bcd and Cad. This was done in all possible combinations of model inputs, either considering one of these special mechanisms per time or both at once. Both mechanisms were of particular interest for gt-1, since it was not clear how it becomes activated due to the lack of TFBS for Bcd and Cad. On the other hand, it was unlikely that these mechanisms could solve the problem, since they depend on the presence of sites for these two main activators. Co-activation did not improve the fits for neither of the CREs and the model predicts only minor (Figure 38) or no (Figure 35) activating contributions from Hb sites. Hence the model suggests, in contrast to what happens in the *eve* MSE2 (Small et al. 1991), that this mechanism is not functioning in *gt*. Bcd-cooperativity is a very unlikely mechanism for a posterior enhancer like gt-3, because the gradient fades away in the posterior and on the other hand it could provoke additional expression in the anterior. Segal et al. (2008) claimed that Bcd-cooperativity drives the anterior domain of gt-1. Taken together all my observations, no conclusive evidence for Bcd-cooperativity was provided. It could create a temporal conflict for gt-1, because Bcd is contributed maternally. Hence, it could trigger expression at C12 or earlier, but gt-1 arises at C13 only.

Giant auto-regulation
Besides elucidating the influences of certain mechanisms, I also needed to clarify the role of Gt on itself. Usually, when Gt was set as a repressor, it did not appear as repressing contributor. Only in a few special cases the fitting got stuck in a solution allowing for the unlikely scenario of auto-repression. Since the model usually underestimates the expression levels, auto-repression would even further aggravate this discrepancy. The opposite effect, Gt auto-activation, was the only mechanism able to achieve the observed levels in both enhancers. There is hardly any experimental evidence for it, apart from the observation made by Eldon and Pirrotta (1991): “embryos carrying strong *gt* mutations still express the protein and show a very similar evolution of the pattern as the wild-type but the intensity of the antibody staining does not increase, suggesting that functional Gt protein may stimulate its own expression”. During the computational identification of the *gt* CREs, Schroeder et al. (2004) did not find any contribution from Gt binding sites in its own enhancers, except for the tip CRE gt-6. Their Gt PWM was built from six footprints and resulted in no score or low values compared to the other input factors in the other newly identified CREs.
The models presented here tend to find two or three Gt sites and its addition as an activating factor resulted in improved fits in all cases. For gt-3 it solves the issue of the levels, but on the other hand repression by Kni is suddenly observed and Kr repression is diminished (Figure 35). This is a rather unlikely scenario based on the mutant experiments (see Jaeger 2011 for review). The main activator of gt-3 is definitely Cad, and auto-activation might only play a negligible role at later stages. In contrast, the model suggests that activation by Bcd and Cad is not sufficient for gt-1 and that auto-activation is necessary to trigger transcription in the anterior (Figure 37). This is biologically reasonable for a late element, because otherwise we would observe expression at earlier stages. Based on the outputs for the combined CRE (Figure 40), we can hypothesize that Gt auto-activation might be reaching over the entire fragment, but that Cad is the main positive input onto gt-3. Even though the fits for both CREs improve with auto-activation, this result has to be interpreted with care. It is not surprising, that inclusion of a pattern that looks very similar to the target pattern as activating input achieves these

improvements *in silico*. Such an activator triggers expression at the desired region and probably less repressing input from other factors is required to refine the boundaries. Therefore, inclusion of such a factor could lead the algorithm to get stuck in solutions solely considering the contribution of this activator. Subsequently, for cases, where auto-activation is actually operating *in vivo*, it will always be very difficult to assess this mechanism with a model.
Several negative control-runs were executed in order to validate the suggested mechanism of Gt auto-activation (Figure 45). I hoped that certain differences between the input and output pattern would become reflected by the model. In particular, the posterior gap gene expression domains shift towards the anterior during C14 and hence, there is a delay between mRNA and protein positioning (Figure 31). Additionally, there are differences in the timing of the peak expression. Depending on the spatio-temporal resolution of the dataset and on the complexity of the mathematical model, it might not be possible to observe such fine nuances. Control runs without Bcd and Cad, but including auto-activation and all other TFs, as well as runs considering only Gt auto-activation and no other input factor at all were performed. Not surprisingly for the latter, no early expression is achieved for gt-3, an anterior domain arises, the posterior stripe is shifted to the posterior and there are issues with the levels (Figure 45A). In contrast, the model fits gt-1 quite well, besides a very subtle shift of the posterior (Figure 45C). The prediction for gt-3 improves when considering all the repressing inputs, but no Bcd or Cad (Figure 45B). The anterior domain becomes abolished and the boundaries are more precise, but a discrepancy between very early and very late levels appears. Also the positioning of the gt-1 boundaries is perfected with the help of the repressors (Figure 45D).

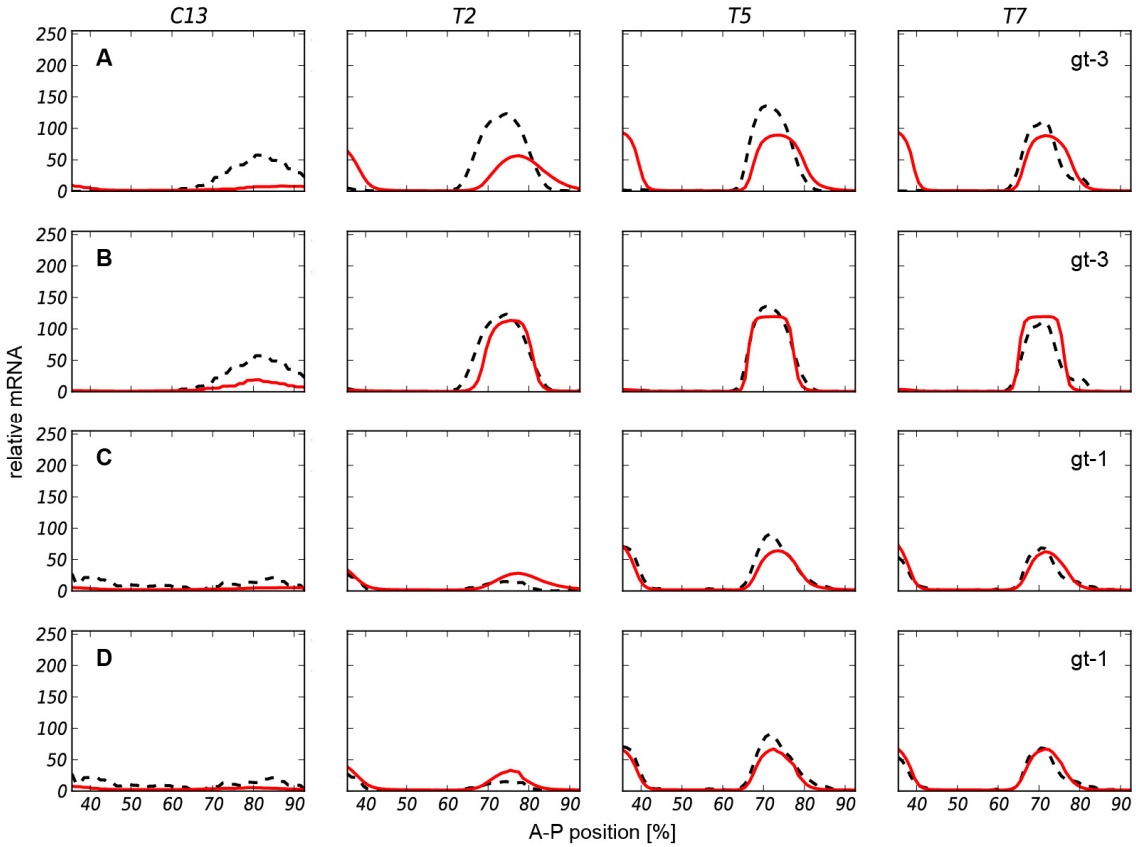


Figure 45: Validation of Gt auto-activation via exclusion of other factors.
(A) Model fit to gt-3 considering only Gt auto-activation and no other input factor (run 25) or **(B)** considering Gt auto-activation but excluding Bcd and Cad (run 27). **(C)** Model fit to gt-1 considering only Gt auto-activation and no other input factor (run 26) or **(D)** considering Gt auto-activation but excluding Bcd and Cad (run 28).

Taken together all the observations from different modeling attempts, I conclude, that the main activator of gt-1 is Gt, Cad might have minor influence and Bcd is probably not required. In contrast, gt-3 is activated by Cad, although it cannot be excluded that slight input from Gt adjusts the levels. The negative control runs (Figure 45), were able to show the delay between mRNA and protein due to the shift and thanks to the temporal resolution of my datasets. Interestingly, in the case of gt-1 the model fit is almost perfect without Bcd and Cad, whereas for gt-3, Cad is definitely required to correct several defects in the output pattern. The hypothesis of auto-activation had to be evaluated *in vivo*, either via mutagenesis of the Gt binding sites or by testing the CRE in a *gt* mutant background.

4.5 Experimental evaluation of Giant auto-activation

In order to evaluate Gt auto-activation, I generated mutated CREs without Gt binding sites. The strategy was to knock-down all predicted sites, even the weak ones, by introducing point mutations but without deleting or inserting bases in order to maintain the length of the CRE and the distances between other TFBS. Thereby, particular care was taken not to destroy or create sites for other TFs of the A-P or dorsal-ventral system and *zelda* (Harrison et al. 2011). A model found 15 sites in gt-1 with the Selex matrix and 86 bases were mutated (Figure 46A). In the case of gt-3, 8 sites were mutated, thereby introducing 28 point mutations (Figure 47A). No obvious differential enrichment of *gt* sites in one or the other CRE could be detected in general (see section 6.4 of Materials and methods for details about TFBS and alignments).
Mutagenesis of gt-1 results in a severely altered expression pattern (Figure 46B, C). The anterior domain gets completely abolished and only the ectopic vector expression remains. It can easily be distinguished from the WT anterior, because it is not a broad domain starting to separate into two stripes and its posterior boundary is located much more to the anterior. The enzymatic *in-situ* hybridization of the mutated gt-1 was performed in parallel with the WT enhancer and both staining reactions were stopped at the same time. I took advantage of the ectopic anterior patch and used it as an internal control. It is in general independent of the CRE, the target line (Figure 26, Figure 27) and even the integration method (Zinzen et al. 2006). Usually, this expression feature is rather weak and does not always come up in all embryos of a staining. If it arises, we can consider the signal to noise ratio sufficiently high and therefore, the ectopic anterior expression can be useful as calibration. In the *in-situ* hybridization of the mutated gt-1, this head patch is visible and hence, the staining was successful. The intensity of the posterior domain is similar to the ectopic anterior and therefore, lower than in the WT enhancer.

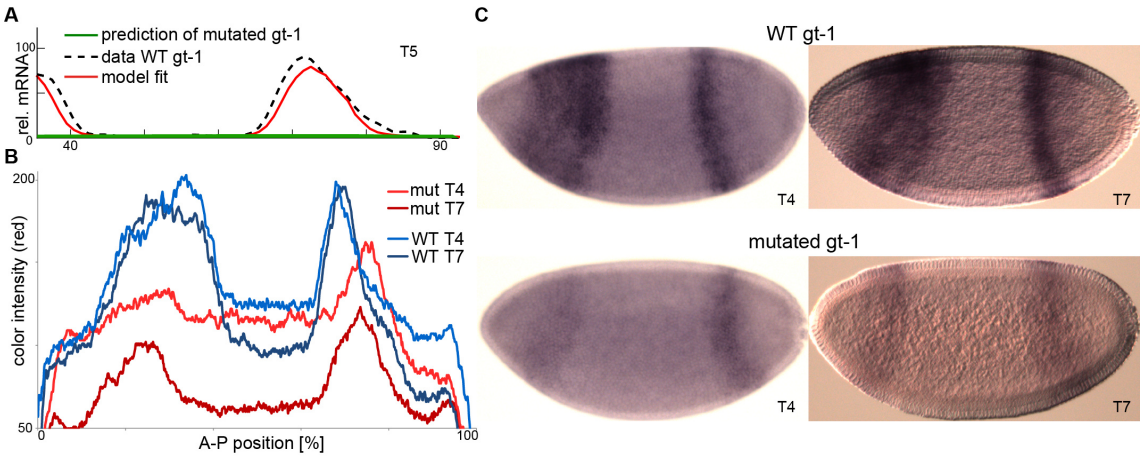


Figure 46: Mutagenesis of Giant binding sites in gt-1 leads to altered expression.
(A) Shown is the relative mRNA concentration (in a.u.) at T5. A model (run 17_1) which had been fit to gt-1 (red) predicts that the mutated gt-1 sequence drives no expression at all (green). The black dashed line is the quantified lacZ mRNA of the WT CRE. **(B)** The lacZ mRNA expression driven by the WT and the mutated (mut) gt-1 sequence was quantified from the enzymatic *in-situs* with the FlyGUI. Shown is the intensity of the color red from the RGB color space, which is a measure for the purple staining. **(C)** LacZ mRNA expression driven by WT gt-1 (upper panel) and by gt-1 with mutated Gt binding sites (lower panel). The depicted embryos are mid- (T4) and late- (T7) cycle 14 and the pictures were taken in BF or DIC. The stainings were done in parallel and stopped at the same time. The expression driven by the mutated gt-1 is reduced in the posterior compared to the WT and the anterior domain becomes abolished.

We can draw the conclusion that Bcd has absolutely no influence on the anterior domain of gt-1, because it is exclusively driven by auto-activation. The posterior domain also relies primarily on Gt, but there might be some contribution from the detected weak Cad sites.
In contrast, the mutated gt-3 sequence drives a posterior domain at the same position and with the same intensity as the WT enhancer (Figure 47B). Additionally, the ectopic anterior expression from the vector can be appreciated. This confirms that gt-3 is essentially activated by Cad, although we cannot entirely exclude minor auto-activation, since such subtle differences in expression levels would be extremely difficult to prove experimentally.

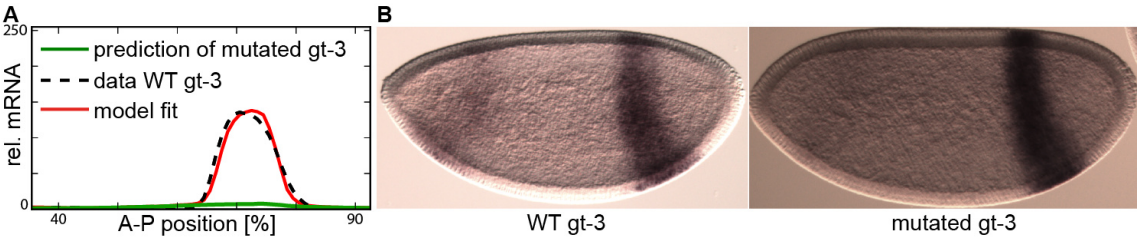


Figure 47: Mutagenesis of Giant binding sites in gt-3 does not change expression.
(A) A model (red) which had been fit to gt-3 considering Gt auto-activation (run 17_1 with the Selex matrix) was used to predict expression from the mutated gt-3 sequence (green). Shown is the relative mRNA concentration (in a.u.) at T5. The black dashed line is the quantified lacZ mRNA of the WT CRE.
(B) The enzymatic *in-situ* hybridization against lacZ mRNA driven by the mutated gt-3 show an intense posterior domain. The depicted embryos are mid-blastoderm stage (DIC images).

4.6 Model prediction of mutants and their experimental evaluation

I took advantage of the models fitted to the CREs in the wild-type background to predict mutant gene expression. I used the parameter values from the optimization runs and the concentrations of the TFs from previously published quantitative gap gene expression data in *Kr* and *tll* mutants (Janssens, Crombach, Richard Wotton, et al. 2013, Surkova et al. 2013) to calculate the pattern of the CREs in these mutants. For this purpose, I selected a subset of the previously shown optimizations and I will focus on the ones including auto-activation and additionally include the one without regulation by Gt in the case of gt-3. In order to evaluate the model predictions experimentally, I crossed the CREs into the mutant backgrounds, either via conventional crosses or by meiotic recombination (see Materials and methods).

4.6.1 Prediction and evaluation of the *Krüppel* mutant

In the *Kr* mutant (Surkova et al. 2013), the anterior Gt protein domain is not affected, whereas the posterior is shifted and expanded to the anterior (Figure 48). The general shift of the posterior gap gene domains over time towards the anterior of the embryo is still taking place and Gt gains on Kni, until they are both expressed at the same position. There is premature reduction of the posterior Gt domain starting at T4, which can be appreciated until T8. No further alteration of the Gt pattern was observed in a *Krüppel-knirps* double mutant.

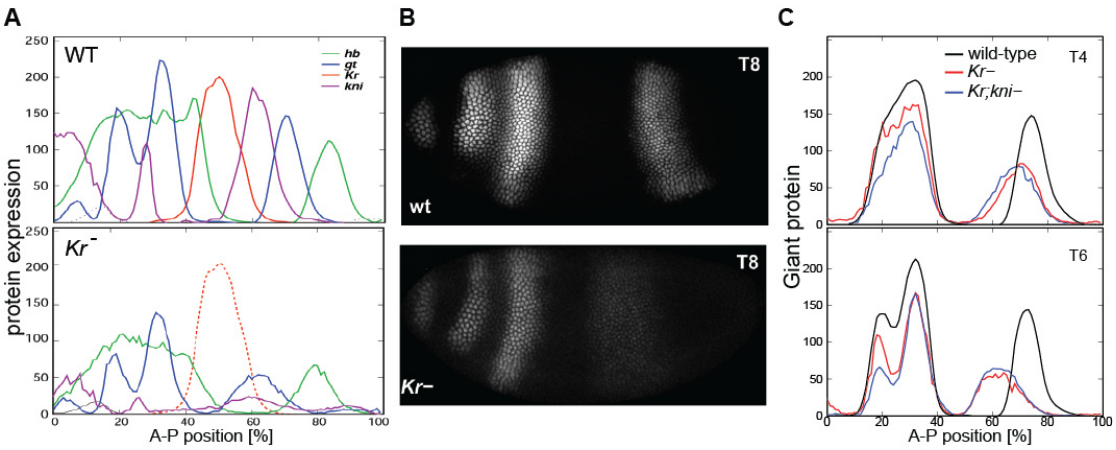


Figure 48: Gap gene expression in the *Krüppel* mutant.
(A) Protein expression of the gap genes Hb, Gt, Kr and Kni in wild-type and in the *Kr* mutant, taken from (Kozlov et al. 2012). **(B)** Fluorescent protein staining of Gt in the WT and in the *Kr* mutant at T8. **(C)** Expression of Gt protein at T4 and T6 in WT (black), *Kr* mutant (red) and *Kr; kni* (blue) double mutant. Taken from Surkova et al. (2013).

Expression of gt-3 in the *Kr* mutant

For gt-3, the model correctly predicts the observed anterior shift and broadening of the posterior domain, disregarding whether it was optimized without input from Gt or with auto-activation (Figure 49). Nevertheless, the issue with the levels observed during the optimizations logically becomes reflected in the predictions. The model considering auto-activation, which was able to correctly resemble the expression levels during the optimization, now qualitatively predicts the reduced level of the posterior domain in the *Kr* mutant, because the reduced Gt concentrations were plugged-in (Figure 49B). In contrast, the model without auto-regulation suggested levels lower than in the data during the fitting procedure and no further reduction is observed in the mutant (Figure 49A).

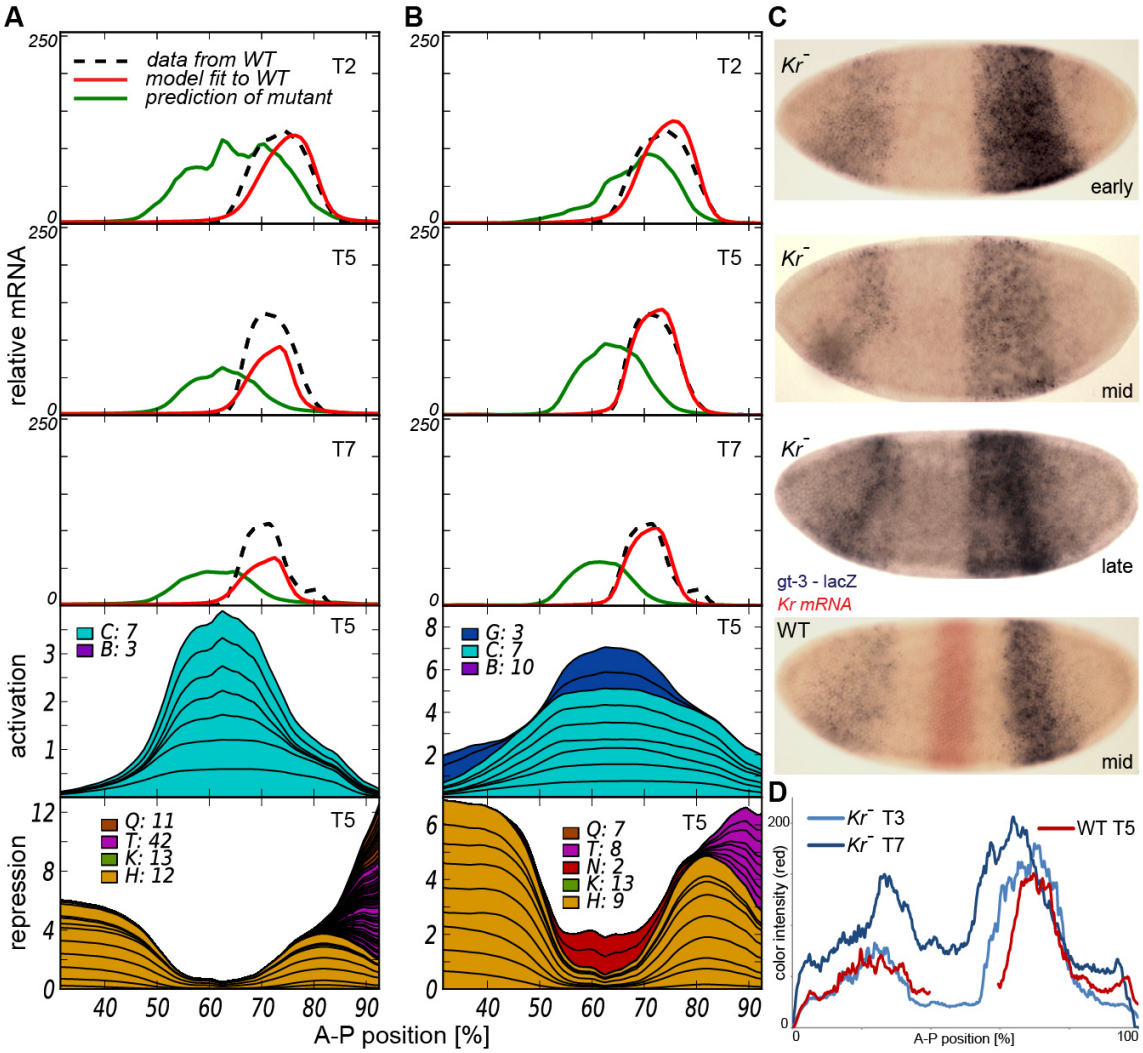


Figure 49: Prediction and evaluation of gt-3 in the *Kr* mutant.
(A and B) Prediction for gt-3 in the *Kr* mutant (green) with models fit to wild-type (red) and quantified expression of gt-3 in the WT for comparison (dashed black line), as well as the activating and repressing contributions. For detailed explanation of the graphs see Figure 34. **(A)** Prediction with a model optimized to gt-3 without considering Gt auto-regulation (run 132_1). **(B)** Prediction with a model optimized to gt-3 considering Gt auto-activation (run 133_2). **(C)** Expression driven by gt-3 expands towards the anterior in the *Kr* mutant. Enzymatic *in-situ* hybridizations against *lacZ* mRNA driven by gt-3 (blue) in the *Kr* mutant and in the wild-type. The *Kr* mRNA staining in red is only visible in the WT. Shown are bright field images of an early, mid and late C14 embryo for the *Kr* mutants and a mid C14A embryo for the WT. **(D)** The lacZ mRNA expression was quantified from the enzymatic *in-situs* with the FlyGUI. Shown is the intensity of the color red from the RGB color space, which is a measure for the purple staining. In the case of the wild-type, the data points between 40 to 60% A-P position were deleted because the *Kr* mRNA staining was picked up in this region.

It is hard to tell from the *in-situ* hybridizations of the CRE-reporter construct, whether there is in fact a reduction in the levels as quantified for Gt protein (Figure 49C). Note that the collected embryos are a pool of genotypes either homozygous or heterozygous for the reporter-cassette. In a mutant background, the ectopic anterior expression is less reliable as an internal control, since we do not know how it is regulated and how it changes in a mutant. The only possible candidate for causing the reduced level of the posterior Gt is Kni, but in the *Kr; kni* double mutant (Surkova et al. 2013) the levels are not altered any further. On the other hand, posterior Hb expands further to the anterior in the double mutant compared to the *Kr* mutant and might take over the role for reducing the levels from Kni.

Expression of gt-1 in the *Kr* mutant

Gt-1 drives expression in three stripes with severe derepression in between them in the *Kr* mutant background (Figure 50B). The additional stripe emerges in the middle of the embryo at the position of Kr in the WT. This was an unexpected result, because the Gt protein pattern does not show such a leakage between the two domains. A hypothesis was, that it is caused by the polymorphisms observed in the gt-1 sequence (see Materials and methods): I amplified the fragment form different fly lines, but all contained point mutations and gaps compared to the published sequence (Schroeder et al. 2004). Nevertheless, they drive the same expression pattern. Although these polymorphisms do not manifest in the WT, they might trigger a difference in a mutant background. In order to exclude this theory, I crossed the original fly line (Schroeder et al. 2004), which contains the same gt-1 sequence as the *D.melanogaster* genome release, into the *Kr* mutant, but it also drives the same unexpected pattern (Figure 50B). There are three possible explanations for the observed three-striped pattern triggered by gt-1 in the *Kr* mutant. A silencer element might be present in the *gt* locus, capable of abolishing the leakage via non-additive interactions. Connected to this is also one of the experimental issues. Although the integration locus did not show an influence on the pattern in the WT, its effect might change in a mutant background. Finally, also the inserted reporter-cassette itself could interfere with other factors, which become derepressed in a *Kr* mutant at later stages.

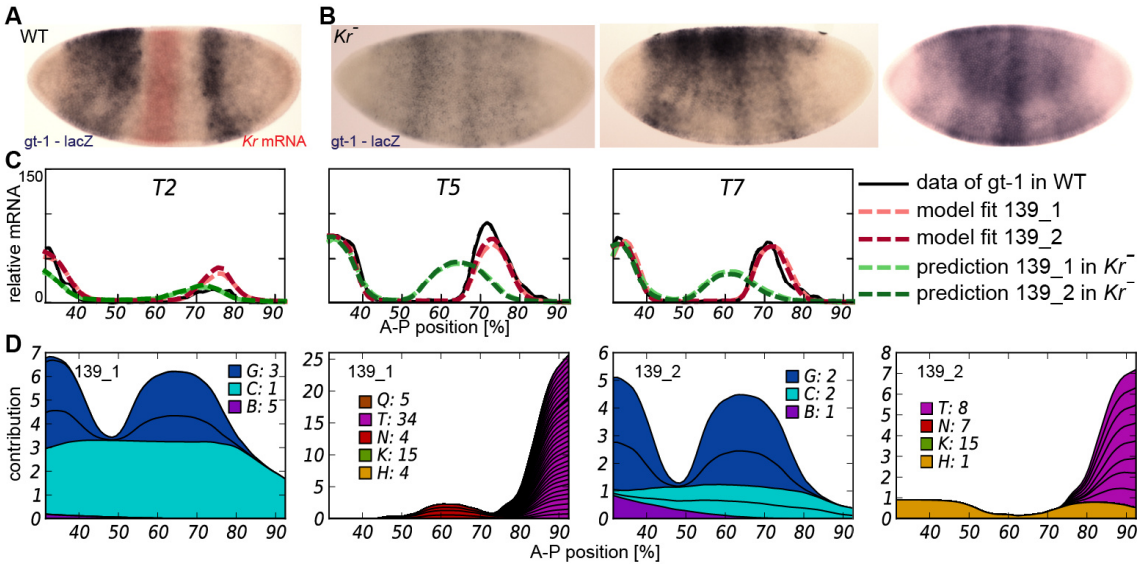


Figure 50: Prediction and evaluation of gt-1 in the *Kr* mutant.
(A and B) Bright field images of embryos after *in-situ* hybridization for *lacZ* mRNA expression driven by gt-1, with *lacZ* in blue and *Kr* mRNA in red (only in the wild-type). Shown is the expression from gt-1 in the WT background **(A)** and a mid and a late C14 embryo for the *Kr* mutant **(B)**, and the last embryo derives from the original fly line from Schroeder et al. (2004). **(C)** Prediction with a model optimized to gt-1 considering Gt auto-activation (run 139_1 in lighter color and 139_2 in darker color). The quantified expression of gt-1 in the wild-type for comparison (black line). The model optimizations are shown as dashed red lines and the predictions for the *Kr* mutant in dashed green lines. **(D)** Regulatory contributions from activators and repressors in the *Kr* mutant at T5, predicted from the runs 139_1 and 139_2.

Models optimized considering auto-activation (run 139), predict a shift and expansion of the posterior domain towards the anterior in the *Kr* mutant (Figure 50C). This coincides with the observed Gt protein pattern but not with the mRNA expression driven by the gt-1 reporter. On the other hand, models optimized excluding auto-activation and either with or without co-activation, predict one broad domain reaching from the anterior edge of the trunk region considered for the model until approximately 80% A-P position (data not shown). It is very interesting, that two model solutions (139_1 and 139_2) differing substantially in their regulatory inputs (Figure 37), attained the same prediction for the *Kr* mutant (overlapping green

dashed lines in Figure 50C). Only 139_2 showed repression from Kr in the WT, which would be a plausible explanation for the shift. In fact, the only possible agent to set the boundaries in these models is Gt itself (Figure 50D). Hb would be able to account for the anterior boundaries of the posterior domain, but the Hb repression considered in 139_2 appears to be rather weak. In 139_1 only some Kni input emerges, but it overlaps with the posterior gt-1 domain in the *Kr* mutant. These observations raise the question, why the fit 139_2 in the WT needed the repression by Kr at all. Moreover, 139_1 gives a higher emphasize on Cad activation compared to auto-activation, which is supposed to account for the boundary positions and expression levels.

Due to the severe alteration of the expression pattern in the gt-1 fly lines, it is not possible to draw any conclusions about the expression level of the posterior domain. In principle, auto-activation could be a possible explanation for the reduction of posterior Gt, because the Cad protein gradient is not altered in gap mutants and hence, only the interaction with Kni is a plausible mechanism. Kr keeps Gt in place in the WT and prevents an intersection with the posterior Kni domain, but the Kr mutant makes this overlap between them possible. Minor repressive contributions from Kni onto the posterior Gt domain could manifest at later stages, when this effect intensifies via the lack of Gt auto-activation.

Expression of the combined CRE in the *Kr* mutant

The combined CREs drives a normal anterior domain as well as a posterior domain expanded towards the anterior in the *Kr* mutant (Figure 51 C). At early stages, the two domains are clearly separated from each other and the anterior is still weaker than the posterior, as in the wild-type. Over time, derepression arises in the middle of the embryo, and intensifies at later stages until a third expression region becomes more pronounced. Nevertheless, the final pattern is not very well defined. A simple addition of the patterns of gt-1 and gt-3 is a plausible explanation for the observed expression driven by the combined fragment.

A model optimized to expression data of the combined enhancer (run 91_1) is able to qualitatively predict the shift and broadening of the posterior domain observed in the endogenous Gt protein pattern and in the experiment (Figure 51A). The model suggests that Hb is responsible for setting the boundaries of the posterior domain, and only minor input from Kni is observed (Figure 51B).

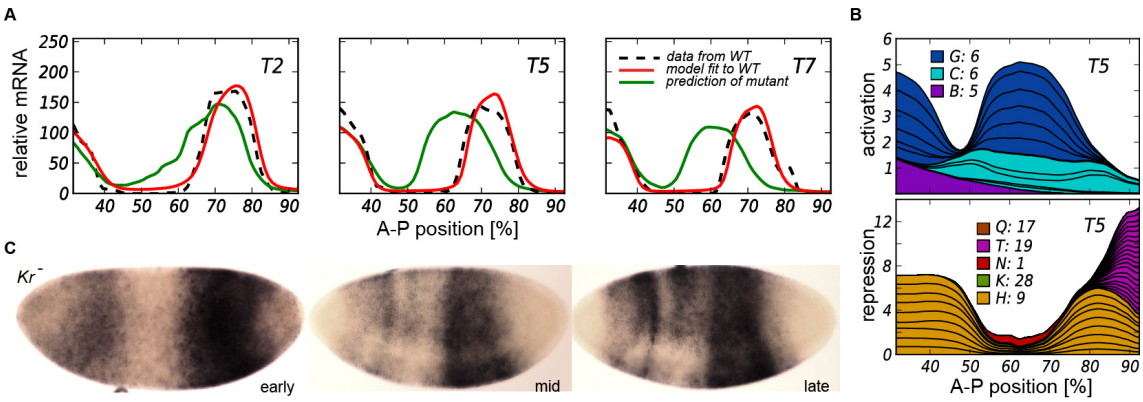


Figure 51: Prediction and evaluation of the combined CRE in the *Kr* mutant.
(A) Prediction for the combined CRE in the *Kr* mutant (green) from a model optimized (red) to the combined CRE considering Gt auto-regulation (run 91_1) at T2, T5 and T7. Quantified expression in the wild-type is shown for comparison (dashed black line). **(B)** Predicted activating and repressing contributions in the *Kr* mutant. **(C)** Enzymatic *in-situ* hybridizations of embryos with *lacZ* mRNA expression driven by the combined CRE in the *Kr* mutant. Shown are BF images at different time points within C14A.

4.6.2 Prediction and evaluation of the *tailless* mutant

Tll is expressed at both poles and the *tll* mutant additionally lacks the posterior Hb domain (Janssens, Crombach, Wotton, et al. 2013). The peak positions of the other gap genes still shift towards the anterior over time, although less far than in the WT (Figure 52). The posterior boundaries of the posterior domains of Gt and Kni are expanded to the posterior. The mutant embryos have a broadened 6th Eve stripe and the 7th stripe is missing or occasionally comes up just before gastrulation.

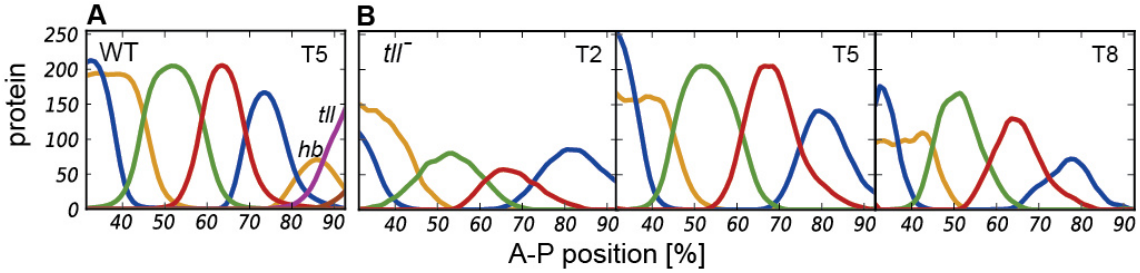


Figure 52: Gap gene expression in the *tailless* mutant.
Protein expression of the gap genes *hunchback*, *giant*, *Krüppel* and *knirps* in wild-type **(A)** at T5 and in the *tailless* mutant **(B)** at T2, T5 and T7 (values from Janssens et al.).

In the *tll* mutant background, the model predicts an expansion and in some cases a slight shift of the posterior domain to the posterior for all three CREs. The prediction from a model fit to gt-3 without auto-regulation (Figure 53A, run 132_1) shows the expansion to the posterior from T3 on (time class not shown). The expression does not reach until the pole, but instead disappears at 90% of the A-P position, and the anterior boundary position does not change. In contrast, a model considering Gt as activator (133_2) predicts an expansion reaching further posteriorly than the former at T1 already (Figure 53B). Additionally, the position of the anterior boundary is located further towards the posterior compared to the WT.
I crossed the gt-3 reporter-construct into the *tll* mutant (Figure 53C) and observed changes similar to the ones predicted by the model without auto-activation (run 132_1): expansion towards the posterior without reaching the pole, while the anterior boundary position is maintained (Figure 53D). Unfortunately, no *tll* mutant fly lines with the reporter constructs for gt-1 and the combined CRE could be obtained.
As explained before, two different modelling solutions for gt-1 considering auto-activation give the same quality of fit during the optimization (run 139_1 and 139_2), but suggest other repressing contributions. Their predictions for gt-1 in the *tll* mutant (Figure 54A and B) now coincide, that the posterior domain expands (T2) and then also shifts (T5, T7) towards the posterior. Interestingly, run 139_1 predicts the disappearing of the anterior domain at T7, although it cannot be deduced from the regulatory contributions of the model which TF could cause this result.

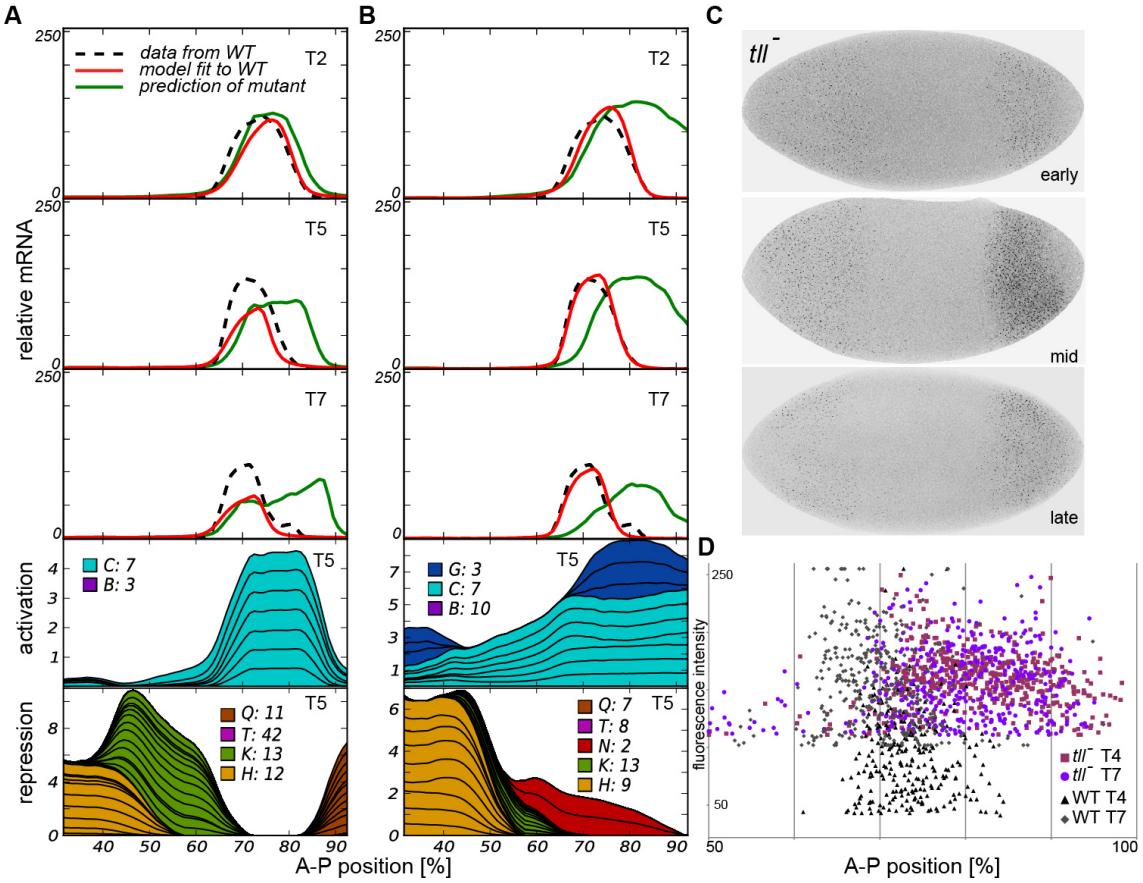


Figure 53: Prediction and evaluation of gt-3 in the *tll* mutant.
(A) Prediction with a model fit to gt-3 without Gt auto-regulation (run 132_1) or **(B)** considering auto-activation (run 133_2). Prediction in the *tll* mutant (green) is compared to the model optimized in the wild-type (red), and the quantified expression of the CRE in the WT (dashed black line) at time classes T2, T5 and T7. The activating and repressing contributions in the *tll* mutant are shown for T5. For detailed explanation of the graphs see Figure 34. **(C)** Fluorescent *in-situ* hybridizations of embryos with *lacZ* expression driven by gt-3 in the *tll* mutant. Shown are inverted images of embryos at different time points of C14A. **(D)** The fluorescent data points from a WT and a *tll* mutant embryo at T4 and T7 show a significant expansion of the gt-3 – driven lacZ expression towards the posterior pole.

The anterior domain is not altered in the predictions from the model fit to the combined CRE including auto-activation (run 91_1). It predicts an expansion of the posterior domain that reaches much further towards the posterior pole than the region considered for the modelling. At later stages, the anterior boundary is again located further posterior compared to the WT. This effect is caused by the inclusion of Gt auto-activation in the models, because the posterior Gt protein domain shifts less towards the anterior in the *tll* mutant compared to the WT (Figure 52). Hence, this change in the extent of activating input gets reflected in all predictions with models considering Gt auto-activation, independent of the CRE.

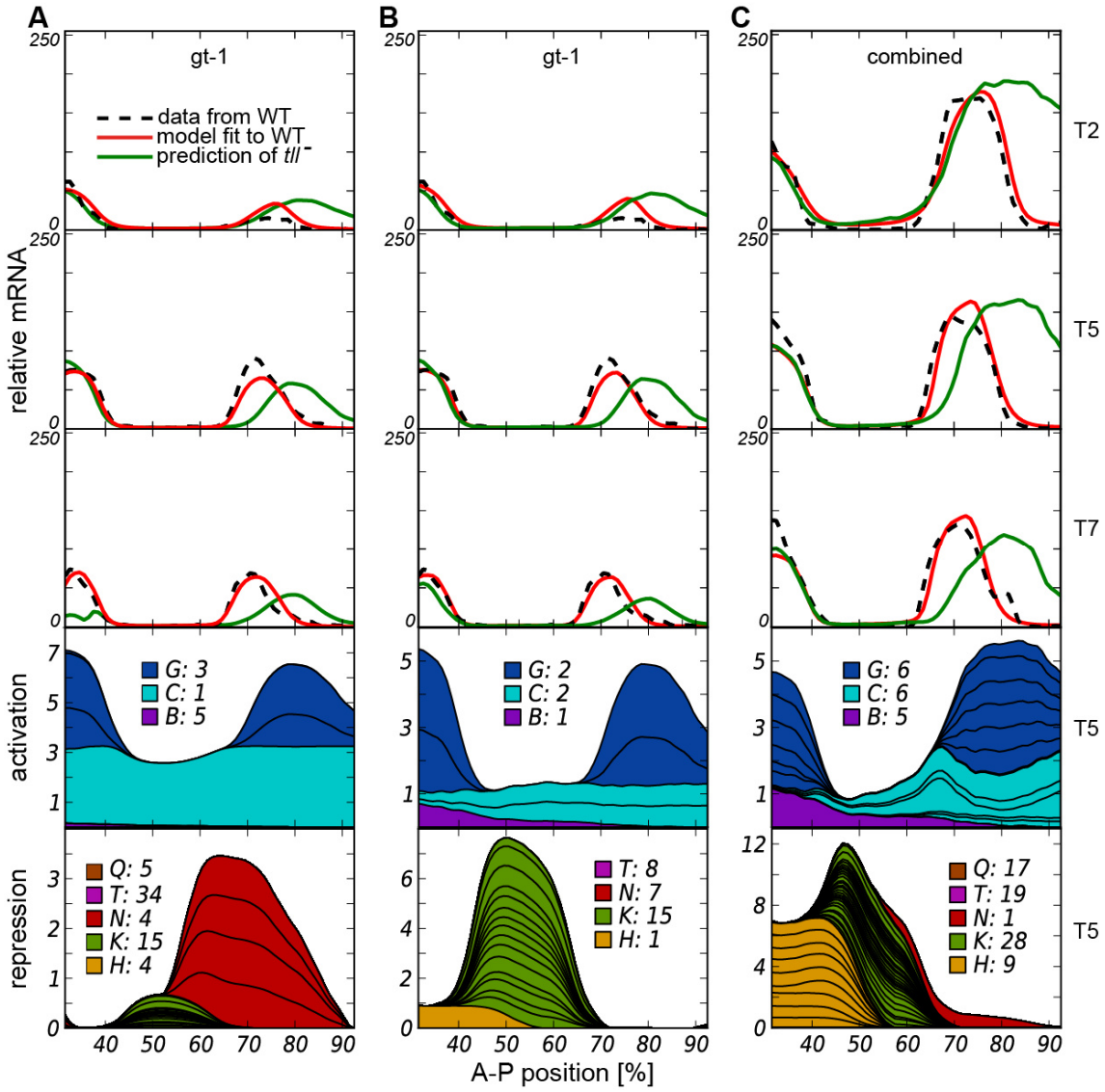


Figure 54: Prediction of gt-1 and the combined CRE in the *tll* mutant.
(A and B) Prediction with a model fit to gt-1 considering Gt auto- activation (**A**: run 139_1 and **B**: 139_2). **(C)** Prediction with a model fit to the combined CRE considering auto-activation (run 91_1). Prediction in the *tll* mutant (green) is compared to the model optimized in the wild-type (red), and the quantified expression of the CRE in the WT (dashed black line) at time classes T2, T5 and T7. The activating and repressing contributions in the *tll* mutant are shown for T5. For detailed explanation of the graphs see Figure 34.

4.7 Expression from *giant* CREs in other mutants

4.7.1 *Hunchback* mutants

Additionally to the zygotic Hb component, the mother provides *hb* mRNA ubiquitously (Tautz et al. 1987), which is subsequently translationally repressed by Nanos (Irish et al. 1989), establishing an anterior Hb protein gradient. According to the literature, the posterior *gt* domain expands in zygotic, as well as in maternal & zygotic *hb* mutants (Eldon and Pirrotta 1991, Struhl et al. 1992). The anterior *gt* stripe 3 is shifted anteriorly in zygotic *hb* mutants (Eldon and Pirrotta 1991). The posterior CRE gt-3 has much more higher scoring Hb sites than gt-1 and the model predicts that Hb has major repressing input on gt-3 at early stages, whereas its influence on gt-1 is much less. We would expect expansion of the posterior driven by gt-3 in a maternal and zygotic *hb* mutant. Since gt-1 is primarily triggered by auto-activation, the differences seen in gt-3 should be reflected in gt-1.

Zygotic *hb* mutants carrying gt-3 or gt-1 were established via conventional crosses or the recombination approach, respectively. Maternal *hb* mutants were achieved by transgenic RNAi (Staller et al. 2013) using a maternal Gal4 driver, triggering a short hairpin (sh) against *hunchback* via the upstream activating sequence (UAS-sh-hb). Embryos carrying the corresponding CRE but lacking any Hb at all, were generated by inducing the UAS-sh-hb in a zygotic *hb* background and crossing to males harboring the CRE in the zygotic *hb* mutant. In principle, all collected embryos should lack the maternal *hb* contribution, and they will be either heterozygous or homozygous for the zygotic *hb* mutation. In order to facilitate this classification, the embryos were co-stained for Hb and Eve protein (Figure 55). The hb^{12} allele contains a premature stop codon before the first finger domain (W256). This truncated peptide can still partially be detected by the anti-Hb antibody in the zygotic *hb* mutants. Nevertheless, in a pool of maternal & zygotic *hb* mutants, the homozygous zygotic mutants can easily be distinguished from the maternal mutants still containing one copy of zygotic *hb*.

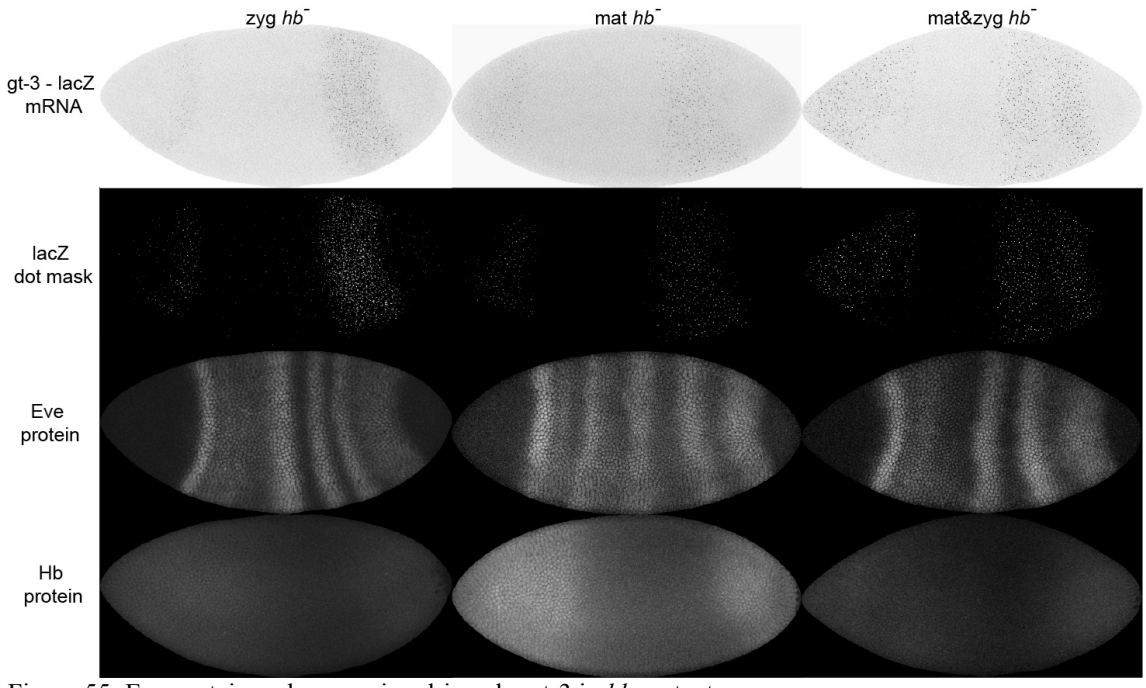


Figure 55: Eve protein and expression driven by gt-3 in *hb* mutants.
Fluorescent stainings of lacZ mRNA driven by gt-3 (inverted images), lacZ dot mask after thresholding, as well as Eve and Hb protein in maternal, zygotic and maternal & zygotic *hb* mutants at T5.

The maternal mutants show six Eve stripes, with the second stripe being weaker and fused to stripe one. Embryos lacking any Hb protein have only five Eve stripes, although a very faint additional (second) stripe can occasionally be appreciated at late stages. In zygotic *hb* mutants, there is derepression between the first and third Eve stripe and the last stripe is very broad and hence, probably a fusion between the sixth and seventh stripe. The eve pattern of the *hb* mutants potentially has higher variability than in the WT and is therefore more difficult to classify.

gt-3 in *hb* mutants

The lacZ mRNA expression driven by gt-3 is altered in all three mutant categories (see Figure 55 for fluorescent and Figure 56A for enzymatic stainings). In particular, it expands to the anterior in the maternal *hb* mutant and slightly towards both sides in the zygotic mutant (Figure 56B). As expected, the domain broadens significantly in the maternal & zygotic *hb* mutant until reaching from approximately 55 to 85% A-P position.

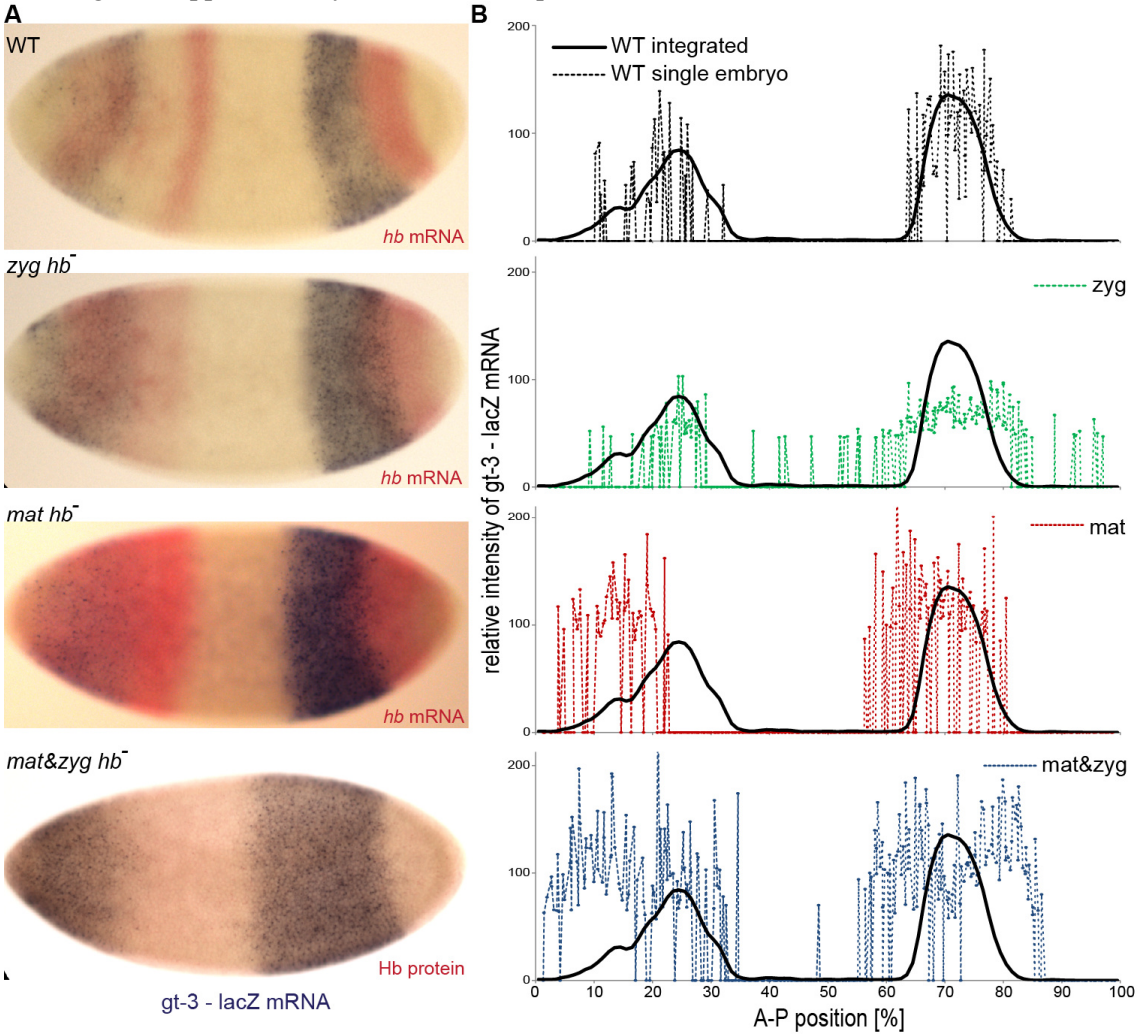


Figure 56: Expression driven by gt-3 expands in *hb* mutants.
(A) Enzymatic *in-situs* against lacZ mRNA (blue) driven by gt-3 in *hb* mutants, co-stained for *hb* mRNA or Hb protein (red) at mid C14A. The maternally contributed *hb* mRNA is detected in the zygotic mutant, and the zygotic *hb* contribution is picked up in the maternal mutant. No Hb protein expression was detected in the maternal & zygotic mutant. The lacZ and *hb* expression domains in the posterior slightly overlap in the maternal and in the zygotic *hb* mutant, whereas they do not overlap in the WT. **(B)** Quantification of lacZ mRNA in zygotic, maternal and maternal & zygotic *hb* mutants. The dotted lines are the fluorescent signals from a single embryo (same as in Figure 55) at time class T5 and the black line is the registered and integrated expression (n=91) in the wild-type background for comparison.

The effect on the anterior boundary of the gt-3 domain could be partially indirect via Kr, which depends on Hb in a concentration-dependent manner (Struhl et al. 1992, Schulz and Tautz 1994). Interestingly, the posterior domain starts to subdivide into two stripes at later stages in the zygotic, as well as in the zygotic & maternal mutants. This was also observed in the endogenous *gt* mRNA and in the posterior domain driven by gt-1 (see below). The ectopic anterior expression from the target lines broadens and intensifies in the zygotic & maternal mutants.

gt-1 in *hb* mutants

The *hb* mutant embryos carrying the gt-1-reporter were co-stained for endogenous *gt* mRNA instead of Hb protein (Figure 57). The expression driven by gt-1 follows the *gt* mRNA in the mutants and differs from the profile of gt-1 – lacZ of the WT (Figure 58). In mutants lacking any Hb at all (Figure 58B), the posterior domain of gt-1 and of the endogenous *gt* mRNA expands into both directions to similar extends as in gt-3 (Figure 56B). In the zygotic *hb* mutant (Figure 58D), the posterior domain driven by gt-1, as well as the endogenous *gt* mRNA expand to the anterior, whereas in the maternal *hb* mutant (Figure 58C), they shift to the anterior. Note that Figure 56 and Figure 58 compare the fluorescent lacZ-signals from single embryos to the registered and averaged profile from at least ten WT embryos. It is possible to observe qualitative changes, but there is certain embryo-to-embryo variability of the boundary positions.

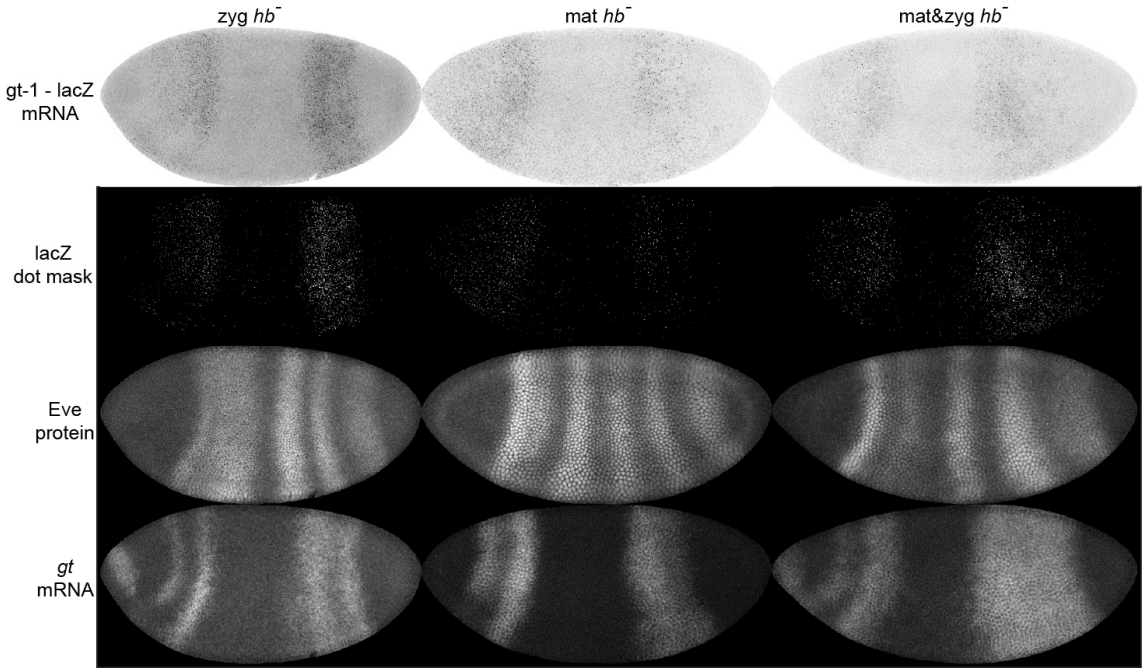


Figure 57: Endogenous *gt* mRNA and expression driven by gt-1 in *hb* mutants.
Fluorescent stainings of lacZ mRNA driven by gt-1 (inverted images), lacZ dot mask after thresholding endogenous *gt* mRNA and Eve protein in maternal, zygotic and maternal & zygotic *hb* mutants at T5.

A refinement of the posterior domain into two stripes for the endogenous *gt* mRNA, as well as for gt-1, can be observed to different degrees in the *hb* mutants. It is most pronounced for the endogenous *gt* mRNA in the zygotic mutant, whereas in the maternal & zygotic *hb* mutant, only the emergence of a subtle separation in the middle of the *gt* mRNA domain can be noticed.
Expression driven by gt-1 in these mutants tends to be weaker in the second half of the posterior domain.

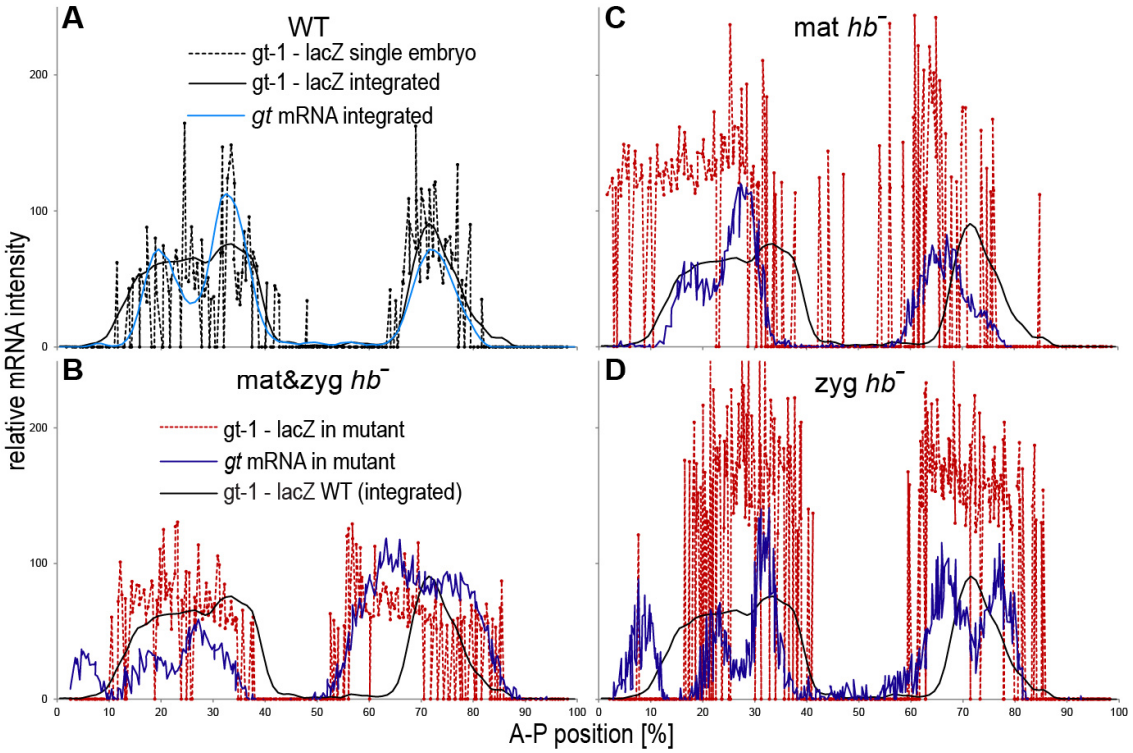


Figure 58: Endogenous *gt* mRNA and expression driven by gt-1 expand in *hb* mutants.
(A) Quantified fluorescent signals in the wild-type at time class T5. Shown is the lacZ mRNA driven by gt-1 from a single embryo (black dotted line) as well as the registered and integrated profiles of gt-1-lacZ (black line, n=41) and the endogenous *gt* mRNA (light blue, n=91). **(B, C, D)** Fluorescent signals of lacZ mRNA driven by gt-1 (red dotted lines) and endogenous *gt* mRNA (dark blue lines) from a single embryo of the different *hb* mutants at T5. The black line is the registered and integrated profile of gt-1-lacZ in the wild-type for comparison (as in A). Shown are the maternal & zygotic **(B)**, maternal **(C)** and zygotic **(D)** *hb* mutants.

The anterior domain is hardly affected in the *hb* mutants. As in the wild-type background, the subdivision of the anterior domain into two stripes in the *hb* mutants is less defined in the gt-1 – driven expression compared to the endogenous *gt* mRNA. The posterior boundary of the anterior *gt* mRNA domain appears to be further to the anterior in all *hb* mutant versions compared to the WT and also the anterior expression of gt-1 tends to follow this trend.
In summary, these results confirm that Hb does not activate gt-1, because the lacZ expression would be abolished in the *hb* mutant background. Models fit to gt-1 considering auto-activation fall into different scenarios, either accounting for repression by Kr with only minor influence from Hb or showing some Kni contribution (Figure 37). Hence, the derepression observed in the Hb mutants is not caused by missing Hb in this case, but by expanded Gt auto-activation and retracted Kr repression.

4.7.2 *Knirps* mutant

Surkova *et al.* (2013) quantified the protein expression of the gap genes Hb and Kr in the *kni* mutant, but unfortunately only qualitative information is available for Gt (Figure 59). The position of its domains is not altered but the level of the posterior is reduced significantly. In the *kni* mutant a similar series of effects like in the *Kr* mutant is provoked: Kni keeps the posterior Hb domain in place in the WT and prevents an intersection with the posterior Gt domain, but the lack of Kni makes this overlap between them possible. The position of Kr is not affected and hence Gt also stays where it was in the WT. One hypothesis was, that due to the increased intersection between Gt and Hb, minor repressive contributions from Hb onto Gt from the posterior manifest at later stages, when this effect intensifies via the lack of Gt auto-activation.

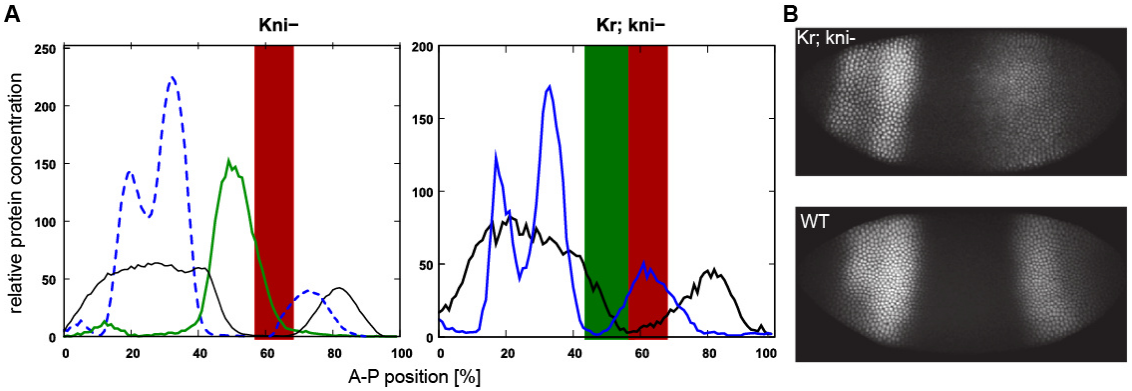


Figure 59: Protein expression in the *kni* and *Kr; kni* double mutant.
Taken from Surkova (2013). **(A)** Superimposed integrated patterns of gap genes from time class T7 for *kni*¯ and *Kr;kni* ¯ embryos. The broken blue line shows the wild-type Gt protein expression with the diminished posterior domain, because in *kni* mutants it is not displaced as compared with wild-type (Kraut and Levine, 1991b). Vertical color bars illustrate Kr (green) and Kni (red) positions in the WT. Hb is shown in black. **(B)** Gt protein expression at T4 in the *Kr; kni* ¯ double mutant compared to WT.

I monitored the expression driven by the CREs gt-3 and gt-1 in the same *kni* deficiency mutant. The Eve protein pattern shows derepression between the 3rd and 7th stripe in the homozygous *kni* mutants and a decrease in the intensity of the 5th stripe in the heterozygous embryos (Figure 60B and C).

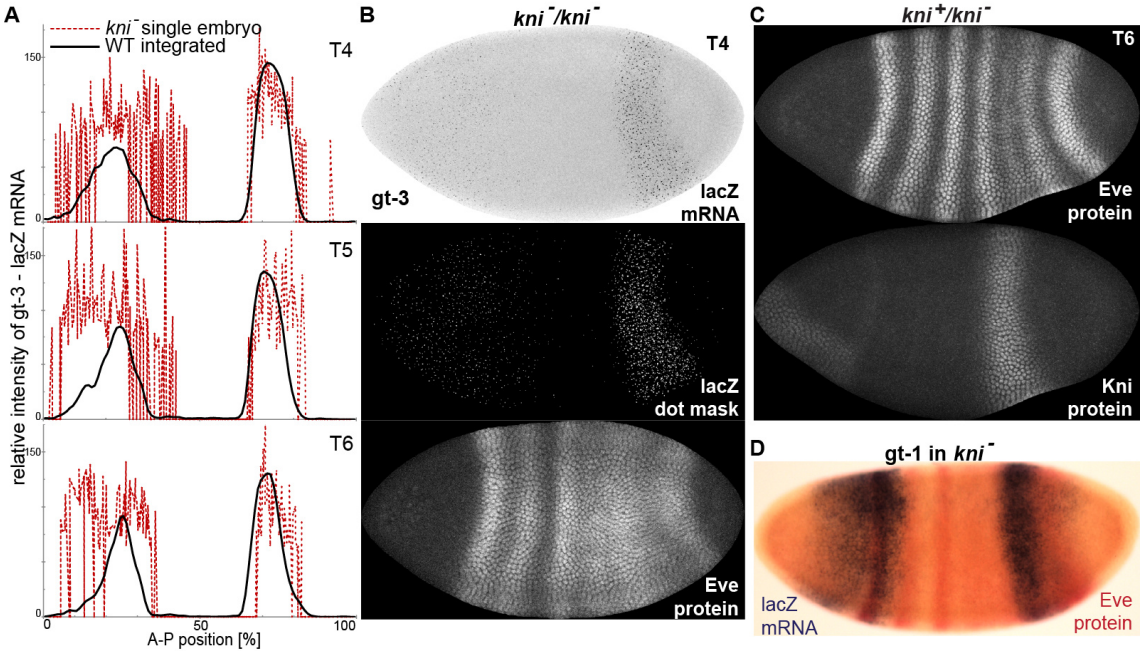


Figure 60: Expression driven by gt-3 and gt-1 is not altered in the *kni* mutant.
(A) Quantification of gt-3 – driven lacZ mRNA in a deficiency mutant lacking *kni*. The red dotted lines are the lacZ fluorescent signals from a single embryo at T4, T5 and T6, and the black line is the registered and integrated expression in the WT for comparison. **(B)** Fluorescent stainings of lacZ mRNA driven by gt-3 (inverted image), lacZ dot mask after thresholding, as well as Eve protein at T4. The embryo was co-stained for Kni protein, but no expression was observed (data not shown). **(C)** Heterozygous embryos were distinguished from homozygous based on the Eve pattern and on the Kni protein stain. The 5th Eve stripe in the embryos heterozygous for the deficiency is weaker. **(D)** Enzymatic *in-situ* against lacZ mRNA (blue) driven by gt-1 at T7. Homozygous *kni* mutants were identified based on the Eve protein pattern (red).

The gap protein patterns are in general linear read-outs from their respective mRNA (Becker et al. 2013). Since the boundaries of the Gt protein domains are not altered, but the level of the posterior is reduced, a similar scenario was expected for the lacZ mRNA. In fact, no changes in the positions of the domains driven by gt-1 and gt-3 were observed, but also no obvious decrease in expression levels could be detected (Figure 60A and D). This result raises the question whether there is post-transcriptional repression acting in the mutant but not in the wild-type or whether another CRE in the locus is responsible for adjusting the levels via repressive inputs.

4.7.3 *Bicoid* and *caudal* mutants

I attempted to test gt-1 in a *bcd* null mutant background, but it was not possible to obtain homozygous viable females to set up the crosses. However, it would not have brought any direct insights, because abolishment of the anterior domain was expected in any case, since lack of Bcd would have led to the loss of the anterior expression driven by the CRE gt23 and subsequently no anterior Gt protein would be available to auto-activate gt-1 in the anterior region. Since Segal et al. (2008) claimed that gt-1 depends on cooperative binding to weak Bcd sites, I tested this CRE in a *bcd* cooperativity mutant containing the point mutation S35T in the homeobox (Lebrecht et al. 2005), but no obvious differences in the anterior domain were observed (data not shown). Unfortunately, the other *bcd* cooperativity mutant line, which was shown to abolish the anterior domain of the endogenous *gt* mRNA, had lost its point mutation at K57R.

In a *cad* mutant, a similar effect as for *bcd* would apply to the posterior domain: without Cad, gt-3 cannot be activated and no endogenous Gt protein would be available to auto-activate gt-1 in the posterior. Unexpectedly, also the anterior domain from gt-1 and from the combined CRE was abolished in the maternal and zygotic *cad* mutant. Since the introduction of the zygotic *cad* mutant into the CRE fly lines resulted in weak stocks, I assume that subsequently, the combination with the maternal *cad* mutant lead to a genetic background with too many alterations and secondary effects.

5 Conclusions and future perspectives

This PhD thesis has shed light onto the mechanisms governing transcriptional regulation of the gap gene *giant* via a combination of experimental and modelling approaches. To our knowledge, it is the first report of such a modeling-based in-depth analysis of CREs for a gap gene, because similar earlier studies mainly focused on stripe enhancers of the pair-rule gene *eve*. Although the expression pattern of *gt* is simpler compared to *eve*, it is the most complicated among the gap genes, because it refines from two broad domains into four separate stripes within cleavage cycle 14. Additionally, it used to cause troubles when modeling the entire gap gene network in the trunk region of the embryo due to incompatible regulatory requirements for the anterior and posterior domains. The datasets presented here monitor the dynamical expression driven by distinct CREs and the modelling and experimental results show that they are differentially regulated.

5.1 Differential regulation of two adjacent CREs

Three distinct CREs are exclusively driving one particular part of the pattern, which are the anterior tip, the anterior or the posterior domain. An additional CRE, gt-1, is capable of driving the anterior and the posterior simultaneously and hence was of particular interest due to this duality. This study focused on the posterior *gt* domain, which is driven by the CREs gt-3 and gt-1. The latter is located directly upstream of the core promoter and adjacent to gt-3. It has to be kept in mind that the borders of these elements were defined based on the cut-off value of an arbitrary motif score. Hence, the first objective was to quantify their expression patterns in order to see if they are able to reproduce the exact endogenous expression over time. The expression boundaries of both CREs coincide with the pattern of the endogenous mRNA within C14A, but there are differences in the levels and in the time of emergence of expression.

5.1.1 Early vs. late regulation

The observed differences in the timing of expression driven by the distinct CREs, reflects varying underlying regulation for early versus late expression. The anterior domain driven by the CRE gt23 arises in C11 due to activation by Bcd. It does not harbor Hb binding sites, because it needs to avoid influences from such a repressor with a broad overlapping domain.
Gt-3 also arises at C11 and contains several high scoring Hb sites, which assure repression in the anterior region of the embryo. The two expression domains driven by gt-1 emerge in C13 only and this CRE contains fewer and weaker Hb sites than gt-3. Additionally, it lacks sites for the maternal activators Bcd and Cad. This is biologically reasonable for a late element, because otherwise we would observe expression at earlier stages. Hence, initially *gt* expression is driven by separate CREs for each domain, whereas at later stages, gt-1 could function as a booster element required for maintenance.

5.1.2 Activation via maternal gradients vs. auto-activation

The gap genes are activated by the maternal protein gradients Bcd and Cad and also some evidence for auto-activation exists. Expression driven by the element gt-3 arises earlier and is activated by Cad, whereas the element gt-1 was of particular interest, since it does not contain Bcd sites and only a few very weak sites for Cad. Two hypotheses were that it depends on Gt auto-activation or that the bimodal factor Hb triggers the expression. If the latter was true, we would expect that expression arises as early as from the other two CREs, because Hb is contributed maternally. Both scenarios were tested *in silico* and experimentally. Models including Hb co-activation were not capable of fitting the gt-1 data properly, because they did not reach the observed levels and the actual contributions from Hb were negligible (run 141). If

Hb would be activating gt-1, we would anticipate reduced or abolished expression in *hb* mutants. I tested the CRE in the maternal, zygotic and maternal & zygotic *hb* mutant and observed expansion of the posterior domain similar to the one in gt-3 (section 4.7.1). Based on this, I can exclude Hb co-activation and additionally the observed behavior would be consistent with the theory of gt-1 auto-activation.
Models considering Gt as an activator improved the fits substantially for all three datasets, but this result needs to be treated very carefully. If a factor expressed at almost the same position as the target pattern, is included as activator, it will of course always lead to the improvement of the fit. Even if the activating input, in this case the Gt protein, trails behind the target, the *gt* mRNA, over time, the repressors in the model could adjust for this temporal shift. The model fit to gt-3 considering only Gt auto-activation without input from any factor is not capable of placing the boundaries correctly due to this discrepancy between the mRNA and protein time courses (Figure 45). Interestingly, this shift is hardly detectable in the case of gt-1 under the same modelling conditions. Nevertheless, these control runs are not a definite proof that Gt is auto-activating in gt-1. I mutated all predicted Gt binding sites in the reporter constructs for gt-3 and gt-1 with particular care in order to not to destroy or generate TFBS for other factors involved in blastoderm patterning (section 4.5). The expression driven by gt-3 did not show any obvious differences compared to the WT CRE, whereas the anterior domain of gt-1 was abolished and the posterior reduced. In summary, the model and the experiments strongly suggest that gt-1 depends on Gt auto-activation, whereas gt-3 is activated by Cad.

5.1.3 Context-dependent activator/repressor switches

I tried to find out which motif could be responsible for the switch of Gt from a repressor to an activator. Some other TFs have a bimodal role and can function as both, but how such switch is achieved at the molecular level is not entirely clear. It tends to depend on the sequence context, meaning on nearby binding sites for another particular factor. Hb was shown to function as a repressor in MSE3 and as an activator in MSE2 due to nearby Bcd sites (Small et al. 1991, 1996). Snail was considered a dedicated repressor until recently, when it was shown to potentiate Twist-mediated activation of mesodermal CREs (Rembold et al. 2014). Comparison of gene expression levels between *snail* mutants and WT revealed 52 Snail-activated and 50 repressed CREs. The search for differentially enriched motifs pointed to a *Tll*-like motif, which upon mutagenesis led to the reduction of reporter gene expression *in vivo*.
I checked if binding sites for other factors are differentially enriched in either gt-1 or gt-3. The candidate list (Kenneth Barr, personal communication) included 23 factors expressed in the early embryo and I added Zld, Bcd, Cad, Hb and Gt itself. I used a distance matrix, as well as the statistical tool *iTFs*[8] for finding sequence signatures of interacting transcription factors (Kazemian et al. 2013). It checks for a possible distance and orientation bias of two PWMs on the input sequence. No major differences were observed, apart from minor cooperative binding of Pangolin and Pannier within 50 bp in gt-1. In principle, it is possible to test candidates that trigger co-activation of Gt with this model, but it requires the concentrations of such factors as input.

5.1.4 Repressing contributions

The model requires repressive input from Hb for gt-3, even when co-activation was included (Figure 35), in order to avoid leakage in the anterior region. The opposite is the case for gt-1: repression by Hb is negligible in models fit to gt-1 considering auto-activation (Figure 37), because it would interfere with the activation of the anterior domain of this CRE. In order to differentiate between the contributions of maternal and zygotic Hb, I generated fly lines with CREs for both *hb* mutant backgrounds, as well as a maternal & zygotic mutant (section 4.7.1). The posterior domain of the endogenous *gt* mRNA, as well as of gt-3 and gt-1 expands to a similar extend in the maternal and in the zygotic *hb* mutant. It is positioned a bit more to the

[8] http://veda.cs.uiuc.edu/cgi-bin/iTFs/search

anterior in the maternal *hb* mutant compared to the WT. In the maternal & zygotic *hb* mutant, the posterior domain of the two CREs and of *gt* expands even much further until being at least twice as broad compared to the WT. Moreover, a subdivision of the posterior domains into two stripes was observed in the *hb* mutants. This was most pronounced for gt-3 in the maternal & zygotic mutant, as well as for gt-1 and the endogenous *gt* mRNA in the zygotic mutant. The anterior domain of gt-1 and of the endogenous *gt* mRNA is not affected in the *hb* mutants, apart from a slight shift of the posterior boundary towards the anterior. The posterior boundary position of the gt-3 domain depend on the repressive strength of the remaining posterior Hb expression, whereas the effect on the anterior boundary could be partially indirect via Kr, which depends on Hb in a concentration-dependent manner (Struhl et al. 1992, Schulz and Tautz 1994). Although according to the modelling solutions, gt-1 is not directly regulated by Hb, we observe a similar result for the posterior domain as with gt-3 in all three *hb* mutant backgrounds, because gt-1 depends on auto-activation. Its domain still expands since the expression of endogenous Gt protein is broadened due to gt-3 misregulation. So far, no datasets for the *hb* mutants are publicly available. The lab of Angela DePace has generated an expression atlas of the most common genes in a maternal *hb* mutant, but used a different staging scheme (Max Staller, personal communication). It would be interesting to use their dataset to plug-in the protein concentrations of the TFs into my models and predict the expression of the CREs in this *hb* mutant.

Most model fits suggest Kr as the major repressor in the middle region of the embryo, as previously reported (Kraut and Levine 1991a) and all selected models predicted a broadening and a shift of the posterior domain towards the anterior in the *Kr* mutant (section 4.6.1). This was confirmed experimentally for gt-3 and also for the other overlapping posterior CRE CE8001 (Berman et al. 2002). However, unexpectedly, a three-striped pattern was observed for gt-1, which might be a locus effect and/or experimental issue (as discussed in section 4.6.1).

Subsequently, the expression driven by the combined fragment looks like an overlay of the expanded domain from gt-3 and the three stripes from gt-1. Interestingly, models for gt-1 considering auto-activation fall into two different scenarios: they show either repression by Kr or by Kni. Hence, we would expect two different solutions from their predictions in the *Kr* mutant. However, both models predict exactly the same expression profile in the mutant, which suggests that the setting of the boundaries does not depend on the repressors but on Gt itself.

Repression by Tll from the pole is found in all model outputs and has been observed for endogenous *gt* in *tll* mutants (Kraut and Levine 1991b). The selected models predict an expansion of the posterior domain to the posterior in the *tll* mutant for all three CREs, which was confirmed experimentally in the case of gt-3 (section 4.6.2). The positions of the endogenous Gt protein domains are not altered in *kni* mutants (Surkova et al. 2013). Usually, the models do not suggest contribution by Kni and no changes in the expression patterns driven by gt-3 and gt-1 were observed in the *kni* mutant (Figure 60).

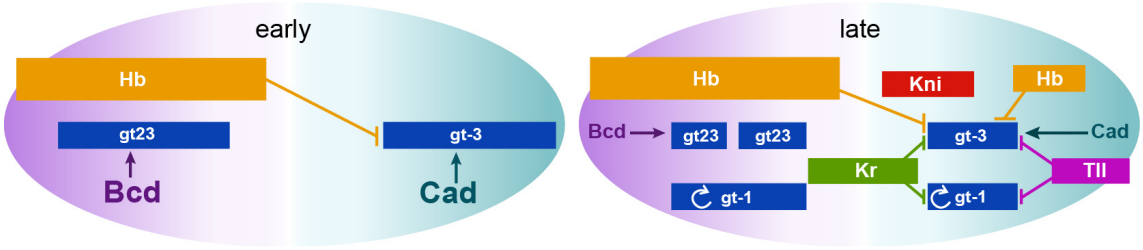


Figure 61: Regulation of *gt* in the *Drosophila* embryo.
The distinct *gt* CREs receive different regulatory inputs, which define their temporal dynamics.

5.2 Comparison with similar transcriptional models of *Drosophila* segmentation

In contrast to most modeling approaches in the field of transcriptional regulation, this study does not aim at predicting expression patterns on a genome-wide scale. Instead it is an in-depth analysis of the regulatory mechanisms of a single gene with high spatio-temporal resolution. The model is fit to data from the gene of interest only (as in Janssens et al. 2006, Kim et al. 2013), whereas other studies fit to a training set of about 40 CREs (Segal et al. 2008, He et al. 2010, Kazemian et al. 2010). I fitted the model to ten time points instead of only one at mid-blastoderm stage, which introduces the aspect of expression dynamics, lacking in most previous studies (see Janssens et al., 2006, for an exception). These are probably the two main reasons why this model fits the data more closely than the models of the other studies (Figure 14). Another main difference lies in the implementation of the mechanisms and in particular, the concept of adaptor factors (Reinitz et al. 2003). Finally, the suggested mechanisms were confirmed by several experiments and the model is capable of predicting the expression patterns of mutants. Similar studies with transcriptional models of *Drosophila* segmentation do not present such an extensive experimental validation.
As one example, I want to compare my conclusions to the results for *gt* of the model from Segal et al. (2008). The criticisms about their model, apart from fitting the PWMs, and sampling instead of optimizing parameter values, are that the model does not fit the data very well and several mechanisms in the model contradict genetic evidence. Their model predicted repression from Kni for gt-1, gt-3, gt-10. In contrast, such a contribution was not suggested by the models in this thesis and I have shown that the expression of gt-1 and gt-3 is not altered in the *kni* mutant (Figure 60). Moreover, they claimed that Bcd cooperativity is the driving force for gt-1 in the anterior region (Figure 12). In my models for gt-1, the incorporation of cooperative binding between nearby Bcd sites was not able to achieve sufficient expression in the anterior. Additionally, I have shown that upon mutation of the Gt sites in this CRE, the anterior domain becomes abolished. This demonstrates that our modeling results are more accurate and specific than previous efforts.

5.3 Limitations of the approach

5.3.1 Experimental limitations

In general, transcriptional modelling requires the concentrations of the TFs as input. In the case of the *Drosophila* blastoderm, such datasets for factors expressed in the anterior region are missing, and hence, the model cannot account for anterior regulation. Also the generation of a dataset for a CREs is labor-intensive and time-consuming due to the necessity for the lateral orientation of the embryos on the slide, the minimum of ten embryos per time class, the precise manual staging and the registration procedure. Unfortunately, lacZ turned out to be an inappropriate reporter gene that only allows extracting binary information for each nucleus. Substantial experimental efforts were taken to avoid positional effects from the integration site, but nevertheless orientation-dependent differences were observed. Other uncertainties concern post-transcriptional regulation or additional effects in the mutants that are silent in the WT.

5.3.2 Limitations of the modelling

Auto-regulation is a critical mechanism for the model to deal with, because it has an elevated compensatory potential that can result in high quality of fit despite wrong underlying regulatory contributions. Subsequently, it will be difficult to assess this mechanism with a model for biological systems, where auto-activation is indeed functioning. Additionally, the model is static and the shift of the protein patterns cannot be reproduced over time. A dynamical model would

account for time-dependent changes in the system and might be able to resolve issues with auto-activation. Moreover, the model cannot deal properly with two distinct roles for a TF at the same time, as in the case of Gt, which is either neutral or auto-activating. For this reason, the simultaneous optimizations were not able to ascertain the correct TF contributions. Also predictions of the expression of the combined CRE with models fit to the separate elements failed due to this differential behavior of the distinct CREs. This is not necessarily a short-come of the model, but rather another example that predictions of enhancer expression in batch need to be seen more critically.
Prediction of the inter-enhancer sequences of the *gt* locus was expected not to show any expression. The model optimized to gt-3 or the combined fragment (runs 132_1 and 91_1) gave only some baseline expression for the region in between gt-3 and gt-6, but the models optimized to gt-1 (runs 139) resulted in saturation mainly due to Cad sites. For the region between gt-6 and gt-10, as well as the 5 kb upstream of gt-10, the selected models predicted a posterior domain or even complete saturation over the entire trunk region. This is probably an accumulative effect, because the model was trained on sequences of 1 to 2 kb, whereas these inter-enhancer sequences span 3 to 5 kb. Moreover, they are presumably affected by chromatin structure and introduction of nucleosome positioning sequences into the model might help to predict non-expressing regions correctly (Raveh-Sadka et al. 2009, Wasson and Hartemink 2009, Wilczynski et al. 2012). We assume an under-representation of nucleosome positioning sequences in the CREs of developmental genes expressed at early stages (Papatsenko et al. 2009). The fact that reporter constructs with CREs are able to drive expression after insertion into a random locus is a sign that this information must be intrinsic on the enhancer sequence.

5.4 Future perspectives

5.4.1 Alternative experimental validation of Gt auto-activation

It has to be kept in mind that the strategy was to achieve the knock-down of every potential Gt site including the very weak ones, which led to the introduction of 28 and 86 point mutations in gt-3 and gt-1, respectively. Since this is an immense amount, unwanted effects might have been introduced. A construct with only the three strongest Gt sites mutated could help to resolve these doubts. Such an experiment could have three possible outcomes: it might lead to the same result as with all sites mutated, which would be a perfect confirmation of the auto-activation. If interactions were altered in the 1st experiment, it could give a different result. If the pattern looks like the expression driven by the WT CRE, it would show that three Gt sites are not sufficient to auto-activate.
In order to achieve a more accurate and quantitative measurement of absolute transcript levels, single molecule FISH (smFISH) protocols were proposed for yeast, *C. elegans* and *Drosophila* (Raj et al. 2008, Little et al. 2011, Trcek et al. 2012). This method uses fluorescently labeled 20mer oligos which simultaneously bind to the nascent transcripts, which are then imaged and counted if they were detected in consecutive z-stacks. I tried out this method applying one set of 48 oligos to the entire lacZ ORF, labeled with Fluorescein495 and two sets hybridizing to one half of the ORF each, labeled with TAMRA557. Only with the second approach, a weak signal was obtained, but photo-bleaching complicated the imaging. Additionally, the signal did not appear as the fine spots reported in other publications, but it rather looked like the lacZ aggregates usually seen in the normal FISH.
An alternative experimental approach to monitor the role of the TF Gt on its own enhancers would be a cell line assay. Transfection or electroporation of *Drosophila* S2 or Kc167 cells are usually applied for this purpose. The CREs need to be placed in front of a leaky promoter in order to drive some baseline expression of a bioluminescent or fluorescent reporter such as luciferase or GFP. Addition of purified Gt protein or its co-expression under a constitutive promoter would then either increase, reduce or not alter the baseline levels.

5.4.2 Is gt-1 essential or redundant?

It is not entirely clear, if the seemingly redundant element gt-1 is required for viability. It could be that this CRE acts as a booster for later stages or that it confers robustness under difficult environmental conditions or if perturbations of the system occur. A new homologous recombination approach in *Drosophila* allows for precise modification of an endogenous locus (Baena-Lopez et al. 2013). The fragment of interest is replaced with an attP site and two markers flanked by loxP sites. It would be interesting to see if such a knock-out of gt-1 gives viable flies.

BAC recombineering is another method to investigate the consequences of a missing CRE, while the rest of the locus is still intact (Venken et al. 2009). A BAC clone containing the entire *gt* locus needs to be modified (Perry et al. 2010, 2011). The *gt* CDS can be replaced with a *yellow*-Kanamycin cassette, because this pigment gene is only expressed in the adult fly and can be used with an intronic probe to monitor nascent transcripts. The CRE of interest can be replaced with an Ampicillin cassette, which is supposed to be a "neutral" sequence, in order to keep certain spacing and to enable selection of positive clones after recombineering. Subsequently, the modified BAC is integrated into the genome and the expression driven by the *gt* locus without gt-1 is quantified and compared to the unmodified BAC. Different conditions, such as temperature and distinct genetic backgrounds can be assessed. The question, if gt-1 and gt-3 are additive, could also be tackled with this method.

6 Materials and methods

6.1 Generation of reporter fly lines

Reporter-constructs harboring a *gt* CRE, the endogenous *gt* core promoter (Juven-Gershon et al. 2008), the lacZ gene and the alpha-tubulin 3´UTR were integrated into predefined attP target lines via recombinase-mediated cassette exchange with the ΦC31 integrase (Bateman et al. 2006).

Cloning
The endogenous *giant* core promoter is 80bp long and includes a TATA box, an initiator (at A +1 bp) and a downstream core promoter element (DPE) at +28 to +32 bp (Juven-Gershon et al. 2008). A fragment including the core promoter, the 5´UTR and the first six codons of the *gt* CDS was amplified from *Dmel* OregonR gDNA using primers with restriction sites for *HindIII*, AvrII and *BamHI* (Table A. 1).

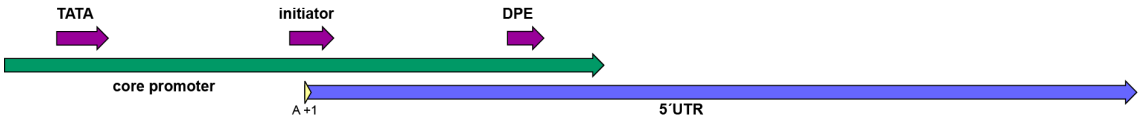


Figure 62: Endogenous *giant* core promoter.
The TFIID complex binds cooperatively to the TATA box, initiator and downstream promoter element.

The plasmid piB-GFP (Bateman et al. 2006) was digested with *HindIII* and *BamHI* in order to leave over the vector backbone with the attB sites. Subsequently, the promoter fragment was ligated with the attB-backbone via *HindIII* and *BamHI*. LacZ, starting with the 8th codon and including the alpha-tubulin 3´UTR, was amplified from the plasmid attB-hZ (gift from Miki Fujioka). It was ligated into the attB-backbone featuring the promoter via *SalI* and *HindIII* to generate the empty plasmid attB-Pgt-Z, which was then digested with *AvrII* and *SpeI* to clone the CREs. Additionally, controls carrying the *hsp70* or the *eve* basal promoter were amplified from attB-hZ and attB-eZ (gift from Miki Fujioka) and included the first six codons of the corresponding CDS.
The CREs gt-1, gt-3, gt1, gt23 and CE8001, as well as the combined fragment gt-1-gt-3 were amplified from OregonR gDNA using primers including restriction sites for *AvrII* and *SpeI*. After subcloning into pGEM-T (Promega), they were excised and ligated into attB-Pgt-Z. Interestingly, the amplified sequence of gt-1 (Appendix) carried several point mutations and gaps compared to the published *D.melanogaster* genome and the original CRE sequence (Schroeder et al. 2004). Therefore, I additionally used gDNA of the target line 37B and the *white* mutant line w^{1118} as templates, which showed the same polymorphisms and binding site turnover for Hb. All restriction enzymes were purchased from Fermentas.

Injections
For the injections, the plasmids were isolated with a Midiprep kit (Qiagen) from DH5alpha, EtOH precipitated and dissolved in water. The attP target lines (y$^-$w$^-$, nanos - phiC31 - y$^+$; attP w$^+$ attP) harboured a *Dmel* codon usage optimized phiC31 integrase on the X chromosome (Bischof et al. 2007) and the target site marked by *white* in the cytological positions 37B7, 38F1 (both II chromosome) or 89B8 (III chromosome) (Bateman et al. 2006). Injections (Table 6) were performed with a microinjector (Eppendorf FemtoJet) following the general protocol (Fish et al. 2007) or contracted with BestGene Inc. (Chino Hills, USA) or Rainbow Transgenic Flies, Inc. (Camarillo, USA). The empty plasmid attB-Pgt-Z was used as negative control in the injections. In some cases, high and low DNA concentrations were injected.

construct	line	conc. [ng/µL]	embryos	larvae	adults	fertile crosses	pos. vials	injected by
Pgt - gt-3	37B		124	27	22	16	2	H.J.
Pgt - gt-3	89B		112	22	12	11	3	A.H.
neg. ctrl Pgt	89B		143	48	20	16	7	A.H.
neg. ctrl Pgt	37B		108	23	21	17	7	A.H.
Peve - gt-3	37B	20	41	15	11	9	3	A.H.
Peve - gt-3	37B	200	100	59	33	21	12	A.H.
neg. ctrl Peve	37B	20	67	17	12	8	3	A.H.
neg. ctrl Peve	37B	250	81	28	13	6	3	A.H.
Phsp - gt-3	37B	260	99	19	7	4	2	A.H.
neg. ctrl Phsp	37B	33	120	57	42	34	21	A.H.
Pgt - gt-1	89B	15	71	61	32	26	3	A.H.
gt-3 - gt-1 combined	89B	30	113	12	4	4	1	A.H.
gt-1_mutated	38F		200	155	100	68	8	R.F.
gt-3_mutated	38F		200	160	82	43	1	R.F.
gt-1_mutated	89B		200	110	44	14	2	R.F.

Table 6: Injections of reporter-constructs into site-specific target lines.
Constructs with the endogenous *giant* (Pgt), the *even-skipped* basal (Peve) and the heat-shock protein 70 (Phsp) promoters were tested. The CREs gt-3, gt-1 and the combined fragment were injected, as well as empty vectors without any CRE (negative control) and mutated CREs. The fly lines from Bateman et al. (2006) with integration sites at 37B, 38F (II. chromosome) and 89B (III. chromosome) were used. Different DNA concentrations were tested. The number of injected embryos, as well as survived larvae and adults are indicated. Crosses with single flies were set up and the fertile ones were counted. Positive vials (highlighted in grey) are fertile crosses giving rise to transgenic offspring. Injections were performed by Astrid Hoermann (A.H.), Hilde Janssens (H.J.) or Rainbow Facilities (R.F.).

The adults were crossed to *white*$^-$ double balancers for the II or the III chromosome, respectively (BS5439, BS2537, BS3720 or Bl/CyO; TM2/TM6) and the offspring was screened for the loss of *white* (Table 7). During the crosses I got rid of the integrase by selecting the males and finally the stocks were homozygoused.

construct	line	efficiencies %					
		larvae / embryos	adult / larvae	fertile / embryos	fertile / adult	pos. vials / embryos	pos. vials / fertile
Pgt - gt-3	37B	22	81	13	73	2	13
Pgt - gt-3	89B	20	55	10	92	3	27
neg. control Pgt	89B	34	42	11	80	5	44
neg. control Pgt	37B	21	91	16	81	6	41
Peve - gt-3	37B	37	73	22	82	7	33
Peve - gt-3	37B	59	56	21	64	12	57
neg. control Peve	37B	25	71	12	67	4	38
neg. control Peve	37B	35	46	7	46	4	50
Phsp - gt-3	37B	19	37	4	57	2	50
neg. control Phsp	37B	48	74	28	81	18	62
Pgt - gt-1	89B	86	52	37	81	4	12
gt-3 - gt-1 combined	89B	11	33	4	100	1	25
gt-1_mutated	38F	78	65	34	68	4	12
gt-3_mutated	38F	80	51	22	52	1	2
gt-1_mutated	89B	55	40	7	32	1	14

Table 7: Efficiencies, survival rates and fertility after injection.
For comparison of efficiencies, the integration efficiencies are used, which represent the positive vials containing transgenic offspring per fertile cross (highlighted in grey).

Orientation PCR
Genomic DNA was isolated either from an over-night egg collection or from 1 or 2 adult flies using Chelex beads (Biorad). The samples were crushed in 100µL Chelex (1,25mg/25mL in H_2O) with 5µL ProteinaseK (600U/ml, Fermentas) and heated at 55ºC for 2h. After mixing with the vortex, the samples were heated to 99ºC for 10min, mixed and centrifuged at max. speed for 3min. The supernatant was transferred to a new tube and centrifuged before usage. In order to know in which orientation the reporter cassette was integrated into the target site, two PCRs using the primer Hind-gtCDS-F with either the primer LL5 or LL3 (see Appendix) were performed with Advantage 2 polymerase (Clonetech). In case of 5´orientation a band with the size of the corresponding CRE plus 372 bp was expected using primer LL5 and at the same time primer LL3 should not give any PCR product.

6.2 *In-situ* hybridization and immuno-staining

Riboprobes
The plasmid BSKSII$^+$-lacZ (gift from Steve Small) was digested with *PstI* and reverse transcription was performed with T3 in Biotin-, DIG- or FITC-labeling mix (Roche) to generate lacZ RNA. Subsequently the probes were fragmented in Carbonate buffer for 40 minutes and purified with Mini Quick Spin Column (Roche). The plasmid pFLC.1-Dmel Gt (BDGP DGCR RE29225) was digested with *SmaI*, which cuts in the CDS, and reverse transcription was performed with T3 in DIG- or FITC-labeling mix (Roche) to give a 790 bp *giant* probe.

Embryo fixation and stainings
After 0-4 h or 1-5 h of egg laying at room-temperature on apple juice plates, the embryos were dechorionated in 50% bleach, fixed in 5% formaldehyde and devitellinized with methanol. Fluorescent *in-situ* hybridization (FISH) was performed according to slightly modified standard protocols (Hughes and Krause 1998, Kosman et al. 1998, Wu et al. 2001) using acetone permeabilization. First, the mRNA is hybridized with the corresponding RNA probe, then the label of the probe is recognized by the complementary primary antibody, which will finally be tagged with the secondary fluorophore-conjugated antibody (Table 8). The embryos were also stained for the *Eve* protein, necessary for time classification and data registration, as well as for the nuclei with Hoechst34580. The quantitative datasets were generated by FISH with lacZ and *gt* riboprobes. Initially, eight different antibody-combinations were tested on embryos from the line 37B/ Pgt- gt-3-lacZ and the best one was used for the datasets.

	lacZ mRNA	**endogenous *gt* mRNA**	**Eve protein**
riboprobe	Biotin	FITC	
1º antibody	mouse anti-Biotin	rabbit anti-FITC	guinea-pig anti-Eve
2º antibody	anti-mouse-Alexa488	anti-rabbit-Alexa647	anti-guinea-pig-Alexa555

Table 8: Labeling and antibody combination of the quantitative datasets.

Enzymatic stainings were performed according to modified protocols (Crombach et al. 2012) but using Acetone instead of ProteinaseK permeabilization. AP-conjugated anti-DIG or anti-FITC antibodies (Roche) were applied, correspondingly. NBT/BCIP (Roche) and in case of double stainings also FastRed (Roche) were used as substrates. If necessary, nuclei were stained with DAPI. Embryos carrying CREs in mutant backgrounds were stained either fluorescently or enzymatically. In case of fluorescent staining, the usual lacZ combination was used and additionally, either the mRNA or the protein of the corresponding mutant was stained. If necessary, rabbit-anti-Eve combined with anti-rabbit-Alexa647 was applied instead of the usual combination.

guinea pig - α- Eve or rabbit - α- Eve
rabbit - α- Gt
guinea pig - α- Hb
guinea pig - α- Kni
guinea pig - α- Tll or rabbit - α- Tll

Table 9: Antibodies used in this study.
The polyclonal antisera used in this study are from David Kosman (Kosman et al. 1998), except rabbit-α-Eve, which was provided by Manfred Frasch.

6.3 Confocal microscopy and image processing

For the quantitative datasets the aim was to scan at least 10 laterally oriented embryos per time class. Based on the expression patterns of the *giant* mRNA and the *Eve* protein it was decided if the embryo was lateral, dorsal or ventral.

Microscope settings
The fluorescently stained embryos were mounted on a microscope slide in Prolong (Invitrogen). Images were taken on a Leica TCS SP5 confocal microscope using a 20x glycerol immersion objective (HCX PL APO lambda blue 20.0x0.70 IMM UV). Two z-positions with approximately 1 µm distance were defined by focusing on the lacZ signal and the nuclear layer. Note that due to the axial chromatic aberration, not all 4 channels (Table 10) are perfectly aligned and hence, the pictures for the endogenous *gt* mRNA display sections which are slightly deeper in the embryo. Photo-multiplier gain and offset were adjusted for each channel separately according to the brightest embryo on each microscope slide. The four channels were scanned sequentially with a speed of 400 Hz, at a resolution of 1024x1024 pixels and applying a digital zoom of 1.3x. The line-accumulation was set to 2 and the frame-average to 4. Additionally, a membrane picture was taken in DIC mode with the same objective, but it had to be refocused on the midsagittal plane. For DIC imaging we used a resolution of 4096x4096 pixels and a line-average of 4.

channel	fluorophore	excitation [nm]	emission [nm]
1	Hoechst	405	417 - 480
2	Alexa488	488	499 - 530
3	Alexa555	543	553 - 618
4	Alexa647	633	652 - 700

Table 10: Excitation and emission wavelengths of the four data channels

Image processing
The image processing pipeline (Surkova, Myasnikova, et al. 2008) was implemented in MatLab by Johannes Jaeger and adapted by Damjan Cicin-Sain. First, an embryo mask is created from the confocal microscope images in order to crop and align the embryos with anterior to the left and dorsal up (Figure 63A). The two optical sections for each channel are averaged in the case of the nuclei and maximized in the case of the other channels. During image segmentation, a binary nuclear mask is created based on the nuclear stain using either the watershed (for C14 embryos) or the threshold method (for earlier embryos). The Shen-Castan edge detection algorithm further refines the regions to give the precise boundaries of the nuclei (Figure 63D). Afterwards, the x and y positions of the centroids of the nuclei can be calculated and the fluorescence intensity is averaged over all pixels in a nucleus. In the case of Eve protein, edge detection was performed to consider only nuclear expression (Figure 63C), whereas this step was omitted for lacZ and *gt* mRNA in order to capture also the cytoplasmic staining (Figure 63B). Since the centroids of the nuclei for Eve and the centroids of the regions for *gt* and lacZ

do not precisely coincide, the algorithm searches for the nuclei within the regions and deletes the ones which are left over, which tend to be the ones at the periphery.

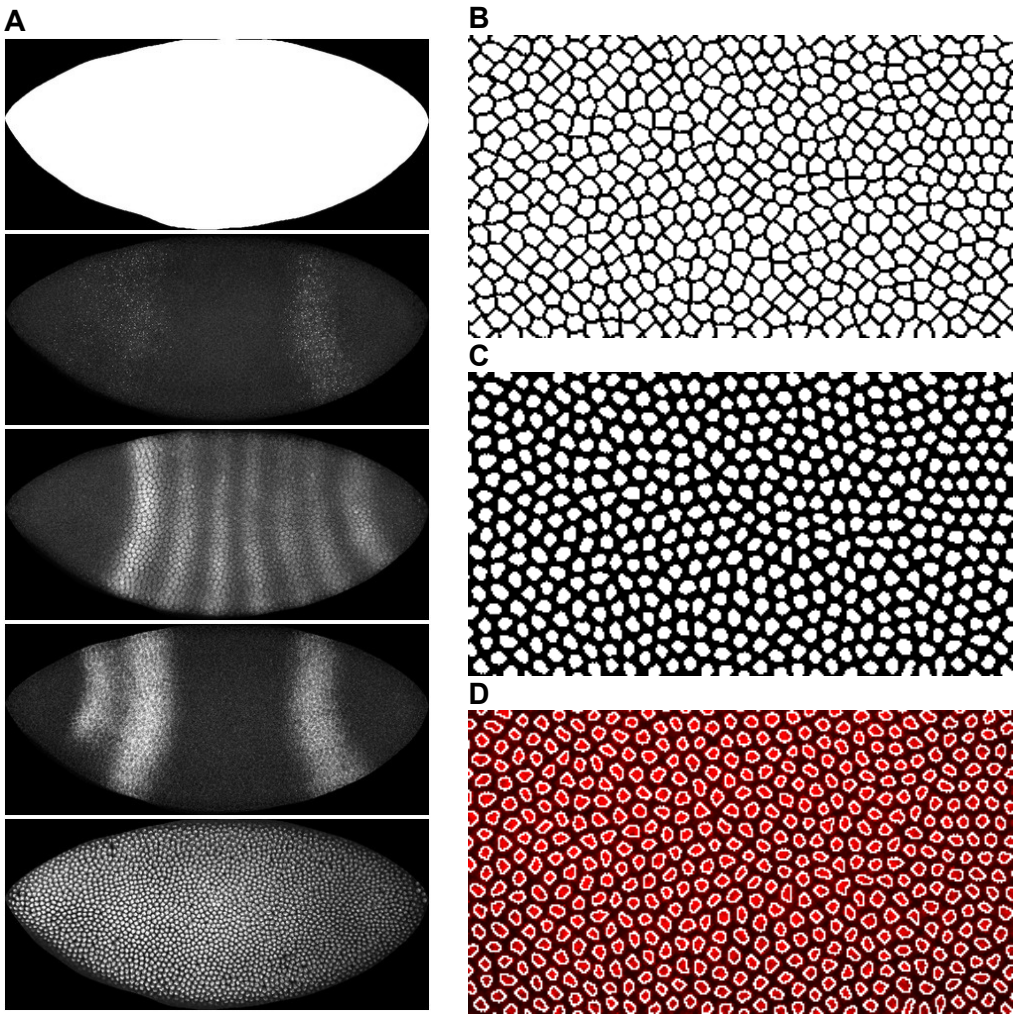


Figure 63: Image segmentation.
(A) Embryo mask created from images of FISH with lacZ mRNA driven by gt-1, *Eve* protein, endogenous *giant* mRNA and the nuclear Hoechst stain. **(B)** Binary nuclear mask after watershed without edge detection. **(C)** Nuclear mask after edge detection. **(D)** Overlay of the Hoechst stain and the nuclear mask with edge detection.

The observed pattern of the lacZ mRNA is spotty, not diffuse, such as the *giant* mRNA. Furthermore, these small, bright dots (or aggregates) do not always lie inside the nucleus (Figure 64A). This was a problem, because the quantification algorithm calculates the average fluorescence intensity from all pixels inside a region, which resulted in very low values for lacZ. In order to solve this problem, a manual histogram threshold had to be set in order to cut off all the low intensity values (Figure 64B). Only the pixel values above this threshold were taken into account for the calculation of the average intensity in the segmented region. This leads to higher signal in the processed data and, at the same time, removes non-specific background staining (Figure 64F).

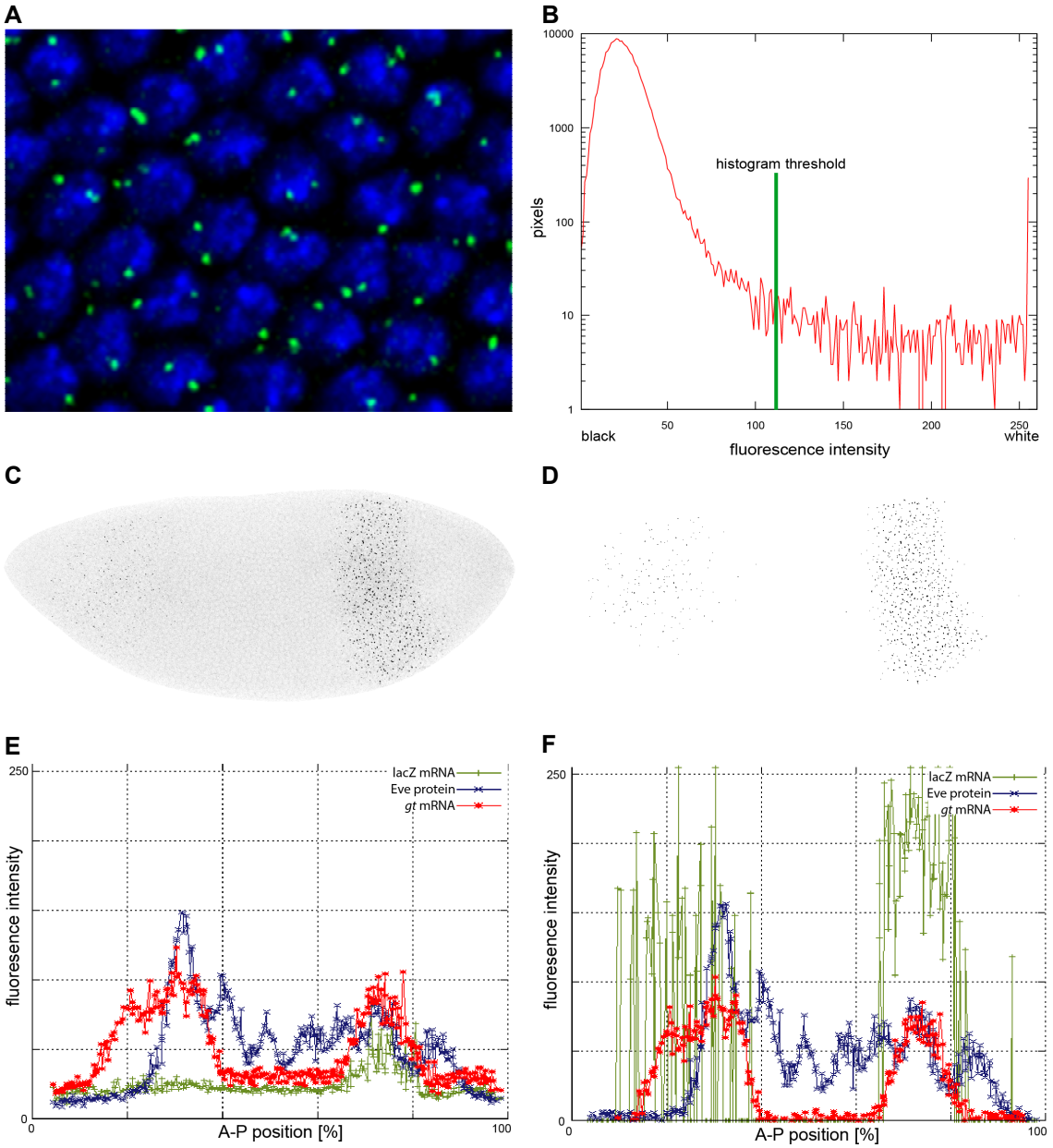


Figure 64: Modified quantification procedure for lacZ.
(A) 40x magnification of nuclei (blue) and lacZ mRNA staining (green). **(B)** Intensity histogram for the lacZ channel showing number of pixels on the y-axis and fluorescence intensity ranging from black (0) to white (255) on the x-axis. The threshold was manually set to 110 in this case. **(C)** Example of an original fluorescent image of lacZ mRNA and **(D)** the corresponding dot mask after the thresholding step. Both images were inverted. **(E)** LacZ quantified with the original quantification procedure, while in **(F)** lacZ was processed with the modifications and the background was removed from Eve and *gt*. Shown are expression data for gt-3-lacZ (green), endogenous *giant* (red) and Eve protein (blue) from a T4 embryo.

Staging and further data processing

The embryos were assigned to cleavage cycles based on the number of detected nuclei. Cycle 14A was additionally subdivided into time classes T1-T8 (Figure 65) by visual inspection of the *Eve* pattern and the membrane morphology (Surkova, Myasnikova, et al. 2008). The data were further processed with BREReA, a package for background removal, registration and averaging of quantitative gene expression data (Kozlov et al. 2009). Non-specific background due to different antisera was removed from *gt* mRNA and Eve protein in batch per slide (Myasnikova et al. 2005). This step was omitted for lacZ mRNA since its background removal was already

inherent to the histogram thresholding. For the last steps, the middle 10% strip of the embryo between 45 and 55% of the DV-axis is taken into account. During data registration, individual expression features are aligned based on the *Eve* protein pattern using a spline-based approach, in order to remove embryo-to-embryo variability (Myasnikova et al. 2001).

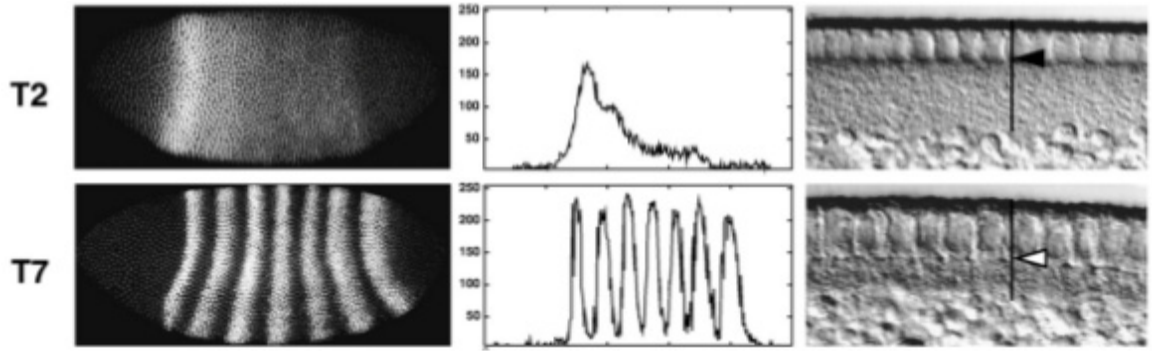


Figure 65: Time classification based on Eve pattern and membrane morphology.
Taken from (Surkova, Kosman, et al. 2008). Shown are the Eve protein expression (left), the segmented expression pattern from the central 10% strip (middle), a high magnification DIC image of the blastoderm (right) from time class T2 and T7 of cycle 14A. In the DIC images, vertical black lines indicate the cortical cytoplasm, the black arrow in time class T2 indicates the elongation of the nuclei, and the white arrow in time classes T7 shows the position of the membrane front.

Afterwards, the data were integrated, meaning grouping the nuclei from different embryos of the same time class into positional bins along the A-P-axis and calculating the average concentration. 100 bins were used in the case of C14, 50 bins for C13 and 25 bins for C12, but subsequently integrated lacZ concentrations of C13 had to be duplicated and C12 had to be quadruplicated, because the model requires 100 bins as input. Finally, lacZ mRNA was smoothened by applying a Gaussian filter (Table 11, implemented in Matlab by Damjan Cicin-Sain), and certain non-expressing regions and negative values were manually set to zero for the model. In the case of C12 of the combined fragment, the pattern had an unusual high baseline. In order to solve this problem, the profile was pulled down to the x-axis by subtracting 35. This procedure can be considered a background removal, similar to the one described above.

	C12	**C13**	**C14**
sigma	3	2	1
filter matrix	[10,1]	[5,1]	[3,1]

Table 11: Parameters for smoothing.
A filter matrix of [x,y], means that x adjacent values in space and y values in time will be considered for the smoothing.

The *Kr* mutant dataset was obtained from the lab of Maria Samsonova (Kozlov et al. 2012, Surkova et al. 2013). Only the protein concentrations of Gt, Hb and Kni for the time classes T1 until T8 were available. The values had to be scaled to the wild-type data (scaling factors: Hb 1.26, Gt 1.46, Kni 2) in order to adjust the quite low levels and to facilitate comparison between data sets. For the other ligands, the values were taken from the WT and all concentrations were smoothed as described above.

6.4 Mutagenesis of Giant sites

The Gt sites to be mutated derived from a model (run 17_1), which was fitted to gt-1 and gt-3 simultaneously over all 10 time points. This optimization run considered Gt as an activator and allowed for Bcd cooperativity, but not for co-activation of Hb. For gt-3, the model found 8 sites with the Gt Selex (Kim et al. 2013) and 3 sites with the B1H matrix (Noyes et al. 2008), which all overlapped with Selex sites (Figure 66 and Figure 67). Mutagenesis of these 8 sites resulted in 28 point mutations. In the case of gt-1, 12 Selex and 3 B1H sites were predicted and again, all the B1H sites coincided with Selex sites. Four additional B1H sites were found with PATSER (Hertz and Stormo 1999), and one of them overlapped with a Selex site predicted by the model. In total, 15 sites were mutated in gt-1, thereby introducing 86 point mutations.

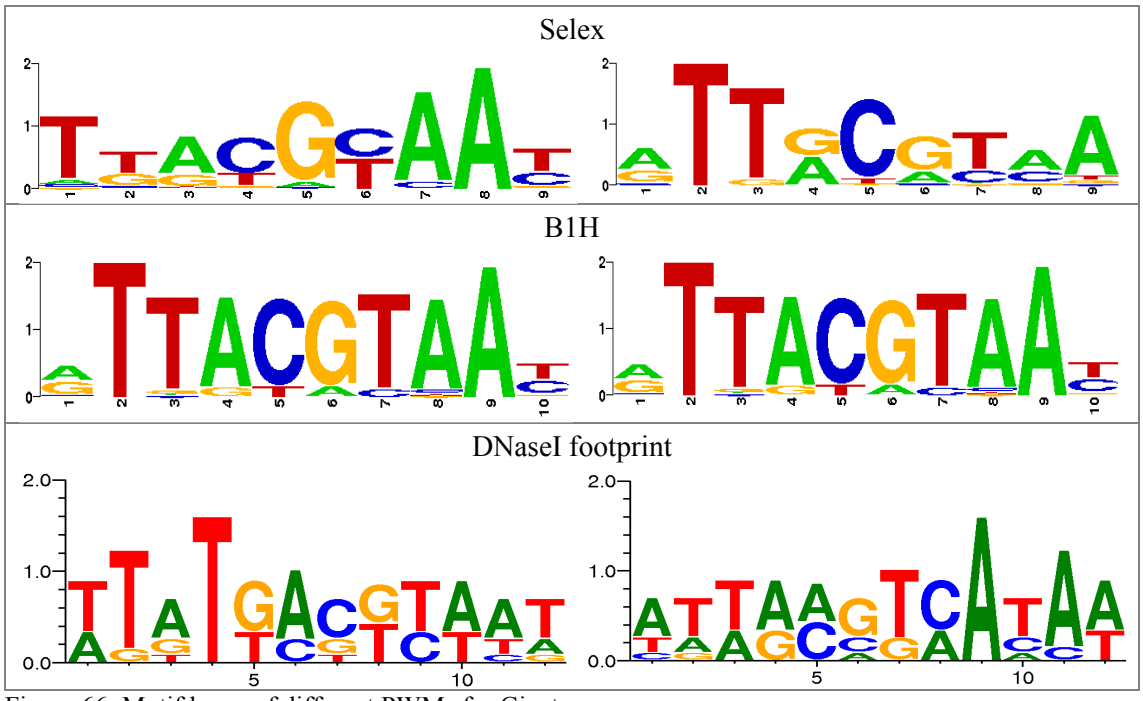


Figure 66: Motif logos of different PWMs for Giant.
Shown is also the reverse complement motif. Note that the B1H matrix is a palindrome. Earlier studies (including Segal et al., 2007 and Schroeder et al., 2004) used matrices built from DNaseI footprints.

It was not possible to make a grounded statement about differential enrichment in the CREs. The top scores are slightly higher in gt-3 compared to gt-1 with both matrices and all the top 20 scores with the Selex matrix are above the optimized threshold from the selected model from Kim et al. (2013), which was 0.6 or lower. Note that the scores in the model output are Patser scores. The rhomboid enhancer served as a control-CRE for DV factors.
The mutated sequences were synthesized by GeneArt® (life technologies, Madrid, Spain), cloned into attB-Pgt-Z and injected into the target lines 38F and in the case of the mutated gt-1 also into 89B.

Model-output for Gt binding sites (run 17_1):

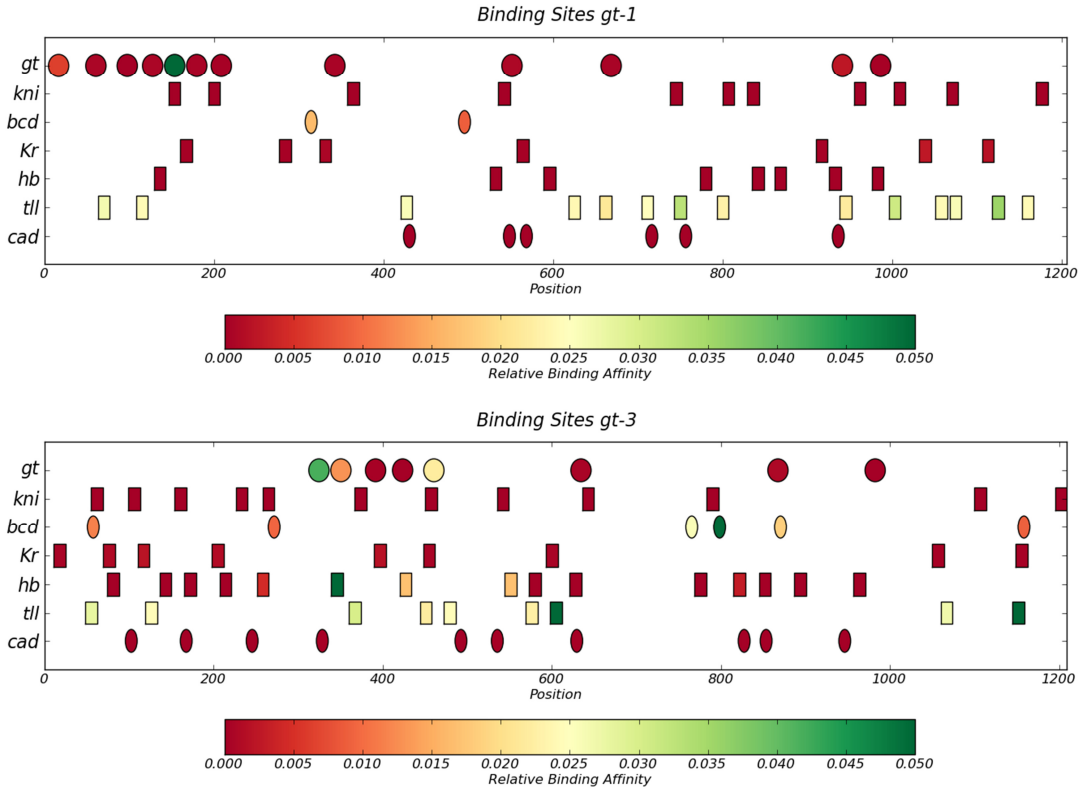


Figure 67: TFBS predicted by the model.
Output from run 17_1, showing the positions of the predicted binding sites for all relevant TFs on the DNA sequence of gt-1 and gt-3. The relative binding affinity is indicated in green for strong sites and in red for weak sites. The Selex matrix was used for Gt.

gt-3:
8 Selex sites
<BindingSite index="16" name="gt" m="-897" n="-874" score="7.7509648" k="0.041958782"
<BindingSite index="17" name="gt" m="-871" n="-848" score="6.5629658" k="0.013061503"
<BindingSite index="18" name="gt" m="-830" n="-807" score="2.3937379" k="0.00021741902"
<BindingSite index="19" name="gt" m="-798" n="-775" score="2.0005621" k="0.00014776073"
<BindingSite index="20" name="gt" m="-761" n="-738" score="7.0944648" k="0.02201627"
<BindingSite index="21" name="gt" m="-587" n="-564" score="3.1109799" k="0.0004398359"
<BindingSite index="22" name="gt" m="-354" n="-331" score="3.822952" k="0.00088518808"
<BindingSite index="23" name="gt" m="-239" n="-216" score="2.0793388" k="0.00015964933"
3 B1H sites (all overlap with Selex sites)
order of strength (strong to weak): 16, 17, 18
<BindingSite index="16" m="-898" n="-875" score="6.4831388" k="0.049306593"
<BindingSite index="17" m="-799" n="-776" score="2.9195822" k="0.02416025"
<BindingSite index="18" m="-761" n="-738" score="2.1830772" k="0.020848373"

gt-1:
12 Selex sites
<BindingSite index="76" m="-1202" n="-1179" score="5.8226472" k="0.0063118252"
<BindingSite index="77" m="-1158" n="-1135" score="3.5243775" k="0.0006601708"
<BindingSite index="78" m="-1121" n="-1098" score="2.0255253" k="0.00015142896"
<BindingSite index="79" m="-1091" n="-1068" score="3.647739" k="0.00074522147"
<BindingSite index="80" m="-1065" n="-1042" score="7.9294505" k="0.04999992"
<BindingSite index="81" m="-1039" n="-1016" score="3.220241" k="0.00048967075"
<BindingSite index="82" m="-1010" n="-987" score="2.8586172" k="0.00034326261"
<BindingSite index="83" m="-876" n="-853" score="3.1125828" k="0.00044052903"
<BindingSite index="84" m="-667" n="-644" score="3.8668827" k="0.00092422457"
<BindingSite index="85" m="-550" n="-527" score="3.244402" k="0.00050143178"

<BindingSite index="86" m="-277" n="-254" score="4.830892" k="0.0023825811"
<BindingSite index="87" m="-232" n="-209" score="2.2843048" k="0.00019525882"
3 B1H sites (all overlap with Selex sites)
order of strength (strong to weak): 72, 74, 73
<BindingSite index="72" m="-1065" n="-1042" score="5.0692176" k="0.037152108"
<BindingSite index="73" m="-667" n="-644" score="2.5594103" k="0.022479637"
<BindingSite index="74" m="-277" n="-254" score="4.3754096" k="0.032334499"
4 additional B1H sites with Patser, 1 of them overlaps with Selex:
B1: >gt-1 position= 523 score= 1.42 ln(p-value)= -5.70 sequence= ATAGCATAAG
B2: >gt-1 position= 649 score= 0.89 ln(p-value)= -5.49 sequence= GTAATACAAC
B3 / S82: >gt-1 position= 203 score= 0.55 ln(p-value)= -5.31 sequence= TTTCCATAAT
B4: >gt-1 position= 285 score= 0.49 ln(p-value)= -5.27 sequence= ATTTCGAAAC

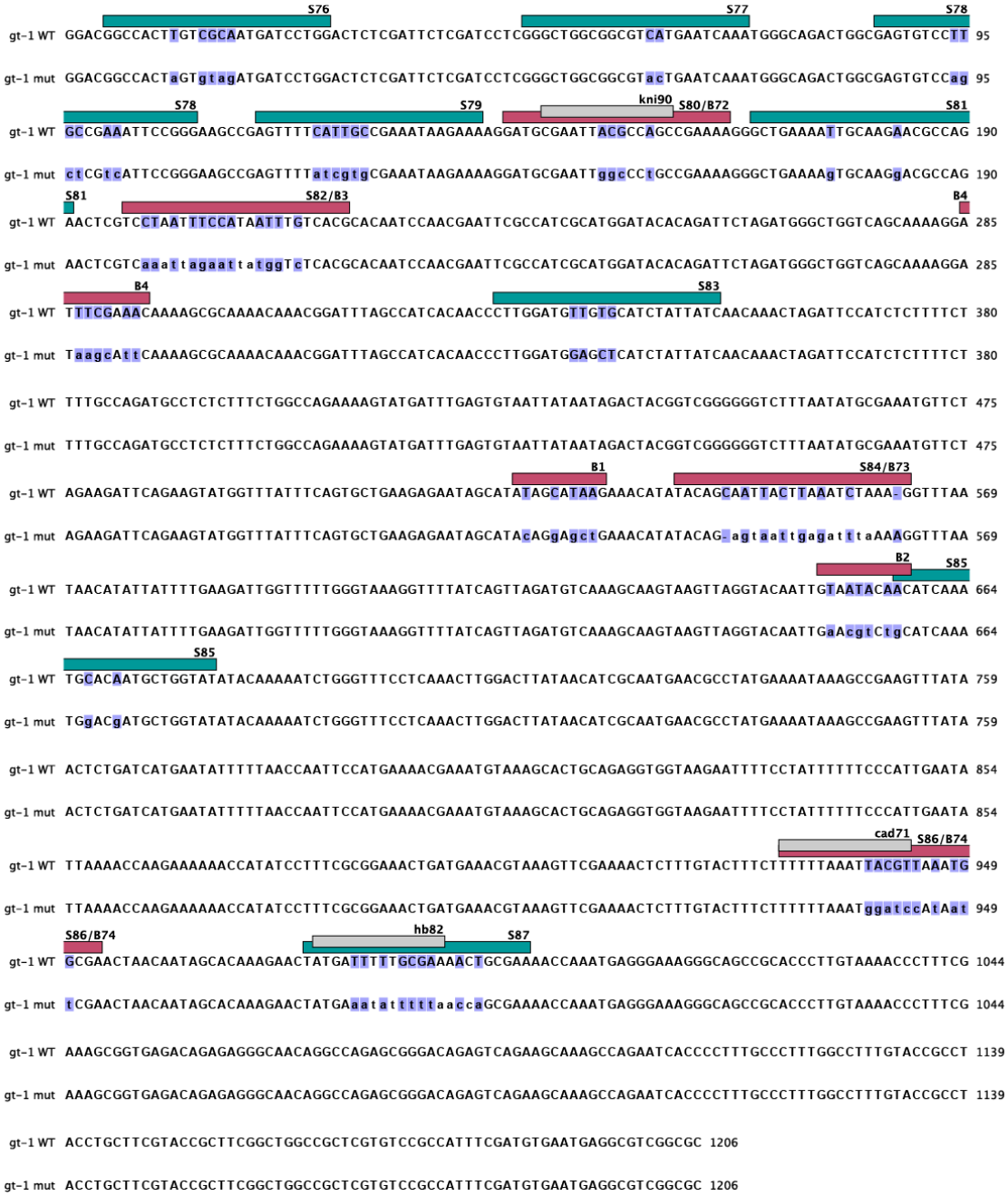


Figure 68: Alignment of WT and mutated gt-1 sequence.
Selex (S) sites are indicated in green, B1H (B) sites in red and overlapping sites for other TFs are shown in gray. Mutated bases are highlighted in blue.

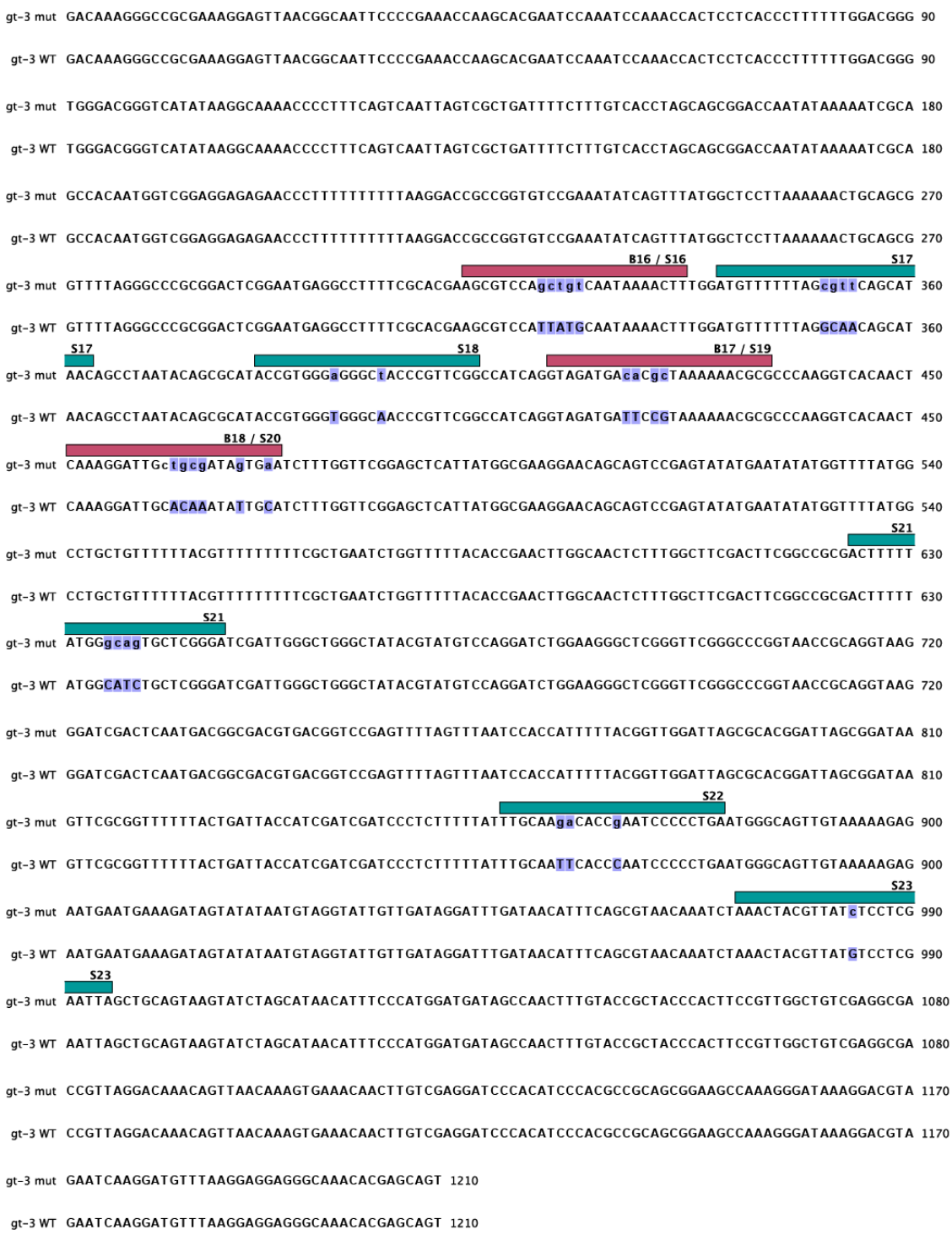


Figure 69: Alignment of WT and mutated gt-3 sequence.
Selex (S) sites are indicated in green, B1H (B) sites in red and mutated bases are highlighted in blue.

6.5 Crosses of CREs into mutant backgrounds

The CRE lines were combined with mutant backgrounds via different approaches, depending if the corresponding factor is maternal, zygotic or both. Furthermore, it depends on which chromosome the gene and the CREs reside. All the maternal and gap gene mutants are homozygous lethal and such crosses need to be propagated over several generations. Hence, most of these intermediate mutants tend to be weak and unfortunately not all methods were successful. See Appendix for a list of fly lines used in this study. If not otherwise stated, the lines homozygous for the CRE on chromosome III (target line 89B) were used for the crosses.

Conventional crosses with mutants

Giant: y^1 sc^1 gt^{X11}/FM6 (BS 1529)
Virgins of the *gt* mutant with kidney-shaped eyes were crossed directly in a cage to homozygous males carrying the CRE on chromosome III. The desired genotype should appear in a ratio of ¼ within the pool of eggs, which were stained fluorescently for endogenous *gt*, lacZ, Eve and Hoechst.

♀ gt^{X11}/FM6 x ♂ CRE/CRE → ♂ gt^{X11} ;; CRE/+

Krüppel: cn^1 bw^1 Kr^1/SM6a, bw^{k1} (BS 3494)
Eggs carrying the CRE in the *Kr* mutant background should appear in a ratio of 3/16, and the ratio decreases to 1/16 in the case of eggs homozygous for the CRE. Embryos were stained enzymatically for lacZ and *Kr* mRNA.

1^{st} cross: Kr^1/SM6a x CRE/CRE → Kr^1/+ ; CRE/+

2^{nd} cross in a cage: Kr^1/+ ; CRE/+ x Kr^1/+ ; CRE/+ → Kr^1/Kr^1 ; CRE/CRE

Since *hb*, *kni* and *tll* are located on chromosome III, these conventional crosses were only possible with gt-3, for which lines with the cassette integrated into chromosome II in 5´orientation were available (w^-; 37B-Phsp70-gt-3/CyO, stock number VII.1.2 and/or w^-; 37B-Peve-gt-3/CyO, stock number V.2.1). The resulting stocks are homozygous for the CRE and the mutant is kept stable over a third chromosome balancer. One out of four eggs will be homozygous for the mutant and the CRE in the cage.

Hunchback: hb^{12} st^1 e^1/TM3, Sb^1 (BS 1755)

Cross 1a brings a II^{nd} chromosome balancer into the mutant line:
hb^{12} st^1 e^1/TM3, Sb^1 x Sco/Cyo,Dfd-YFP ; Sb/TM6B → +/CyO ; hb^{12} st^1 e^1/TM6B
Screen for *ebony* and against *stubble*.

Cross 1b brings III^{rd} chromosome balancer into the CRE line:
CRE/CRE x CyO/Bl ; TM2/TM6B → CRE/Bl ; +/TM6B

Cross 2 combines the mutant and the CRE:
+/CyO ; hb^{12} st^1 e^1/TM6B x CRE/Bl ; +/TM6B → CRE/CyO ; hb^{12} st^1 e^1/TM6B
Screen for CyO and *ebony* and against *bancal*.

Cross 3 makes the CRE homozygous:
CRE/CyO ; hb^{12} /TM6B x CRE/CyO ; hb^{12} /TM6B → CRE/CRE ; hb^{12} /TM6B

The eggs were stained enzymatically for lacZ and *hb* mRNA.

Knirps: Df(3L)ri-79c/TM3, Sb^1 (BS 3127)

Cross 1 brings a II^{nd} chromosome balancer into the mutant line:
Kni^-/TM3, Sb^1 x Bl/Cyo ; TM2/TM6B → +/Bl ; Kni^-/TM6B
Screen for *humeral* and against *ebony* and CyO.

Cross 2 combines the mutant and the CRE (from the hb cross):
+/Bl ; Kni^-/TM6B x CRE/CyO ; hb^{12} st^1 e^1/TM6B → CRE/Bl ; Kni^-/TM6B
Screen for *bancal* and *humeral* and against CyO and *ebony*.

Cross 3 makes the CRE homozygous:
CRE/Bl ; Kni^-/TM6B x CRE/Bl ; Kni^-/TM6B → CRE/CRE ; Kni^-/TM6B

The eggs were stained fluorescently for lacZ mRNA, Eve and Kni protein.

Tailless: Df(3R)tll^g, ca^1/TM3, Sb^1, Ser^1 (BS 2599)

Cross 1 brings a II^{nd} chromosome balancer into the mutant line:
Tll^g/TM3, Sb^1 x Bl/Cyo ; TM2/TM6B → +/Bl ; Tll^g/TM6B
Screen for *humeral* and against *ebony* and CyO.

Cross 2 combines the mutant and the CRE (from the hb cross):
+/Bl ; Tll^g/TM6B x CRE/CyO ; hb^{12} st^1 e^1/TM6B → CRE/Bl ; Tll^g/TM6B
Screen for *bancal* and *humeral* and against CyO and *ebony*.

Cross 3 makes the CRE homozygous:
CRE/Bl ; Tll^g/TM6B x CRE/Bl ; Tll^g/TM6B → CRE/CRE ; Tll^g/TM6B

The eggs were stained fluorescently for lacZ mRNA, Eve and Tll protein.

Meiotic recombination

The recombination approach takes advantage of the fact that if a mutation or integrated cassette is not kept over a balancer, then multiple crossovers with the other allele are possible. Hence, the lines homozygous for the CRE on chromosome III were used for these recombination crosses with *hb*, *kni* and *tll*, which are all on chromosome III too.

Cross 1: mutant x CRE
hb^{12}/TM3 x CRE/CRE → hb^{12}/CRE
(screen against *stubble* and take virgins only)

Cross 2: recombination might take place
♀ hb^{12}/CRE x ♂ CyO/Bl ; TM2/TM6B → hb^{12}, CRE/TM6B (if recombined)

Cross 3: set up 20 crosses to balancer to establish candidate stocks
hb^{12}, CRE/TM6B x CyO/Bl ; TM2/TM6B → +/CyO ; hb^{12}, CRE/TM6B

gDNA was isolated from the male parent with Chelex and a control PCR for lacZ was performed. If there are flies in a stock without a marker, then this stock does not carry the mutation, because it is homozygous viable. Cuticle preparations or *in-situ* hybridizations were carried out with the remaining lacZ positive and homozygous lethal lines, in order to find out, if they harbor the gap mutant in fact.

Transgenic RNA interference

In order to generate maternal mutants, I used the RNAi approach. Transgenic RNAi takes advantage of the GAL4/UAS-system to drive localized and timed expression of a RNA fragment complementary to the target gene (Ni et al. 2009, Staller et al. 2013). I used the maternal-tubulin Gal4 driver (from D. St. Johnston and F. Wirtz-Peitz), homozygous for Gal4 on the IInd and IIIrd chromosome, in order to achieve expression in the maternal germ line. Different UAS lines expressing either long or short hairpins (sh, 400 - 600bp long, (Ni et al. 2011)) are available and I decided on the UAS-shRNA-hb, expressing a short hairpin against *hunchback* from the Valium22 vector. It carries *vermillion* as a marker and was integrated into the site attP40 on the IInd chromosome (gift from Max Staller).

Hunchback

Cross 1 brings a IIIrd chromosome balancer into the UAS-shRNA-hb line:
UAS-hb/ UAS-hb x Bl/CyO; TM2/TM6 → UAS-hb/Bl; +/TM6

Cross 2 combines the UAS-shRNA-hb line with the hb^{12} mutant:
UAS-hb/Bl; +/TM6 x +/CyO; hb^{12}, e^{-}/TM6 → UAS-hb/CyO; hb^{12}, e^{-}/TM6
Select for CyO and *ebony* and against *bancal*. The line was made homozygous for UAS-hb.

Cross 3 incorporates the Gal4-driver:
UAS-hb/UAS-hb; hb^{12}/TM6 x Gal4/Gal4; Gal4/Gal4 → UAS-hb/Gal4; hb^{12}/Gal4
Collect virgins without *humeral*.

Cross 4 in a cage: virgins providing eggs with Gal4-induced short hairpins against *hb* are crossed to males carrying the CRE in a zygotic *hb* mutant background.
♀ UAS-hb/Gal4; hb^{12}/Gal4 x ♂ CRE/CRE; hb^{12}/TM6 → CRE/UAS-hb; hb^{12}/ hb^{12}
or CRE/Gal4; hb^{12}/ hb^{12}
The embryos will be heterozygous for the CRE and the zygotic *hb* allele will occur homozygous in a ratio of ¼ in the pool of genotypes. The eggs were stained fluorescently for lacZ mRNA, Eve and Hb protein.

Germ line clones

Caudal

CREs in a <u>zygotic *cad* mutant</u> background were generated via conventional crosses:

Cross 1a brings a IIInd chromosome balancer into the mutant line:
cad FRT/CyO x CyO/Bl; TM2/TM6 → cad FRT/CyO; +/TM6
Screen for CyO and against *bancal*.

Cross 1b brings IInd chromosome balancer into the CRE line:
CRE/CRE x CyO/Bl; TM2/TM6 → +/Bl; CRE/TM2 and +/CyO; CRE/TM2
+/CyO; CRE/TM2 x +/Bl; CRE/TM2 → Bl/CyO; CRE/TM2
The line was made homozygous for the CRE.

Cross 2 combines the mutant and the CRE:
cad FRT/CyO; +/TM6 x Bl/CyO; CRE/CRE → cad FRT/CyO; CRE/TM6
Screen for CyO and *humeral* and against *bancal*.

Cross 3 makes the CRE homozygous:
cad FRT/CyO; CRE/TM6 x cad FRT/CyO; CRE/TM6 → cad FRT/CyO; CRE/CRE

To generate <u>maternal *cad* mutants</u>, the following cross was set up in at least 10 tubes:
♀ cad FRT/CyO x ♂ y^{-} w^{-}, hs-flp; ovo^{D1} FRT w^{+}/CyO

The parents were flipped to a new tube every day in order to get eggs which were laid in a small time frame. The 3rd instar larvae were heat shocked twice during 2 h at 38°C on 2 consecutive days in order to express the flipase (flp). Virgins without CyO were collect and they should have the following genotype in the maternal germ line if the crossover has taken place:

→ $y^- w^-$, hs-flp/w^- ; cad FRT/cad FRT

For maternal *cad* mutants, the following cross was set up in a cage and the eggs were collected and stained enzymatically for lacZ and *cad* mRNA:

♀ $y^- w^-$, hs-flp/w^- ; cad FRT/cad FRT x ♂ CRE/CRE → cad FRT/+ ; CRE/+

For maternal and zygotic mutants, the virgins collected after the heat shock, were crossed to the CREs with the zygotic *cad* mutant background:

♀ cad FRT/cad FRT x ♂ cad FRT/CyO; CRE/CRE → cad FRT/cad FRT; CRE/+

Hunchback

In a similar way, *hb* germ line clones were generated, but unfortunately resulted in infertility.

Appendix

1 Sequences and plasmids

1.1 Primers

Forward primers are indicated with an F and reverse primers with an R. Restriction sites, additional bases or other special features are separated by a hyphen. Mutated sites are indicated in upper case. Primers were purchased from Thermo Fisher Scientific or Sigma.

name	sequence
sal-tub-F	ataat-gtcgac-ttttgcctaattgtttcagatttatg
lacZ-Hind-R	attat-aagctt-ctggccgtcgttttacaacg
Hind-gtCDS-F	ataat-aagctt-ctcgtgcattagcatggtg
gtProm-Avr-Bam-R	attat-ggatccctagg-ttttcggtataaatgcaggg
Avr-CE8001-F	ataat-cctagg-tgcccattcaggggattgg
CE8001-Spe-R	attat-actagt-gaaacttaccatcacttcgag
Avr-gt1small-F	ataat-cctagg-cggaacggatgcgctgcccag
gt1small-Spe-R	attat-actagt-gattcccctgcattacgtcaaac
Avr-gt23small-F	ataat-cctagg-tctgccctgccctgctctg
gt23small-Spe-R	attat-actagt-ggcgacttggatcgtgagctg
Avr-gt1gaul-F	ataat-cctagg-gcgccgacgcctcattcac
gt1gaul-Spe-R	attat-actagt-ggacggccacttgtcgcaatg
Avr-gt3gaul-F	ataat-cctagg-actgctcgtgtttgccctcc
gt3gaul-Spe-R	attat-actagt-gacaaagggccgcgaaagg
Avr-gt6gaul-F	ataat-cctagg-cgttttgggccattgttcc
gt6gaul-Spe-R	attat-actagt-tctgtcgcctgctatttatttattaaatg
Avr-gt10gaul-F	ataat-cctagg-tcgcaggatccttgtccagg
gt10gaul-Spe-R	attat-actagt-cacgtggcgacttggatcgtg
Hind-Peve-F	ataat-aagctt-ggttcggtatccgtgcat
Hind-hsp70-F	ataat-aagctt-gattccaatagcaggcat
Peve-Avr-Bam-R2	attat-ggatcc-ctagggagcgcagcggtataaaaggg
hsp70-Avr-Bam-R2	attat-ggatcc-ctaggaagagcgccggagtataa
Bam-Xho-kni-F	aatta-ggatcc-ctcgag-tgtgcacggagctccgcgag
kni-Bgl-Xba-R	taata-tctaga-atat-agatct-aaccgctttagtcccgcc
Bgl-gt-3-F	ttaat-agatct-actgctcgtgtttgccctcc
gt-3-Xba-R	ttaat-tctaga-gacaaagggccgcgaaagg
Avr-gt-1-F	attaa-cctagg-gcgccgacgcctcattcac
gt-1-Xho-R	attaa-ctcgag-ggacggccacttgtcgcaatg

Table A. 1: Primers for creating the reporter constructs.

name	sequence
lacZ1877-F	tgctgatatggttgatgtc
lacZ4504-R	cagttatctggaagatcagg
lacZ2330-F	tgtctgacaatggcagatc
LL5	actgtgcgttaggtcctgttcattgtt
LL3	ccttagcatgtccgtggggtttgaat
attR	gatgggtgaggtggagtacg
gt-3-middle-R	catcaggtagatgattcc
seq gt-1 F	gatgtgaatgaggcgtcg
seq gt down R	agtacttaaatgcgagcg
seq gt-3 F	gatagccaactttgtacc
seq y begin R	tcagggtcacaaggatcc
seq CE8001 F	gagatgaaagtgcggagg
seq gt-1 R	tcggctggcgtaattcgc
bcd-coop-F	tgttgtcctgcatgatg
bcd-coop-R	actgcatgtgcatgtgac
gt-HA-F	ctagatcaccagtctatatagc
gt-HA-R	tgacccaaaaactggacatacg
89B8-5 F	agggcggaaggcatggtaca
89B8-3 R	cgatgacaataaccaatcgatggcc
89B-3R2	tctactcacattggattccgtc
phiC31-dm-F	atggacacgtatgccggtgctt
phiC31-dm-R	taggccgctacgtcttcggtg
37B-F	actcgcgagcacacacgcacac
37B-R	acacgatgttggcagcatagc
38F-F	aacgaagacctagtgttagg
38F-R	acattggtgctcttctcgc

Table A. 2: Primers for controls.

name	sequence
gt-3 mut B1H16-F	gaagcgtccaGCTGTcaataaaactttggatgt
gt-3 mut B1H16-R	ccaaagttttattgACAGCtggacgcttcgtgcga
gt-3 mut B1H17-F	caggtagatgaCAcGCtaaaaaacgcgcccaa
gt-3 mut B1H17-R	gcgcgtttttaGCgTGtcatctacctgatggccg
gt-3 mut B1H18-F	ctcaaaggattgCTGCGatattgcatctttggttcg
gt-3 mut B1H18-R	caaagatgcaatatCGCAGcaatcctttgagttg
gt-1 mut B1H72-F	aaggatgcgaattGGCccTgccgaaaagggctga
gt-1 mut B1H72-R	ccttttcggcAggGCCaattcgcatccttttctta
gt-1 mut B1H73-F	gaaacatatacagAGTAATTGAGaTtTAaaaggtttaataaca
gt-1 mut B1H73-R	ttattaaacctttTAaAtCTCAATTACTctgtatatgtttctt
gt-1 mut B1H74-F	gtactttctttttaaatGGATCCaTaATTcgaactaacaatagc
gt-1 mut B1H74-R	attgttagttcgAATtAtGGATCCatttaaaaaagaaagtacaa
gt-1 mut B1H73a-F	gaaacatatacagcGaAtTcGtGaTtcAaaaggtttaataaca
gt-1 mut B1H73a-R	ttattaaacctttTgaAtCaCgAaTtCgctgtatatgtttctt
gt-1 mut B1H74a-F	gtactttctttttaaatGacgCtaTaAgTcgaactaacaatagc
gt-1 mut B1H74a-R	attgttagttcgAcTtAtaGcgtCatttaaaaaagaaagtacaa

Table A. 3: Primers for mutagenesis.

1.2 Plasmids

plasmid	insert	source	clone	mut
pGEMT-gtProm	*gt* promoter	OregonR	1	
pGEMT-CE8001	enhancer	OregonR	1	
pGEMT-gt1	enhancer	OregonR	1	
pGEMT-gt23	enhancer	OregonR	1	
pGEMT-gt-1	enhancer	OregonR	2	
pGEMT-gt-3	enhancer	OregonR	2	
pGEMT-gt-6	enhancer	OregonR	2	
pGEMT-gt-10	enhancer	OregonR	2	
pBS-attB-gtProm	attB sites	Jack Bateman	1	
pBS-attB-gtProm-lacZ-tub	lacZ + α-tubulin-3'UTR	Miki Fujioka	3	0
pBS-attB-gtP-Z-CE8001	CE8001	OregonR	1	2
pBS-attB-gtP-Z-gt-3	gt-3	OregonR	1	0
pBS-attB-gtP-Z-gt1	gt1	OregonR	6, 7	
pBS-attB-gtP-Z-gt-1	gt-1	OregonR	15	gaps
pJet-CE8001	CE8001	OregonR	1	
pJet-gt-1	gt-1	OregonR	1, 2, 3	gaps
pJet-CE8001 w1118	CE8001	w1118	3	
pBS-attB-gtP-Z-gt23	gt23	OregonR	1	6
pJet-gt23-oregon	gt23	OregonR	6	
pJet-gt23-w1118	gt23	w1118	2	
pJet-gt23-37B-3	gt23	target line 37B	3	2
pBS-attB-gtP-Z-gt23	gt23	target line 37B	3	2
pBS-attB-Peve-Z	*eve* promoter	attBeZ SR105	1	
pBS-attB-Peve-Z-gt-3	*eve* promoter	attBeZ SR105	1	
pBS-attB-Phsp70-Z	*hsp70* promoter	attBhZ SR105	3	
pBS-attB-Phsp70-Z-gt-3	*hsp70* promoter	attBhZ SR105	1	
pBS-attB-gt-1-gt-3-gtP-Z	combined inverted	OregonR	5	
pBS-attB-gt-3-gt-1-gtP-Z	combined	OregonR	45	
pBS-attB-spacer-gtP-Z	kni spacer 340 bp	cDNA clone	1	0
pBS-attB-gtP-Z-gt-1-mut72	B1H 72 mutated	attB-gtP-Z-gt-1	1	4
pJet-gt-1_mut_72_74	B1H 72, 74 mutated	OE-PCR	11	10
pMK-RQ-gt-1_mutated	all Gt sites mutated	GeneArt		86
pBS-attB-gtP-Z-gt-1-mut	all Gt sites mutated		2	86
pMA-T-gt-3_mutated	all Gt sites mutated	GeneArt		29
pBS-attB-gtP-Z-gt-3-mut	all Gt sites mutated		17	29
pTV-gt-3	3'homology arm (gt-3)		6	
pTV-gt-3-gtCDS	5'homology arm (gt CDS)		2	

Table A. 4: Plasmids.
The vector backbones are either pGEMT (Promega), pJet (Fermentas) or pBS (BluesSkript). Source denotes from which fly line or plasmid the insert was amplified or who provided starting material. Number of mutated sites (mut) is indicated if applicable. All plasmids are Ampicillin resistant apart of pMK-RQ-gt-1_mutated, which is Kanamycin resistant. The CREs with all Gt sites mutated were synthesized by GeneArt® (life technologies, Madrid, Spain) and provided as pMK-RQ-gt-1_mutated and pMA-T-gt-3_mutated. The pTV vector (Baena-Lopez et al. 2013) serves for homologous recombination in the genome. Overlap-extension PCR (OE-PCR) was used to mutate sites in gt-1.

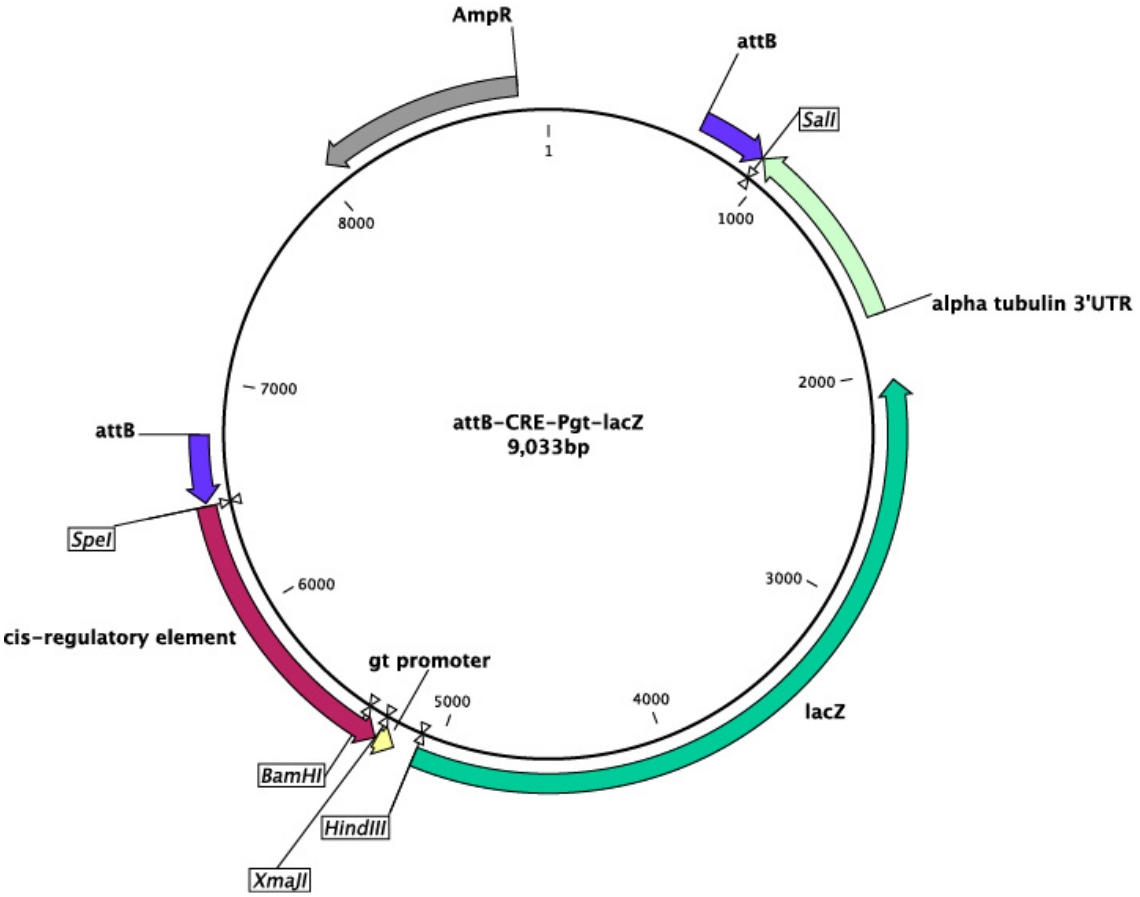


Figure 70: Plasmid attB-CRE-Pgt-lacZ.
The insertion cassette harbours a CRE, the endogenous *giant* core promoter, the lacZ reporter gene and the alpha tubulin 3´UTR, and is flanked by attB site for recombination in attP target lines.

1.3 Sequences of the *giant* CREs

>gt_-1_construct
gcgccgacgcctcattcacatcgaaatggcggacacgagcggccagccgaagcggtacgaagcaggtaggcggtacaaaggccaaagggcaaaggggtgattctggctttgcttc
tgactctgtcccgctctggcctgttgccctctctgtctcaccgctttcgaaagggttttacaagggtgcggctgccctttccctcatttggtttttcgcagttttcgcaaaaatcatagttctttgtgctat
tgttagttcgccatttaacgtaattttaaaaagaaagtacaaagagttttcgaactttacgtttcatcagtttccgcgaaaggatatggtttttcttggttttaatattcaatgggaaaaaatagga
aaattcgtaccacctctgcagtgctttacatttcgttttcatggaattggttaaaaatattcataatcaatatagagttataaacttcggctctattttcataggcgttcattgcgatgttataagtcca
agtttgaggaaaccccagattttgtatataccagcattgtgcatttgatttttattacaattgtacctaacttacttgctttgacatccaactgataaaaccctacccaaaaaccaatcttaaaa
atcttaagttattaaaccttagatttaagtaactgctgtatatattgctttcaaatataaagaataaaccgcaattaaacatcgtaggctattctcttcagcactgaaataaaccatacttctgaa
tcttctagaacatttcgcatatcaaagacccccgaccgtagactattataattacactcaaatcatacttttctggccagaaagagaggcatctggcagaagaaaagagatggaatcta
gtttgttgataatagatgcacaacatccaagggttgtgatggctaaatccgtttgttttgcgcttttgtttcgaaatccttttgctgaccagcccatctagaatctgtgtatccatgcgatggcgaatt
cgttggattgtgcgtgacaaattatggaaattaggacgagttctggcgttcttgcaattttcagccctttcggctggcgtaattcgcatccttttcttatttcggcaatgaaaactcggcttcccgg
aattttggcaaggacactcgccagtctgcccatttgattcatgacgccgccagcccgaggatcgagaatcgagagtccaggatcattgcgacaagtggccgtcc

>gt_-3_construct
actgctcgtgtttgccctcctccttaaacatccttgattctacgtcctttatccctttggcttccgctgcggcgtgggatgtgggatcctcgacaagttgtttcactttgttaactgtttgtcctaacggt
cgcctcgacagccaacggaagtgggtagcggtacaaagttggctatcatccatgggaaatgttatgctagatacttactgcagctaattcgaggacataacgtagtttagatttgttacgct
gaaatgttatcaaatcctatcaacaatacctacattatatactatctttcattcattctcttttttacaactgcccattcagggggattgggtgaattgcaaataaaaagagggatcgatcgatggt
aatcagtaaaaaaccgcgaacttatccgctaatccgtgcgctaatccaaccgtaaaaatggtggattaaactaaaactcggaccgtcacgtcgccgtcattgagtcgatcccttacctgc
ggttaccgggcccgaacccgagcccttccagatcctggacatacgtatagcccagcccaatcgatcccgagcagatgccataaaaagtcgcggccgaagtcgaagccaaagagtt
gccaagttcggtgtaaaaaccagattcagcgaaaaaaaaacgtaaaaaacagcaggccataaaaccatatattcatatactcggactgctgttccttcgccataatgagctccgaacc
aaagatgcaatatttgtgcaatcctttgagttgtgaccttgggcgcgttttttacggaatcatctacctgatggccgaacgggttgcccacccacggtatgcgctgtattaggctgttatgctgtt
gcctaaaaaacatccaaagttttattgcataatggacgcttcgtgcgaaaaggcctcattccgagtccgcgggccctaaaaccgctgcagttttttaaggagccataaactgatatttcgga
caccggcggtccttaaaaaaaaaaggttctctcctccgaccattgtggctgcgatttttatattggtccgctgctaggtgacaaagaaaatcagcgactaattgactgaaaggggttttgc
cttatatgacccgtcccacccgtccaaaaaagggtgaggagtggtttggatttggattcgtgcttggtttcggggaattgccgttaactcctttcgcggccctttgtc

>gt_-6_construct
cgttttgggccattgttccaactcggagcactgagatcggtatctatttgggtgagaaccgcttgcggtacagttatgtgcggaaagtgtgaaacaattctaatagccggcgcaaagaccct
tattcccgataatgccctttcaatgctcacttaattttatgcattttactgttgattagctgccatcagccacgacgcccattgtcagcggccgagcaccgttaggggtttttccggagttaatcaa
accgccgttcgaagagcccacataaaaagagttatgatccgcccccgaccatcgggatttcgagcccactcgatctatcaagcggcccttgaaattttcacacctttttgcgggagtgtcc
ttgctgctgcactccttttgtccttttgccattgttcaaatgcaagccaaagccgcctacgaaaccccgttagaatggctactgttagccccttcctctcgagaagcgaaattcaaaattcca
gatcccagatcacattgttggccaggacaattgctgtcaaaatggcattgcagcgatttgcggatttggccgtcctggtaattgtgtcaattgtcagttcgcaccatctcttaactgccatcagc
atataaagcatatccatccagcacgatctccagcggaacggatgcgctgcccagatacaatcgggatcggcagttagggaccgggcgagaacggtagatcggaggaaatcctgga
acagagaacggaactgaactgaacggaactcaatgaggtcctaagcctcattccgcataactcttttttcatagagtgactcctggccagctcgtaatatgagcatttaaccccgtaaagtc
ttcgacccggccacgcccattgtatccgagttccgttccgttctgtactcatctcctttgagaagcacagttgccttttgcctcgagttttttcacatgaaactaaaaacgactgaaatccatgag
attaaacggtctttaagcggtttaagcccttttgtgcacgcagttttgtaggtggcaaaatgtgccattgtcaccgagttctgtcaaaaatatgaactagcatatcatttcataattttcagtttcatct
ttaaaagtgcatttttttgcttctcggaaggagaacacaaatattgttggaagctcagtgggccgtagggttccgggtgatccgacaggtgaccaggtacagctctctcgatcctccaatcctc
catgctaatcctgactatctggctgcctcaagttaatctcgcataatattgcgtgttccgaacagcgtaatcccctgaaaacaagtgtccagatccgcagctctccacttttaagccggcaat
ggattctagatgacttgactaactgtcgagctaggctcaacgaacagatttctctgagcctggtcgcaacagtaatacgtttgacgtaatgcaggggaatcaaaaaggtcgaaaaatgtg
gctaaaaactggcaagtaaggaactagagttaagtaagctaagcttgacaaaactattcttactatattgctgatttcttactgtttggacagtttaccattccgtagatcactgaaagttgtgtc
caaaaaatgacgggattttagaaaaggacctaaagttaatcttctgaactagaacttgtctcaaggtcgcataattcagttccgtcgccatgttacgttacatcctccattgccagggattgc
gccatttttggcgtttataatcgaggtcctgctacgaaatcgaagactctgccgaactccgggccatgcaataaagtagagctcgctcgcaggctcggccaatcagaggcagaaaccac
agctggaggggatttcgaaaaaccaccaccacaagccgctgcgacttcattttgtgcgtttggtgggaaactggttctctgggcaataaatccccgcaacttcattaaagcctggagatat
agtactgaatatagggagttggaggaagcggtgaaggtggaggtggaggtggggatggagatggacgtgcagggagtgggtgacggatccagagatccagggcttagggaaaat
gtgacagctgtgccagctgacccagtccattgtccgatgtccattgtgtccatccggaggcgagggagtctcctgcatttaataaataaatagcaggcgacaga

>gt_-10_construct
tcgcaggatccttgtccaggcgccaatcgagcacctcttaagcctttacagctgcccaacggaagtgaagtgaagtgaagtaaaggagttgctcaaagtggcctcacgttcgccgaggt
caatgggtgtgttttcgaagtccgcactgtaatgtcactcttttggttgaaaaacatttggaaagtctgcctctagttccctggacaccagatacatccataagaaagcataaggtaaattat
gaattggcaatcaattgaaataaacagatttcccttcgctctacagaaatttggttctcacacaactttgagtccctttttgcaggcgcacacaaaggtgtgtggataaagtttttgagagaa
aaaaactcattacgagttgcattaagctcataattgattgacaaacgggccgacgagctggtggaactgatggccaagacgtaccatagtgcagggtccacgctctggattagaccaat
tttaaggggaattgctatctgctcggttcacacatagctgatctgccctgccctgctctgaattgcaggacgttagattctgggccgagaattttgagccgtgggcgtgtgcgggccatggatt
tcacacaccgatctcaggaggatcctggaacaatcccttgcagcatgacgaaaggtaaatttgtaaacaattgtgctaattaagccagcccaataaaaatcggctctcggaggcttgag
aggctcgagcggcctgagtggcctgagagcccggggcggatggccgaatggcctgagatctcccggggagatccaggcgagcacttgaactaaatccggtagcttaaacgagttaa
aggaacttgaacggaacttaatcccataaatttgggaaaaatacaattttaatcaggttaattcgtgtgtagcgcgactgttgcatggtttcgggatctgggccacgggatcaggggccag
gtagtcggtcgagcgaaaggattaggcctctggatctcggggatttagcgcaaacagaggtgagtccgtccagcacaggcccaccgcccatcgccttcctgcgacattttaatttaatctc
agattacgagcgaacgaggagacgagatcgcggatcgggaacgggatcggcatcgggatcagtccacagtccaccagattaggccactccgctcgtatattgcaaaataatggtcg
agttccaggggccgccagttcgattgcaaaagtttcacgtaaatccgtaagattagatggctagatccccagattttcagattccctcccgcgcttagcccaattgttaggtcaattgccaaa
tcacttgagcggggaatggaaaactgggccatctcatcgtggaggtaatcccagtcgcagcagatctaggtagatccttgaaagccactggtcatgcccactaaagctagcagtgcctc
cgatgatcgcattcccctgatccccacacaagtccaaagtagtcctctggactagatgcgttgactttcgaaccctgggaatactacagctacaaatagaaatgctacaaaacttaatta
taaataatgtttagattgcagaaacaaatattctagaaactgcacatatttacgaattgtacatcaagattgtgaaagaataaatcaatgattaaaatcaaatatcgtttctctttctaccgaca
aaattcacgggaaaaagtgcgaaaaatatcagctcacgatccaagtcgccacgtg

>gt_CE8001
tgcccattcaggggattgggtgaattgcaaataaaaagagggatcgatcgatggtaatcagtaaaaaaccgcgaacttatccgctaatccgtgcgctaatccaaccgtaaaaatggt
ggattaaactaaaactcggaccgtcacgtcgccgtcattgagtcgatcccttacctgcggttaccgggcccgaacccgagcccttccagatcctggacatacgtatagcccagcccaat
cgatcccgagcagatgccataaaaagtcgcggccgaagtcgaagccaaagagttgccaagttcggtgtaaaaaccagattcagcgaaaaaaaaacgtaaaaaacagcaggcca
taaaaccatatattcatatactcggactgctgttccttcgccataatgagctccgaaccaaagatgcaatatttgtgcaatcctttgagttgtgaccttgggcgcgttttttacggaatcatctacc
tgatggccgaacgggttgcccacccacggtatgcgctgtattaggctgttatgctgttgcctaaaaaacatccaaagttttattgcataatggacgcttcgtgcgaaaaggcctcattccgag
tccgcgggccctaaaaccgctgcagttttttaaggagccataaactgatatttcggacaccggcggtccttaaaaaaaaaaggggttctctcctccgaccattgtggctgcgatttttatattgg
tccgctgctaggtgacaaagaaaatcagcgactaattgactgaaaggggttttgccttatatgacccgtcccacccgtccaaaaaagggtgaggagtggtttggatttggattcgtgcttgg
tttcggggaattgccgttaactcctttcgcggccctttgtcttaatgggttaacagtcgcatattggaatccttagcttcgaatggcccctgtcgtaaacgtaagttgaaagcaacttatcacatat
taggactgccatttttgtttctgagaatctgtattcccgctttttaccctaaaaaaccctagtttgggataagcaaagatcagcgctctcctccgcactttcatctcgaagtgatggtaagtttc

>gt_gt23
tctgccctgccctgctctgaattgcaggacgttagattctgggccgagaattttgagccgtgggcgtgtgcgggccatggatttcacacaccgatctcaggaggatcctggaacaatccctt
gcagcatgacgaaaggtaaatttgtaaacaattgtgctaattaagccagcccaataaaaatcggctctcggaggcttgagaggctcgagcggcctgagtggcctgagagcccggggc
ggatggccgaatggcctgagatctcccggggagatccaggcgagcacttgaactaaatccggtagcttaaacgagttaaaggaacttgaacggaacttaatcccataaatttgggaaa
aatacaattttaatcaggttaattcgtgtgtagcgcgactgttgcatggtttcgggatctgggccacgggatcaggggccaggtagtcggtcgagcgaaaggattaggcctctggatctcg
gggatttagcgcaaacagaggtgagtccgtccagcacaggcccaccgcccatcgccttcctgcgacattttaatttaatctcagattacgagcgaacgaggagacgagatcgcggatc
gggaacgggatcggcatcgggatcagtccacagtccaccagattaggccactccgctcgtatattgcaaaataatggtcgagttccaggggccgccagttcgattgcaaaagtttcacg
taaatccgtaagattagatggctagatccccagattttcagattccctcccgcgcttagcccaattgttaggtcaattgccaaatcacttgagcggggaatggaaaactgggccatctcatc
gtggaggtaatcccagtcgcagcagatctaggtagatccttgaaagccactggtcatgcccactaaagctagcagtgcctccgatgatcgcattcccctgatccccacacaagtccaaa
gtagtcctctggactagatgcgttgactttcgaaccctgggaatactacagctacaaatagaaatgctacaaaacttaattataaataatgtttagattgcagaaacaaatattctagaaa
ctgcacatatttacgaattgtacatcaagattgtgaaagaataaatcaatgattaaaatcaaatatcgtttctctttctaccgacaaaattcacgggaaaaagtgcgaaaaatatcagctca
cgatccaagtcgcc

>gt_gt1
cggaacggatgcgctgcccagatacaatcgggatcggcagttagggaccgggcgagaacggtagatcggaggaaatcctggaacagagaacggaactgaactgaacggaactc
aatgaggtcctaagcctcattccgcataactctttttcatagagtgactcctggccagctcgtaatatgagcatttaaccccgtaaagtcttcgacccggccacgcccattgtatccgagttcc
gttccgttctgtactcatctcctttgagaagcacagttgccttttgcctcgagtttttcacatgaaactaaaaacgactgaaatccatgagattaaacggtctttaagcggtttaagccctttgtgc
acgcagttttgtaggtggcaaaatgtgccattgtcaccgagttctgtcaaaaatatgaactagcatatcatttcataattttcagtttcatctttaaaagtgcatttttttgcttctcggaaggagaac
acaaatattgttggaagctcagtgggccgtagggttccgggtgatccgacaggtgaccaggtacagctctctcgatcctccaatcctccatgctaatcctgactatctggctgcctcaagtta
atctcgcataatattgcgtgttccgaacagcgtaatcccctgaaaacaagtgtccagatccgcagctctccacttttaagccggcaatggattctagatgacttgactaactgtcgagctag
gctcaacgaacagatttctctgagcctggtcgcaacagtaatacgtttgacgtaatgcaggggaatc

>gt+36
cgatgcatccgtgtcatggttggcatgtccctactgcaagagaagaaataccaatgtaatctaggtagtctaggtaacggagtgccgggcttataatcccaatacatgtgccttttttcccact
aagagctcttaacttggccaaaaattgaaatatttaagagtggcatacatttttaaactcctaattatacgtttaattgtgaggctttacaatttttttgatttgaaaaaaaaattttttttttcgaaatttt
aaatttttgtaaggggtaccatcacaaaaatttttaaaatttcgaaaaaaatttttttttttaactttgttaatatttaaatacatgttttttgggttagtcaaaacgaccaaaaagaagtattcggttca
caatttgggcaagttttgggtgagatatggcacgctgagtgataaagcagacggaatatattgaatttaattaggttttttttttttgcgatttccaacagttaacagcttggtgccattatatatatat
agaacagattagtctggatgatttttcggatctggttggcaaatcgggtatgcccggttaggtccttggtccttggcagctggctgccaaacaaggggatcagagcgacgcggcttccgat
ctggtttcgttttttcccattcttctggtgtccttcggtttttgggaactcttccagacaatttcgaggcagaaaaacaacggcaacaaatgaaggggcgagctggcaaacaaagtgcagga
caaagcaggaaatttgcaaattgttcttgcggcaaacgagttcctcgccaggtccttcgtacgtcgctgtccttgcaaatttcaggtttgcatttaaattgaaattaaacatatgctaatttccct
aacagatttctgcgctgcggttgctaaccataaaattgagcgggtcctttagacaaggagtcttaaaggatgctctatggctctacggtctgagatctgcgatttgcgttttcttctcggtgtgtg
agtgtgagtttgtgtgtgtgcgtgtgtacgcaacaattacgccccagcataatgatgccatttttgcaaaaatgcggcgatcggcaaagccaaaaggggccttaaaggatcgtcccaaag
atcctggcctttgtctctcacgaagagtgagacagaaaatgggttttcgtgtaatccataaacgtcaccccagcccaagaataagatgaaaagccgggcgaaactacacaaggaaac
gcagccggtaaaggcgagcaaaaaaatattttctttccagcccaagatcctttgatgaatgtgcgtgtgtgtgtgtgacttgtgcccggaggattagctcggggtggtaggaaaaaaga
agttggagaaggaaaagaaaaaaaaaacacagcaacaaattaaatgtattagagaacgcaccacccctttttactgtgggtaccaagatctggatcgggatcgctgtctctagggggtt
agggacttacgtcaattcgtcccttccttggctaaattaaaattagagctagagcaaagacaaacacggctacctcatgggtgcgatcgtaattgtttttttccaggtgtgccgccgattcg
gaaaccgagcttatcgagcgatctacagatcgatcgatctaaatggacgtatcgatccatttctctgcacagattgtatgtcttttggggtatggacttacctggtaacttgatctttttgcgacta
cgtcgcctggcgaactgc

Mutated *gt* CREs

Mutated bases are in upper case.

>gt-1_mutated
ggacggccactAgtGTAGatgatcctggactctcgattctcgatcctcgggctggcggcgtACtgaatcaaatgggcagactggcgagtgtccAGCTcgTCattccgggaagcc
gagttttATCGTGcgaaataagaaaaggatgcgaattGGCccTgccgaaaagggctgaaaaGtgcaagGacgccagaactcgtcAAATTAGAATTATGGtCtcac
gcacaatccaacgaattcgccatcgcatggatacacagattctagatgggctggtcagcaaaaggatAAGCaTTcaaaagcgcaaaacaaacggatttagccatcacaaccctt
ggatggagctcatctattatcaacaaactagattccatctcttttcttttgccagatgcctctctttctggccagaaaagtatgatttgagtgtaattataatagactacggtcgggggtctttaat
atgcgaaatgttctagaagattcagaagtatggtttatttcagtgctgaagagaatagcataCagGaGCTgaaacatatacagAGTAATTGAGATTTAaaaggtttaataac
atattattttgaagattggtttttgggtaaaggttttatcagttagatgtcaaagcaagtaagttaggtacaattgAaCGTcTGcatcaaatgGacGatgctggtatatacaaaaatctgg
gtttcctcaaacttggacttataacatcgcaatgaacgcctatgaaaataaagccgaagtttataactctgatcatgaatatttttaaccaattccatgaaaacgaaatgtaaagcactgcag
aggtggtaagaattttcctatttttttcccattgaatattaaaaccaagaaaaaaccatatcctttcgcggaaactgatgaaacgtaaagttcgaaaactctttgtacttttcttttttaaatGGAT
CCaTaATTcgaactaacaatagcacaaagaactatgaAATATTTTTAACCAgcgaaaaccaaatgagggaaagggcagccgcaccctgtaaaaccctttcgaaagc
ggtgagacagagagggcaacaggccagagcgggacagagtcagaagcaaagccagaatcacccctttgcccttggccttgtaccgcctacctgcttcgtaccgcttcggctggcc
gctcgtgtccgccatttcgatgtgaatgaggcgtcggcgc

>gt-3_mutated
gacaaagggccgcgaaaggagttaacggcaattccccgaaaccaagcacgaatccaaatccaaaccactcctcacccttttttggacgggtgggacgggtcatataaggcaaaac
ccctttcagtcaattagtcgctgattttctttgtcacctagcagcggaccaatataaaaatcgcagccacaatggtcggaggagagaacccttttttttttaaggaccgccggtgtccgaaata
tcagtttatggctccttaaaaaactgcagcggttttagggcccgcggactcggaatgaggccttttcgcacgaagcgtccaGCTGTcaataaaacttggatgtttttagCGTTcagc
ataacagcctaatacagcgcataccgtgggAgggcTacccgttcggccatcaggtagatgaCAcGCtaaaaaacgcgcccaaggtcacaactcaaaggattgCTGCGata
GtgAatctttggttcggagctcattatggcgaaggaacagcagtccgagtatatgaatatatggttttatggcctgctgtttttacgttttttttcgctgaatctggttttacaccgaacttggcaa
ctctttggcttcgacttcggccgcgacttttatggGCAGtgctcgggatcgattgggctgggctatacgtatgtccaggatctggaagggctcgggttcgggcccggtaaccgcaggtaa
gggatcgactcaatgacggcgacgtgacggtccgagttttagtttaatccaccatttttacggttggattagcgcacggattagcggataagttcgcggtttttactgattaccatcgatcgat
ccctcttttttatttgcaaGAcaccGaatccccctgaatgggcagttgtaaaaagagaatgaatgaaagatagtatataatgtaggtattgttgataggatttgataacatttcagcgtaaca
aatctaaactacgttatCtcctcgaattagctgcagtaagtatctagcataacatttcccatggatgatagccaacttttgtaccgctacccacttccgttggctgtcgaggcgaccgttaggac
aaacagttaacaaagtgaaacaacttgtcgaggatcccacatcccacgccgcagcggaagccaaagggataaaggacgtagaatcaaggatgtttaaggaggagggcaaacac
gagcagt

Figure 71: ChIP-on-chip at the *giant* genomic locus.
Generated with the UCSC genome browser[9] on the *D.melanogaster* assembly (BDGP R5/dm3) for relevant factors.

[9] https://genome.ucsc.edu/

2 Fly stocks

genotype	line	chrom.	reference
lacZ reporters for *gt*			
pP[gt_CE8001]/CyO	1765	II	Berman BP, 2002
pP[gt_CE8001]/TM6B, Hu, Tb	1766	III	Berman BP, 2002
pP[gt_gt1]	G11	X	Ochoa-Espinosa A, 2005
pP[gt_gt1]	G12	II	Ochoa-Espinosa A, 2005
pP[gt_gt23]	G13	II	Ochoa-Espinosa A, 2005
pP[gt_gt23]	G14	III	Ochoa-Espinosa A, 2005
pP[gt_-1_construct]	MI 22 (3)	III	Schroeder MD, 2004
pP[gt_-1_construct]	MI 23 (5)	III	Schroeder MD, 2004
pP[gt_-3_construct]	MI 27 (5)	X	Schroeder MD, 2004
pP[gt_-3_construct]	MI 28 (2)	II	Schroeder MD, 2004
pP[gt_-3_construct]	MI 29 (3)	II	Schroeder MD, 2004
pP[gt_-6_construct]	MI 33 (2)	II	Schroeder MD, 2004
pP[gt_-6_construct]	MI 34 (3)	III	Schroeder MD, 2004
pP[gt_-6_construct]	MI 35 (4)	III	Schroeder MD, 2004
pP[gt_-10_construct]	MI 39 (1)	X	Schroeder MD, 2004
pP[gt_-10_construct]	MI 41 (2)	X	Schroeder MD, 2004
site-directed integration			
y-w-, nanos - phiC31 - y+; attP w+ attP 37B7	phiC31; 37B	X; II	Basler, Bateman
y-w-, nanos - phiC31 - y+; attP w+ attP 38F1	phiC31; 38F	X; II	Basler, Bateman
y-w-, nanos - phiC31 – y+ ; ; attP w+ attP 89B8	phiC31; 89B	X; III	Basler, Bateman
y1 w*; P{attP.w+.attP}JB37B7	37B	II	Bateman JR, 2006
y1 w*; P{attP.w+.attP}JB38F1	38F	II	Bateman JR, 2006
y1 w*; P{attP.w+.attP}JB89B8	89B	III	Bateman JR, 2006
y w M{eGFP.vas-int.Dm}ZH-2A; +; Sb/TM6B; +	vas phi ZH2A	I; III	Bischof J, 2007
y1 M{vas-int.Dm}ZH-2A w*; M{3xP3-RFP.attP}ZH-51D	BS24483	I; II	Bischof J, 2007
balancer stocks			
w ; TM3, Sb, Ser / Tm6B, Tb	BS2537	I ; III	
y[1] w[*] ; TM3, Sb[1]/TM6B, Tb[+]	BS3720	I ; III	
w[1118]; In(2LR)Gla, wg[Gla-1] Bc[1]/CyO	BS5439	I ; II	
FM7a	BS785	I	
w ; CyO / Bl ; TM2 / TM6B	MS32	I ; II ; III	
Sp / CyO ; delta(2-3), Sb / TM6, Ubx G	transposase	II, III	

Table A. 5: Fly lines with lacZ reporters for *gt*, target lines for site-specific integration and balancers. BS refers to Bloomington Stock.

allele / def.	mutagen	class	description
gt^{X11}	X ray	amorphic	abdominal segments 5 to 8 frequently affected, 6 & 7 always missing
Kr^{1}	spontaneous	amorphic	thorax and anterior abdomen is abnormal in homozygous embryos
hb^{12}	EMS	amorphic	W256 replaced by premature stop codon before the 1st finger domain
bcd^{6} (bcd^{E1})	EMS	amorphic	aa 156-494 replaced by 55 out-of-frame aa; homeobox affected
cad^{264}	EMS		lack of maternal cad leads to weak posterior segmentation defect
kni: Df(3L)ri-79c			computed breakpoints include 77B7-77B9;77F1-77F5
tll: Df(3R)tll-g	X ray		computed breakpoints include 99F1-99F2;100B5

Table A. 6: Mutant alleles and deficiencies.
All mutants are homozygous lethal before the end of embryonic stage. EMS is ethyl methane-sulfonate.

genotype	description	source	chrom.
mutants			
cn[1] bw[1] Kr[1]/SM6a, bw[k1]	*Kr* allele	BS3494	
hb[12] st[1] e[1]/TM3, Sb[1]	*hb* allele	BS1755	
Df(3R)tll-g, ca[1]/TM3, Sb[1] Ser[1]	*tll* deficiency	BS2599	III
Df(3L)ri-79c/TM3, Sb[1]	*kni* deficiency	BS3127	III
y[1] sc[1] gt[X11]/FM6	*gt* allele	BS1529	
Df(1)62g18, y[1]/Dp(1;2;Y)w[+] & C(1)DX, y[1] f[1]/Dp(1;2;Y)w[+]	*gt* deficiency	BS1517	
th1 st1 kniri-1 bcd6 rnroe-1 pp/TM3, Sb1	*bcd*E1	BS 3630	III
w1118, bcd K57R/CyO	*bcd* cooperativity	Lebrecht	II
w1118, bcd S35T / Tm3, Sb	*bcd* cooperativity	Lebrecht	III
y[1] P{y[+mDint2] w[BR.E.BR]=SUPor-P}KG01741/FM7c	transposon insertion	BS14395	I
germ line clones (GLC) and transgenic RNAi			
w^{*}; P{neoFRT}82B P{ovoD1-18}3R/st^{1} βTub85D^{D} ss^{1} e^{s}/TM3, Sb^{1}	GLC	BS2149	
w* ; P{w+ ovoD1} P{neo FRT} 40A/T(1,2)OR64/SM6a	GLC	Irion	I ; II
y*w*, P{hs-flp}122 ; If/CyO, hs-hid	GLC	Irion	I ; II
y*w*, P{hs-flp}122 ; P{w+ ovoD1} P{neo FRT} 40A/T(1,2)OR64/CyO, hs-hid	GLC		I ; II
hb[12] FRT 82B [neo+] e[1] / TM3 Sb	hb GLC	DePace	III
cad(2L-264-12-3)FRT40A/CyO, hs-hid	cad GLC	Irion	II
P{otu-GAL4::VP16.R}1, w^{*}; P{GAL4-nos.NGT}40; P{GAL4::VP16-nos.UTR}CG6325^{MVD1}	MTD-Gal4	BS 31777	I, II, III
y w; P(mat-tub-Gal4)mat67; P(mat-tub-Gal4)mat15	mat-tub-Gal4		II, III
y sc v ; {v+ y+ UAS-shRNA-hb} attP40	hb short-hairpin	Staller	II

Table A. 7: Fly lines for mutants, germ line clones and transgenic RNAi.
BS refers to Bloomington Stock.

gene	arm	cp	gene	arm	cp
gap genes			**maternal genes**		
hb	3R	85A5	*bcd*	3R	84A5
Kr	2R	60F5	*cad*	2L	38E9
kni	3L	77E3	*nos*	3R	91F4
gt	X	3A3	*stau*	2R	55B5
tll	3R	100A6	**pair-rule genes**		
hkb	3R	82A4	*eve*	2R	46C10
nub	2L	33F1	*h*	3L	66D10
cas	3R	83C1	*slp*	2L	24C6
odt	X	7F10	*run*	X	19E2
btd	X	8F10	*odd*	2L	24A1
ems	3R	88A2	*ftz*	3R	84A6
segment polarity genes			*prd*	2L	33C3
en	2R	47F17	*opa*	3R	82D8
wg	2L	27F1	*lilli*	2L	23C1

Table A. 8: Cytogenic positions of relevant segmentation genes.
Arm refers to the chromosome arm. cp denotes the cytogenic position.

genotype	stock	chrom.	ori	description
lacZ reporter				
w-; 37B-Pgt-gt-3	I.1.2	II-37B7	3'	*gt* core promoter – gt-3
y1 w*; ; 89B-Pgt-gt-3	II.2.3	III-89B8	5'	*gt* core promoter – gt-3
y1 w*; ; 89B-Pgt-neg.ctrl.	III.4.1	III-89B8	3'	empty cassette att-Pgt-Z
y1 w*; ; 89B-Pgt-neg.ctrl.	III.5.4	III-89B8	5'	empty cassette att-Pgt-Z
w-; 37B-neg.ctrl.	IV.2.8	II-37B7	3'	empty cassette att-Pgt-Z
w-; 37B-neg.ctrl.	IV.6.1	II-37B7	5'	empty cassette att-Pgt-Z
w-; 37B-Peve-gt-3/CyO	V.1.1	II-37B7	3'	*eve* basal promoter – gt-3
w-; 37B-Peve-gt-3/CyO	V.2.1	II-37B7	5'	*eve* basal promoter – gt-3
w-; 37B-Peve-neg.ctrl./CyO	VI.1.5	II-37B7	5'	empty cassette att-Peve-Z
w-; 37B-Phsp70-gt-3/CyO	VII.1.2	II-37B7	5'	*hsp70* promoter – gt-3
w-; 37B-Phsp70-gt-3/CyO	VII.1.4	II-37B7	3'	*hsp70* promoter – gt-3
w-; 37B-Phsp70-neg.ctrl./CyO	VIII.2.1	II-37B7	5'	empty cassette att-Phsp-Z
y+, w-; ; 89B-Pgt-gt-1	IX.3.1.	III-89B8	5'	*gt* core promoter – gt-1 (gaps)
y+, w-; ; 89B-Pgt-gt-1	IX.3.3	III-89B8	3'	*gt* core promoter – gt-1 (gaps)
y1 w*; ; 89B-Pgt-gt-1-gt-3	X.1.1	III-89B8	5'	combined CRE (gt-1 and gt-3)
y1 w*; ; 89B-Pgt-gt-1_mut	1.2	III-89B8	5'	gt-1 with all Gt sites mutated
w-; 38F-Pgt-gt-1_mut	2.1	II-37B7	5'	gt-1 with all Gt sites mutated
w-; 38F-Pgt-gt-3_mut /CyO	1.1	II-37B7	5´	gt-3 with all Gt sites mutated
mutants				
Kr / CyO ; 89B-Pgt-gt-1-gt-3	BS3494, X.1.1	II, III	5'	combined CRE in *Krüppel* mutant
Kr / CyO ; 89B-Pgt-gt-3 / TM6	BS3494, II.2.3	II, III	5'	gt-3 in *Krüppel* mutant
37B -Phsp gt-3 ; hb12/ TM6	VII.1.2, BS1755	II, III	5'	gt-3 in *hunchback* mutant
37B -Phsp gt-3 ; Kni / TM6	VII.1.2, BS3127	II, III	5'	gt-3 in *knirps* mutant
37B -Peve gt-3 ; Kni / TM6	V.2.1, BS3127	II, III	5'	gt-3 in *knirps* mutant
37B -Phsp gt-3; Tll / TM6	VII.1.2, BS2599	II, III	5'	gt-3 in *tailless* mutant
y1 sc1 gtX11 /FM6 ;; 89B-Pgt- gt-3	BS1529, II.2.3	I, III	5'	gt-3 in *giant* mutant
cad, FRT / CyO ; 89B-Pgt-gt-3	cad, II.2.3	II, III	5'	gt-3 in *caudal* mutant
cad, FRT / CyO ; 89B-Pgt-gt-1	cad, IX.3.1	II, III	5'	gt-1 in *caudal* mutant
cad, FRT / CyO ;89B-Pgt-gt-1-gt-3	cad, X.1.1	II, III	5'	combined CRE in *caudal* mutant
89B-Pgt-gt-3, hb12 / TM6	24	III	5´	gt-3 recombined with *hb* mutant
hb12, 89B-Pgt-gt-1 / TM6	12	III	5´	gt-1 recombined with *hb* mutant
Kni, 89B-Pgt-gt-1 / TM3	11	II, III	5´	gt-1 recombined with *kni* mutant
bcd6, Bcd S35T / TM2	5	III		Bicoid cooperativity rescue
UAS-sh-hb; hb12 / TM6B		II, III		transgenic RNAi

Table A. 9: Fly lines generated during this study.
Ori refers to the orientation of the integrated reporter-cassette. BS refers to Bloomington Stock.

3 Model

3.1 PWMs

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Bcd Selex	A	83	74	108	48	6	381	379	5	0	6	72	61	65	68
	C	114	159	127	149	1	0	0	0	383	340	136	174	166	158
	G	106	72	114	11	0	2	4	4	0	3	132	60	52	49
	T	80	78	34	175	376	0	0	374	0	34	43	88	100	108
Cad B1H	A	9	12	3	4	12	38	0	4	22	1				
	C	10	6	3	0	0	0	0	0	0	8				
	G	4	4	3	0	2	0	0	7	15	10				
	T	11	16	29	34	24	0	38	27	1	1				
Hb Selex	A	53	2	0	2	0	0	0	281	31	20				
	C	6	6	2	0	2	3	2	0	43	100				
	G	224	3	0	0	0	0	0	3	78	109				
	T	7	279	288	288	288	287	288	6	138	61				
Kr Selex	A	17	187	158	0	1	0	8	0	2	44				
	C	73	5	39	194	194	197	22	2	34	109				
	G	6	0	0	1	0	0	6	0	2	15				
	T	101	5	0	2	2	0	161	195	159	29				
Kni B1H	A	19	25	16	5	0	21	0	17	1	0	25	5		
	C	1	1	0	9	4	0	0	0	3	26	0	12		
	G	2	0	0	6	1	5	26	8	18	0	1	7		
	T	4	0	10	6	21	0	0	1	4	0	0	2		
Gt Selex	A	86	12	776	8	83	0	1020	1106	15					
	C	62	108	25	762	19	556	88	0	378					
	G	19	359	275	65	996	0	1	0	85					
	T	942	630	33	274	11	553	0	3	631					
Tll footprint Rajewsky 2002	A	12	1	1	5	2	11	1	0	0					
	C	8	2	2	1	3	1	17	2	3					
	G	0	2	1	0	15	5	0	1	2					
	T	0	15	16	14	0	3	2	17	15					
Hkb B1H	A	3	8	0	0	0	0	0	0	24					
	C	0	0	0	0	32	0	0	0	3					
	G	18	24	32	32	0	32	3	31	2					
	T	11	0	0	0	0	0	29	1	3					

Table A. 10: PWMs used in the model.
Matrices as used in Kim et al. (2013), except Hkb.

3.2 Optimization

Optimization runs were repeated 10 times at the beginning and then reduced to 3, since the outputs tend to fall into similar solutions, regarding the quality of fit and the contributions of the TFs. One optimization run took from several hours to several days (depending on the input) on the machines of the lab or on the cluster of the CRG (using 20 processors, Intel® Xeon® E5-2680, 2.70GHz).

parameter	value
starting temperature	100000
gain	3
interval	100
lambda memu	0.2
lambda memv	100
initial moves	100000
tau	100
freeze count	5
criterion	0.1
end temperature	1

Table A. 11: Parameters for simulated annealing.

		91_1	132_1	139_1	139_2			91_1	132_1	139_1	139_2
T_a	Bcd	2,879	4,830	0,693	4,946	A_a	Bcd	0,145	0,00001	1,217	1,630
	Cad	4,992	4,389	4,930	4,773		Cad	0,055	0,062	3,081	0,035
	Gt	4,458	n.a.	3,884	4,867		Gt	0,064	n.a.	0,060	0,054
	Hb	3,880	1,096	1,645	4,041		Hb	0,052	0,055	0,021	3,270
	Kr	0,032	0,016	0,129	0,292		Kr	0,009	0,063	0,001	0,036
	Kni	4,522	4,015	2,477	1,824		Kni	1,279	0,888	0,177	0,156
	Tll	4,162	0,020	1,288	4,496		Tll	0,014	0,006	0,077	0,076
	Hkb	1,094	0,082	3,060	4,872		Hkb	0,697	0,199	0,510	0,592
λ_a	Bcd	0,538	1,142	0,912	1,060	E^A	Bcd	19,750	0,407	17,643	19,277
	Cad	0,500	5,000	3,386	4,798		Cad	3,854	0,852	3,281	1,033
	Gt	4,999	n.a.	1,613	1,728		Gt	0,830	n.a.	2,502	2,778
	Hb	4,999	2,113	0,674	1,074	E^Q	Hb	0,177	0,999	0,431	0,455
	Kr	4,992	3,846	4,965	4,141		Kr	0,997	0,994	0,962	0,0003
	Kni	1,763	2,689	4,358	0,567		Kni	0,997	0,867	0,000	0,010
	Tll	4,980	4,659	4,875	4,951		Tll	0,999	0,998	0,091	0,455
	Hkb	0,584	4,997	0,582	0,733		Hkb	0,879	0,999	0,302	0,653
θ		5,000	5,000	7,767	6,005						

Table A. 12: Parameter values of the selected models.
The number of the run is indicated in the first row. Shown are the PWM threshold T_a, the PWM scaling factor λ_a and the protein scaling factor A_a for each TF and the activation threshold θ for each run, as well as the activation coefficient E^A for Bcd, Cad and Gt and the quenching coefficient E^Q for Hb, Kr, Kni, Tll and Hkb. All values are in arbitrary units.

nr.	CRE	Gt	coact	dist	coop	threshold	bsize	direct	max	AP%	notes
1	gt-3	R	yes	10	yes	B, H fix	24	on	adjust	35-92	test-run. T1-T8. dataset not yet cleaned-up.
2	gt-3	R	yes	10	no	B, H fix	24	on	adjust	35-92	test-run. T1-T8. dataset not yet cleaned-up.
3	gt-3	off	no		no	B, H fix	24	on	adjust	35-92	test-run. dataset not yet cleaned-up.
4	gt-3	off	yes	10	yes	B, H fix	24	on	adjust	35-92	test-run. dataset not yet cleaned-up.
5	gt-3	off	no		no	B, H fix	24	on	adjust	35-92	
6	gt-3	off	no		no	B, H fix	24	on	adjust	35-92	
7	gt-3	A	yes	10	yes	B, H fix	24	on	adjust	35-92	
8	gt-3	A	no		no	B, H fix	24	on	adjust	35-92	
9	gt-1	off	no		no	B, H fix	24	on	adjust	35-92	
10	gt-1	A	no		no	B, H fix	24	on	adjust	35-92	
11	gt-1	off	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
12	gt-1	A	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
13	gt-3, gt-1	A	no		no	B, H fix	24	on	adjust	35-92	Hkb excluded
14	gt-3, gt-1	A	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
15	gt-3, gt-1	off	no		no	B, H fix	24	on	adjust	35-92	Hkb excluded
16	gt-3, gt-1	off	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
17	gt-3, gt-1	A	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
18	gt-3	A	no		yes	B, H fix	24	on	adjust	35-92	Hkb excluded
19	gt-1	A	no		no	B, H fix	24	on	adjust	35-92	Hkb excluded
20	gt-3, gt-1	A	yes	10	yes	B, H fix	24	off	adjust	35-92	Hkb excluded
21	gt-3, gt-1	A	no		no	B, H fix	24	off	adjust	35-92	Hkb excluded
22	gt-3, gt-1	A	no		no	B, H fix	14	off	adjust	35-92	Hkb excluded
23	gt-3, gt-1	A	no		yes	B, H fix	24	on	adjust	35-92	
24	gt-3, gt-1	A	no		yes	B, H fix	24	off	adjust	35-92	
25	gt-3	A	no		no	B, H fix	24	off	adjust	35-92	only Gt as input TF
26	gt-1	A	no		no	B, H fix	24	off	adjust	35-92	only Gt as input TF
27	gt-3	A	no		no	B, H fix	24	off	adjust	35-92	Bcd and Cad excluded
28	gt-1	A	no		no	B, H fix	24	off	adjust	35-92	Bcd and Cad excluded
29	gt-3, gt-1	A	no		no	B, H fix	24	off	adjust	35-92	Bcd and Cad excluded
30	gt-3, gt-1	R	no		no	B, H fix	14	off	adjust	35-92	
31	gt-3, gt-1	R	yes	10	yes	B, H fix	14	off	adjust	35-92	
32	gt-3	A	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1
33	gt-1	A	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1
34	gt-3	off	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1

nr.	CRE	Gt	coact	dist	coop	threshold	bsize	direct	max	AP%	notes
35	gt-1	off	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1
36	gt-3	A	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1. Bcd and Cad excluded
37	gt-1	A	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1. Bcd and Cad excluded
38	gt-3	off	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1. Bcd and Cad excluded
39	gt-1	off	no		no	B, H fix	14	off	adjust	35-92	C12, C13, T1. Bcd and Cad excluded
40	gt-3, gt-1	off	no		no	B, H fix	14	off	adjust	35-92	WLS
41	gt-3, gt-1	A	no		no	B, H fix	14	off	adjust	35-92	WLS
42	gt-3, gt-1	R	no		no	B, H fix	14	off	adjust	35-92	WLS
43	gt-3, gt-1	off	yes	10	yes	B, H fix	14	off	adjust	35-92	WLS
44	gt-3, gt-1	A	yes	10	yes	B, H fix	14	off	adjust	35-92	WLS
45	gt-3, gt-1	R	yes	10	yes	B, H fix	14	off	adjust	35-92	WLS
46	gt-3, gt-1	off	no		yes	B, H fix	14	off	adjust	35-92	WLS
47	gt-3, gt-1	A	no		yes	B, H fix	14	off	adjust	35-92	WLS
48	gt-3, gt-1	R	no		yes	B, H fix	14	off	adjust	35-92	WLS
49	gt-3, gt-1	off	yes	10	no	B, H fix	14	off	adjust	35-92	WLS
50	gt-3, gt-1	A	yes	10	no	B, H fix	14	off	adjust	35-92	WLS
51	gt-3, gt-1	R	yes	10	no	B, H fix	14	off	adjust	35-92	WLS
52	gt-3, gt-1	off	no		no	adjust all	14	off	adjust	35-92	
53	gt-3, gt-1	A	no		no	adjust all	14	off	adjust	35-92	
54	gt-3, gt-1	R	no		no	adjust all	14	off	adjust	35-92	
55	gt-3, gt-1	off	yes	10	yes	adjust all	14	off	adjust	35-92	
56	gt-3, gt-1	A	yes	10	yes	adjust all	14	off	adjust	35-92	
57	gt-3, gt-1	R	yes	10	yes	adjust all	14	off	adjust	35-92	
58	gt-3, gt-1	both	no		no	adjust all	14	off	255	35-92	Gt co-activated by zld
59	gt-3, gt-1	both	yes	150	yes	adjust all	14	off	255	35-92	Gt co-activated by zld
60	gt-3, gt-1	off	no		no	adjust all	14	off	255	35-92	all B1H
61	gt-3, gt-1	A	no		no	adjust all	14	off	255	35-92	all B1H
62	gt-3, gt-1	R	no		no	adjust all	14	off	255	35-92	all B1H
63	gt-3, gt-1	off	yes	150	yes	adjust all	14	off	255	35-92	all B1H
64	gt-3, gt-1	A	yes	150	yes	adjust all	14	off	255	35-92	all B1H
65	gt-3, gt-1	R	yes	150	yes	adjust all	14	off	255	35-92	all B1H
66	gt-3, gt-1	off	no		no	adjust all	14	off	255	31-92	
67	gt-3, gt-1	A	no		no	adjust all	14	off	255	31-92	
68	gt-3, gt-1	R	no		no	adjust all	14	off	255	31-92	
69	gt-3, gt-1	off	yes	150	yes	adjust all	14	off	255	31-92	

nr.	CRE	Gt	coact	dist	coop	threshold	bsize	direct	max	AP%	notes
70	gt-3, gt-1	A	yes	150	yes	adjust all	14	off	255	31-92	
71	gt-3, gt-1	R	yes	150	yes	adjust all	14	off	255	31-92	
72	gt-3, gt-1	off	no		no	adjust all	14	off	255	31-92	T1-T8
73	gt-3, gt-1	A	no		no	adjust all	14	off	255	31-92	T1-T8
74	gt-3, gt-1	R	no		no	adjust all	14	off	255	31-92	T1-T8
75	gt-3, gt-1	off	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
76	gt-3, gt-1	A	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
77	gt-3, gt-1	R	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
78	combined	off	no		no	adjust all	14	off	255	31-92	T1-T8
79	combined	A	no		no	adjust all	14	off	255	31-92	T1-T8
80	combined	R	no		no	adjust all	14	off	255	31-92	T1-T8
81	combined	off	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
82	combined	A	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
83	combined	R	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
84	combined	off	no		no	adjust all	14	off	255	31-92	
85	combined	A	no		no	adjust all	14	off	255	31-92	
86	combined	R	no		no	adjust all	14	off	255	31-92	
87	combined	off	yes	150	yes	adjust all	14	off	255	31-92	
88	combined	A	yes	150	yes	adjust all	14	off	255	31-92	
89	combined	R	yes	150	yes	adjust all	14	off	255	31-92	
90	combined	off	no		no	adjust all	14	off	255	31-92	
91	combined	A	no		no	adjust all	14	off	255	31-92	
92	combined	R	no		no	adjust all	14	off	255	31-92	
93	combined	off	yes	150	yes	adjust all	14	off	255	31-92	
94	combined	A	yes	150	yes	adjust all	14	off	255	31-92	
95	combined	R	yes	150	yes	adjust all	14	off	255	31-92	
96	combined	off	no		no	adjust all	14	off	255	31-92	Kni excluded
97	combined	A	no		no	adjust all	14	off	255	31-92	Kni excluded
98	combined	R	no		no	adjust all	14	off	255	31-92	Kni excluded
99	combined	off	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
100	combined	A	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
101	combined	R	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
102	all 3	off	no		no	adjust all	14	off	255	31-92	T1-T8
103	all 3	A	no		no	adjust all	14	off	255	31-92	T1-T8
104	all 3	R	no		no	adjust all	14	off	255	31-92	T1-T8

nr.	CRE	Gt	coact	dist	coop	threshold	bsize	direct	max	AP%	notes
105	all 3	off	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
106	all 3	A	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
107	all 3	R	yes	150	yes	adjust all	14	off	255	31-92	T1-T8
108	all 3	off	no		no	adjust all	14	off	255	31-92	Kni excluded
109	all 3	A	no		no	adjust all	14	off	255	31-92	Kni excluded
110	all 3	R	no		no	adjust all	14	off	255	31-92	Kni excluded
111	all 3	off	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
112	all 3	A	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
113	all 3	R	yes	150	yes	adjust all	14	off	255	31-92	Kni excluded
114	all 3	off	no		no	adjust all	14	off	255	31-92	
115	all 3	A	no		no	adjust all	14	off	255	31-92	
116	all 3	R	no		no	adjust all	14	off	255	31-92	
117	all 3	off	yes	150	yes	adjust all	14	off	255	31-92	
118	all 3	A	yes	150	yes	adjust all	14	off	255	31-92	
119	all 3	R	yes	150	yes	adjust all	14	off	255	31-92	
120	combined	off	no		no	adjust all	14	off	255	31-92	Hb excluded
121	combined	A	no		no	adjust all	14	off	255	31-92	Hb excluded
122	combined	R	no		no	adjust all	14	off	255	31-92	Hb excluded
123	combined	off	no		yes	adjust all	14	off	255	31-92	Hb excluded
124	combined	A	no		yes	adjust all	14	off	255	31-92	Hb excluded
125	combined	R	no		yes	adjust all	14	off	255	31-92	Hb excluded
126	gt-3, gt-1	off	no		no	adjust all	14	off	255	31-92	Hb excluded
127	gt-3, gt-1	A	no		no	adjust all	14	off	255	31-92	Hb excluded
128	gt-3, gt-1	R	no		no	adjust all	14	off	255	31-92	Hb excluded
129	gt-3, gt-1	off	yes	150	yes	adjust all	14	off	255	31-92	Hb excluded
130	gt-3, gt-1	A	yes	150	yes	adjust all	14	off	255	31-92	Hb excluded
131	gt-3, gt-1	R	yes	150	yes	adjust all	14	off	255	31-92	Hb excluded
132	gt-3	off	no		no	adjust all	14	off	255	31-92	
133	gt-3	A	no		no	adjust all	14	off	255	31-92	
134	gt-3	R	no		no	adjust all	14	off	255	31-92	
135	gt-3	off	yes	150	yes	adjust all	14	off	255	31-92	
136	gt-3	A	yes	150	yes	adjust all	14	off	255	31-92	
137	gt-3	R	yes	150	yes	adjust all	14	off	255	31-92	
138	gt-1	off	no		no	adjust all	14	off	255	31-92	
139	gt-1	A	no		no	adjust all	14	off	255	31-92	

nr.	CRE	Gt	coact	dist	coop	threshold	bsize	direct	max	AP%	notes
140	gt-1	R	no		no	adjust all	14	off	255	31-92	
141	gt-1	off	yes	150	yes	adjust all	14	off	255	31-92	
142	gt-1	A	yes	150	yes	adjust all	14	off	255	31-92	
143	gt-1	R	yes	150	yes	adjust all	14	off	255	31-92	
144	gt-3	off	no		no	adjust all	14	off	255	31-92	all B1H
145	gt-3	A	no		no	adjust all	14	off	255	31-92	all B1H
146	gt-3	R	no		no	adjust all	14	off	255	31-92	all B1H
147	gt-3	off	yes	150	yes	adjust all	14	off	255	31-92	all B1H
148	gt-3	A	yes	150	yes	adjust all	14	off	255	31-92	all B1H
149	gt-3	R	yes	150	yes	adjust all	14	off	255	31-92	all B1H
150	gt-1	off	no		no	adjust all	14	off	255	31-92	all B1H
151	gt-1	A	no		no	adjust all	14	off	255	31-92	all B1H
152	gt-1	R	no		no	adjust all	14	off	255	31-92	all B1H
153	gt-1	off	yes	150	yes	adjust all	14	off	255	31-92	all B1H
154	gt-1	A	yes	150	yes	adjust all	14	off	255	31-92	all B1H
155	gt-1	R	yes	150	yes	adjust all	14	off	255	31-92	all B1H
156	gt-3	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by bcd, cad. all B1H
157	gt-1	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by bcd, cad. all B1H
158	gt-3	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by cad. all B1H
159	gt-1	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by cad. all B1H
160	gt-3	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by bcd. all B1H
161	gt-1	both	no		no	adjust all	14	off	255	31-92	Gt co-activated by bcd. all B1H
162	gt-3	off	no		no	adjust all	14	off	255	31-92	T1-T8
163	gt-1	off	no		no	adjust all	14	off	255	31-92	T1-T8
164	gt-3	A	no		no	B, H fix	14	off	255	31-92	

Table A. 13: Optimization runs and their corresponding inputs.

CRE indicates which datasets were used as input. Gt refers to the role of Giant protein, which was either set as a repressor (R), an activator (A), turned off (off), or co-activated (both) by the TF indicated in notes. Coact indicates if co-activation (coact) of Hb by Bcd and Cad was considered and the co-activation distance for Cad was adjusted within the value stated in the column dist and 200bp. Coop shows if cooperation between Bcd sites was included.The threshold was either adjusted for all TF or fixed for Bcd and Hb to 1.71 and 0.63, respectively. The size of the binding sites (bsize) was either set to 14bp for all TF or to 24bp for Gt. Direct repression (direct) was either turned on or off and the maximum rate (max) was fixed to 255 or adjusted. An A-P range of 31-92% or 35-92% was considered in the models. All runs were performed as OLS, unless WLS is stated in notes. C12, C13, and T1-T8 of C14A were considered, unless other time points are stated in notes. All TF were included, unless otherwise stated in notes. The same PWMs as in Kim et al. (2013) or B1H matrices for all TF (all B1H) were used. From run 11 on, the expression of Tll and Hkb was set to zero in the middle of the embryo. From run 90 on, smoothed CRE datasets were used.

4 Expression dynamics of *giant* CREs

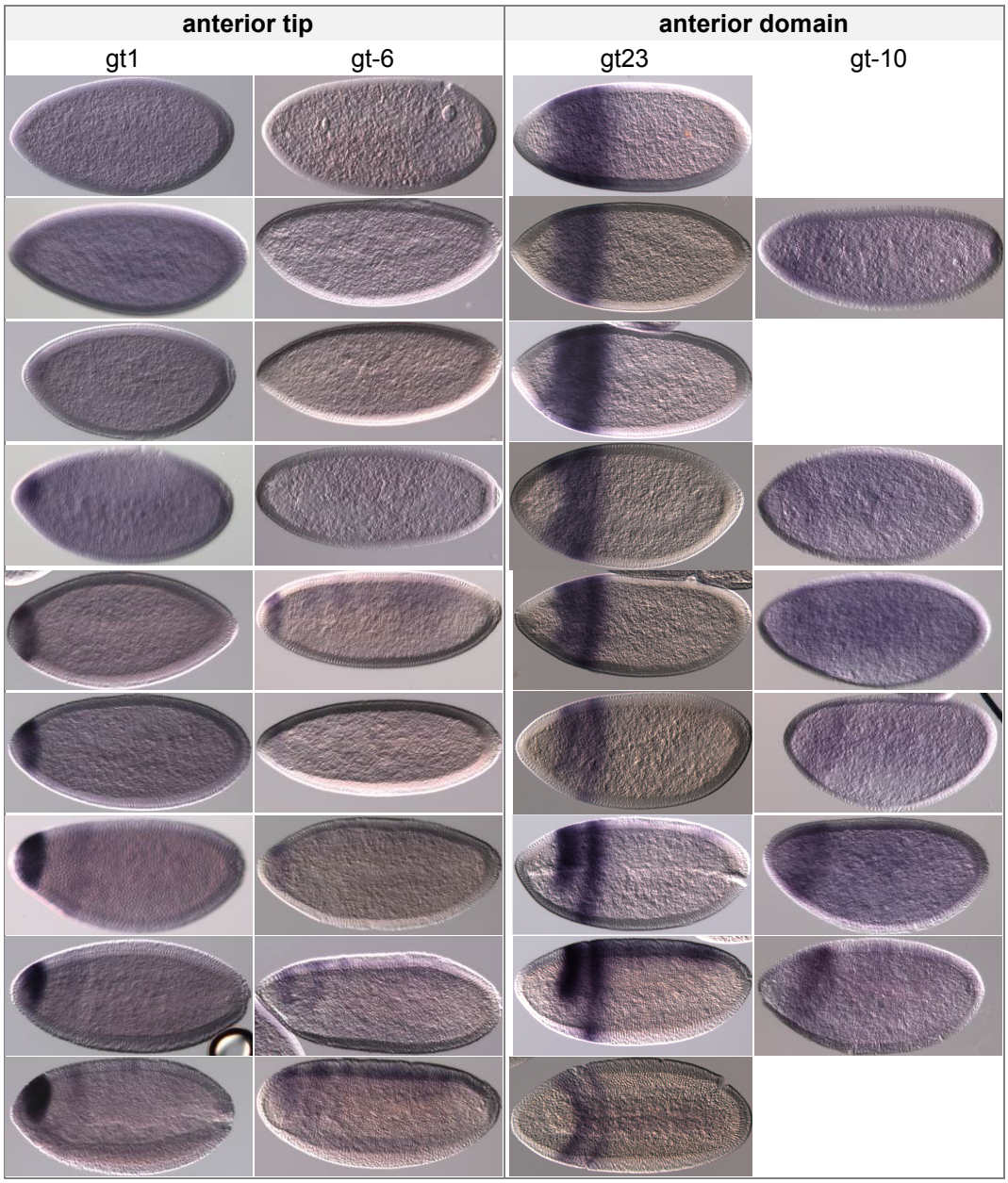


Figure 72: Previously available reporter-fly lines with *gt* CREs driving expression in the anterior. Enzymatic *in situ* hybridizations with a DIG-labeled lacZ probe in cleavage cycle 14, starting with the earliest and ending at gastrulation. Stages were estimated based on the membrane morphology. Anterior to the left, dorsal up. All embryos are lateral, except the gastrulating one of gt23, which is dorsal.

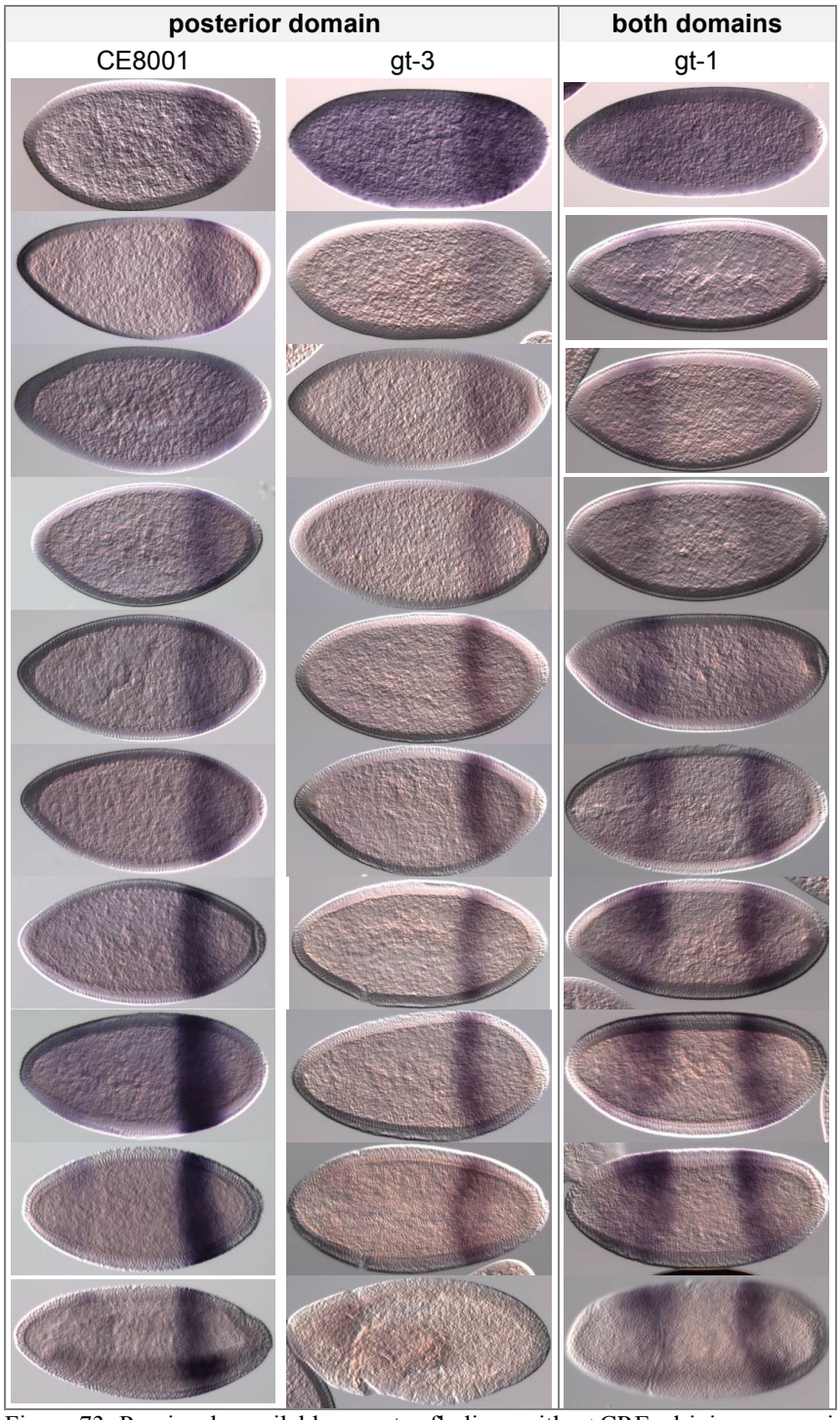


Figure 73: Previously available reporter-fly lines with *gt* CREs driving expression in the posterior. Enzymatic *in situ* hybridizations with a DIG-labeled lacZ probe at cleavage cycle 14. The first row shows embryos at C13 and the last row at gastrulation. Stages are estimated based on the membrane morphology.

Bibliography

Akam, M. 1987. The molecular basis for metameric pattern in the Drosophila embryo. Development. 101: 1–22.

Arnold, C. D., D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark. **2013**. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 339: 1074–7.

Arnosti, D. N., and M. M. Kulkarni. **2005**. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. 94: 890–8.

Ashworth, J., E. J. Wurtmann, and N. S. Baliga. **2012**. Reverse engineering systems models of regulation: discovery , prediction and mechanisms. Curr. Opin. Biotechnol. 23: 598–603.

Ashyraliyev, M., K. Siggens, H. Janssens, J. Blom, M. Akam, and J. Jaeger. **2009**. Gene circuit analysis of the terminal gap gene huckebein. PLoS Comput Biol. 5: e1000548.

Baena-Lopez, L. A., C. Alexandre, A. Mitchell, L. Pasakarnis, and J.-P. Vincent. **2013**. Accelerated homologous recombination and subsequent genome modification in Drosophila. Development. 140: 4818–25.

Bateman, J. R., A. M. Lee, and C. Wu. **2006**. Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. Genetics. 173: 769–777.

Becker, K., E. Balsa-canto, D. Cicin-sain, A. Hoermann, and H. Janssens. **2013**. Reverse-Engineering Post-Transcriptional Regulation of Gap Genes in Drosophila melanogaster. PLoS Comput Biol. 9: 37–40.

Berger, S. L., W. D. Cress, a Cress, S. J. Triezenberg, and L. Guarente. **1990**. Selective inhibition of activated but not basal transcription by the acidic activation domain of VP16: evidence for transcriptional adaptors. Cell. 61: 1199–208.

Berger, S. L., B. Piña, N. Silverman, G. a Marcus, J. Agapite, J. L. Regier, S. J. Triezenberg, and L. Guarente. **1992**. Genetic isolation of ADA2: a potential transcriptional adaptor required for function of certain acidic activation domains. Cell. 70: 251–65.

Berleth, T., M. Burri, G. Thoma, D. Bopp, S. Richstein, G. Frigerio, M. Noll, and C. Nüsslein-Volhard. **1988**. The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. EMBO J. 7: 1749–1756.

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. **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, M. B. Eisen, and S. E. Celniker. **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.

Bischof, J., R. K. Maeda, M. Hediger, F. Karch, and K. Basler. **2007**. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A. 104: 3312–3317.

Blackwood, E. M., and J. T. Kadonaga. **1998**. Going the Distance: A Current View of Enhancer Action. Science. 281: 60–63.

Bridges, C. B., and E. Gabritschevsky. **1928**. The giant mutation inDrosophila melanogaster. Z. Indukt. Abstamm. Vererbungsl. 46: 231–247.

Broenner, G., Q. Chu-LaGraff, C. Q. Doe, B. Cohen, D. Weigl, H. Taubert, and H. Jaeckle. **1994**. Sp1/egr-like zinc-finger protein required for endoderm speciication and germ-layer formation in Drosophila. Nature. 369: 664–668.

Broenner, G., and H. Jaeckle. **1991**. Control and function of terminal gap gene activity in the posterior pole region of the Drosophila embryo. Mech. Dev. 35: 205–211.

Burz, D. S., and S. D. Hanes. **2001**. Isolation of mutations that disrupt cooperative DNA binding by the Drosophila bicoid protein. J Mol Biol. 305: 219–230.

Burz, D. S., R. Rivera-Pomar, H. Jäckle, and S. D. Hanes. **1998**. Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. EMBO J. 17: 5998–6009.

Bussemaker, H. J., H. Li, and E. D. Siggia. **2000**. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc. Natl. Acad. Sci. U. S. A. 97: 10096–100.

Campos-Ortega, J.-A., and V. Hartenstein. **1985**. The Embryonic Development of Drosophila melanogaster. Springer, Heidelberg, Germany.

Capovilla, M., E. D. Eldon, and V. Pirrotta. **1992**. The giant gene of Drosophila encodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes. Development. 114: 99–112.

Carey, M. 1998. The Enhanceosome and Transcriptional Synergy. 92: 5–8.

Chen, H., Z. Xu, C. Mei, D. Yu, and S. Small. **2012**. A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes. Cell. 149: 618–29.

Courey, a J., and S. Jia. **2001**. Transcriptional repression: the long and the short of it. Genes Dev. 15: 2786–96.

Couronne, O., A. Poliakov, N. Bray, T. Ishkhanov, D. Ryaboy, E. Rubin, L. Pachter, and I. Dubchak. **2003**. Strategies and tools for whole-genome alignments. Genome Res. 13: 73–80.

Crombach, A., M. a García-Solache, and J. Jaeger. **2014**. Evolution of early development in dipterans: Reverse-engineering the gap gene network in the moth midge Clogmia albipunctata (Psychodidae). Biosystems.

Crombach, A., K. R. Wotton, D. Cicin-Sain, M. Ashyraliyev, and J. Jaeger. **2012**. Efficient reverse-engineering of a developmental gene regulatory network. PLoS Comput. Biol. 8: e1002589.

Dunipace, L., A. Ozdemir, and A. Stathopoulos. **2011**. Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Development. 138: 4075–84.

Eldon, E. D., and V. Pirrotta. **1991**. Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. Development. 111: 367–378.

Erceg, J., T. E. Saunders, C. Girardot, D. P. Devos, L. Hufnagel, and E. E. M. Furlong. **2014**. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. PLoS Genet. 10: e1004060.

Fakhouri, W. D., A. Ay, R. Sayal, J. Dresch, E. Dayringer, and D. N. Arnosti. **2010**. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. Mol. Syst. Biol. 6: 341.

Fish, M. P., A. C. Groth, M. P. Calos, and R. Nusse. **2007**. Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. Nat Protoc. 2: 2325–2331.

Foe, V. E. 1989. Mitotic domains reveal early commitment of cells in Drosophila embryos. Development. 107: 1–22.

Foe, V. E., and B. M. Alberts. **1983**. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. J Cell Sci. 70: 31–70.

Fuchs, A., L. S. Cheung, E. Charbonnier, S. Y. Shvartsman, and G. Pyrowolakis. **2012**. Transcriptional interpretation of the EGF receptor signaling gradient. Proc. Natl. Acad. Sci. U. S. A. 109: 1572–7.

Galas, D. J., and A. Schmitz. **1978**. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 5 : 3157–3170.

Gergen, J. P., and E. F. Wieschaus. **1986**. Localized requirements for gene activity in segmentation of Drosophila embryos: analysis of armadillo, fused, giant and unpaired mutations in mosaic embryos. Roux's Arch. Dev. Biol. 195: 49–62.

Ghavi-Helm, Y., F. a. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, and E. E. M. Furlong. **2014**. Enhancer loops appear stable during development and are associated with paused polymerase. Nature.

Gilbert, S. F. 2000. Developmental Biology, 6th ed. Sinauer Associates, Sunderland (MA).

Goto, T., P. Macdonald, and T. Maniatis. **1989**. Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell. 57: 413–22.

Gray, S., and M. Levine. **1996**. Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. Genes Dev. 10: 700–710.

Gray, S., P. Szymanski, and M. Levine. **1994**. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes Dev. 8: 1829–1838.

Green, M. R. 2005. Eukaryotic transcription activation: right on target. Mol. Cell. 18: 399–402.

Groth, A. C., M. Fish, R. Nusse, and M. P. Calos. **2004**. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. Genetics. 166: 1775–1782.

Haecker, A., M. Bergman, C. Neupert, B. Moussian, S. Luschnig, M. Aebi, and M. Mannervik. **2008**. Wollknäuel is required for embryo patterning and encodes the Drosophila ALG5 UDP-glucose : dolichyl-phosphate glucosyltransferase. Development. 135: 1745–1749.

Haecker, A., D. Qi, T. Lilja, B. Moussian, L. P. Andrioli, S. Luschnig, and M. Mannervik. **2007**. Drosophila Brakeless Interacts with Atrophin and Is Required for Tailless-Mediated Transcriptional Repression in Early Embryos. PLoS Biol. 5.

Han, K., M. S. Levine, and J. L. Manley. **1989**. Synergistic Activation and Repression of Transcription by Drosophila Homeobox Proteins. Cell. 56: 573–583.

Harding, K., T. Hoey, R. Warrior, and M. Levine. **1989**. Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. EMBO J. 8: 1205–1212.

Harrison, M. M., X.-Y. Li, T. Kaplan, M. R. Botchan, and M. B. Eisen. **2011**. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. 7: e1002266.

He, X., M. A. H. Samee, C. Blatti, and S. Sinha. **2010**. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol. 6.

Hertz, G. Z., and G. D. Stormo. **1999**. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15: 563–77.

Hong, J.-W., D. A. Hendrix, and M. S. Levine. **2008**. Shadow enhancers as a source of evolutionary novelty. Science. 321: 1314.

Hughes, S. C., and H. M. Krause. **1998**. Single and double FISH protocols for Drosophila, pp. 93–101. *In* Paddock, S.W. (ed.), Methods Mol Biol. Humana Press.

Irish, V., R. Lehmann, and M. Akam. **1989**. The Drosophila posterior-group gene nanos functions by repressing hunchback activity. Nature. 338: 646–648.

Jaeckle, H., D. Tautz, R. Schuh, E. Seifert, and R. Lehmann. **1986**. Cross-regulatory interactions among the gap genes of Drosophila. Nature. 324: 668–670.

Jaeger, J. 2011. The gap gene network. Cell. Mol. Life Sci.

Jaeger, J., and N. A. M. Monk. **2010**. Reverse engineering of gene regulatory networks, pp. 9–34. *In* Lawrence, N.D. (ed.), Learn. Inference Comput. Syst. Biol. MIT Press, Cambridge, Massachusetts, USA.

Jaeger, J., D. H. Sharp, and J. Reinitz. **2007**. Known maternal gradients are not sufficient for the establishment of gap domains in Drosophila melanogaster. Mech Dev. 124: 108–128.

Jaeger, J., S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. **2004**. Dynamic control of positional information in the early Drosophila embryo. Nature. 430: 368–371.

Janssens, H., A. Crombach, K. Richard Wotton, D. Cicin-Sain, S. Surkova, C. Lu Lim, M. Samsonova, M. Akam, and J. Jaeger. **2013**. Lack of tailless leads to an increase in expression variability in Drosophila embryos. Dev. Biol. 377: 305–17.

Janssens, H., A. Crombach, K. R. Wotton, D. Cicin-Sain, S. Surkova, C. L. Lim, M. Samsonova, M. Akam, and J. Jaeger. **2013**. Lack of tailless leads to an increase in expression variability in Drosophila embryos. Dev. Biol. 377: 305–17.

Janssens, H., S. Hou, J. Jaeger, A.-R. Kim, E. Myasnikova, D. Sharp, and J. Reinitz. **2006**. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat Genet. 38: 1159–1165.

Jiang, J., T. Hoey, and M. Levine. **1991**. Autoregulation of a segmentation gene in Drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element. Genes Dev. 5: 265–277.

Jiménez, G., a Guichet, a Ephrussi, and J. Casanova. **2000**. Relief of gene repression by torso RTK signaling: role of capicua in Drosophila terminal and dorsoventral patterning. Genes Dev. 14: 224–31.

Johnston, S. D., and C. Nüsslein-Volhard. **1992**. The origin of pattern and polarity in the Drosophila embryo. Cell. 68: 201–219.

Junion, G., M. Spivakov, C. Girardot, M. Braun, E. H. Gustafson, E. Birney, and E. E. M. Furlong. **2012**. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell. 148: 473–86.

Jürgens, G., E. Wieschaus, C. Nüsslein-Volhard, and H. Kluding. **1984**. Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster. II. Zygotic loci on the third chromosome. Roux's Arch. Dev. Biol. 193: 283–295.

Juven-Gershon, T., J.-Y. Hsu, and J. T. Kadonaga. **2008**. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. Genes Dev. 22: 2823–2830.

Kaufman, T. C. ., S. E. Tasaka, and D. T. Suzuki. **1973**. The interaction of two complex loci, zeste and bithorax in Drosophila melanogaster. Genetics. 57: 299–321.

Kazemian, M., C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky, and S. Sinha. **2010**. Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. PLoS Biol. 8.

Kazemian, M., H. Pham, S. a Wolfe, M. H. Brodsky, and S. Sinha. **2013**. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res. 41: 8237–52.

Kim, A.-R., C. Martinez, J. Ionides, A. F. Ramos, M. Z. Ludwig, N. Ogawa, D. H. Sharp, and J. Reinitz. **2013**. Rearrangements of 2.5 kilobases of noncoding DNA from the Drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. PLoS Genet. 9: e1003243.

Kosman, D., S. Small, and J. Reinitz. **1998**. Rapid preparation of a panel of polyclonal antibodies to Drosophila segmentation proteins. Dev. Genes Evol. 208: 290–4.

Kozlov, K. N., E. Myasnikova, A. A. Samsonova, S. Surkova, J. Reinitz, and M. Samsonova. **2009**. GCPReg package for registration of the segmentation gene expression data in Drosophila. Fly. 3: 151–156.

Kozlov, K., S. Surkova, E. Myasnikova, J. Reinitz, and M. Samsonova. **2012**. Modeling of gap gene expression in Drosophila Kruppel mutants. PLoS Comput. Biol. 8: e1002635.

Kraut, R., and M. Levine. **1991a**. Mutually repressive interactions between the gap genes giant and Kruppel define middle body regions of the Drosophila embryo. Development. 111: 611–621.

Kraut, R., and M. Levine. **1991b**. Spatial regulation of the gap gene giant during Drosophila development. Development. 111: 601–9.

Kvon, E. Z., T. Kazmar, G. Stampfel, J. O. Yáñez-Cuna, M. Pagani, K. Schernhuber, B. J. Dickson, and A. Stark. **2014**. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature.

Kvon, E. Z., G. Stampfel, J. O. Yáñez-Cuna, B. J. Dickson, and A. Stark. **2012**. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev. 26: 908–13.

Labalette, C., Y. X. Bouchoucha, M. A. Wassef, P. A. Gongal, J. Le Men, T. Becker, P. Gilardi-Hebenstreit, and P. Charnay. **2011**. Hindbrain patterning requires fine-tuning of early krox20 transcription by Sprouty 4. Development. 138: 317–26.

Lagha, M., J. P. Bothma, and M. Levine. **2012**. Mechanisms of transcriptional precision in animal development. Trends Genet. 28: 409–16.

Lam, J., and J.-M. Delosme. **1988a**. An Efficient Simulated Annealing Schedule: Derivation. Yale Electrical Engineering Department.

Lam, J., and J.-M. Delosme. **1988b**. An Efficient Simulated Annealing Schedule: Implement. Eval. Yale Electrical Engineering Department.

Latchman, D. S. 1997. Transcription Factors: an overview. Int. J. Biochem. Cell Biol. 29: 1305–1312.

Lawrence, P. A. 1992. The making of a Fly. Blackwell Scientific Publications, Oxford, U.K.

Lebrecht, D., M. Foehr, E. Smith, F. J. P. Lopes, C. E. Vanario-Alonso, J. Reinitz, D. S. Burz, and S. D. Hanes. **2005**. Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. Proc Natl Acad Sci U S A. 102: 13176–13181.

Lemon, B., and R. Tijian. **2000**. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14: 2551–2569.

Li, L. M., and D. N. Arnosti. **2011**. Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. Curr. Biol. 21: 406–12.

Lifanov, A. P., V. J. Makeev, A. G. Nazina, and D. A. Papatsenko. **2003**. Homotypic regulatory clusters in Drosophila. Genome Res. 13: 579–588.

Little, S. C., G. Tkačik, T. B. Kneeland, E. F. Wieschaus, and T. Gregor. **2011**. The Formation of the Bicoid Morphogen Gradient Requires Protein Movement from Anteriorly Localized mRNA. PLoS Biol. 9: 17.

Löhr, U., H.-R. Chung, M. Beller, and H. Jäckle. **2009**. Antagonistic action of Bicoid and the repressor Capicua determines the spatial limits of Drosophila head gene expression domains. Proc. Natl. Acad. Sci. U. S. A. 106: 21695–700.

Ma, X., D. Yuan, K. Diepold, T. Scarborough, and J. Ma. **1996**. The Drosophila morphogenetic protein Bicoid binds DNA cooperatively. Development. 122: 1195–206.

Mannervik, M., Y. Nibu, H. Zhang, and M. Levine. **1999**. Transcriptional Coregulators in Development. Science. 284: 606–609.

Markstein, M., C. Pitsouli, C. Villalta, S. E. Celniker, and N. Perrimon. **2008**. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat Genet. 40: 476–483.

Martinez, C., J. S. Rest, A.-R. Kim, M. Ludwig, M. Kreitman, K. White, and J. Reinitz. **2014**. Ancestral Resurrection of the Drosophila S2E Enhancer Reveals Accessible Evolutionary Paths through Compensatory Change. Mol. Biol. Evol. 1–14.

Meng, X., M. H. Brodsky, and S. a Wolfe. **2005**. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat. Biotechnol. 23: 988–94.

Mlodzik, M., and W. J. Gehring. **1987**. Expression of the caudal Gene in the Germ Line of Drosophila : Formation of an RNA and Protein Gradient during Early Embryogenesis. Cell. 48: 465–478.

Mohler, J., E. D. Eldon, and V. Pirrotta. **1989**. A novel spatial transcription pattern associated with the segmentation gene, giant, of Drosophila. EMBO J. 8: 1539–1548.

Myasnikova, E., A. Samsonova, K. Kozlov, M. Samsonova, and J. Reinitz. **2001**. Registration of the expression patterns of Drosophila segmentation genes by two independent methods. Bioinformatics. 17: 3–12.

Myasnikova, E., M. Samsonova, D. Kosman, and J. Reinitz. **2005**. Removal of background signal from in situ data on the expression of segmentation genes in Drosophila. Dev. Genes Evol. 215: 320–6.

Näär, A. M., B. D. Lemon, and R. Tjian. **2001**. Transcriptional coactivator complexes. Annu. Rev. Biochem. 70: 475–501.

Narachi, M. A., and J. B. Boyd. **1985**. The giant (gt) mutants of Drosophila melanogaster alter DNA metabolism. Mol Gen Genet. 199: 500–506.

Ni, J.-Q., L.-P. Liu, R. Binari, R. Hardy, H.-S. Shim, A. Cavallaro, M. Booker, B. D. Pfeiffer, M. Markstein, H. Wang, C. Villalta, T. R. Laverty, L. a Perkins, and N. Perrimon. **2009**. A Drosophila resource of transgenic RNAi lines for neurogenetics. Genetics. 182: 1089–100.

Ni, J.-Q., R. Zhou, B. Czech, L.-P. Liu, L. Holderbaum, D. Yang-Zhou, H.-S. Shim, R. Tao, D. Handler, P. Karpowicz, R. Binari, M. Booker, J. Brennecke, L. a Perkins, G. J. Hannon, and N. Perrimon. **2011**. A genome-scale shRNA resource for transgenic RNAi in Drosophila. Nat. Methods. 8: 405–7.

Nibu, Y., and M. S. Levine. **2001**. CtBP-dependent activities of the short-range Giant repressor in the Drosophila embryo. Proc Natl Acad Sci U S A. 98: 6204–6208.

Noyes, M. B., X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky, and S. A. Wolfe. **2008**. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res. 36: 2547–2560.

Nüsslein-Volhard, C., H. G. Frohnhöfer, and R. Lehmann. **1987**. Determination of anteroposterior polarity in Drosophila. Science. 238: 1675–81.

Nüsslein-Volhard, C., and E. Wieschaus. **1980**. Mutations affecting segment number and polarity in Drosophila. Nature. 287: 795–801.

Nüsslein-Volhard, C., E. Wieschaus, and H. Kluding. **1984**. Mutations affecting the pattern of the larval cuticle inDrosophila melanogaster. I. Zygotic loci on the second chromosome. Roux's Arch. Dev. Biol. 193: 267–282.

Oberstein, A., A. Pare, L. Kaplan, and S. Small. **2005**. Site-specific transgenesis by Cre-mediated recombination in Drosophila. Nat Methods. 2: 583–585.

Ochoa-Espinosa, A., G. Yucel, L. Kaplan, A. Pare, N. Pura, A. Oberstein, D. Papatsenko, and S. Small. **2005**. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A. 102: 4960–4965.

Oliphant, A. R., C. J. Brandl, and K. Struhl. **1989**. Defining the Sequence Specificity of DNA-Binding Proteins by Selecting Binding Sites from Random-Sequence Oligonucleotides: Analysis of Yeast GCN4 Protein. Mol. Cell. Biol. 9: 2944–2949.

Papatsenko, D., Y. Goltsev, and M. Levine. **2009**. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res. 37: 5665–5677.

Park, J. M., B. S. Gim, J. M. Kim, J. Ho, H. Kim, J. Kang, Y. Kim, J. I. N. M. O. Park, B. S. O. O. Gim, J. M. O. Kim, and J. H. O. Yoon. **2001**. Drosophila Mediator Complex Is Broadly Utilized by Diverse Gene-Specific Transcription Factors at Different Types of Core Promoters. Mol. Cell. Biol. 21.

Parkhurst, S. M. 1998. Groucho: making its Marx as a transcriptional co-repressor. Trends Genet. 14: 130–132.

Payankaulam, S., and D. N. Arnosti. **2009**. Groucho corepressor functions as a cofactor for the Knirps short-range transcriptional repressor. Proc. Natl. Acad. Sci. U. S. A. 106: 17314–9.

Perry, M. W., A. N. Boettiger, J. P. Bothma, and M. Levine. **2010**. Shadow enhancers foster robustness of Drosophila gastrulation. Curr Biol. 20: 1562–1567.

Perry, M. W., A. N. Boettiger, and M. Levine. **2011**. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc Natl Acad Sci U S A. 108: 13570–13575.

Perry, M. W., J. P. Bothma, R. D. Luu, and M. Levine. **2012**. Precision of hunchback expression in the Drosophila embryo. Curr. Biol. 22: 2247–52.

Petschek, J. P., N. Perrimon, and A. P. Mahowald. **1987**. Region-Specific Defects in l(1)giant Embryos of Drosophila. Dev. Biol. 119: 175–189.

Pfeiffer, B. D., A. Jenett, A. S. Hammonds, T.-T. B. Ngo, S. Misra, C. Murphy, A. Scully, J. W. Carlson, K. H. Wan, T. R. Laverty, C. Mungall, R. Svirskas, J. T. Kadonaga, C. Q. Doe, M. B. Eisen, S. E. Celniker, and G. M. Rubin. **2008**. Tools for neuroanatomy and neurogenetics in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 105: 9715–20.

Pisarev, A., E. Poustelnikova, M. Samsonova, and J. Reinitz. **2009**. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. Nucleic Acids Res. 37: D560–D566.

Poustelnikova, E., A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz. **2004**. A database for management of gene expression data in situ. Bioinformatics. 20: 2212–21.

Raj, A., P. Van Den Bogaard, S. A. Rifkin, A. Van Oudenaarden, and S. Tyagi. **2008**. Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods. 5: 877–879.

Rajewsky, N., M. Vergassola, U. Gaul, and E. D. Siggia. **2002**. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics. 3: 30.

Raveh-Sadka, T., M. Levo, and E. Segal. **2009**. Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res. 19: 1480–96.

Reinitz, J., S. Hou, and D. H. Sharp. **2003**. Transcriptional Control in Drosophila. Complexus. 1: 54–64.

Reinitz, J., and M. Levine. **1990**. Control of the initiation of homeotic gene expression by the gap genes giant and tailless in Drosophila. Dev. Biol. 140: 57–72.

Rembold, M., L. Ciglar, J. O. Yáñez-Cuna, R. P. Zinzen, C. Girardot, A. Jain, M. a Welte, A. Stark, M. Leptin, and E. E. M. Furlong. **2014**. A conserved role for Snail as a potentiator of active transcription. Genes Dev. 28: 167–81.

Rivera-Pomar, R., X. Lu, N. Perrimon, H. Taubert, and H. Jaeckle. **1995**. Activation of posterior gap gene expression in the Drosophila blastoderm. Nature. 376: 253–256.

Samee, M. A. H., and S. Sinha. **2014**. Quantitative modeling of a gene's expression from its intergenic sequence. PLoS Comput. Biol. 10.

Sanders, K. 1976. Specification of the Basic Body Pattern in Insect Embryogenesis, pp. 126–228. *In* Adv. In Insect Phys. Academic Press, Freiburg, Germany.

Saurin, a J., Z. Shao, H. Erdjument-Bromage, P. Tempst, and R. E. Kingston. **2001**. A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. Nature. 412: 655–60.

Sayal, R., S.-M. Ryu, and D. N. Arnosti. **2010**. Optimization of reporter gene architecture for quantitative measurements of gene expression in the Drosophila embryo. Fly (Austin). 5: 47–52.

Schroeder, M. D., C. Greer, and U. Gaul. **2011**. How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation. Development. 138: 3067–78.

Schroeder, M. D., M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia, and U. Gaul. **2004**. Transcriptional Control in the Segmentation Gene Network of Drosophila. PLoS Biol. 2: 1396–1410.

Schulz, C., and D. Tautz. **1994**. Autonomous concentration-dependent activation and repression of Krüppel by hunchback in the Drosophila embryo. Development. 120: 3043–9.

Schüpbach, T., and E. Wieschaus. **1986**. Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo. Roux's Arch. Dev. Biol. 195: 302–317.

Schwartz, M. B., R. B. Imberski, and T. J. Kelly. **1984**. Analysis of Metamorphosis in Drosophila melanogaster: Characterization of giant, an Ecdysteroid-Deficient Mutant '. Dev. Biol. 103: 85–95.

Segal, E., T. Raveh-sadka, M. Schroeder, U. Unnerstall, and U. Gaul. **2008**. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 451: 535–540.

Simpson-Brose, M., J. Treisman, and C. Desplan. **1994**. Synergy between the hunchback and bicoid morphogens is required for anterior patterning in Drosophila. Cell. 78: 855–65.

Sinha, S., Y. Liang, and E. Siggia. **2006**. Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Res. 34: 555–559.

Small, S., a Blair, and M. Levine. **1992**. Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J. 11: 4047–57.

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: 314–24.

Small, S., R. Kraut, T. Hoey, R. Warrior, and M. Levine. **1991**. Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. 5: 827–839.

Spitz, F., and E. E. M. Furlong. **2012**. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13: 613–26.

Staller, M. V, D. Yan, S. Randklev, M. D. Bragdon, Z. B. Wunderlich, R. Tao, L. a Perkins, A. H. Depace, and N. Perrimon. **2013**. Depleting gene activities in early Drosophila embryos with the "maternal-Gal4-shRNA" system. Genetics. 193: 51–61.

Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht. **1982**. Use of the "Perceptron" algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. 10 : 2997–3011.

Struhl, G., P. Johnston, and P. a Lawrence. **1992**. Control of Drosophila body pattern by the hunchback morphogen gradient. Cell. 69: 237–249.

Strunk, B., P. Struffi, K. Wright, B. Pabst, J. Thomas, L. Qin, and D. N. Arnosti. **2001**. Role of CtBP in Transcriptional Repression by the Drosophila giant Protein. Dev. Biol. 239: 229–240.

Surkova, S., E. Golubkova, Manu, L. Panok, L. Mamon, J. Reinitz, and M. Samsonova. **2013**. Quantitative dynamics and increased variability of segmentation gene expression in the Drosophila Krüppel and knirps mutants. Dev. Biol. 376: 99–112.

Surkova, S., D. Kosman, K. Kozlov, Manu, E. Myasnikova, A. A. Samsonova, A. Spirov, C. E. Vanario-Alonso, M. Samsonova, and J. Reinitz. **2008**. Characterization of the Drosophila segment determination morphome. Dev Biol. 313: 844–862.

Surkova, S., E. Myasnikova, H. Janssens, K. Kozlov, A. Samsonova, J. Reinitz, and M. Samsonova. **2008**. Pipeline for acquisition of quantitative data on segmentation gene expression from confocal images. Fly. 2.

Tamkun, J. W., R. Deuring, M. P. Scott, M. Kissinger, a M. Pattatucci, T. C. Kaufman, and J. a Kennison. **1992**. brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. Cell. 68: 561–72.

Tautz, D. 1988. Regulation of the Drosophila segmentation gene hunchback by two maternal morphogenetic centres. Nature. 332: 281–284.

Tautz, D. 2000. Evolution of transcriptional regulation. Genet. Dev. 10: 575–579.

Tautz, D., R. Lehmann, H. Schnurch, R. Schuh, E. Seifert, A. Kienlin, K. Jones, and H. Jackle. **1987**. Finger protein of novel structure encoded by hunchback, a second member of the gap class of Drosophila segmentation genes. Nature. 327: 383–389.

Tomancak, P., A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. **2002**. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. 3.

Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin. **2007**. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol. 8: R145.

Trcek, T., J. a Chao, D. R. Larson, H. Y. Park, D. Zenklusen, S. M. Shenoy, and R. H. Singer. **2012**. Single-mRNA counting using fluorescent in situ hybridization in budding yeast. Nat. Protoc. 7: 408–19.

Tuerk, C., and L. Gold. **1990**. Systematic Evolution of Ligands by Exponential Enrichment : RNA Ligands to Bacteriophage T4 DNA Polymerase. Science. 249: 505–510.

Venken, K. J. T., J. W. Carlson, K. L. Schulze, H. Pan, Y. He, R. Spokony, K. H. Wan, M. Koriabine, P. J. de Jong, K. P. White, H. J. Bellen, and R. A. Hoskins. **2009**. Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. Nat Methods. 6: 431–434.

Vinson, C., P. Sigler, and S. McKnight. **1989**. Scissors-grip model for DNA recognition by a family of leucine zipper proteins. Science. 246: 911–916.

Wasson, T., and A. J. Hartemink. **2009**. An ensemble model of competitive multi-factor binding of the genome. Genome Res. 19: 2101–2112.

Wieschaus, E., C. Nüsslein-Volhard, and G. Jürgens. **1984**. Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster. III. Zygotic loci on the X-chromosome and fourth chromosome. Wilhelm Roux's Arch. Dev. Biol. 193: 296–307.

Wilczynski, B., Y.-H. Liu, Z. X. Yeo, and E. E. M. Furlong. **2012**. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. PLoS Comput. Biol. 8.

Wu, X., V. Vasisht, D. Kosman, J. Reinitz, and S. Small. **2001**. Thoracic Patterning by the Drosophila Gap Gene hunchback. Dev. Biol. 237: 79–92.

Yáñez-Cuna, J. O., E. Z. Kvon, and A. Stark. **2013**. Deciphering the transcriptional cis-regulatory code. Trends Genet. 29: 11–22.

Yu, D., and S. Small. **2008**. Precise registration of gene expression boundaries by a repressive morphogen in Drosophila. Curr. Biol. 18: 868–76.

Zinzen, R. P., C. Girardot, J. Gagneur, M. Braun, and E. E. M. Furlong. **2009**. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature. 462: 65–70.

Zinzen, R. P., K. Senger, M. Levine, and D. Papatsenko. **2006**. Computational models for neurogenic gene expression in the Drosophila embryo. Curr. Biol. 16: 1358–65.