

The Evolution of Bicoid Regulated Genes in Insects

Thesis submitted for the degree of Doctor of Philosophy at
the University of Leicester



by
Alistair P. McGregor (B.Sc.)
Department of Genetics
University of Leicester

January 2002

UMI Number: U149252

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U149252

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

The Evolution of Bicoid Regulated Genes in Insects

Alistair P. McGregor B.Sc.

Department of Genetics

University of Leicester

A network of interactions between transcription factors and *cis*-regulatory sequences controls the expression of developmental genes. Changes in either the *cis*- or *trans*-acting components of a developmental interaction can have consequences for both the output of the interaction and for the greater network of interactions. Thus, the evolution of regulation is considered to be a major force in the evolution of morphological diversity. To investigate the evolution of an interaction, this thesis has compared the Bicoid-dependent regulation of *hunchback* and *orthodenticle* expression in the Dipterans, *Drosophila melanogaster*, *Musca domestica*, *Calliphora vicina*, *Lucilia sericata* and *Megaselia abdita*.

hb genes were isolated from *Calliphora* and *Lucilia* and these encode a number of domains that are conserved in *hb* from other species such as *Drosophila*, *Musca* and *Tribolium*. In contrast to the coding sequences, the *hb* promoters from *Calliphora* and *Lucilia* differ from each other and from the *Drosophila* and *Musca* *hb* promoters in terms of the number, sequence, orientation and spacing of Bcd-binding sites that they contain.

Analysis of intra-specific variation in the *M. domestica* *hb* gene demonstrated that both coding and non-coding sequences are subject to slippage generated turnover of simple motifs and that the extent of this turnover is dependent upon region specific constraints. This suggests that mechanisms of turnover are responsible for the different *hb* promoter configurations observed in the Dipterans.

To investigate any functional consequences of the differences in both *bcd* and *hb* between *Drosophila*, *Musca* and *Megaselia*, transcription assays were carried out using homogeneous and heterogeneous combinations of these two components in yeast. The results of these assays suggest that differences in Bcd and the *hb* promoters between these species may have co-evolved to maintain the interaction. Therefore, to investigate how other Bcd-regulated genes have evolved in *Musca*, the expression patterns and coding sequences of the *otd* gene were characterized in *Musca* and compared to those of *Drosophila otd*.

List of Abbreviations

A	adenine
BSA	bovine serum albumin
C	cytosine
cDNA	complementary DNA
DEPC	diethylpyrocarbonate
DNaseI	deoxyribonuclease I
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
G	guanine
GST	glutathione-S-transferase
HA	hemagglutinin
HTH	helix-turn-helix
MY	million years
MYA	million years ago
NMR	nuclear magnetic resonance
OD	optical density
ONPG	o-nitrophenyl β -D-galactopyranoside
PAGE	polyacrylamide gel electrophoresis
PNK	polynucleotide kinase
PNTM	Protran [®] nitrocellulose transfer membrane
RNAi	RNA interference
SSC	sodium saline citrate
T	thymine
TAE	Tris-acetate EDTA
TE	Tris EDTA
TBE	Tris-borate EDTA
UTR	untranslated region

Acknowledgements

I wish to thank the following people for their help over the last three years or so. Without some of them I wouldn't be finished yet and without others I would have been finished months ago, with more money, fewer hangovers but a lot less laughter.

My supervisor Gabby for making me think about evolution and for supporting my experiments, oh and for money when my studentship finished! Phil Shaw for advice with experiments, constant encouragement and the importance of the pmol.

John Hancock for the simplicity stuff. Steven Hanes for the yeast plasmids and Michael Stauber for the *Megaselia* plasmid. All the people who kindly donated smelly flies and Joachim Reischl for giving me the REAL *Drosophila otd* sequence.

Everyone else who has passed through the genetics department over the last few years: Barry (Dr Star), Audrey, Ian, Jon, Zoë, Elena, Turi, Mark, Tony, Matt, Colin, Seth, Neale F, Neil B, Mike, Alistair, Ray, Claire, Stu, Rich, Ruth, Andy, Ben, Terry, Sam, Ezio, Bambos, Jenny, Carol, John S, Hilda, Tim, Simon, Marcus (scissors kick). I would especially like to thank Fred for lots of encouraging scientific chats and constant sarcasm and Raymond for being there when Celtic won the treble. Pat made the lab a great place to work and taught me how to cheat at squash. I would also like to thank Knoxy, Keith, Matt, Gavin, Amin, Bill and Jit for lots of drinks and laughs.

I would like to thank my parents for their constant emotional and financial support and as I did this for them they had better read it. My Gran for scones, eggs, wine bottles, tenners and vinegar bottles. Finally I would like to thank Naomi for just being Naomi.

Communications

Some of the work carried out for this thesis was published in the following articles:

McGregor, A. P., Shaw, P. J. and Dover, G. A. (2001). Sequence and expression of the *hunchback* gene in *Lucilia sericata*: a comparison with other Dipterans. *Dev Genes Evol* **211**, 315-318.

McGregor, A. P., Shaw, P. J., Hancock, J. M., Bopp, D., Hediger, M., Wratten, N. S. and Dover, G. A. (2001). Rapid restructuring of bicoid-dependent *hunchback* promoters within and between Dipteran species: implications for molecular co-evolution. *Evol Dev* **3**, 397-407.

Shaw, P. J., Salameh, A., McGregor, A. P., Bala, S. and Dover, G. A. (2001). Divergent structure and function of the *bicoid* gene in Muscoidea fly species. *Evol Dev* **3**, 251-262.

Shaw, P. J., Wratten, N. S., McGregor, A. P. and Dover, G. A. Co-evolution in *bicoid*-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. *Evol Dev* (submitted).

Contents

Chapter 1 General introduction

1.1 The Evolution of Development	1
1.2 The evolution of conserved genes: duplication, co-option and heterochrony	1
1.3 Modular <i>cis</i> -regulatory sequences	3
1.4 The consequences of <i>cis</i> -regulatory change	5
1.5 How can we study the evolution of an interaction?	6
1.6 The Dipterans and the emergence of <i>bicoid</i>	6
1.7 Binding of the Bcd homeodomain to DNA	8
1.8 The role of <i>bcd</i>	9
1.9 The evolution of <i>bcd</i>	10
1.10 The expression and role of <i>hb</i> in the <i>Drosophila</i> embryo	11
1.11 Bcd-dependent activation of <i>hb</i>	12
1.12 Evolution of the Bcd- <i>hb</i> interaction	14
1.13 <i>bcd</i> in other Dipterans	16
1.14 Aims of this thesis	17

Chapter 2 Materials and Methods

2.1 Materials	
2.1.1 Media	19
2.1.2 Organisms	19
2.1.3 Plasmids	21
2.1.4 Oligonucleotides	22
2.2 Methods	
2.2.1 Standard molecular biology techniques	
2.2.1.1 DNA precipitation and phenol-chloroform extraction	22
2.2.1.2 Restriction digests	22
2.2.1.3 Gel extraction	22
2.2.1.4 Ligation of DNA fragments	22
2.2.1.5 Transformation of <i>E. coli</i>	23
2.2.1.6 Preparation of plasmid DNA	23
2.2.1.7 Agarose gel electrophoresis	24
2.2.1.8 Denaturing polyacrylamide (sequencing) electrophoresis	24
2.2.1.9 Southern analysis	25
2.2.1.10 DNA sequencing	26
2.2.2 Extraction of genomic DNA	26

2.2.3 DNA amplification by the polymerase chain reaction	27
2.2.4 Construction of suppression-PCR libraries	27
2.2.5 mRNA extraction	28
2.2.6 5' and 3' Rapid Amplification of cDNA Ends (RACE) - PCR	28
2.2.7 DNaseI footprinting	
2.2.7.1 Primer end-labelling	28
2.2.7.2 PCR	29
2.2.7.3 Protein synthesis	29
2.2.7.4 Binding reaction and DNaseI digestion	29
2.2.8 Yeast techniques	
2.2.8.1 Transformation of yeast	30
2.2.8.2 β -galactosidase assays	30
2.2.8.3 Protein extraction from yeast cultures	31
2.2.8.4 Electrophoresis and Western analysis of protein samples	31
2.2.9 <i>In situ</i> hybridisations of whole mount embryos	32
2.2.9.1 <i>In vitro</i> transcription for synthesis of riboprobes	32
2.2.9.2 Dechoriation	33
2.2.9.3 Fixation	33
2.2.9.4 Pre-treatment of embryos for <i>in situ</i> hybridisation	33
2.2.9.5 Prehybridisation and hybridisation	34
2.2.9.6 Pre-immunoreaction and immunoreaction	34
2.2.9.7 Colour staining	35
2.2.9.8 Permanent mounting, microscopy and photography	35
2.2.10 Computer analysis	35

Chapter 3 Characterisation of *hb* genes in *Calliphora*, *Lucilia* and *Musca*

3.1 Introduction

3.1.1 Aims	37
3.1.2 Experimental overview	37
3.1.3 suppression PCR (sPCR)	37

3.2 Results

3.2.1 Cloning of <i>hb</i> from <i>Calliphora</i>	38
3.2.2 Cloning of <i>hb</i> from <i>Lucilia</i>	39
3.2.3 Southern analysis of <i>Calliphora hb</i> and <i>Lucilia hb</i>	40
3.2.4 <i>hb</i> protein sequence comparison and analysis	42
3.2.5 Walking from the <i>hb</i> coding region into the 5' UTR in <i>Calliphora</i> and <i>Lucilia</i>	42
3.2.6 Mapping the 5' end of the <i>Lucilia hb</i> mRNA using RACE PCR	43

3.2.7 Cloning of the <i>Lucilia hb</i> putative P2 promoter	44
3.2.8 Cloning of the <i>Calliphora hb</i> putative P2 promoter	45
3.2.9 <i>hb</i> transcript structure in <i>Musca</i> , <i>Calliphora</i> and <i>Lucilia</i>	46
3.2.10 Characterisation of <i>hb</i> expression patterns in <i>Calliphora</i> and <i>Lucilia</i>	47
3.3 Discussion	
3.3.1 <i>Calliphora</i> and <i>Lucilia</i> both encode Hb proteins with conserved functional domains	48
3.3.2 Conserved and diverged expression patterns of <i>hb</i> in <i>Calliphora</i> and <i>Lucilia</i> suggest that some aspects of <i>hb</i> regulation in these species have changed	50

Chapter 4 Characterisation of the *hb* promoters in *Calliphora* and *Lucilia*

4.1 Introduction	
4.1.1 Characterisation of the putative <i>Calliphora</i> and <i>Lucilia hb</i> promoter regions	54
4.2 Results	
4.2.1 Expansion of the known <i>bcd</i> sequences in <i>Lucilia</i> and <i>Calliphora</i>	54
4.2.2 Characterisation of the Bcd-binding sites in the <i>Lucilia</i> and <i>Calliphora hb</i> P2 promoters using DNaseI footprinting	56
4.2.3 Analysis of the Bcd-protected regions in the <i>Calliphora</i> and <i>Lucilia hb</i> promoters	57
4.3 Discussion	
4.3.1 Evolution of the Bcd protein in the Dipterans	59
4.3.2 Bcd binding to consensus and non-consensus sites	60
4.3.3 Co-operative Bcd-binding	63
4.3.4 Spacing of Bcd-binding sites	64
4.3.5 Do other transcription factors bind to the <i>Calliphora</i> and <i>Lucilia hb</i> promoters	66

Chapter 5 Analysis of intra-specific and inter-specific variation in *hb*

5.1 Introduction	67
5.2 Methods	
5.2.1 The analysis of simplicity	69
5.2.2 Sequencing of three regions of <i>hb</i> from six strains of <i>Musca</i>	70
5.3 Results	
5.3.1 Intra-specific polymorphisms in <i>Musca domestica hb</i>	71
5.3.2 <i>Musca hb</i> sequence simplicity analysis	73

5.3.3 Analysis of the <i>Calliphora</i> and <i>Lucilia hb</i> promoter sequences	73
5.4 Discussion	
5.4.1 Patterns of <i>hb</i> polymorphisms within <i>Musca domestica</i>	74
5.4.2 Intra-specific and inter-specific variation in the <i>hb</i> promoter	76

Chapter 6 Functional analysis of the Bcd-*hb* interaction

6.1 Introduction	
6.1.1 Have <i>bcd</i> and the <i>hb</i> P2 promoters co-evolved in <i>Drosophila</i> and <i>Musca</i>	79
6.1.2 A yeast system to investigate Bcd-dependent transcription	79
6.2 Materials and Methods	
6.2.1 Construction of <i>bcd</i> expression vectors	81
6.2.2 Construction of <i>hb</i> promoter <i>lacZ</i> reporter vectors	82
6.3 Results	
6.3.1 Comparison of the <i>Drosophila</i> , <i>Musca</i> and <i>Megaselia</i> Bcd transcriptional activities in yeast	83
6.3.2 Comparison of the <i>Drosophila</i> and <i>Musca hb</i> P2 promoters transcriptional activities in yeast	84
6.3.3 Analysis of the <i>Calliphora</i> and <i>Lucilia hb</i> promoter regions transcriptional activities in yeast	86
6.3.4 Western analysis of <i>bcd</i> expression in yeast	87
6.4 Discussion	
6.4.1 Incompatibilities between components of the Bcd- <i>hb</i> interaction from <i>Drosophila</i> and <i>Musca</i>	90
6.4.2 <i>Megaselia</i> Bcd function?	91
6.4.3 Bcd-dependent transcription from the <i>Calliphora</i> and <i>Lucilia hb</i> promoters	92
6.4.4 Measuring promoter sensitivity in yeast	93

Chapter 7 Characterisation of *orthodenticle* in *Musca*

7.1 Introduction	94
7.2 Results and discussion	
7.2.1 Amplification of the <i>Musca otd</i> homeodomain using degenerate PCR	95
7.2.2 5' RACE PCR to map the <i>Musca otd</i> transcription start site	95
7.2.3 3' RACE PCR to amplify the 3' region of the <i>Musca otd</i> transcript	95
7.2.4 Southern analysis of <i>Musca otd</i>	96
7.2.5 <i>Musca otd</i> transcript and protein structure	96

7.2.6 Analysis of <i>otd</i> expression patterns in <i>Musca</i>	98
7.2.7 Conclusions and future work	99

Chapter 8 General discussion

8.1 Results summary	101
8.2 Implications of these data	
8.2.1 The evolution of <i>bcd</i> and Bcd-dependent regulation	103
8.2.2 Modelling the Bcd- <i>hb</i> interaction: predicting the consequences of change	105
8.2.3 Co-evolution of <i>bcd</i> and Bcd-dependent promoters?	106
8.2.4 The Bcd gradient and embryo size	107
8.3 Future work	108
References	110

Appendices

Chapter 1

General introduction

1.1 The Evolution of Development

Development can be considered as a complex set of genetic interactions and variations in such interactions ultimately generate the range of species-specific morphologies observed throughout the animal and plant kingdoms. Therefore, a paradox exists between the maintenance of genetic interactions, which will give rise to a particular species ontogeny and the evolution of such interactions to generate the diversity of species morphologies.

It has been proposed that a major mechanism employed in the evolution of genetic novelty is the repeated, differential utilisation of existing components rather than the invention of new ones (Wilson *et al.*, 1974; Jacob 1977). In the last decade it has become apparent that the evolution of development is driven in part by the redeployment of existing genes and genetic interactions rather than by the evolution of completely new genes and developmental programs (Palopoli and Patel 1996; Duboule and Wilkins 1998; Carroll *et al.*, 2001; Davidson 2001). Thus, studies of the evolution of the *cis*-regulatory regions of genes, in and between species, are the key to understanding how genetic interactions evolve and therefore, how development evolves.

1.2 The evolution of conserved genes: duplication, co-option and heterochrony

Despite the apparently conserved role of Hox genes in different animal lineages (for reviews see Carroll 1995; Raff 1996; Holland 1999), duplications of Hox clusters, or of individual Hox genes and the subsequent loss of some genes, have been common during the evolution of these genes (Carroll *et al.*, 2001). For example, the vertebrates have four Hox clusters resulting from a possible tetraploidization event in this lineage. It is thought that the duplication and divergence of Hox genes has allowed these genes to assume new functions (Akam 1989; Holland *et al.*, 1994). Duplication of genes represents a piece in the puzzle of conserved genes and diverse morphologies, but how does a duplicated gene assume a novel role in development?

The three *Drosophila* genes *paired*, *gooseberry* and *gooseberry neuro* are related genes encoding transcription factors, which are thought to have arisen by two duplication events. An elegant series of experiments has shown that the proteins encoded by these

three genes are functionally interchangeable and that their individual functions are determined by their *cis*-regulatory sequences (Li and Noll 1994; Xue and Noll 1996). Therefore, the duplication and divergence of *cis*-regulatory regions can allow the expression of genes in novel temporal and spatial patterns. Indeed, it has been suggested that the duplication of Hox gene clusters in vertebrates was the 'permissive' step, for morphological elaboration resulting from subsequent regulatory divergence (Holland *et al.*, 1994). It must be remembered that a given regulatory gene is part of a regulatory network and therefore when it is expressed in a novel domain the downstream interactions regulated by this gene may now also be expressed in this new domain. If this ectopic expression is selected, then a co-option of this part of the network can result. It is thought that such a co-option has resulted in the deployment of the Hox genes in vertebrate limbs (Shubin *et al.*, 1997).

While gene duplication and subsequent regulatory re-wiring may have played a role in the evolution of Hox genes in the vertebrates, interestingly, all four classes of arthropod (chelicerates, crustaceans, myriapods and insects) and the onychophora share the same suite of Hox genes despite 540 million years (MY) of divergence. However, these animals also display diversity in their segmentation patterns (Averof and Akam 1993; Carroll 2000). It has been hypothesised that subtle changes in the timing and magnitude of Hox gene expression are responsible for changes in segment morphology in arthropods (Akam 1998, 2000) and in vertebrates (Belting *et al.*, 1998). For example, differences in abdominal expression of *Ultrabithorax* (*Ubx*) between Dipterans and Lepidopterans have resulted in the larvae of the latter displaying prolegs (Carroll 1994). In vertebrates, while *Hoxc6* is expressed at the cervical-thoracic boundary in both mice and chickens, this transition in identity is positioned at different places along the anterior-posterior axis of these two species (Burke *et al.*, 1995). Furthermore, *Hoxc8* expression is driven further to the posterior in chickens than in mice, as a result of a heterochronic shift in the expression of this Hox gene. Importantly, this shift in expression between these species is caused by a few differences within transcription factor binding sites of the early *Hoxc8* enhancer (Belting *et al.*, 1998; Shashikant *et al.*, 1998).

The general feature which emerges from the examples described above, is that the re-wiring of genetic regulatory circuits, based on the interactions between transcription factors and the *cis*-regulatory regions of their target genes is a major force in developmental evolution (Arnone and Davidson 1997). What are the features of the *cis*-regulatory regions of genes that facilitate such developmental differences?

1.3 Modular *cis*-regulatory sequences

It is the modular structure of genetic *cis*-regulatory regions that allows gene expression in spatially and temporally independent domains. Each discrete module acts analogously to a logic switch in a circuit by controlling gene expression in response to the local concentrations of different transcription factors. Whether the expression of a given gene is switched on or off in a particular domain depends on the signature of transcription factor binding sites in the modules that control the expression of the gene. This can be tested *in vivo* by placing *cis*-regulatory modules upstream of a reporter gene and re-generating the natural expression pattern. For example, figure 1.1 shows part of the *cis*-regulatory regions of the *Drosophila* segmentation genes *knirps* (*kni*), *even-skipped* (*eve*) and *Ubx*. The *kni* module reads the local expression of maternal and gap gene encoded transcription factors to drive a band of *kni* expression in the posterior of the *Drosophila* embryo (figure 1.1A). This *kni* promoter illustrates the composition of a typical module; characteristically containing multiple transcription factor binding sites that either activate or repress *kni* expression (Rivera-Pomar *et al.*, 1995). The expression of *eve* stripe 2 is driven by an independent module from those modules that are responsible for generating the other stripes of *eve* expression (figure 1.1B, Small *et al.*, 1992, 1996). This is further illustrated by the structure of the *Ubx* BRE element (figure 1.1C), which in combination with the PBX element is responsible for *Ubx* expression in parasegments (PS) 6, 8, 10 and 12; whereas *Ubx* expression in PS 5, 7, 9, 11 and 13 is determined by the ABX module (Qian *et al.*, 1993). Indeed, modular promoters are a common feature of all genes that display distinct temporal and spatial expression patterns, from the Endo16 system in the sea urchin to the human β -globin locus (Arnone and Davidson 1997).

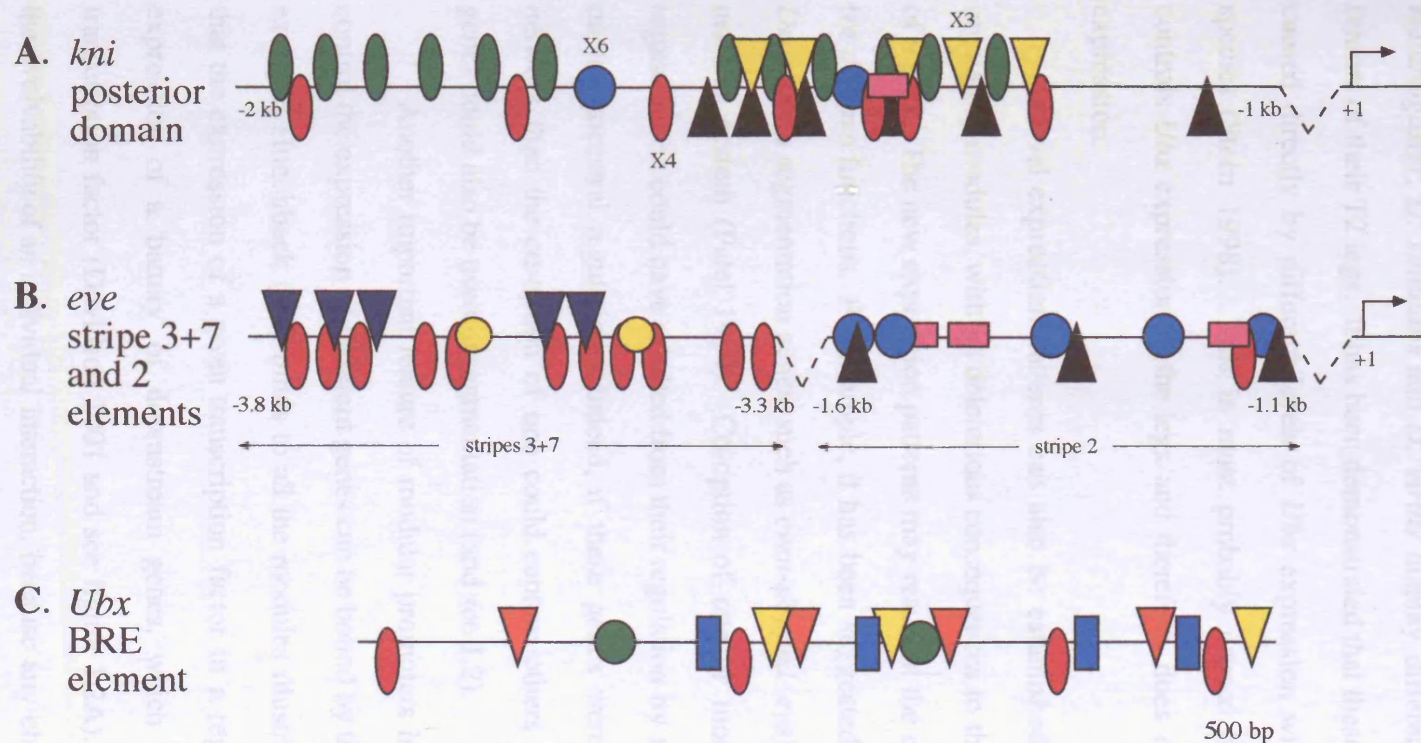


Figure 1.1 *Cis*-regulatory modules responsible for the expression of the *Drosophila* segmentation genes *kni* (A), *eve* (B) and *Ubx* (C). Each shape represents an individual binding site for the following transcription factors: Hunchback (red ovals), Bicoid (blue circles), Tailless (yellow triangles), Krüppel (black triangles), Caudal (green ovals), Giant (pink rectangles), Knirps (yellow circles), Twist (orange triangles), Engrailed (green circles), Fushi tarazu (blue rectangles) and D-Stat (blue triangles). The size of symbols is not representative of binding site sizes. Multiple binding sites are labelled with the number of sites present, for example X6. The scale in A and B is given in kb upstream from the transcription start site, which is represented by black arrows positioned at +1. In B, the two independent modules are indicated by the horizontal arrows and are labelled according to the stripes of *eve* expression they generate. Parts A and B adapted from Arnone and Davidson 1997, figure 1. The *eve* promoter was characterised by Small *et al.*, 1992, 1996 and the *kni* promoter by Rivera-Pomar *et al.*, 1995. The *Ubx* BRE element drives expression in parasegments 6, 8, 10 and 12 (Qian *et al.*, 1993; Zhang *et al.*, 1991).

Modular promoters mean that changes to transcription factor binding sites in one module will not affect the output of other modules and therefore, different domains of the expression of an individual gene can evolve independently. This can again be illustrated by differential *cis*-regulation of *Ubx* between closely related *Drosophila* species. *D. melanogaster*, *D. simulans* and *D. virilis* display different patterns of trichomes on the femurs of their T2 legs. It has been demonstrated that these differences in morphology are caused directly by different levels of *Ubx* expression within these appendages in each species (Stern 1998). This is most probably caused by changes in the module that controls *Ubx* expression in the legs and therefore, does not affect other aspects of *Ubx* expression.

Novel expression patterns can also be established under the control of new *cis*-regulatory modules, without deleterious consequences to the ancestral expression domains of a gene. The new expression patterns may result in the co-option of that particular gene for *de novo* functions. For example, it has been suggested that the ancestral role of many *Drosophila* segmentation genes, such as *even-skipped* (*eve*), was in the development of the nervous system (Patel 1994). Co-option of one or more of these genes for roles in segmentation could have resulted from their regulation by new modules without any affect on the ancestral regulation. Indeed, if these genes were already part of an integrated network then the co-option of one could capture others and therefore, eventually these genes could also be used in segmentation (and see 1.2).

Another important feature of modular promoters is that independent modules that control the expression of different genes can be bound by the same transcription factor; for example, Hunchback (Hb) binds to all the modules illustrated in figure 1.1. This means that the expression of a given transcription factor in a regulatory network can affect the expression of a 'battery' of downstream genes, which contain binding sites for that transcription factor (Davidson 2001 and see figure 1.2A). This will place constraints on the evolvability of an individual interaction, because any changes to the *cis* or *trans* acting components could have knock-on effects upon the wider regulatory networks.

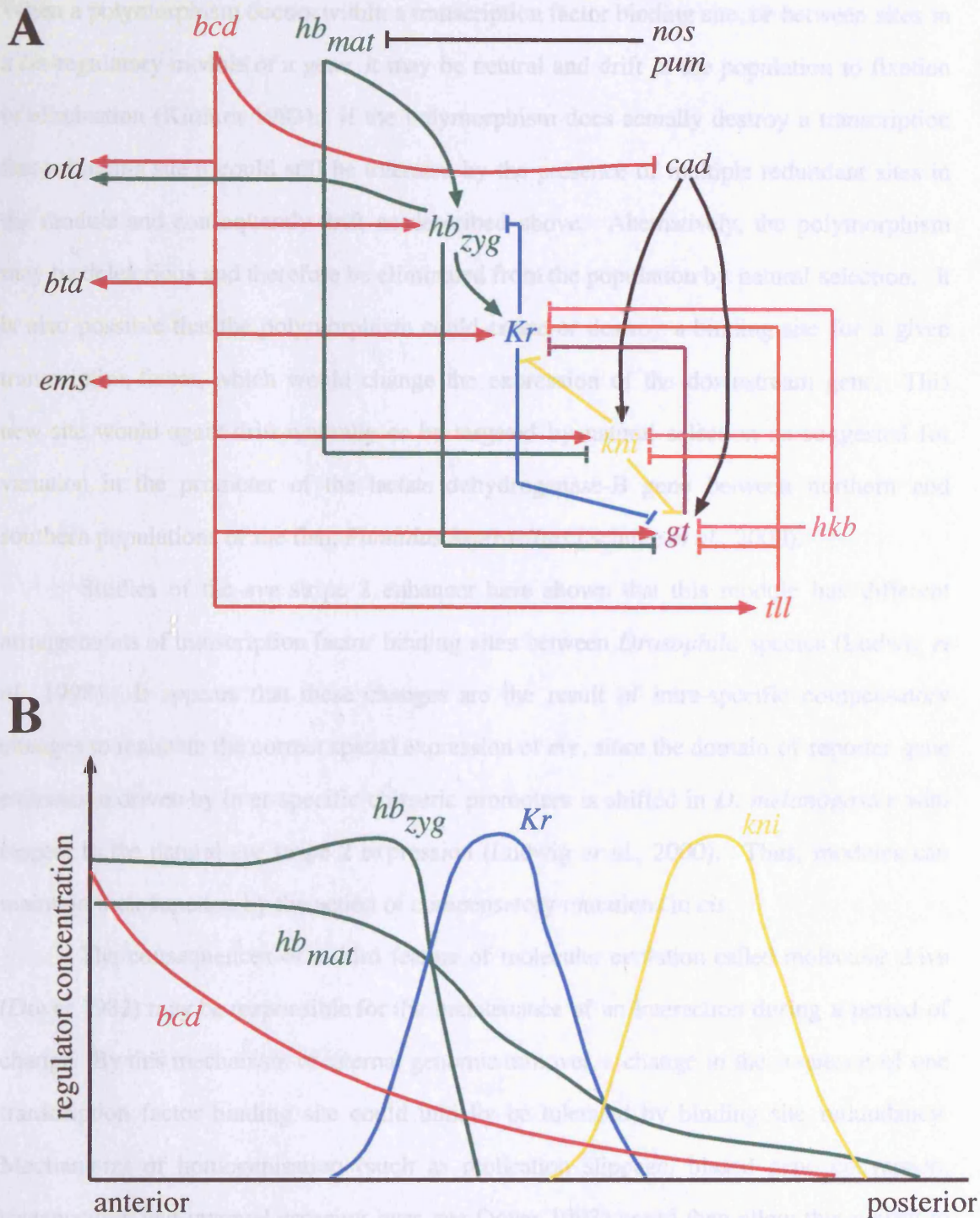


Figure 1.2 The Bcd-dependent regulatory network (A) and control of *Kr* and *kni* expression (B)
A. Network of maternal and zygotic interactions controlling anterior-posterior development in *Drosophila*. Arrowheads indicate activation and truncated lines represent repression (adapted from Sauer *et al.*, 1996).

B. The *Kr* and *kni* expression domains are controlled by the concentration gradients of Bcd and maternal Hb. Bcd activates expression of both genes and Hb sets the anterior borders of *Kr* and *kni* expression by repression. The posterior border of *Kr* expression is determined by Hb activation. Zygotic Hb also represses *Kr* in the anterior (see 1.10 for further details and references).

1.4 The consequences of *cis*-regulatory change

When a polymorphism occurs within a transcription factor binding site, or between sites in a *cis*-regulatory module of a gene, it may be neutral and drift in the population to fixation or elimination (Kimura 1983). If the polymorphism does actually destroy a transcription factor binding site it could still be tolerated by the presence of multiple redundant sites in the module and consequently drift as described above. Alternatively, the polymorphism may be deleterious and therefore be eliminated from the population by natural selection. It is also possible that the polymorphism could create or destroy a binding site for a given transcription factor, which would change the expression of the downstream gene. This new site would again drift neutrally or be targeted by natural selection as suggested for variation in the promoter of the lactate dehydrogenase-B gene between northern and southern populations of the fish, *Fundulus heteroclitus* (Schulte *et al.*, 2000).

Studies of the *eve* stripe 2 enhancer have shown that this module has different arrangements of transcription factor binding sites between *Drosophila* species (Ludwig *et al.*, 1998). It appears that these changes are the result of intra-specific compensatory changes to maintain the correct spatial expression of *eve*, since the domain of reporter gene expression driven by inter-specific chimeric promoters is shifted in *D. melanogaster* with respect to the natural *eve* stripe 2 expression (Ludwig *et al.*, 2000). Thus, modules can maintain their function by the action of compensatory mutations in *cis*.

The consequences of a third feature of molecular evolution called molecular drive (Dover 1982) may be responsible for the maintenance of an interaction during a period of change. By this mechanism of internal genomic turnover, a change in the sequence of one transcription factor binding site could initially be tolerated by binding site redundancy. Mechanisms of homogenisation (such as replication slippage, biased gene conversion, transposition and unequal crossing over, see Dover 1993) could then allow this variant to spread and replace the original transcription factor binding sites. Since DNA turnover occurs at a higher rate than point mutations, this would allow time for the selection of transcription factor variants capable of binding the new sequence. This mechanism has been called molecular co-evolution (Dover and Flavell 1984; Ohta and Dover 1984; Dover

2000) and would result in the maintenance of a developmental interaction between species, but incompatibilities in the *trans* and *cis* acting components between species.

1.5 How can we study the evolution of an interaction?

The introduction to this thesis up to this point has described the central importance of the *cis*-regulatory regions of genes to the evolution of developmental interactions. The general aim of this work was to investigate how a *cis*-regulatory module has evolved and at the same time to determine whether molecular co-evolution (see above) plays a role in the evolution of a regulatory interaction involving that module. Therefore, a well understood interaction had to be chosen that has a similar role between the species of choice. In this way developmental noise can be reduced to a minimum (Carroll 1994). However, species must be chosen that have diverged to such an extent that there has been enough time for changes in regulatory components to have occurred.

1.6 The Dipterans and the emergence of *bicoid*

The Dipterans all exhibit the long germ band mode of development as opposed to short and intermediate germ band insects such as *Schistocerca* and *Tribolium* respectively (see figure 1.3). These definitions are based on differences in the relative timing of segmentation (reviewed by Nagy 1994). For example, in *Drosophila* all the segments are specified by the end of the blastoderm stage, whereas in extreme short germ band insects most segments are added sequentially after gastrulation. However, studies of parasitic wasps (Hymenoptera) have demonstrated that such diverse modes of development can evolve over relatively short periods of time (approximately 50 MY) as a consequence of the evolving life styles of these animals (Grbic and Shand 1998).

The regulatory cascade, which controls embryogenesis in *Drosophila* is summarised in figure 1.2 and figure 1.4 (for reviews see Lawrence 1992 and Rivera-Pomar and Jäckle 1996). In *Drosophila*, the principle determinant of anterior-posterior polarity is the anteriorly localised product of the *bicoid* (*bcd*) gene, which activates the expression of genes such as *hb* (figure 1.2A and see below). However, this may be a derived

characteristic since it is thought that the posteriorly localised gene *caudal* (*cad*) may regulate the expression of *bcd* target genes such as *hb* in beetles and grasshoppers (Wolff *et al.*, 1998; Patel *et al.*, 2001; Dearden and Akam 2001).

Despite having last shared a common ancestor approximately 100 MYA (figure 1.3; Beverley and Wilson 1984), *Musca* has both a similar morphology and early embryogenesis to *Drosophila*. Indeed, this is true for the Dipterans in general (Sommer and Tautz 1991a, references therein). The only difference between the early embryos of *Drosophila* and *Musca* is in the mitotic behaviour of blastoderm nuclei, whose divisions are slightly asynchronous in the posterior of *Musca* embryos (Sommer and Tautz 1991b).

It has been demonstrated that *Musca* orthologs of *bcd*, *hb*, *Krüppel* (*Kr*), *kni*, *tailless* (*ttl*), *hairy*, *engrailed* (*en*), and *Ubx*, representing all levels of the *Drosophila* regulatory hierarchy, are expressed in similar patterns to the *Drosophila* genes (Sommer and Tautz 1991a). Comparisons of the Bcd-*hb* interaction between *Drosophila* and *Musca* have provided substantial insights into the evolution of such regulatory interactions (see below; Schröder and Sander 1993; Bonneton *et al.*, 1997; Hancock *et al.*, 1999). Are the differences in interactive components seen between *Drosophila* species (Acalyptratae) and *Musca* also observed in other members of the Calyptratae (figure 1.3)? Preliminary investigations of the blowflies, *Calliphora vicina* and *Lucilia sericata*, have isolated regions of *bcd* and *hb* in these species (see below) and so, it is also possible to ask questions concerning the interaction of these genes in these species.

It is thought that *bcd* arose as the anterior determinant in the Cyclorrhapha (figure 1.3 and see 1.9), which are a monophyletic group with distinct taxonomic features such as long germ-band eggs that have very little extra-embryonic tissue. On the other hand, non-Cyclorrhaphans are polyphyletic and characteristically have longer embryonic development and eggs with enlarged extra-embryonic tissue; for example *Clogmia* (figure 1.3). Therefore, Cyclorrhaphan species represent an appropriate group in which to study how the Bcd-dependent regulation has continued to evolve.

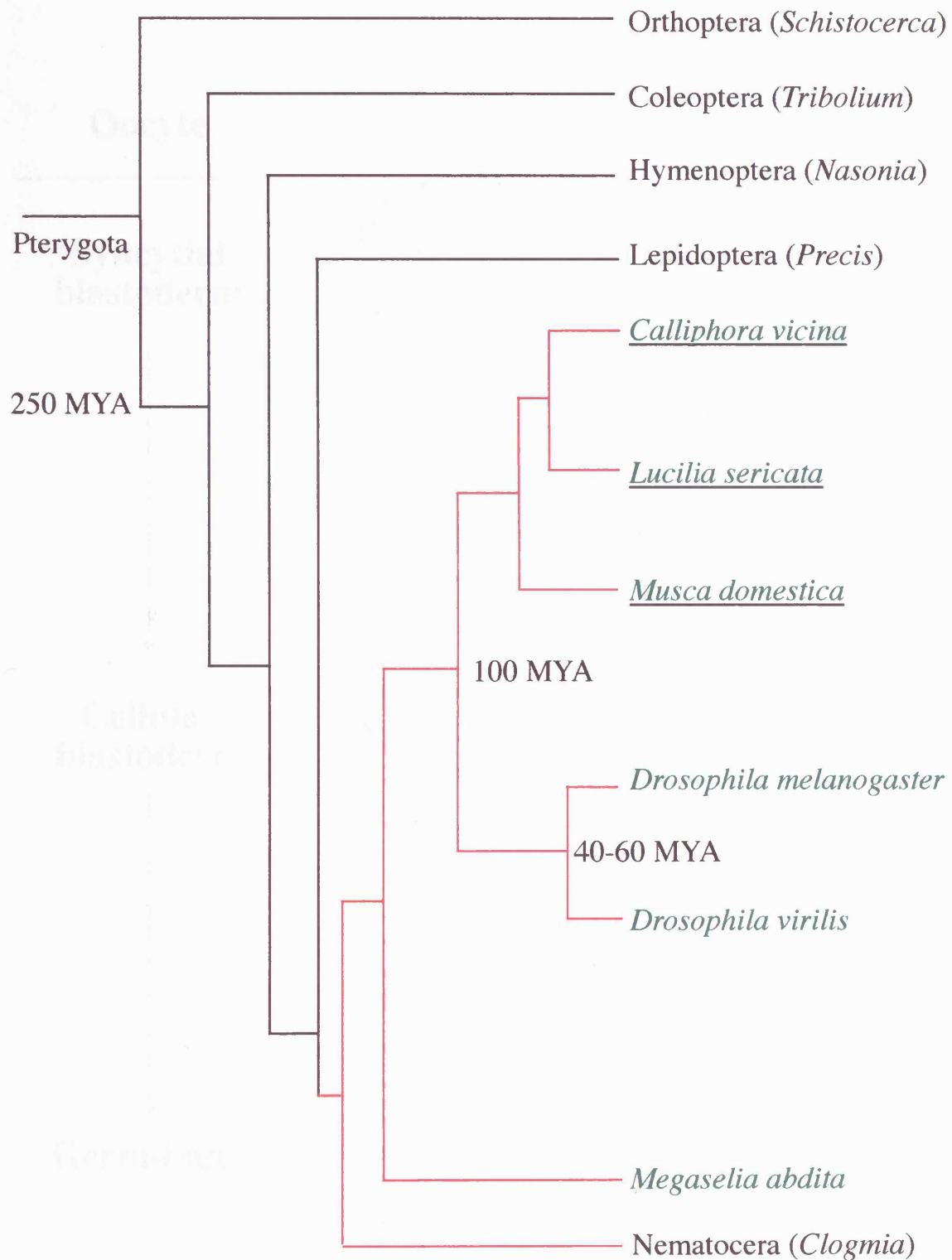


Figure 1.3 Insect phylogeny and relationships within the Diptera

The tree shows the relationships of insect groups as distant as Orthoptera to the Diptera (indicated by the red lines). The divergence times are taken from estimates by Beverley and Wilson 1984, and in Richards and Davies 1977. Tree is not to scale and not all branches are shown. Species of the Calyptratae are underlined and Cyclorrhaphan species from which *bcd* genes have been isolated are shown in green fonts.

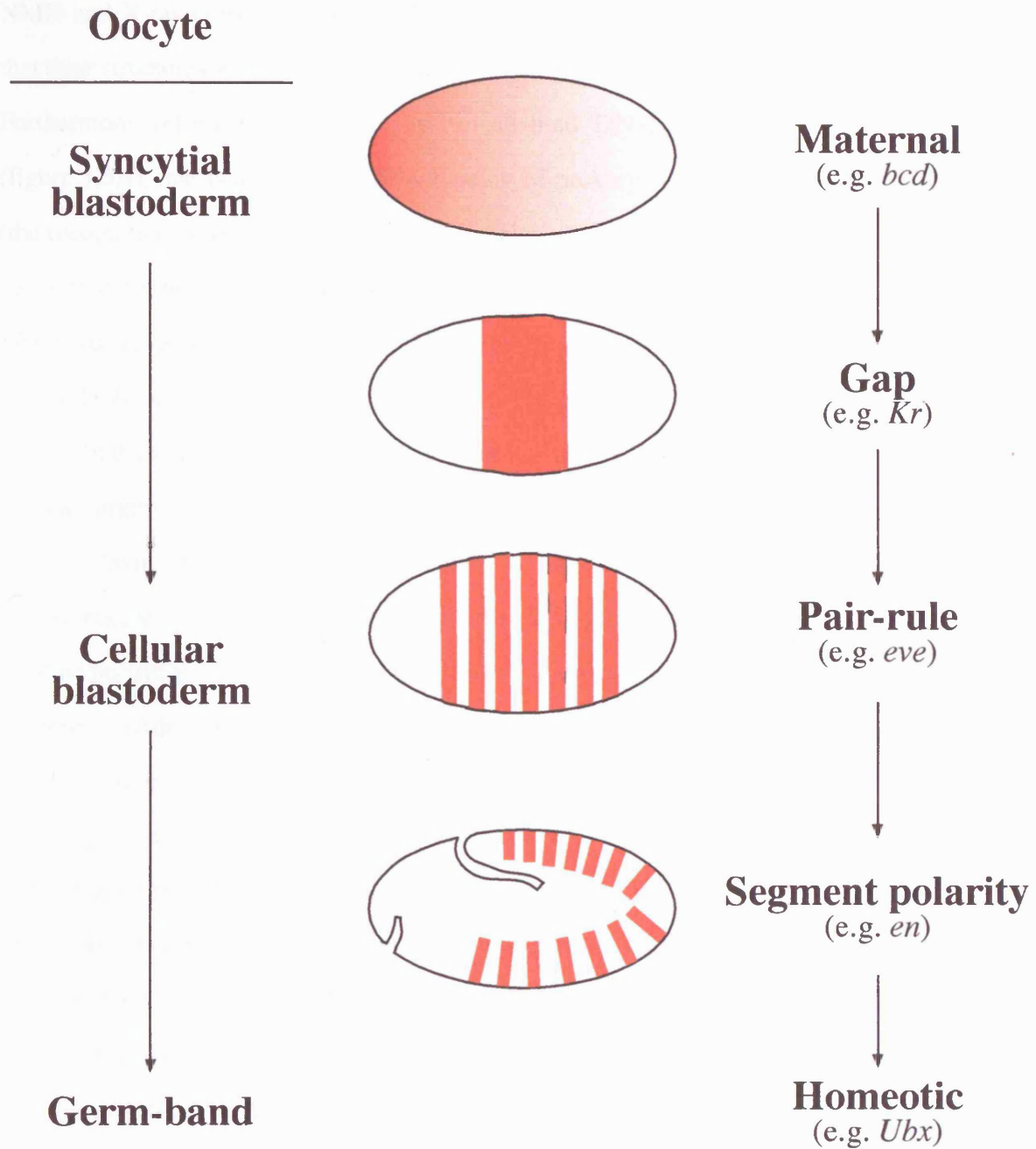
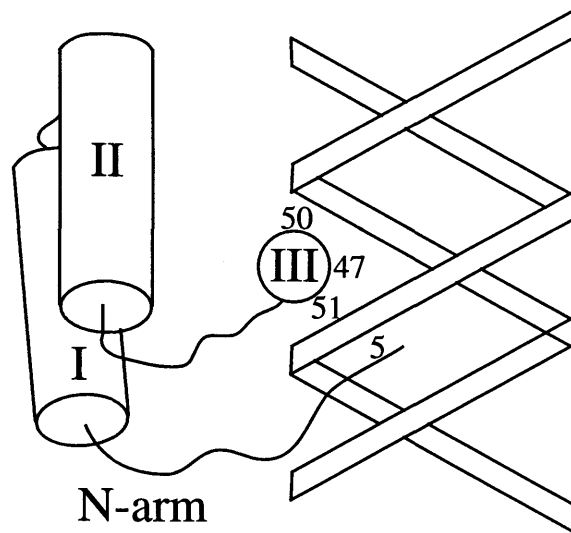
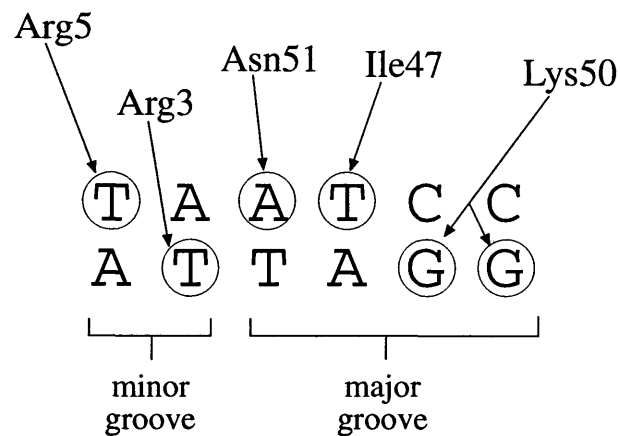


Figure 1.4 The regulatory hierarchy controlling segmentation in *Drosophila*
 The left handside of the diagram indicates the direction of ontogeny of the embryo concurrent with the expression patterns in the centre. Protein distribution is represented by the red colour in embryos orientated with the anterior to the left and dorsal up. The group of genes expressed at each stage is shown on the right and an example of each class of gene is also listed.

1.7 Binding of the Bcd homeodomain to DNA

In *Drosophila*, Bcd binds to DNA of the consensus sequence TAATCC (Driever and Nüsslein-Volhard 1989) using a 60 amino acid homeodomain (Hanes and Brent 1989). NMR and X-ray crystallography studies of several different homeodomains have shown that their structures are universally arranged in three alpha-helices and an N-terminal arm. Furthermore, different homeodomains can all bind DNA using a similar mechanism (figure 1.5A), that is akin to the DNA binding of prokaryotic HTH proteins. Helix III (the recognition helix) inserts into the DNA major groove and the N-terminal arm contacts bases in the minor groove (Hanes and Brent 1991; Otting *et al.*, 1990; Kissinger *et al.*, 1990; Tucker-Kellogg *et al.*, 1997; Wolberger *et al.*, 1991; reviewed in Gehring *et al.*, 1994a, 1994b).

In the minor groove of the DNA, at the consensus TAATCC binding site sequence, residues arginine-3 and arginine-5 of the homeodomain N-terminal arm contact the bp 2 and bp 1 thymines respectively (figure 1.5B). In the major groove residues asparagine-51 and isoleucine-47 of the recognition helix (helix III) contact adenine at bp 3 and thymine at bp 4 respectively. In addition, lysine-50 forms hydrogen bonds with the bp 5 and bp 6 guanines (Ades and Sauer, 1994, 1995; Tucker-Kellogg *et al.*, 1997) (figure 1.5B). Residues tryptophan-48 and phenylalanine-8 are conserved in many homeodomains including Bcd and have important structural roles in stabilising the homeodomain conformation for DNA binding (Subramaniam *et al.*, 2001). It has been demonstrated in *Drosophila* that residue 50 helps to specify the DNA sequences which are bound by the homeodomain (Treisman *et al.*, 1989; Hanes and Brent 1989, 1991; Hanes *et al.*, 1994). Lysine-50 class homeodomains such as Bcd preferentially bind to the *Drosophila hb* Bcd-binding sequence TAATCC (Wilson *et al.*, 1996). When this sequence is changed to TAATTC it is no longer recognised by a K50 class homeodomain, but by the Antennapedia (Antp) homeodomain which has a glutamine in place of a lysine at homeodomain position 50 (Percival-Smith *et al.*, 1990). Similarly, Bcd with a glutamine at residue 50 specifically recognises the Antp optimal binding sequence (Hanes and Brent 1989). The hydrogen bonds formed between the side chain of a lysine at position 50 and

A**B****Figure 1.5** Homeodomain-DNA binding

A. Cartoon representing the binding of the homeodomain to DNA. The three helices are numbered and the N-terminal arm is labelled. Helix III lies perpendicular to the other helices and contacts the DNA in the major groove. The N-terminal arm contacts minor groove. Residues important for sequence recognition are numbered.

B. Shown are the specific lysine-50 class homeodomain amino acid contacts with the optimal binding site sequence in the DNA major and minor grooves (Tucker-Kellogg *et al.*, 1997). Contacts between residues and specific bases (circled) are represented by arrows. Numbering refers to the position of each residue in the homeodomain.

Adapted from Ades and Sauer 1995, figure 2.

the guanines of bps 5 and 6 (figure 1.5B) may establish a stronger, more specific, interaction than by the Van der Waals attractions generated by a glutamine at this position in contact with the thymine at bp 6 of TAATTA (Tucker-Kellogg *et al.*, 1997; Ades and Sauer 1994).

There are variations in the bases contacted by different homeodomains and these are likely to be in part responsible for the specificity of a given homeodomains interaction with DNA. For example, arginine-3 in the N-terminal arm of the En, Eve and Antp homeodomains can contact different bases depending on the trajectory of the N-terminal arm. Similarly, the Eve homeodomain can make different DNA contacts with its glutamine side chain (Hirsch and Aggarwal 1995).

1.8 The role of *bcd*

The transcription factor encoded by *bcd* is required for the development of head and thoracic structures in *Drosophila* (Frohnhofer *et al.*, 1986; Berleth *et al.*, 1988).

Transcription of *bcd* takes place in the nurse cells and the mRNAs are subsequently transported to the oocyte where they are localised at the anterior pole by the *swallow* and *exuperantia* gene products (Berleth *et al.*, 1988). Upon translation, Bcd proteins diffuse through the embryo and are rapidly degraded, thus forming a concentration gradient that is highest at the anterior of the embryo and detectable up to 30% egg length in the posterior (Driever and Nüsslein-Volhard 1988a). Bcd acts as a classic morphogen to pattern the embryo along the anterior-posterior axis by the activation and repression of target genes in a concentration dependent manner (Driever and Nüsslein-Volhard 1988b). For example, Bcd activates the expression of *orthodenticle* (*otd*) and *kni* in the anterior and posterior of the embryo respectively (Finkelstein and Perrimon 1990; Rivera-Pomar *et al.*, 1995) and represses *ill* and *cad* (translational repression) in the anterior (figure 1.2A; Pignoni *et al.*, 1992; Dubnau and Struhl 1996).

Cytoplasmic transfer experiments have shown that when *bcd* is transplanted in to other regions of the embryo it affects the development of ectopic anterior structures (Frohnhofer *et al.*, 1986). Increasing the maternal dosage of *bcd* by up to 6 copies results

in a posterior shift in the expression of target genes such as *hb* and anterior markers such as the cephalic furrow (Driever and Nüsslein-Volhard 1988b; Namba *et al.*, 1997). However, these embryos develop into viable adults by the actions of increased cell death (Namba *et al.*, 1997).

It has been suggested that the specific configurations of Bcd-binding sites in the promoters of genes whose expression is regulated by Bcd may determine where in the Bcd gradient the downstream gene is expressed (Driever 1993).

1.9 The evolution of *bcd*

bcd was first isolated in *D. melanogaster* (Berleth *et al.*, 1988) and since then orthologs have been found and sequenced in a number of other Dipterans including *D. pseudoobscura*, *Musca*, *Calliphora*, *Lucilia*, *Megaselia* and *Phormia* (Seeger and Kaufman 1990; Schröder and Sander 1993; Stauber *et al.*, 1999; Shaw *et al.*, 2001).

It has been proposed that *bcd* is a sister gene of *zerknüllt* (*zen*) and that they are both derived from the duplication, early in the Dipteran lineage, of an ancestral Hox class 3 gene (Stauber *et al.*, 1999). This is further supported by the linkage of *bcd* and *zen* in both *Calliphora* and *Lucilia* (Brown *et al.*, 2001). However, this region of the Hox cluster has been fully sequenced in flour beetle *Tribolium* and although an independent duplication of *zen* was observed in this species, no *bcd*-like gene was discovered (Brown *et al.*, 2001). Indeed, only *zen*-like genes have been reported in the non-Cyclorrhaphans *Haematopota* (Tabanidae) and *Clogmia* (Stauber *et al.*, 2002). Therefore, to date, no *bcd*-like gene has been isolated from species other than the Cyclorrhaphan Dipterans. It is possible that a highly diverged *bcd*-like gene is present at a different locus in non-Cyclorrhaphans, which would mean that techniques such as degenerate PCR would be unable to isolate it. Intriguingly, the *Tribolium hb* promoter region and *cad* 3' UTR are both regulated by *bcd* when they are placed in *Drosophila* (Wolff *et al.*, 1998), although this should be treated with caution since factors other than *bcd* may regulate *cad* by using a similar mechanism in *Tribolium* (Gibson 2000). Alternatively, *bcd* may have taken over the role of anterior determinant in Dipterans from the ancestral insect system (Dearden and Akam 1999;

Brown *et al.*, 2001). This would mean that *bcd* regulation of target genes, such as *hb* and *Kr*, must have been wired in to extant genetic regulatory circuits controlled by an ancestral factor or factors.

1.10 The expression and role of *hb* in the *Drosophila* embryo

hb encodes a zinc-finger transcription factor and was first isolated in *Drosophila*, where its role as a gap gene is essential for early development of the thoracic and abdominal segments (Lehmann and Nüsslein-Volhard 1987; Tautz *et al.*, 1987).

In *Drosophila* there are two *hb* transcripts, of 3.2 and 2.9 kb respectively, which encode identical proteins spliced to different 5' UTRs. These two mRNAs are transcribed from the P1 and P2 promoters respectively (Tautz *et al.*, 1987). Expression of *hb* is both maternal and zygotic and these domains are partially redundant, since zygotic *hb* mutations are lethal to the embryo, while the absence of maternal expression can be tolerated. However, *hb_{mat}* and *hb_{zyg}* double mutants have more severe defects than the effects of the individual mutations. This is exemplified by a mirror image duplication of abdominal segments in the double mutant, contrasting with gnathal and thoracic defects, and/or a deletion of abdominal segments A7 and A8 observed in the zygotic mutant (Lehmann and Nüsslein-Volhard 1987; Irish *et al.*, 1989; Hülkamp *et al.*, 1989; Simpson-Brose *et al.*, 1994).

Maternal *hb* mRNAs transcribed from the P1 promoter are localised throughout the early *Drosophila* embryo, but their translation is restricted to the anterior half of the embryo by negative post-transcriptional regulation in the posterior. Pumilio (Pum) binds to the 3' UTR of the *hb* mRNA and recruits Nanos (Nos) which results in deadenylation of the *hb* mRNA and translational repression (Wharton and Struhl 1991; Murata and Wharton 1995; Wrenden *et al.*, 1997). This is possibly the only embryonic function of Nos since *nos/hb_{mat}* double mutants develop normally (Hülkamp *et al.*, 1989; Irish *et al.*, 1989; Struhl *et al.*, 1989).

Zygotic *hb* expression in *Drosophila* occurs in two phases. Early zygotic expression consists of Bcd-dependent expression of P2 transcripts in the anterior half of

the embryo up to approximately 55% egg length (Tautz 1988; Schröder *et al.*, 1988; Driever and Nüsslein-Volhard 1989). Later zygotic *hb* expression patterns of both P1 and P2 transcripts are characterised by an anterior stripe located at PS4, which is necessary for T2 development and a posterior stripe required for development of A8 (Tautz 1988; Hülskamp *et al.*, 1994; Margolis *et al.*, 1995; Wu *et al.*, 2001). PS4 expression is dependent on autoregulation by Hb and its position is also possibly in part determined by Kr repression at its posterior border (Treisman and Desplan 1989). The posterior stripe of *hb* expression is dependent on Tll activation and Hucklebein (Hkb) repression mediated through *cis*-regulatory sequences located upstream of the P1 transcription start site (Margolis *et al.*, 1995). After gastrulation *hb* P1 transcripts persist in the CNS, where they are involved in blastomere fate determination (Kambadur *et al.*, 1998 and see 3.3.2).

Repression of *hb* translation in the posterior of the embryo results in a gradient of Hb protein with the highest concentration in the anterior. This gradient determines the spatial expression of genes including *Kr*, *kni* and *gt*, in a concentration dependent manner (Hülskamp *et al.*, 1990, 1994; Struhl *et al.*, 1992; Wu *et al.*, 2001). For example, different Hb concentrations contribute to setting the anterior boundaries of *kni* and *Kr* expression domains by transcriptional repression of these genes. However, at a lower concentration Hb activation of *Kr* transcription sets the posterior boundary of *Kr* expression (figure 1.2B). The maintenance of these expression domains may involve the Polycomb group genes such as *Enhancer of zeste* (Pelegri and Lehmann 1994). It is thought that the high levels of *hb* zygotic expression in the anterior are required not only to pattern the head region, but to specify T2 by repression of *Kr* and *Ubx* and activation of *Antp* (Harding and Levine 1988; Irish *et al.*, 1989; Wu *et al.*, 2001).

1.11 Bcd-dependent activation of *hb*

In *Drosophila*, Bcd activates early zygotic *hb* expression by binding to a 300 bp promoter located upstream of the P2 transcription start site. This was determined by the ability of this sequence to drive reporter gene expression in the natural anterior *hb* expression domain (Schröder *et al.*, 1988; Driever and Nüsslein-Volhard 1989; Struhl *et al.*, 1989).

DNaseI footprinting has been used to determine that there are 7 Bcd-binding sites of variable affinity contained in this promoter (Driever and Nüsslein-Volhard 1989; Ma *et al.*, 1996). It is thought that the configuration of these binding sites, in terms of spacing, number, orientation and sequence, are responsible for the sharp threshold of expression at approximately 55% egg length. Disruption of these binding sites results in reduced transcription, a more shallow threshold and an anterior shift in expression (Driever *et al.*, 1989a; Struhl *et al.*, 1989; Yuan *et al.*, 1999; Ma *et al.*, 1999; Zhao *et al.*, 2000). Therefore, the threshold position at the posterior boundary of early *hb* expression is determined both by the concentration of Bcd (see above) and by the configuration (or signature) of binding sites in the P2 promoter (Gibson 1996).

There is also evidence that Bcd and Hb act synergistically with components of the TFIID transcriptional complex to drive high levels of *hb* expression in the anterior of the embryo (Simpson-Brose *et al.*, 1994; Sauer *et al.*, 1995a and 1995b). This synergy may also contribute to the threshold width of early *hb_{zyg}* expression and its position at 55% egg length. Simpson-Brose and co-workers (1994) demonstrated that greater expression from a reporter gene was observed *in vivo* when the upstream promoter contained binding sites for both Bcd and Hb, rather than promoters containing sites for just one factor or the other. Indeed, the *Drosophila hb* P2 promoter contains two Hb binding sites (Treisman and Desplan 1989). It was subsequently shown that the glutamine-rich and alanine-rich domains of Bcd could interact with TAF_{II}110 (TATA-binding protein associated factor) and TAF_{II}60 respectively, and that Hb could also interact with TAF_{II}60. These two TAFs are components of the TFIID complex, along with TBP (TATA-binding protein), TAF_{II}250 and at least 5 other TAFs (Goodrich and Tjian 1994). *In vitro* transcription and DNaseI footprinting experiments have suggested that synergy between Bcd and Hb and their interactions with the above TAFs gave higher levels of transcription by favouring the recruitment of TFIID to the TATA box (Sauer *et al.*, 1995a, 1995b). However, recent analysis has suggested that the interactions between Bcd and Hb and components of the TFIID complex may be redundant or not required *in vivo*. Schaeffer and co-workers (1999) determined that Bcd Δ QAC (Bcd without either the glutamine-rich, alanine-rich or

C-terminal domains) could rescue *Drosophila bcd* mutants and that dominant negative TAF_{II}110 and TAF_{II}60 mutants had no effect on this rescue. Therefore, the details of *hb* transcriptional activation by Bcd and Hb and their interactions with the basal transcription machinery still remain to be understood. It is possible that Bcd and Hb do have overlapping functions since many Bcd target gene promoters also contain Hb binding sites (e.g. Hoch *et al.*, 1991; Simpson-Brose *et al.*, 1994).

1.12 Evolution of the Bcd-*hb* interaction

The expression patterns of *hb* are very similar between *D. melanogaster* and *D. virilis*, which diverged approximately 60 MYA. Although, there appears to be a broadening of the PS4 stripe in *D. virilis*, the Bcd-dependent domain of expression is the same in these two species (Treier *et al.*, 1989). Interestingly, there is a deletion in the *D. virilis hb* P2 promoter with respect to the *D. melanogaster* promoter, which has resulted in the removal of binding sites x1 and x2 (Treier *et al.*, 1989). Nevertheless, the *D. virilis* promoter is able to drive natural *hb* expression when placed in *D. melanogaster* (Lukowitz *et al.*, 1994). Therefore, it seems that sites x1 and x2 are not needed in the *D. virilis* promoter, possibly due to compensatory mutations that have generated other binding sites, resulting in a configuration of binding sites that gives the same response as the *D. melanogaster hb* promoter configuration *in vivo* (see 4.3.4).

The *D. pseudoobscura bcd* gene codes for a protein that is 81% similar to *D. melanogaster* Bcd (Seeger and Kaufman 1990). Furthermore, these two Bcd proteins have identical homeodomains and similar activation domains, although the glutamine-rich domain is expanded in *D. pseudoobscura* (see figure 4.1). Interestingly, while the sequences of the *bcd* 3' UTRs have diverged in a range of *Drosophila* species they can still form very similar secondary structures, which allow the correct localisation of *bcd* in *D. melanogaster*, despite divergence times of up to 60 MY (Macdonald and Struhl 1988; Macdonald 1990; Luk *et al.*, 1994). Indeed, the *bcd* 3' UTR of the housefly, *Musca domestica*, can also form a similar secondary structure to that of *Drosophila bcd*, despite their sequences being unalignable (Shaw *et al.*, 2001) and the 100 MY lapse since these

species last shared a common ancestor (Beverley and Wilson 1984). However, comparisons of the *Drosophila* Bcd homeodomain with that of *Musca* have revealed 5 amino acid differences between these species and 4 of these differences are also present in the *Lucilia* and *Calliphora* Bcd homeodomains (Sommer and Tautz 1991a; Schröder and Sander 1993; Shaw *et al.*, 2001). These differences are somewhat surprising since the amino acid sequences of homeodomains are known to be conserved over large phylogenetic distances. For example, the putative human ortholog of *Antp* exhibits only a single amino acid difference from the homeodomain coded for by this gene in *Drosophila*, despite these species last sharing a common ancestor more than 500 MYA (for a review see Gehring *et al.*, 1994a). It has been suggested that the differences between the *Drosophila* and *Musca* Bcd homeodomains may have allowed subtle differences in their respective binding site sequence preferences (Bonneton *et al.*, 1997; Shaw 1998).

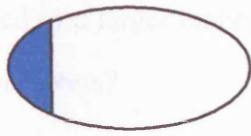
The *Musca hb* promoter was sequenced and was found to be unalignable with the *D. melanogaster hb* P2 promoter sequence (Bonneton *et al.*, 1997). Characterisation of Bcd-binding sites in the *Musca* promoter revealed that there were 10 sites spread over 525 bp. Therefore, the *Drosophila* and *Musca hb* promoters have diverged in the number, spacing orientation and sequence of binding sites, despite high conservation of the *hb* coding sequence between these species (Bonneton *et al.*, 1997). However, the *Musca hb* promoter was able to rescue *Drosophila hb* mutant defects in anterior structures and could drive reporter gene expression in the anterior of the *Drosophila* embryo as seen for *Drosophila* promoter constructs (Bonneton *et al.*, 1997; Shaw 1998). Interestingly these experiments revealed that the *Musca* promoter also drove ectopic expression in *Drosophila* embryos, represented by the persistence of Bcd-dependent expression in the anterior tip, which was also *tor* dependent. These experiments demonstrated that *Drosophila* Bcd is able to recognise the *Musca hb* promoter, but that there are possibly subtle differences in this regulatory interaction between these species. It possible that the differences in the Bcd homeodomains between *Drosophila* and *Musca* have co-evolved (see 1.4) with the restructured *hb* promoter sequences between these species (Bonneton *et al.*, 1997; Hancock *et al.*, 1999).

1.13 *bcd* in other Dipterans

The *Megaselia abdita* (see figures 1.3 and 1.6) *bcd* gene has been fully sequenced and its expression patterns characterised. In addition, RNAi has been employed to investigate the function of *bcd* in this species in comparison to *Drosophila* (Stauber *et al.*, 2000). The homeodomain of *Megaselia* Bcd has 18 amino acid differences with respect to the *Drosophila* Bcd homeodomain and this in combination with differences in other domains of the protein may have resulted in functional differences between Bcd in these species (see figure 4.2 and 6.4.2). In addition, *bcd* expression is located further to the posterior in *Megaselia* than in *Drosophila*. The application of RNAi to phenocopy *bcd* in *Megaselia* has shown defects in abdominal segments in addition to the head and thoracic defects observed using this technique in *Drosophila* (Stauber *et al.*, 2000). Using RNAi to phenocopy *bcd* mutations in *Musca* resulted in defects only in head structures (Shaw *et al.*, 2001). These results may be evidence for the role of *bcd* being restricted to more anterior regions of the embryo during the course of Dipteran evolution. Interestingly, *Megaselia* cytoplasm failed to rescue anterior structures in *Drosophila bcd* mutants during cytoplasm transfer experiments (Schröder and Sander 1993; summarised in figure 1.6), as did *Megaselia bcd* in transgenic *Drosophila bcd* mutant embryos (P. Shaw personal communication). More surprisingly, *Calliphora* cytoplasm also failed to rescue *Drosophila bcd* mutants despite Bcd from these species having a similar homeodomain (Schröder and Sander 1993). The expression of *bcd* in *Calliphora* is restricted to the most anterior tip of the embryo, which again supports a more anterior role for *bcd* in this species (Schröder and Sander 1993). *Lucilia* cytoplasm did rescue anterior structures in *Drosophila* embryos and in *Lucilia*, *bcd* is expressed in a similar pattern to *Drosophila* and *Musca bcd* (figure 1.6).

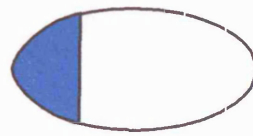
As can be observed in figure 1.6, while the *Megaselia* and *Drosophila* embryos are approximately the same size, those of *Musca* and *Lucilia* are twice as long and the *Calliphora* embryo is even longer. Since *bcd* function is dependent on an anterior-posterior gradient, these differences in size may have implications for the activation of

Drosophila



82%

Megaselia



0%

Musca



45%

Lucilia



25%

Calliphora



0%

0.5 mm

Figure 1.6 Embryo sizes of the higher Dipterans

Illustrated are the embryo sizes of *Drosophila*, *Megaselia*, *Musca*, *Lucilia* and *Calliphora* given approximately to scale. The anterior is to the left and blue shading represents the domain of expression of the *bcd* mRNAs in each species (see text for references). The percentages given to the right of each embryo represents the extent to which anterior cytoplasm from each species rescues anterior structures in *Drosophila* embryos (Schroder and Sander 1993). These percentages are the fraction of cuticles that show some rescue effect.

target gene expression by *bcd* in larger embryos. Is the *Musca hb* promoter (see above) a more sensitive configuration of Bcd-binding sites adapted to read a shallower gradient of Bcd in a larger embryo? Is this structure also evident in the *Calliphora* and *Lucilia hb* promoters?

1.14 Aims of this thesis

The Bcd-*hb* interaction has been characterised in depth in *Drosophila* (see above) where it plays a vital role in the patterning of the early embryo. This interaction appears to be conserved in *Musca*, however there are a number of differences in both the *trans* and *cis*-acting factors between these species. These changes have consequences for the maintenance of genetic regulatory networks of which *bcd* and *hb* are part.

In general terms this thesis sought to explore how changes in the regulatory regions (promoters) of genes occur and how these changes may be tolerated within the genetic regulatory networks of which these promoters are a part. Therefore, I have further investigated the evolution of the Bcd-*hb* interaction in the Dipterans *Drosophila*, *Musca*, *Megaselia* *Calliphora* and *Lucilia* and the specific aims of the project were designed to address the following questions:

- What is the amino acid sequence of the Hb protein in *Lucilia* and *Calliphora* and are the *hb* expression patterns in these species (particularly the putatively Bcd-dependent expression domain) similar to the patterns seen in other insects?
- Is the divergence in the configurations of Bcd-binding sites seen between the *Drosophila* and *Musca hb* promoters reflected in the *hb* promoters of *Lucilia* and *Calliphora*?
- Are inter-specific differences in *hb* a reflection of intra-specific variation? Indeed, what are the mutational mechanisms that have driven the divergence of the *Drosophila* and *Musca hb* promoters?
- Are the re-structured *hb* promoters merely an outcome of compensatory evolution in *cis*, or have they also co-evolved with changes in the Bcd homeodomain (compensatory

evolution in *trans*)? Do species with larger embryos require more sensitive promoters to read the gradient of Bcd?

- How have other Bcd regulated genes, such as *otd*, evolved in *Musca* compared with *Drosophila*?

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Media

LB (Luria broth): 1% (w/v) Bacto-tryptone (Difco), 0.5% (w/v) Bacto-yeast extract (Difco), 1% (w/v) NaCl. LB-agar was made as above but with the addition of 1.5% (w/v) agar (Difco). The antibiotics ampicillin and kanamycin were added when appropriate to LB cultures and LB-agar plates to a working concentration of 50 µg/ml (from stock solutions of 50 mg/ml in ethanol). Tetracycline was used at a working concentration of 12.5 µg/ml (stock solution 12.5 mg/ml in 50% aqueous ethanol).

YPD (Yeast peptone dextrose): 10 g yeast extract, 20 g peptone, 0.1 g NaOH and 2% glucose per litre of water.

SD (synthetic drop-out): 6.7 g yeast nitrogen base, 850ml of water, 50 ml of 40% glucose, 100 ml of amino acid solution lacking appropriate amino acids. For drop out plates 20 g of agar was added per litre.

2.1.2 Organisms

Bacteria

The following strains of *Escherichia coli* were used:

DH5α (Gibco BRL): *supE44 hsdR17 recA1 endA1 gyrA96 thi-1 relA1*.

XL1-Blue (Stratagene): *supE44 hsdR17 recA1 endA1 gyrA46 thi relA1 lac⁻ F' [proAB⁺ lacI^q lacZΔM15 Tn10(tet^r)]*.

XL-1 Blue MRA: *Δ(mcrA)183, Δ(mcrCB-hsdSMR-mrr)173, endA1, supE44, thi-1, gyrA96, relA1, lac*.

Bacterial stocks and stocks transformed with plasmids were maintained at -20°C in equal volumes of overnight LB cultures and glycerol.

Yeast

Saccharomyces cerevisiae strain EGY48 (MAT α , *leu2*, *ura3*, *trp1*, *his3*, *lexAOp-LEU2*; Gyuris *et al.*, 1993) was used.

Musca domestica

Laboratory strains of *Musca* were donated by the following sources: Edinburgh; Prof. D. Saunders, University of Edinburgh. Cardiff; Dr L. Senior, Insect Investigations, University of Cardiff. White and Zurich; Prof. A. Dubendorfer, University of Zurich. Rutgers; Prof. Plapp, University of Arizona. Flies were maintained in cages at 26°C with sucrose, dried milk and water. Larval food was prepared as described in Bonneton *et al.*, 1997.

Calliphora vicina* and *Lucilia sericata

Calliphora vicina were a gift from Prof. D. Saunders, University of Edinburgh and *Lucilia sericata* were donated by Dr Jon Reid, Zoology Department, University of Leicester. Populations of these flies were maintained in cages at 26°C with sucrose and water. Horse blood agar (200 ml oxalated horse blood, 50 g brewers yeast, 10ml nipogen (10%), 20 g agar and 790 ml water) was used as larval food for these two species.

2.1.3 Plasmids

Plasmid	Description	Source
pECBCD	pET42b with the <i>Calliphora bcd</i> homeodomain coding sequence inserted (see figure 4.1).	This work
pELBCD	pET42b with the <i>Lucilia bcd</i> homeodomain coding sequence inserted (see figure 4.1).	This work
pBC103	(2 μ , LEU2) GAL1 promoter vector with HA epitope tag	S. Hanes (Cohen & Brent)
pLR1 Δ 1	(2 μ , URA3) <i>lacZ</i> reporter	S. Hanes (West <i>et al.</i> , 1984)
pRS313-G4ERV16	(CEN-ARS, HIS3) hormone responsive activator	S. Hanes (Louvion <i>et al.</i> , 1993)
pDB1.2	(2 μ , LEU2) pBC103 with <i>D. melanogaster bcd</i>	S. Hanes (Burz <i>et al.</i> , 1998)
pDBhb.19	(2 μ , URA3) Bcd site reporter, <i>hb</i> 230 bp enhancer in pLR1 Δ 1	S. Hanes (Burz <i>et al.</i> , 1998)
pBCMBCD	(2 μ , LEU2) pBC103 with <i>Musca bcd</i>	This work
pMABCD	(2 μ , LEU2) pBC103 with <i>Megaselia bcd</i> cDNA	This work
pMhbP2+	(2 μ , URA3) pLR1 Δ 1 with <i>Musca hb</i> enhancer	This work
pCVP+	(2 μ , URA3) pLR1 Δ 1 with <i>Calliphora hb</i> enhancer inserted 5'-3'.	This work
pCVP-	(2 μ , URA3) pLR1 Δ 1 with <i>Calliphora hb</i> enhancer inserted 3'-5'.	This work
pLSP+	(2 μ , URA3) pLR1 Δ 1 with <i>Lucilia hb</i> enhancer inserted 5'-3'.	This work
pLSP-	(2 μ , URA3) pLR1 Δ 1 with <i>Lucilia hb</i> enhancer inserted 3'-5'.	This work
pCVP5+	(2 μ , URA3) pLR1 Δ 1 with <i>Calliphora hb</i> enhancer inserted 5'-3' (from -786 to -473).	This work
pCVP5-	(2 μ , URA3) pLR1 Δ 1 with <i>Calliphora hb</i> enhancer inserted 3'-5' (from -786 to -473).	This work
pBCDR1	<i>Musca bcd</i> coding region with <i>Eco</i> RI and <i>Hind</i> III flanking sites in pBluescript KS+.	P. Shaw
pMASB	Plasmid containing the <i>Megaselia bcd</i> cDNA.	M. Stauber
pD1	<i>Musca hb</i> P2 <i>Dra</i> I fragment in pBluescript KS+ (<i>Eco</i> RV).	Shaw 1998

Table 2.1 Plasmids used in this work.

2.1.4 Oligonucleotides

All oligonucleotides used were synthesised by Interactiva Biotechnologie and supplied as lyophilised pellets. The sequences are listed in appendix A.

2.2 Methods

2.2.1 Standard molecular biology techniques

2.2.1.1 DNA precipitation and phenol-chloroform extraction

Alcohol precipitation and phenol-chloroform extraction of nucleic acids were done according to Sambrook *et al.*, (1989).

2.2.1.2 Restriction digests

Restriction digests were done according to the manufacturer's recommendations in the buffer supplied with the enzyme. Vector DNA was dephosphorylated by the addition of 2-3 units of Shrimp Alkaline Phosphatase (SAP) (5 units/μl, USB-Amersham) to the restriction digest.

2.2.1.3 Gel extraction

Fragments were run out on standard agarose gels in 1x TAE and gel-purified using a gel extraction kit (Qiagen) according to the manufacturers instructions.

2.2.1.4 Ligation of DNA fragments.

10-50 ng of linearised vector DNA was incubated with an appropriate amount of insert DNA to give a rough molar ratio vector:insert of 1:3. The DNA was then put on ice and T4 ligase buffer (supplied with enzyme) added to 1x concentration (Gibco BRL), with 1-2 units of T4 DNA ligase (1 Weiss unit/μl, Gibco BRL). The reaction was incubated

overnight at 16°C in a final volume of 15 µl. Half of the ligation reaction was transformed into *E. coli* as described below.

PCR products were cloned and transformed into *E. coli* using TOPO kits (Invitrogen) according to the manufacturers instructions.

2.2.1.5 Transformation of *E. coli*

The CaCl₂ method, as described in Sambrook *et al.*, (1989), was used for general plasmid transformations. For transformation of large plasmids (>10 kb), electroporation was used. Electrocompetent cells were prepared as follows: a 0.5 l LB-tetracycline culture of *E. coli* strain XL1-blue was grown to mid-log phase (OD₆₀₀=0.55). Cells were washed sequentially to remove salts as follows: Cells were pelleted by centrifugation in a Sorvall ultracentrifuge GS-3 rotor at 4000 rpm for 10 mins (4°C). The cell pellet was resuspended in 500 ml of ice-cold deionised water. The cell suspension was spun again and the cell pellet resuspended in 250 ml of ice-cold deionised water. The cell suspension was then spun in a SS-34 rotor at 9000 rpm and resuspended in 10 ml of ice-cold 10% (w/v) glycerol. The cell suspension was then spun again and the pellet resuspended in 1 ml of ice-cold 10% (w/v) glycerol. 40 µl aliquots of cell suspension were frozen in dry ice-ethanol. Cell aliquots were stored at -80°C.

Electroporation: plasmid DNA was prepared by ethanol precipitation and resuspended in 10 µl of deionised water. An electrocompetent cell aliquot was thawed on ice and added with the transforming DNA to an electroporation cuvette. An electric pulse was delivered using a slot apparatus unit (GenePulser, Biorad), set at 25 µF and 1.5 kV. Cells were recovered at 37°C for 1 hour in 1 ml of SOC medium (prepared as described in Sambrook *et al.*, 1989). Aliquots of recovered cells were plated out on appropriate agar medium. Typical efficiency: 1 x 10⁸ transformants per µg of DNA.

2.2.1.6 Preparation of plasmid DNA

Plasmid DNA was isolated from bacterial cultures using mini- or maxi-prep kits (Qiagen) according to the manufacturers instructions.

2.2.1.7 Agarose gel electrophoresis

0.5-1.8% (w/v) gels were cast using Seakem LE agarose (Flowgen) dissolved in 1x TAE. 5x loading buffer (5x TBE, 15% (w/v) Ficoll-400 (Pharmacia Biotech.), 0.25% (w/v) bromophenol blue) was added to the DNA samples before loading and gels were run in horizontal perspex slab gel tanks at 1-6 V/cm in the corresponding buffer. DNA was visualised by the addition of ethidium bromide (EtBr) (0.5 µg/ml) to the gel mix before casting and observing the fluorescence at 300 nm UV on a transilluminator. DNA size markers, such as λ HindIII markers (Gibco BRL) (fragments 23130, 9416, 6557, 4361, 2322, 2027, 564 and 125 bp) and/or Φ X174 HaeIII markers (Advanced Biotechnologies) (1353, 1078, 872, 603, 310, 281, 271, 234, 194, 118 and 72 bp), were used to estimate the sizes of DNA fragments. The gel was photographed with a video imaging system. For isolation of small DNA fragments for cloning, 1% (w/v) gels were cast with low-melting agarose in 1x TAE and EtBr (0.5 µg/ml).

2.2.1.8 Denaturing polyacrylamide (sequencing) electrophoresis

Glass plates 21 x 50 cm or 42 x 50 from a Sequi-Gen sequencing gel apparatus set (Biorad) were used. 6% polyacrylamide gels (5% for DNaseI footprinting) were cast using 'Sequagel' (gas-stabilised 19:1 acrylamide:bisacrylamide acrylamide solution in 8.3 M urea, National Diagnostics, Flowgen), in 1x TBE, according to the manufacturer's recommendations. 24 or 48-well 0.4 mm thick teflon sharktooth combs (Biorad) were used to make the wells for sample loading. For routine dideoxy sequencing, gels were run for 2-4 hours in 1x TBE buffer at 43 W or 90W constant power and maximum voltage, which kept the gel at roughly 50°C during the run.

For DNaseI footprinting, gels were run as 'gradient' gels. The top buffer was 0.5x TBE and the bottom buffer 1x TBE. After samples had entered the gel (10-15 minutes after loading), 1/2 the volume of the bottom buffer of 3 M sodium acetate was added to the bottom buffer, which lowers the conductivity of the lower buffer establishing an ionic gradient which creates a more linear rate of migration for the smaller fragments.

After the gel run was complete, gels were fixed in a solution of 10% (v/v) acetic acid, 15% (v/v) methanol for 10 minutes. Gels were dried onto Whatman 3MM paper in a vacuum drier (Biorad model 583) at 80°C for 60-90 minutes. Gels were exposed to X-ray film (Fuji RX100) for 1-7 days at room temperature.

2.2.1.9 Southern analysis

Labelling of the desired probe DNA fragment was done according to Feinberg and Vogelstein (1984). After labelling, the probe was purified by sephadex spin-column chromatography to remove unincorporated nucleotides. The probe was then denatured and pipetted directly into the hybridisation solution.

Approximately 5 µg of genomic DNA were used in each digest and the digests were run out on 0.6% (w/v) 1x TBE agarose gels at 3 V/cm for 6 hours or 1 V/cm overnight. The gel was then capillary blotted onto a Hybond N+ nylon membrane (Amersham) via alkaline transfer in alkaline transfer solution (1.5 M NaCl, 0.25 M NaOH) overnight. The filter was neutralised in a solution of 0.2 M Tris-HCl, (pH 8), 2x SSC and prehybridised in 20 ml of Church-Gilbert buffer (0.5 M sodium phosphate (pH 7.2), 1% (w/v) BSA, 1 mM Na₂EDTA, 7% (w/v) SDS, see Church and Gilbert 1984) at 65°C for a minimum of 4 hours. The prehybridisation buffer was discarded and 15 ml of freshly filtered Church-Gilbert buffer added together with the denatured radioactive probe and hybridised overnight at 65°C. The filter was then washed serially at 65°C in pre-warmed solutions of SSC: 0.1% (w/v) SDS in which the stringency of wash was increased by lowering the concentration of SSC. Typically, washes of 2x, 0.5x and 0.1x SSC were performed. After the final wash, the filter was wrapped in Saran wrap and autoradiographic film was exposed to the filter in an X-ray cassette for 1-7 days at -80°C.

To re-probe filters they were first stripped of radioactive probe by washing at 65°C for a minimum of 2 hours in pre-warmed filter stripping solution (2 mM Tris-HCl (pH 7.5), 0.1% (w/v) SDS, 1 mM Na₂EDTA).

2.2.1.10 DNA sequencing

Plasmid DNA obtained from Qiagen minpreps was denatured by incubation in 0.2 M NaOH, 0.1 mM Na₂EDTA for 20 minutes at 37°C. Denatured DNA was precipitated with ethanol and resuspended in 10 µl of TE. 1-3 µg of denatured plasmid were used per sequencing reaction. 1-3 pmol of sequencing oligonucleotide were annealed to 1-3 µg of denatured double-stranded DNA by heating to 70°C for 3 minutes and cooling slowly to 45°C (1°C/min) in sequencing buffer (40 mM Tris-HCl (pH 7.5), 20 mM MgCl₂, 50 mM NaCl). Labelling and termination reactions were done as described in the Sequenase v2.0 protocol (Amersham). Termination mixes were made according to the T7 sequencing kit (Pharmacia Biotech.). Termination reactions were done in microtitre plates for 4 minutes at 37°C. Samples were denatured by heating for 2 minutes at 80°C just before loading on to gels (see above).

DNA was also sequenced using the automated service provided by PNACL, University of Leicester.

2.2.2 Extraction of genomic DNA.

Extraction of genomic DNA from single adult *M. domestica*, *L. sericata* and *C. vicina* was carried out according to the protocol for *Drosophila* as described by Hamilton *et al.*, 1991.

Larger scale genomic extractions were carried out as follows: approximately ten adults were frozen in liquid nitrogen, to which 5 ml of homogenisation buffer (160 mM sucrose, 80 mM EDTA and 100 mM Tris pH 8) was added. The flies were then homogenised using a polytron electric homogeniser in 6, 10 second pulses, with 20 second rest intervals on ice. RNaseA was then added to 0.1 mg/ml, with incubation at 37 °C for 30 minutes. SDS to 1% (w/v) and proteinase K to 0.08 mg/ml were then added, with incubation at 50°C for 4 hours. The homogenate was then extracted with equal volumes of phenol-chloroform and then chloroform. The phases were mixed gently and separated using Phase Lock Gel tubes (Flowgen). The DNA was precipitated using an equal volume of ethanol, and sodium acetate to 0.3 M. The DNA was then washed in 70% ethanol and air dried before being resuspended in 0.5 ml of TE.

2.2.3 DNA amplification by the polymerase chain reaction.

Reactions were carried out in 25 µl or 50 µl volumes using 50-100 ng of template DNA. PCR buffer was prepared as described in Jeffreys *et al.*, 1990 (as an 11.1X concentrate). Alternatively, the Expand High Fidelity PCR System (Roche) was used according to the manufacturers instructions. A standard primer concentration of 300 nM was used, which was increased to appropriate levels when degenerate primers were used. Reaction conditions varied with the primers and template DNA used.

2.2.4 Construction of suppression-PCR libraries

Suppression-PCR libraries were generated from *M. domestica*, *L. sericata* and *C. vicina* genomic DNA and were used to walk both 5' and 3' into regions of unknown sequence using PCR with an adaptor primer and gene specific primers (Siebert *et al.*, 1995; Devon *et al.*, 1995; Padegimas and Reichert 1998).

Typically 5 µg of genomic DNA was restricted with either blunt cutting enzymes, or sticky ended cutters followed by Klenow mediated end-filling reactions (Sambrook *et al.*, 1989). Agarose gel electrophoresis and Southern transfer of approximately 4 µg of restricted DNA allowed estimates of the size of the fragment of interest and the average fragment size. This allowed calculation of the number of DNA ends to enable efficient adaptor ligation. Adaptor was made by coincidental annealing and phosphorylation of oligonucleotides ol992 and ol993 at 37°C for 1 hour (100 pmol ol992, 100 pmol ol993, 1X PNK forward buffer, 2 mM ATP and 40 units of PNK). The PNK was then denatured at 65°C for 20 minutes, before the adaptor was alcohol precipitated and resuspended in TE to give a concentration of 2 µM. Adaptor was then ligated to each genomic restriction in a ten-fold excess to the approximate concentration of genomic DNA ends, over night at 16°C. The ligations were then diluted 100 fold and 1 µl of these libraries was sufficient template for PCR.

2.2.5 mRNA extraction

mRNA was extracted from adult *Musca* females and from *Musca* and *Lucilia* early embryos using a Stratagene mRNA isolation kit according to the protocols supplied therein. The mRNA concentration of extracts was estimated by comparing the fluorescence of a serial dilution in EtBr, to that of known concentrations of yeast tRNA.

2.2.6 5' and 3' Rapid Amplification of cDNA Ends (RACE) - PCR

5' RACE-PCR was performed using the Gibco BRL 5' RACE System Version 2 and the protocols supplied therein. This method allows the cloning of the 5' end of a specific transcript from a short stretch of known downstream sequence. Basically, a cDNA is synthesised from an mRNA template using a gene specific primer and Reverse Transcriptase (RT). The cDNA then has a cytosine rich tag added using TdT (Terminal deoxynucleotidyl Transferase) and this allows the use of PCR to amplify a specific product using a primer based on the tag sequence (AAP) and a nested gene specific primer.

3' RACE PCR was performed using the Gibco BRL 3' RACE System according to the manufacturers instructions. Using this method the 3' end of a transcript can be cloned based on primers designed in known upstream sequence. Adaptor primer (AP) is annealed to the poly A tails of an mRNA population and cDNAs are then generated using RT. A gene specific primer is then used in combination with another primer based on the AP sequence (AUAP) to amplify a specific product using PCR.

2.2.7 DNaseI footprinting

2.2.7.1 Primer end-labelling

Primers were end-labelled with [³³P] for 30 minutes at 37°C in the following reaction (10 µl): 10 pmol primer, 0.5 µl T4 PNK, (10 units/ul), 5 µl [³³P] γ-ATP, (111 TBq/mmol), 1x PNK forward reaction buffer and water. The reaction was stopped by heating to 65°C for 15 minutes.

2.2.7.2 PCR

Labelled primers were used to generate labelled PCR probes in 50 µl reactions of the following composition: 2 – 5 ng of plasmid template, end labelled primer at 0.1 µM, opposing primer at 0.1 µM, 1.5 mM MgCl₂, 1X React IV PCR buffer, 0.2 mM dNTPs and 0.5 µl *Taq* polymerase (Advanced Biotechnologies). PCR was carried out for 22-25 cycles under appropriate conditions. The product was purified using a Qiagen PCR purification kit and then quantified on a minigel.

2.2.7.3 Protein synthesis

Calliphora and *Lucilia* Bcd homeodomain-GST fusion proteins were synthesised from pECBCD and pELBCD (table 2.1) respectively by P. Shaw, using the method described in McGregor *et al.*, 2001a.

2.2.7.4 Binding reaction and DNaseI digestion

Labelled PCR probe was incubated with protein for 30 minutes at room temperature in the following binding reactions (50 µl): 10 ng DNA, protein (at 100 nM, 10 nM or 1 nM), 100 ng dI:dC and 25 µl of 2X binding buffer (80 mM Tris pH 7.5, 0.2 M NaCl, 40% glycerol, 0.2% Triton X-100, 2mM DTT). In the control reactions either no protein was added or GST tag was added to 0.2 µg/ml. 50 µl of 50 mM MgCl₂/10 mM CaCl₂ was then added and the reactions placed on ice. DNaseI was then added to a final concentration of 0.75 µg/ml and the reactions incubated on ice for 5 minutes. Digestion was stopped by the addition of 90 µl of stop mix (0.1 M EDTA, 1% SDS, 0.2 M NaCl and 0.1 mg/ml yeast tRNA). The reactions were then extracted with an equal volume of phenol/chloroform and the DNA precipitated with two volumes of 100% ethanol. The pellet was washed in 70% ethanol, air dried and resuspended in 3 µl of sequencing gel loading buffer.

Regions of protected sequence were distinguished by the comparison of digestion patterns between samples and controls. When a protein binds to DNA this can cause the

DNA to bend and increase the exposure of nearby sequences to DNaseI. This effect is seen as hypersensitive bands on gels.

2.2.8 Yeast techniques

2.2.8.1 Transformation of yeast

Yeast transformations were carried out as described below using the LiOAc method (Ito *et al.*, 1983).

Yeast cultures were grown to saturation at 30 °C in 20 ml of YPD or appropriate drop-out media. The cells were collected by centrifugation and washed in 20 ml of water before being resuspended in 1 ml of LiOAc (0.1 M). The cells were again pelleted and resuspended in LiOAc to a total volume of 0.5 ml. The cells were then aliquoted into the 50 µl volumes used for each transformation, after which each aliquot was spun down and the supernatant removed. To the cell pellet, 240 µl PEG 4000, 36 µl LiOAc (1 M), 25 µl single-stranded sonicated salmon sperm DNA (2mg/ml) and 1-5 µg of each plasmid were added, in a total volume of 50 µl. Up to three plasmids were transformed into the yeast in a single transformation reaction. Reactions were then vortexed for 30 s and incubated at 30 °C for 30 mins. The cells were then heat-shocked at 42 °C for 20 mins, before being pelleted at 6500 rpm for 15 s in a bench-top centrifuge. The cells were then resuspended in 50 µl of water and different dilutions plated onto appropriate drop-out plates. The plates were incubated at 30 °C for 3 days.

2.2.8.2 β -galactosidase assays

Yeast cultures were grown and harvested essentially as described by Burz *et al.* (1998). Briefly, 5 ml yeast cultures were grown at 30°C in appropriate drop-out media until the cells reached a density of 1.2-1.5 OD₆₀₀. The cultures were then diluted to approximately 0.075 OD₆₀₀ and β -estradiol (Sigma) was then added to induce Bicoid expression and the cells grown to approximately 0.5 OD₆₀₀. 1.5 ml of each culture were harvested by spinning down the cell pellet, washing the cells in 1 ml of Z-buffer (0.06 M

$\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$, 0.04 M $\text{NaH}_2\text{PO}_4 \cdot 2\text{H}_2\text{O}$, 0.01 M KCl and 0.001 M $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$) and resuspending the cells in 300 μl of Z-buffer. The cell density of each culture was recorded after the wash step. The cells were then frozen in liquid nitrogen and stored at -80°C .

β -galactosidase assays were carried out essentially as described in Hanes and Brent 1989. The cells were freeze-thawed three times and 1:10 or 1:5 dilutions of the harvested cells, in a total volume of 100 μl , were used in each assay. To the cells were added 700 μl of Z-buffer/ β -mercaptoethanol and 160 μl ONPG (4 mg/ml) and the reaction was incubated at 30°C until the appearance of the yellow colour. The reaction was then stopped by the addition of 400 μl of NaCO_3 (1 M) and the time recorded. The reactions were then briefly spun down and the OD_{420} recorded. The β -galactosidase activity of each sample was quantified using the following calculation: β -galactosidase units/min = $1000\text{OD}_{420}/\text{VT OD}_{600}$, where V is the dilution factor and T is the time. For each clone three independent cultures were assayed and the results averaged.

2.2.8.3 Protein extraction from yeast cultures

Yeast cultures were grown as described above for the β -galactosidase assays and pelleted by brief centrifugation. The cell pellet was then washed in dH_2O and resuspended in 40 μl of dH_2O before being frozen at -80°C . Each sample was subsequently thawed on ice in the presence of 1x protease inhibitors (Roche) and then brought to a total volume of 100 μl by the addition of 2x Laemmli buffer (0.025 M Tris-HCl pH 6.8, 4% SDS, 20% glycerol, 2% β -mercaptoethanol, 0.002% bromophenol blue). Samples were then vortexed briefly and frozen on dry ice. Prior to loading on SDS-page gels each sample was heated to 100°C for 10 minutes.

2.2.8.4 Electrophoresis and Western analysis of protein samples

SDS-PAGE gels were cast using 'Protogel' gas-stabilised acrylamide solution (37.5:1 acrylamide:bisacrylamide) (National Diagnostics) and a Mini-Protean II gel casting unit (Biorad). The resolving gel was 10% polyacrylamide and the stacking gel 6% polyacrylamide. Pre-stained protein marker (broad range 6-175 kDa, New England

Biolabs.) and samples were denatured by heating to 100°C for 10 minutes prior to loading. Typically, up to 20 µl of the yeast protein samples described above were loaded and gels were then run as described in Sambrook *et al.*, (1989).

Proteins were transferred from SDS-page gels to PNTM membranes as described by Harlow and Lane (1988) and fixed using Ponceaus stain. Membranes were probed with 1/1600 diluted mouse anti-HA primary antibody (Boehringer) and 1/1000 HRP conjugated sheep anti-mouse secondary antibody (Amersham). This was carried according to the protocol supplied with the ECL Western blotting detection and analysis system (Amersham), which was subsequently used for the detection of bands.

2.2.9 *In situ* hybridisation of whole-mount embryos

Whole mount *in situ* hybridisations were carried out on *Musca*, *Lucilia* and *Calliphora* embryos essentially as described by Tautz and Pfeifle (1989), with modifications (Bonneton *et al.*, 1996).

2.2.9.1 *In vitro* transcription for synthesis of riboprobes

Approximately 2 µg of Qiagen-purified plasmid DNA containing a cloned insert of DNA was linearised by restriction digestion. After completion of the digestion the linearised plasmid was purified using a column from a PCR purification kit (Qiagen) and eluted in 30 µl of DEPC-treated (RNase-free) water. Linearised plasmid was then used as template for *in vitro* transcription using DIG DNA-labelling kit components (Roche) in the following reaction: linearised template, 10 µl; 1x component buffer (40 mM Tris-HCl pH 8.0, 6 mM MgCl₂, 10 mM DTT, 2 mM Spermidine, 10 mM NaCl, 0.1 units RNase inhibitor); 1x rNTPs (1 mM rATP, rGTP, rCTP, 0.65 mM DIG-11-UTP); RNasin (Promega), 20 units and T7 or T3 RNA polymerase, 40 units. Transcription was performed at 37°C for 2 hours and was stopped by heating to 65°C for 10 minutes to inactivate the enzyme. RNA was precipitated with LiCl and ethanol to remove unused rNTPs and resuspended in 50 µl DEPC-treated water with 20 units of RNasin (Promega).

The yield and integrity of product was tested by agarose gel electrophoresis and was typically about 8 µg of riboprobe per reaction.

2.2.9.2 Dechoriation

Cat meat or horse blood agar were placed on petri dishes to collect embryos from *Musca*, *Lucilia* and *Calliphora*. The embryos were removed with a brush and transferred to a wire basket, where they were rinsed with distilled water and dechorionated with household bleach (about 5% (w/v) Na(HClO)₃) in a watch-glass for two minutes. The embryos were then rinsed thoroughly with water to remove the bleach.

2.2.9.3 Fixation

Dechorionated embryos were fixed in screw-capped glass vials containing 1.82 ml of DIG-FIX solution, 2 ml heptane, 0.68 ml formaldehyde (~37% solution, stabilised with 10-15% (v/v) methanol, Sigma). Vials were placed on a rotating wheel for 30 minutes at room temperature. Fixed embryos were aspirated from the organic-aqueous interface with pasteur pipette and transferred to a fresh vial containing 2 ml methanol and 1 ml heptane. The vitelline membrane was removed by vortexing on the lowest setting on an electric vortex for 30-60 seconds. De-vitellinised embryos sink into the methanol layer and were collected by aspiration with a pasteur pipette. The embryos were washed once with 1 ml of methanol to ensure dehydration and stored in methanol at -20°C.

2.2.9.4 Pre-treatment of embryos for *in situ* hybridisation

All washes and incubations described in this and subsequent steps carried out in 1 ml volumes of liquid on a rotating wheel at room temperature unless otherwise stated. Fixed embryos were re-hydrated by washing for 3 minutes in methanol:PBT (1:1) then twice in PBT (phosphate buffered saline with tween; 130 mM NaCl, 70 mM Na₂HPO₄, 30 mM NaH₂PO₄ and 0.1% (w/v) Tween). Rehydrated embryos were post-fixed for 20 minutes in PBT: 5% formaldehyde. Post-fixation was stopped by rinsing embryos briefly in PBT,

then embryos were washed twice for 5 minutes each in PBT. The embryos were then washed for 5 minutes in PBT plus 5 µg of proteinase K.

Proteinase K digestion was stopped by rinsing briefly in PBT and then washing twice for 5 minutes in PBT plus 2 mg of glycine. The embryos were then post-fixed a second time for 20 minutes in PBT; 5% (v/v) formaldehyde which was stopped by rinsing briefly in PBT and then washing twice for 5 minutes in PBT.

2.2.9.5 Pre-hybridisation and hybridisation

Pre-treated embryos were washed in 0.5 ml of a 1:1 solution of PBT: Hyb-D (50% (v/v) deionised formamide, 5x SSC, 0.1% (w/v) Tween-20, 1 mg/ml yeast tRNA, 2% (w/v) DIG blocking reagent) for 15 minutes. The embryos were then transferred to 0.5 ml of Hyb-D and incubated for 30 minutes at 55°C, then 65°C for 1 hour to denature endogenous enzymes. The embryos were then returned to 55°C and incubated for 30 minutes.

Embryos were hybridised in 0.1 ml of fresh Hyb-D together with 10-50 ng of riboprobe overnight at 55°C.

2.2.9.6 Pre-immunoreaction and immunoreaction

Embryos were washed to remove unbound riboprobe. Washes in 0.5 ml solution of 4:1, 3:2, 2:4, 1:4 Hyb-D:NTB (150 mM NaCl, 100 mM Tris-HCl (pH 7.5), 0.1% (w/v) Tween-20, 0.2% (w/v) DIG blocking reagent) were carried out for 10 minutes each at 60°C. After a final wash for 10 minutes in 0.5 ml NTB at 60°C, embryos were pre-incubated for 4 hours in 1 ml NTB plus 2% (w/v) goat serum (Boehringer Mannheim) at 4°C.

Embryos were incubated overnight at 4°C in 1 ml NTB, 2% (v/v) goat serum, Anti-DIG-AP antibody (polyclonal Fab fragments conjugated to alkaline phosphatase, Boehringer Mannheim) diluted 1/2000.

2.2.9.7 Colour staining

Embryos were washed three times in PBT for 30 minutes each at 4°C, then washed twice for 10 minutes each at 4°C and finally once at room temperature in 1 ml of colouration solution (CS; 0.1 M NaCl, 0.1 M Tris-HCl (pH 9.5), 50 mM MgCl₂, 0.1% (w/v) Tween-20). Colouration reaction was initiated by incubating the embryos in 1 ml CS + 0.45 µl NBT + 3.5 µl X-phosphate (NBT from the DIG kit: nitroblue tetrazolium salt dissolved in 70% (v/v) dimethylformamide). Staining was checked after 2 hours and stopped after 3-4 hours by washing embryos twice in 1 ml PBT. Embryos were dehydrated by rinsing in 1 ml ethanol:PBT (1:1) then twice in 1 ml absolute ethanol.

2.2.9.8 Permanent mounting, microscopy and photography

Dehydrated embryos were washed in 0.5 ml 1:1 ethanol:Spurr (Spurr: low viscosity embedding medium (hard composition), Sigma) and then 0.3 ml of Spurr. The embryos were left to settle to the bottom of the tube and were taken up in a small volume of Spurr (about 60 µl) and mounted on a glass slide. Slides were incubated overnight at 65°C and analysed using a Nikon Optiphot-2 microscope. Photographs were taken at 200x magnification with a Nikon exposure unit with automatic exposure times, typically 60-120 ms.

2.2.10 Computer analysis

Sequence alignments were made using the Clustal W program (Thompson *et al.*, 1994) or the GCG algorithm PILEUP and dotplots were generated using COMPARE and DOTPLOT; all of these programs are available on the GCG (1994) package, version 8.1.

Consensus sequences for the Bcd-binding sites in the *Lucilia* and *Calliphora hb* promoters were calculated from these alignments by the frequency of bases at each position. If a position did not have one particular base present in 50% or more of the aligned sequences then that position was left ambiguous in the consensus sequence.

PEPTIDESORT (also in GCG) was used to predict the molecular weights of the Bcd proteins from each species. NIH Image 1.61 was used to compare the density of hybridising bands in Western analysis.

The SIMPLE 34 program (Hancock and Armstrong 1994) was used in the analysis of sequence simplicity and this is explained in full in 5.2.1.

Chapter 3
Characterisation of *hb* genes in *Calliphora*,
Lucilia and *Musca*

3.1 Introduction

3.1.1 Aims

Are the roles of the *Calliphora* and *Lucilia* *hb* orthologs functionally equivalent to those in other higher Dipteran species such as *Drosophila* and *Musca*? This can be determined by examining the blowfly Hb protein sequences for functional domains previously characterised in *Drosophila* and analysing the *hb* expression patterns in these species. The *hb* coding region sequences could then be used as a platform to walk into the P2 promoters and investigate the regulation of *hb* in *Calliphora* and *Lucilia*.

3.1.2 Experimental overview

The region of *hb* that codes for the N-terminal zinc-finger domain had previously been sequenced in *Calliphora* and shown to be highly conserved in comparison to the *Drosophila* and *Musca* orthologs (Sommer *et al.*, 1992; accession number LO1591). Genomic libraries were unavailable for either *Calliphora* or *Lucilia*; therefore, the *Calliphora* *hb* zinc-finger sequence was used as the basis for PCR-based strategies to clone and sequence *hb* and its regulatory regions in these species. Professor Bownes (University of Edinburgh) did kindly donate a *Calliphora* genomic library, but by that time the *hb* promoter region had already been sequenced in this species.

DNA and RNA probes were then generated from the *Calliphora* and *Lucilia* *hb* sequences to perform Southern analysis to determine *hb* copy number and *in situ* hybridisation to characterise *hb* expression patterns in these species.

3.1.3 suppression PCR (sPCR)

sPCR is a method that can be used to walk from regions of known sequence in genomic DNA into regions where the sequence is not known (see 2.2.4). This technique was chosen to clone the *hb* genes in *Calliphora* and *Lucilia* because it is quicker than traditional methods such as phage library construction and screening. Since the *hb* genes of other Dipterans only have a single short intron in the 5' UTRs, the complete coding

region of *hb* could be readily amplified in *Calliphora* and *Lucilia* without having to PCR across numerous introns. Successful sPCR depended on the proximity of restriction sites to known sequence at distances conducive to amplification using standard PCR. Therefore, the sPCR libraries were constructed using *Calliphora* and *Lucilia* genomic DNA cut with a range of restriction enzymes reflecting the differing GC contents in the various regions of the gene.

Although it was necessary to adjust sPCR conditions for different primers and walks, it was found that specific products could be obtained from all the libraries. Artifactual smears of high molecular weight were consistently generated by the adaptor primer AOL995 from the *Dra*I, *Eco*RV, *Hinc*II and *Ssp*I *Lucilia* libraries and the *Dra*I, *Hinc*II and *Ssp*I *Calliphora* libraries (these libraries were more concentrated than the others used). These artefacts were probably caused by amplification from restriction fragments incompletely ligated to the adaptor resulting in incomplete suppression, which competes with the amplification of specific products. Therefore, single primer controls and reamplification with nested gene specific primers were routinely performed which allowed specific products to be identified.

3.2 Results

3.2.1 Cloning of *hb* from *Calliphora*

PCR was initially used to extend the known *Calliphora hb* sequence using a specific primer (CZR) based on the zinc-finger sequence and a degenerate primer (ABF) based on the A-box sequence (figure 3.1A and see 3.3.1). A product of approximately 900 bp was amplified and the sequence of the 3' end of this product matched the known *Calliphora hb* sequence (see above).

To sequence the 5' end of the coding region, sPCR was performed using the gene specific primer CALHB (figure 3.1A) and a product of approximately 600 bp was amplified, from a *Bgl*III *Calliphora* library, which overlapped with the ABF-CZR sequence.

The translation start site of *Calliphora hb* was determined by the first in frame AUG and conservation of the first six amino acids with those of *Musca hb* and more extensive conservation with the 5' end of the *Lucilia hb* coding region (see below and figure 3.2). Primers (LSTSF, CALHB, BFHB3, BFHB5, BFBH4 and OCZ) were then used to amplify directly from genomic DNA to confirm the sequences of the above PCR products (figure 3.1A).

To sequence the 3' half of the *Calliphora hb* coding region, sPCR was performed using BFHB1 and the nested primer BFHB2. This resulted in an amplicon of approximately 1.2 kb, from an *SspI* *Calliphora* library, which extended 3' to the C-terminal zinc-finger coding domain. Further sPCR was performed using primers CALFA and CALFB, which resulted in a 650 bp product from a *DraI* *Calliphora* library. This product, when compared to the *Musca hb* gene, was shown to contain the 3' end of the coding region sequence and part of the putative 3' UTR sequence downstream of an in frame stop codon. Sequences were confirmed directly from genomic DNA using primers CALHF, CALHR, CCF1, CCF2, CCR1 and CCR2 in combination with the primers described above (figure 3.1A).

3.2.2 Cloning of *hb* from *Lucilia*

To facilitate the cloning of a *hb* gene from *Lucilia*, PCR was performed using the degenerate primers LSZF and LSZR, whose design was based on *Musca* and *Calliphora hb* N-terminal zinc-finger domain amino acid sequences chosen to minimise degeneracy (figure 3.1B). This resulted in the amplification of a 116 bp product from *Lucilia* genomic DNA. Sequencing of this product and alignment with the *Musca* and *Calliphora hb* zinc finger domain sequences showed that the amplicon was a *hb* zinc finger domain fragment.

With the aim of cloning a larger region of *Lucilia hb*, gene specific primers were designed to amplify between the zinc finger and the N-terminal A-box domain (primers ABLS and LSHZR in figure 3.1B). A 900 bp product was amplified from *Lucilia* genomic DNA using these primers and BLAST searches demonstrated that the sequence

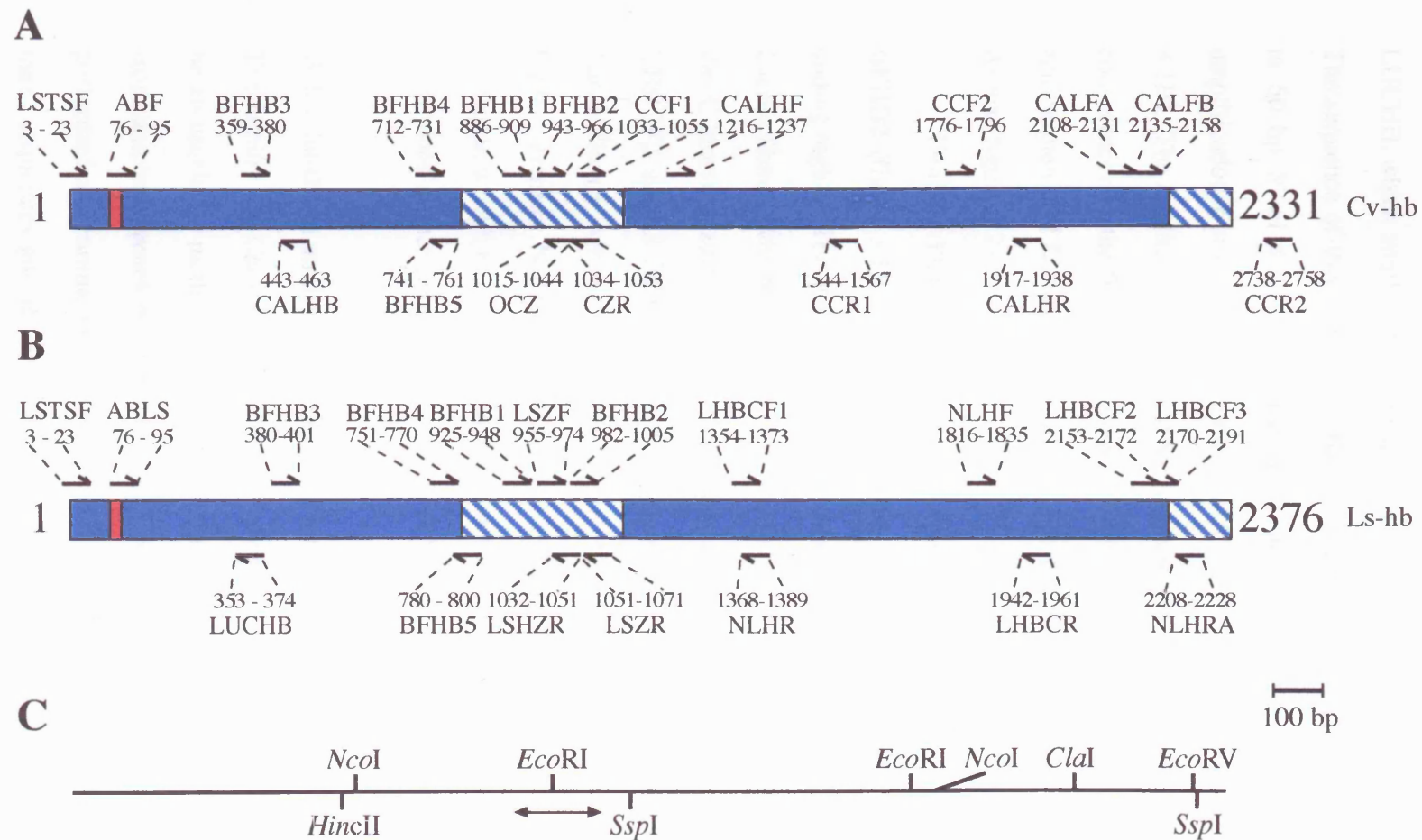


Figure 3.1. Positions of PCR primers used to amplify the *hb* gene in *Calliphora* (A) and *Lucilia* (B) and restriction sites used in Southern analysis (C).

A and B. The blue rectangle represents the entire coding region in both species from the translation start site (1) to the last codon (2231 and 2276 respectively). The hatched boxes represent the two zinc-finger encoding domains and the red box the A-box encoding region. Arrows represent the direction of each primer and the numbers are the primer positions in the coding region sequence (the accession number of the *Lucilia* sequence is AJ301662). The primer sequences are listed in appendix A. **C.** The double headed arrow represents the position of the probe used in Southern analysis. *Calliphora* restriction sites are above the line and *Lucilia* sites are below.

of this product was highly related to *Musca hb* (highest BLAST score; $p = 2e^{-56}$) and thus that it was part of *Lucilia hb*.

To walk further 5' in the *Lucilia hb* gene, sPCR was performed using the primer LUCHB, which amplified a product of approximately 400 bp from a *DraI Lucilia* library. The sequence of this product overlapped with the ABL5-LSHZR sequence and extended to 50 bp 5' of the putative translation start site. This sequence was confirmed by amplification from *Lucilia* genomic DNA using primers LSTS and LUCHB (figure 3.1B). The *Lucilia hb* translation start site was determined by the first in frame AUG and conservation of the first six amino acids with those from *Musca hb*. In addition, the amino acid sequence of *Lucilia Hb* was identical to that of *Calliphora Hb* up to and including the A-box (figure 3.2).

Further sPCR experiments were then performed using the primers BFHB1 and BFHB2 (figure 3.1B) to sequence the *Lucilia hb* gene 3' of the N-terminal zinc-finger coding region. This generated a specific product of approximately 1.2 kb from an *EcoRV Lucilia* library, which overlapped with the known sequence and resulted in a walk as far as the C-terminal zinc finger domain. Based on the sequence of this product, primers LHBCF2 and LHBCF3 were used to generate a 400 bp sPCR product from a *DraI Lucilia* library, which contained the 3' end of the *Lucilia hb* gene and part of the putative 3' UTR. Primers NLHR, NLHF, NLHRA, LHBCF1 and LHBCR were then used in combination with the above primers to subclone regions of the *Lucilia hb* coding region to confirm the sequence by amplification directly from genomic DNA.

3.2.3 Southern analysis of *Calliphora hb* and *Lucilia hb*

The *Lucilia* and *Calliphora hb* sequences were assembled from PCR products, which can be an unreliable method to clone genes since artefacts can be generated by amplification from unrelated genes with similar conserved domains. Therefore, Southern analysis was performed on genomic DNA from both *Calliphora* and *Lucilia* to confirm the accuracy of these sequences and to determine the copy number of *hb* in these species. This was

[illegible]

Figure 3.2 Clustal W alignment of the Hunchback protein sequences from *Drosophila melanogaster* (Dmhb, Tautz *et al.*, 1987), *Musca domestica* (Mdhb, Bonneton *et al.*, 1997), *Calliphora vicina* (Cvhb), *Lucilia sericata* (Lshb, McGregor *et al.*, 2001b) and *Megaselia adita* (Mahb, Stauber *et al.*, 2000). Conserved residues are highlighted in grey. Closed boxes are the zinc-finger domains and dashed boxes are the A, B, C, D, E and F boxes (see text).

important because *hb* has been duplicated in an annelid and one copy is a pseudogene (Savage and Shankland 1996).

Calliphora and *Lucilia* genomic DNA were Southern blotted (see 2.2.1.9) with a 162 bp *hb* probe spanning the N-terminal zinc-finger region from each species (figure 3.1).

Hybridisation patterns for the *Calliphora hb* zinc-finger region are shown in figure 3.3A and the restriction site positions are illustrated in figure 3.1C. The *Cla*I and *Eco*RV digests gave bands of approximately 6.5 kb and 6 kb respectively. This was as expected since these two enzymes were predicted to cut only once in the *Calliphora hb* coding region and not in the probe sequence. *Eco*RI digestion resulted in two bands of approximately 0.7 kb and 6 kb as predicted. This enzyme cut within the probe sequence and also approximately 700 bp further 3'. The *Calliphora hb* sequence contained two *Nco*I restriction sites on either side of the probe sequence and these were also confirmed by probe hybridisation to a product of approximately 1.1 kb.

The *Lucilia hb* zinc-finger hybridisation patterns are shown in figure 3.3B and restriction site positions in figure 3.1C. *Hinc*II was predicted to cut only once in the *hb* coding region (5' of the probe sequence) and the probe hybridised to a band of approximately 2 kb. There were no predicted *Dra*I sites in the coding region; therefore, hybridisation with the zinc-finger probe has produced a band of approximately 2.5 kb containing the entire coding region of *hb* in this species. There is a *Dra*I site approximately 50 bp upstream of the translation start site and another in the 3' UTR, as determined by sequencing products from the *Dra*I sPCR library. The probe also hybridised to an *Ssp*I fragment of approximately 1.2 kb, which confirmed the site for this enzyme immediately 3' of the probe sequence and indicates the presence of another *Ssp*I site further upstream in the intron.

Southern analysis of *Lucilia* and *Calliphora* genomic DNA therefore confirmed that the *hb* gene was single copy in these species and that the sequences generated by PCR were accurate with respect to genomic DNA.

3.3.4 *hb* protein sequence comparison and analysis

Calliphora hb has an ORF of 2131 bp and potentially encodes a protein of 710 amino acids, while *Lucilia hb* has an ORF of 2376 bp long encoding a protein of 791 amino acids. The protein sequences of *Calliphora hb* and *Lucilia hb* are shown aligned with those of other species in figure 3.4. The *Calliphora* and *Lucilia* proteins display 94% identity with each other and 71% identity with the *Drosophila* protein.

The *hb* protein was conserved between these species for example, the zinc finger domains are 85% and 96% conserved between *Calliphora* and *Lucilia* respectively. Conserved domains include the A, C, D, F and F boxes, some of which have been shown to be important for *hb* functions (see 3.3.1); however, the B box, which has been shown to be important for the conserved zinc finger protein domain (Hafiz et al., 1997), is not conserved.

Figure 3.3 shows the results of Southern analysis of *Calliphora* and *Lucilia* genomic DNA digested with *Cla*I, *Eco*R1, *Eco*RV and *Nco*I for *Calliphora* and *Hinc*II, *Dra*I and *Ssp*I for *Lucilia*. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the Southern analysis of *Calliphora* and *Lucilia* genomic DNA digested with *Cla*I, *Eco*R1, *Eco*RV and *Nco*I for *Calliphora* and *Hinc*II, *Dra*I and *Ssp*I for *Lucilia* are shown in figure 3.3.

3.3.5 Walking from the *hb* coding region into the 5' UTR in *Calliphora* and *Lucilia*

The first step in the strategy to clone the *hb* gene was to walk 3' from the coding region into the upstream region. This was done using aPCR. A primer (BLDSR) was designed from part of the *hb* coding region and used for both *Calliphora* and *Lucilia*. When a PCR was performed with the BLDSR primer and a product of approximately 450 bp from both species was obtained.

Figure 3.4 shows the results of the PCR. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

The results of the PCR are shown in figure 3.4. The results show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species. The results also show that the *hb* gene is present in both species and that the *hb* gene is conserved between the two species.

Figure 3.3 Southern analysis of: **A.** *Calliphora* genomic DNA digested with *Cla*I (C), *Eco*R1 (E), *Eco*RV (Ev) and *Nco*I (N). **B.** *Lucilia* genomic DNA digested with *Hinc*II (H), *Dra*I (D) and *Ssp*I (S). Hybridised with a 162 bp probe spanning the *hb* N-terminal zinc finger coding region of each species (892-1053 in *Calliphora hb* coding sequence and 930-1092 in *Lucilia hb*, see figure 3.1). The size markers are given in bp.

3.2.4 *hb* protein sequence comparison and analysis

Calliphora hb has an ORF of 2331 bp and putatively encodes a protein of 777 amino acids, while *Lucilia hb* has an ORF of 2376 bp long encoding a protein of 791 amino acids. The protein sequences of *Calliphora* and *Lucilia* Hb are shown aligned with those of other species in figure 3.2. The *Calliphora* and *Lucilia* proteins display 94% identity with each other and 72% and 74% identity respectively with Hb from *Drosophila*. This alignment illustrates that the functional domains of the Hb protein are conserved between these species; for example, the N- and C-terminal zinc-finger domains are 85% and 96% conserved between the Dipteran species. Other conserved domains include the A, C, D, E and F boxes, some of which have been implicated in Hb functions (see 3.3.1); however, the B box, which has previously been described as a conserved zinc-finger protein domain (Tautz *et al.*, 1987), has diverged between these species.

3.2.5 Walking from the *hb* coding region into the 5' UTR in *Calliphora* and *Lucilia*

The preliminary strategy to clone the blowflies *hb* P2 promoters was to walk 5' from the coding region into the promoter via the leader and intron using sPCR. A primer (BLOS R) was designed from part of the coding region conserved in both *Calliphora* and *Lucilia*. When sPCR was performed with this primer it resulted in products of approximately 450 bp from both *Dra*I and *Ssp*I *Calliphora* libraries, but no products from any of the *Lucilia* libraries. The sequences of these two *Calliphora* products overlapped with the known *Calliphora* sequence and extended 5' across the leader and into the intron. The 3' intron boundary in *Calliphora hb* was determined by the conservation of the 3' splice junction sequence between *Calliphora* and *Musca* (see below and figure 3.6B). Further primers were designed in both *Lucilia* and *Calliphora* to walk 5' across the intron using sPCR; however, these sPCRs were unsuccessful under a range of conditions. This inhibition of PCR may have been caused by runs of As and Ts in the *hb* leader and intron, in combination with a lack of restriction sites, up to 1 or 2 kb upstream, for the enzymes used to generate the sPCR libraries. Indeed, the *hb* intron in *Calliphora* and *Lucilia* may have

been expanded in size in comparison to *Musca*, which would have required multiple rounds of sPCR to cross, or long range amplification conditions.

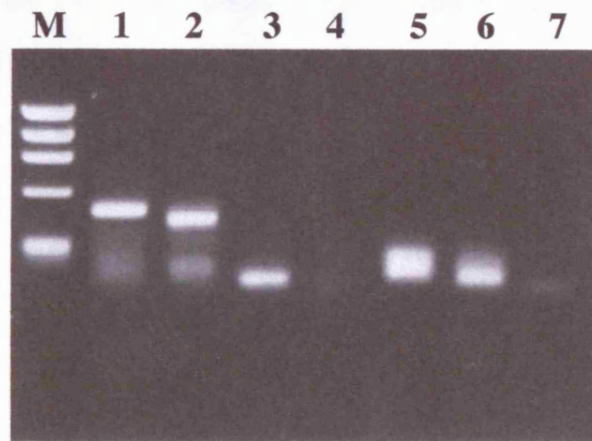
3.2.6 Mapping the 5' end of the *Lucilia hb* mRNA using RACE PCR

A 5' RACE strategy was adopted to determine the end of the *hb* transcript and sequence the 5' UTR of this gene in *Lucilia*. This modus operandi would then permit walking into the *hb* promoter from the leader using sPCR without having to sequence the intron.

mRNA was extracted from approximately 0.3 g of 1-6 hour old *Lucilia* embryos (see 2.2.5). Primers for first strand cDNA synthesis (LSHBRT) and PCR (LSHBRA and JALR) were designed based on the *Lucilia hb* coding region sequence (see appendix A and figure 3.4B). First strand cDNA synthesis was performed on approximately 300 ng of mRNA using the *Lucilia hb* specific primer LSHBRT (see 2.2.6). cDNA products were subsequently purified and TdT-tailed according to the kit protocols. PCR was then performed on the tailed cDNA template using the specific primer LSHBRA and the tail primer AAP. The first round of PCR produced a very faint product of approximately 450 bp. Reamplification with either the first round primers or a nested primer (JALR) in combination with AAP, resulted in specific products of approximately 450 bp (figure 3.4A). The sequence of the 3' end of this product overlapped with the known *Lucilia hb* sequence (see 3.2.2) and was approximately the same length as the *Musca hb* 5' UTR (359 bp and 452 bp respectively). Indeed, the 5' end of the *Lucilia* RACE product was conserved at 28 out of the first 34 bases (82%) compared to the *Musca hb* P2 transcription start site (figure 3.4B), although the leaders were only 49% conserved in total between these species. This was evidence that the *Lucilia* RACE product had the complete 5' end of the proximal (P2) *hb* transcript.

The putative splice site was also identified in the *Lucilia hb* leader by the conservation of 11 bp with the *Musca hb* splice site (figure 3.4B) and conservation of the 3' part of this sequence with the putative *Calliphora hb* 3' splice junction as determined by sPCR (see above). It was important to identify this site to enable primers to be designed in the 5' part of the leader to facilitate walking into the P2 promoter using sPCR, since it had

A



B

ATCAGTTGCA TTCTAGCATC AATACGATAA ACATCTCTCT CTTTAGTATT
 TTCTTAAACG GAATACGAAA AGATTTTAAA AAACATAAAA AATTTATAGT
 GCGGAAAAAG ATTTATAAAA TCTTTTTTTAA ACAAATAAAC TTCCTAGAAG
 TTAAAACATT AACATTTAAA TTTCAACGAA TTTATTTAAA TTATAAAAAA
 ACCAAAACAA AAACAACAAG ACATCACTGT GGATATAATT TAAAAAAAAC
 TGCCAAAGAT CTCCTCTTTT GGTAATTTTT TATACCCAC CTTATACCAA
 ACCCCCCTTT AAACCTAAAC CCCTGAAAAC ACACACACAC ATTCTAATAA
 CTGCCAACAT GCAGAATTGG GACACCATGC AAACCACCGC TAATTATGTG
 GAACACAATA ACTGGTATAA CAATATGTTT GCCGCCA

1 2

Figure 3.4 Mapping of the *Lucilia hb* transcription start site using 5' RACE PCR
A. Results of reamplification of primary PCR and controls with specific primers and AAP. M = markers (1353, 1078, 872, 603 and 310 bp from top to bottom). Lanes 1 and 2, primary reaction reamplified with primers LSHBRA and JALR respectively. Lanes 3 and 4, untailied cDNA control reamplified with primers LSHBRA and JALR respectively. Lane 5, tailed mRNA control and lane 6, mRNA control, both reamplified with LSHBRA. Lane 7, no template control with LSHBRA. **B.** Sequence of the reamplified product (from lane 1 above). Underlined in blue are the regions corresponding to primers 1. LSHBRA and 2. JALR. The translation start site is represented by the black arrow and the transcription start site by the red arrow. Nucleotides boxed in red are conserved in comparison to the *Musca hb* transcription start site. Nucleotides in the hatched red box are conserved in comparison to the *Musca hb* splice site, whose position is represented by the red triangle (see figure 3.6B).

proved problematic to cross both the *Lucilia* and *Calliphora hb* introns using sPCR (see above).

3.2.7 Cloning of the *Lucilia hb* putative P2 promoter region

Primers were designed based on the *Lucilia hb* RACE product sequence 5' of the putative splice site (see above and figure 3.5A). No specific products were amplified initially using primer LRPRO1, but reamplification with primer LRPRO2 resulted in a number of possible specific products from various *Lucilia* sPCR libraries. Based on the size and quantity of product, those of approximately 650 bp and 400 bp from the *FspI* and *SspI* *Lucilia* libraries respectively were selected for further analysis. The sequences of both these products overlapped with the transcription start site sequence and with each other (figure 3.5A). A second round of sPCR was performed using primers LRPRO3 and LRPRO4 and this resulted in a 350 bp product from a *BamHI* library, which overlapped with the previous sPCR product (figure 3.5A). Again primers were designed based on this sequence (LRPRO5 and LRPRO6) and used in sPCR which generated a product of approximately 300 bp from the *SspI* library (figure 3.5A).

To confirm that the sequences of the sPCR products were contiguous with respect to genomic sequence, PCR was performed directly on *Lucilia* genomic DNA with primers LPROF and LPROF2 in combination with the above primers (figure 3.5A). These sequences corroborated the sPCR product sequences; however, neither the genomic sequences nor the sPCR product sequences contained an *FspI* site or a *BamHI* site, from which libraries products were obtained. It appears that the strain of *Lucilia* (origin Bristol) used to make the *FspI* sPCR library contained an *FspI* site but the strain from which genomic DNA was isolated (origin Leicester) to confirm the sequences did not. The latter *Lucilia* strain did have 5 out of 6 bp of the *FspI* restriction site sequence conserved in the corresponding position in its *hb* promoter. Such differences in sPCR products have also been seen between strains of *Musca* (P. Shaw personal communication). There are also no close matches to a *BamHI* site in the sequences. It appears that the *BamHI* library sPCR product was serendipitously amplified by the adaptor primer AOL995, facilitated by the

low annealing temperature used in that particular amplification, which allowed annealing of this primer to a genomic region similar in sequence to the adaptor sequence.

3.2.8 Cloning of the *Calliphora hb* putative P2 promoter region

It was observed that the *hb* transcription start sites were conserved between *Lucilia* and *Musca* (figure 3.4B); therefore, it was probable that the *Calliphora* transcription start site was similarly conserved. sPCR was performed on the *Calliphora* sPCR libraries using primer LRPRO2 (based on the *Lucilia hb* transcription start site sequence). This resulted in a product of approximately 300 bp from the *Calliphora SspI* library (figure 3.5B). The sequence of this product was similar (66% identity) to the *Lucilia* sequence directly upstream of the *hb* transcription start site and the sequence and position of the putative TATA box was conserved in comparison to both the *Lucilia* and *Musca* sequences (figure 3.6A). Further primers were designed based on this *Calliphora* sequence (CVPRO1 and 2) and a second round of sPCR was performed on the *Calliphora* libraries. Products of approximately 200 bp and 1500 bp were amplified from the *DraI* and *HincII* libraries respectively. The sequences of these products overlapped with each other and with the previous *Calliphora* sequence (figure 3.5B). To verify the integrity of the sPCR sequences, PCR primers were designed to amplify these regions directly from genomic DNA in combination with the above primers (CVPROF2, CVPRO4, CVPROF3 and CVPROF5; see appendix A for sequences), as shown by the subclones in figure 3.5B.

To demonstrate that this putative promoter region was contiguous with the *Calliphora hb* coding region a 2-step PCR strategy with an elongation temperature of 65°C was employed to amplify between these regions using primers BFHB7 and BLOSR. This resulted in a product of approximately 1100 bp. Sequencing of the ends of this product demonstrated that it spanned from the transcription start site to the *Calliphora hb* coding region.

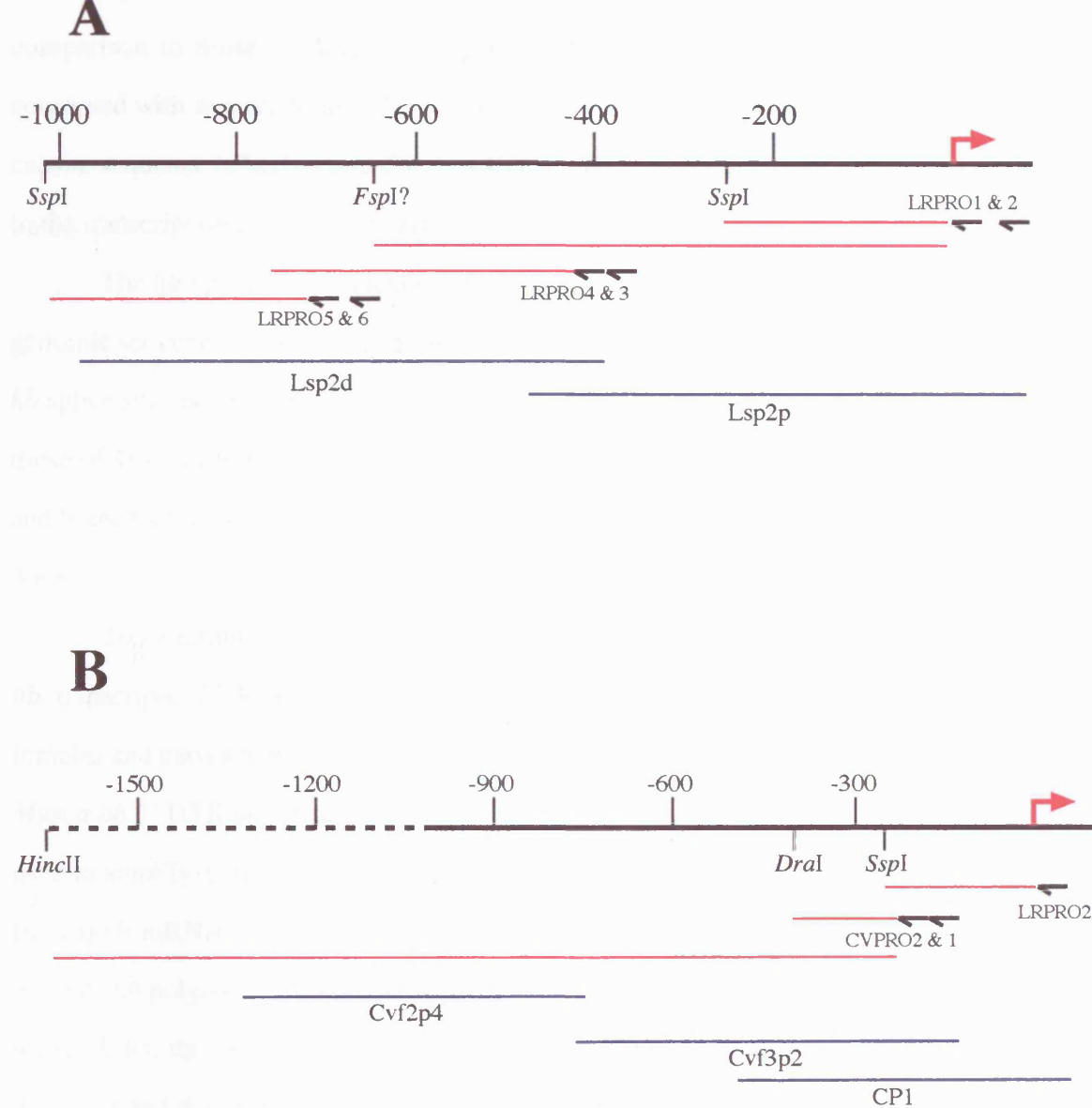


Figure 3.5 Cloning of the putative *hb* P2 promoter regions from *Lucilia* (A) and *Calliphora* (B). The scale represents the distance in bp upstream of the transcription start sites of each species (red arrows). The dashed line scale bar in B illustrates unpublished sequence. The positions of relevant restriction sites are shown beneath the scale. Primer positions are represented by the black arrows (see appendix A for sequences). Red lines represent the sPCR products and blue lines represent the genomic clones generated to confirm the sequences. Accession numbers are AJ319595 (*Calliphora*) and AJ319596 (*Lucilia*).


3.2.9 *hb* transcript structure in *Musca*, *Calliphora* and *Lucilia*

The *Calliphora* and *Lucilia hb* transcript cap site and TATA box sequences are shown in comparison to those of *Musca* in figure 3.6A. The *hb* capsites in the blowflies are conserved with respect to the *Musca hb* capsite and these match the arthropod consensus capsite sequence (Cherbas and Cherbas 1993). The positions of the TATA boxes relative to the transcription start sites are also conserved.

The *hb* splice site sequences of *Calliphora* and *Lucilia* were inferred from the *hb* genomic sequences from these species and the *Lucilia* 5' RACE product (see above). The *hb* splice sites and putative *Calliphora* branch point sequences are similar in comparison to those of *Musca hb* (figure 3.6B). In addition, the consensus *D. melanogaster* splice sites and branch sites (Mount *et al.*, 1992) are also conserved in *Lucilia* and *Calliphora* (figure 3.6B).

To determine the position of the 3' ends of both the maternal and zygotic *Musca hb* transcripts, 3' RACE PCR was performed on *Musca* mRNA isolated from adult females and early embryos (see 2.2.5). Primers (M3R1 and M3R2) were designed in the *Musca hb* 3' UTR upstream of the two putative polyadenylation signals (figure 3.7A) and used to amplify a specific product of approximately 750 bp from the cDNAs synthesised from both mRNA templates (figure 3.7B). The end of this product corresponded to the more distal polyadenylation signal sequence. Sequencing of the other products on this gel revealed that they were rRNA artefacts as shown by BLAST searches. This experiment demonstrated that both the maternal and zygotic *Musca hb* transcripts use only the most distal of the putative polyadenylation signals previously described (Bonneton *et al.*, 1997). This is similar to *Drosophila hb* where only the distal polyadenylation signal of two is used (Tautz *et al.*, 1987) and therefore, this mechanism is probably conserved in all Dipterans including *Calliphora* and *Lucilia*. The position of the polyadenylation signal in *Musca* means that *hb* has a 3' UTR of 927 bp in this species. This suggests that only part of the putative *hb* mRNA 3' UTR has been sequenced in *Calliphora* and *Lucilia* (see above). These sequences are shown aligned with part of the *Musca* sequence in figure 3.8A and illustrate that *Calliphora hb*, like *Musca*, has two Nanos response elements

A

-30 +1 

Musca GTTTAAATATCTGAGG--GTTTTTTGGTTT-GAGC**ATCAGTTGCATTTCAG**

Lucilia GTTTAAATA-CTT-GGCTGATTTTGT-TTTACAGC**ATCAGTTGCATTCTAG**

Calliphora GTTTAAATA-CTC-GTGTGATTTTGT-TTTACAGC**ATCAGTTGCATTCTAG**

Arthropod consensus cap site sequence:

-10....gcaTCAGT....+10

B

D. melanogaster consensus splicing sites:

5' MAG | gtragtw.....wctaty.....tttttyyyttncag | RT 3'

Musca hb splicing sites:

5' AAG | gtaccta.....tgtaatt.....tttttttttttttag | AT 3'

Calliphora hb putative splicing sites:

5' AAG | gtacttt.....ttaataa.....ttatttcatttttag | AT 3'

Lucilia hb putative splicing sites:

5' AAG | | AT 3'

Figure 3.6 Summary of *hb* transcription start sites (A) and splicing sites (B) from Dipterans. A. Alignment of *hb* transcription start sites from *Musca*, *Lucilia* and *Calliphora*. Numbering is with respect to the transcription start site (marked with an arrow). Nucleotides in bold are transcribed. A putative TATA box is shown underlined. Underneath the alignment is shown the arthropod consensus capsite sequence (Cherbas and Cherbas 1993).

B. The putative *Calliphora* and *Lucilia hb* intron splice sites are shown alongside those of *D. melanogaster* and *Musca* (Shaw 1998). The vertical lines represent splice boundaries with exonic sequences in upper case and intronic sequences in lower case. Dots denote intervening intronic sequences. Putative branch points are shown in italics and invariant bases are underlined in the *D. melanogaster* sequence (Mount *et al.*, 1992).

A

```

4561 gttttttttt tttttatttt tggtatcctt tttttcgttg ctttgaattg taaataatta
4621 tccccacccc ccaacaaaaa aattctcagt agcttaggag gggcctgcct ttaggtcacc
4681 ttaagtgaat cgttgtcatg aattgtaaat atgaaaatca acatttagtt ttaagttaaa
4741 tatttttttt ttaaattttt aataatttta agttttaaag gttccttctc caacaacaac
4801 aaaaaattaa taatattttt gtaatatttg tataatctta aaaaacaaa acaatatgac
4861 gtattatata attgatattt ttagcaaaaa acgaacaaac cgaacccatg ccccaaaact
4921 agttaaaatt ttttagtttt aaaaataatt aattttatat tttttttaat ttttaatttt
4981 ctaattttta caaaaacctt gaaaggttta aaaattcttt agtcataagt ttataatttt
5041 ttaaattact tagaacataa aaataattat tttttttatt attattttcc ataataatca
5101 acaaatcaaa aaatttaaaa attctactct ttttatttaa aaaaaatcaa ttttaacttt
5161 aaaacaacca accaaaccaa acaaaactca cgaaaatcaa aatcaaacaa aaaccctga
5221 aacaaagcca agtcattagt ttaggttaaa taataaaata gttattatta ttattattat
5281 ttttataaaa aaaattttat gtaaaccaag ggacatcatc tattcgagag agagaaaaaa
5341 caaagcaaaa aaactatacc aaaattctca acttttgtgt tatattttta gtccatgttt
5401 ttttattaaa aaaaaaaaaa aaaattttat tttacaaata aaaaacaaaa gaaatcacat

```

B

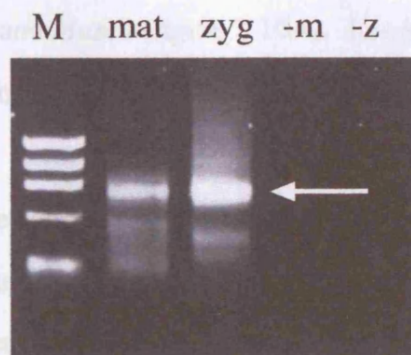


Figure 3.7 *Musca hb* 3' RACE PCR

A. Sequence of the *Musca hb* 3' UTR. Numbering is according to the sequence in Bonneton *et al.*, 1997 (accession number Y13050). The putative polyadenylation signals are underlined in italics. The red arrows represent the positions of primers M3R1 and M3R2 (nested).

B. Results of 3' RACE using *Musca* maternal mRNA (mat) and zygotic mRNA (zyg) derived cDNAs as templates. The negative controls shown contained only mRNA. The white arrow identifies the *hb* specific product (see 3.2.9). Markers (M) are 1353, 1078, 870, 603 and 310 bp from top to bottom.

(NREs) and that the proximal NRE is also present in *Lucilia*. It is probable that *Lucilia* also has the more distal NRE, but the sequence obtained in this work doesn't extend that far 3' in this species. NREs are conserved in many insects (figure 3.8B), which suggests a common mechanism for the regulation of *hb* in these species (see 3.3.2).

This analysis of the *hb* transcript in *Calliphora* and *Lucilia* has determined that it is similar in structure to the zygotic *Musca hb* transcript. A comparison of the *hb* transcripts from the blowflies, *Musca*, *Drosophila* and *Tribolium* is shown in figure 3.9.

3.2.10 Characterisation of *hb* expression patterns in *Calliphora* and *Lucilia*

In situ hybridisations were performed (see 2.2.9) on fixed *Calliphora* and *Lucilia* embryos of mixed ages to determine the expression patterns of the *hb* mRNA in these species. Sense and anti-sense DIG-labelled RNA probes were made for both species corresponding to bases 3-374 in *Lucilia hb* and 1216-2758 in *Calliphora hb* (figure 3.1). The results of the antisense probe hybridisations are illustrated in figures 3.10 (*Lucilia*) and 3.11 (*Calliphora*). The sense probes were used as negative controls since they can be used to distinguish between background staining and gene specific expression. No hybridisation signals above background from embryos of either species were observed with sense probes (data not shown).

Lucilia hb is first expressed maternally where it is evenly distributed throughout the embryo, as in *Drosophila* and *Musca* (figure 3.10A). *Lucilia* also has similar *hb* early zygotic expression patterns to both *Drosophila* and *Musca* and staging of the *Lucilia* embryos was determined by comparison with *Drosophila* embryogenesis. The early zygotic expression patterns are characterised by a broad domain in the anterior half of the embryo and a posterior domain (figure 3.10B). The dynamic expression patterns of *Lucilia hb* continue with retraction from the poles of the embryo at the onset of cellularisation and subsequently, just before gastrulation, it is expressed in one strong and possibly two weaker anterior stripes and two posterior stripes (figure 3.10C and 3.10D). The strongest staining of the anterior stripes corresponds to PS4 in *Drosophila*. This domain of expression appears to be necessary for T2 thoracic development in *Drosophila*

A

```

mdhb3utr_ GACTTG-ACTCACA--ATTTGTAA-ATTGTTTT-TTTTTTTTTTATTTTGTATCCTTT 4591
cvhb3utr_ GAATTG-ATCTAAATCATTT-CAA-A--ATTT-GTAAATTTTTTGT---ATT-TCTTTA 2381
lshb3utr_ GAATTGCACTCAAAA-ACAT-TAACA--ATTTGTAAATTTTTTGT---ATT-TCTTTA 2438

I
mdhb3utr_ TTTTCGTTGCTTTGAATTGTAAATAATTATCCCCA---CCCCAACAAAA-----AAAT 4643
cvhb3utr_ TTTTCGTTGCTTTGAATTGTACATAGAAACCAAAAAAACCCAAACAAAACCCCAAC 2441
lshb3utr_ TTTTCGTTGCTTTGAATTGTATATAGAAA-----CCC--AA-AAAA-----AAAC 2480

mdhb3utr_ TCTC-----AGTAGCTTAGGA--GGGGCC---TGCC---TTTAGGT-CACCTTAA- 4684
cvhb3utr_ CCCCCCTATTTAAGTAGCTTAGGTTCCGGTATCAAATCCCAAATTT-GATGCCCCCTAAT 2500
lshb3utr_ CCTCCCC-----CCTAATTTAAGT--AG--CT---TAGG---TTT-GGT---TTT--- 2504

II
mdhb3utr_ GTGAATCGTTGTCATGAATTGTAAATATGAAAATC--AA-CATTTA-GTTTTAAGTTAAA 4740
cvhb3utr_ GCGAACGTTGTCAGAATTGTAAATATGAAAATTTTAGTCATATAAGTCATATTTTAAAT 2560

mdhb3utr_ T--ATTTTT--TTTTTAA----AT-----TTTAATAATT-TTA--AGTTTTAAAGGTT 4783
cvhb3utr_ TTTATTTTTAACTCTTAAGCAAATGGTTCCTTTTATGATTACTACCAATTATATATTTT 2620

```

B

dm1	GTTGTCCAGAATTGTA
dm2	GTTGTGCGAAAATTGTA
dv1	GTTGTCCAGAATTGTA
dv2	GTTGTGCGAGAATTGTA
md1	GTTGCTTTGAATTGTA
md2	GTTGTCATGAATTGTA
cv1	GTTGCTTTGAATTGTA
cv2	GTTGTCAAGAATTGTA
ls1	GTTGCTTTGAATTGTA
tc	GTTGT-GTGTATTATA
sa	CTTGAACGATATTGAC
lm	GTTGTCTCTCATTGTG

Figure 3.8 **A.** Alignment of the *Musca hb* 3' UTR (md) with the putative *Calliphora hb* (cv) and *Lucilia hb* (ls) 3' UTRs. Numbering of the *Musca* sequence refers to the base positions in the cooper strain sequence (accession number Y13050). The numbers of the *Calliphora* and *Lucilia* sequences are the base positions from the translation start site. Boxes I and II represent the two Nanos response elements (NREs). **B.** Alignment of the NREs from the above species and *D. melanogaster* (dm), *D. virilis* (dv), *Tribolium* (tc), *Schistocerca americana* (sa) and *Locusta migratoria* (lm) (adapted from Patel *et al.*, 2001, figure 3E). Conserved bases are highlighted in grey.

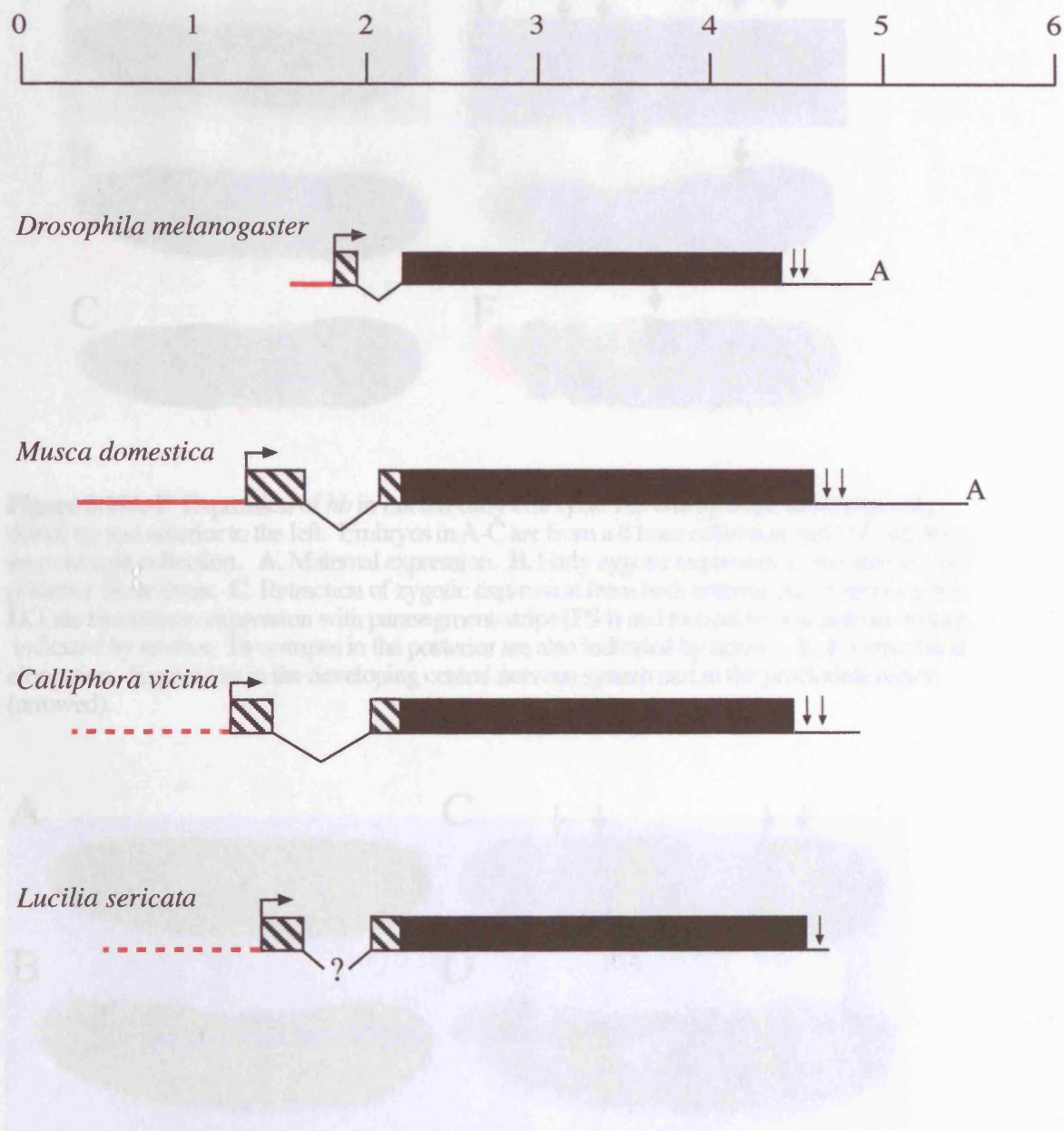


Figure 3.9 *hb* P2 transcript structure in *Drosophila*, *Musca*, *Calliphora* and *Lucilia*
 The Hb protein coding region of each species is represented by the black boxes and the 5' UTRs by the striped boxes. The large black arrows show the positions of the transcription start sites and diagonal lines represent intronic sequences (unknown sequence is represented by ?). The 3' UTRs are illustrated by a horizontal black line and end in an A representing the position of the polyadenylation signal in those species in which this has been determined. The downward arrows represent the positions of NREs in the 3' UTRs. Solid red lines represent the *Drosophila* and *Musca* Bcd-dependent *hb* promoters and dashed red lines the putative Bcd-dependent *hb* promoter regions cloned from *Calliphora* and *Lucilia* (see text). The numbers on the scale bar are given in kilobases.

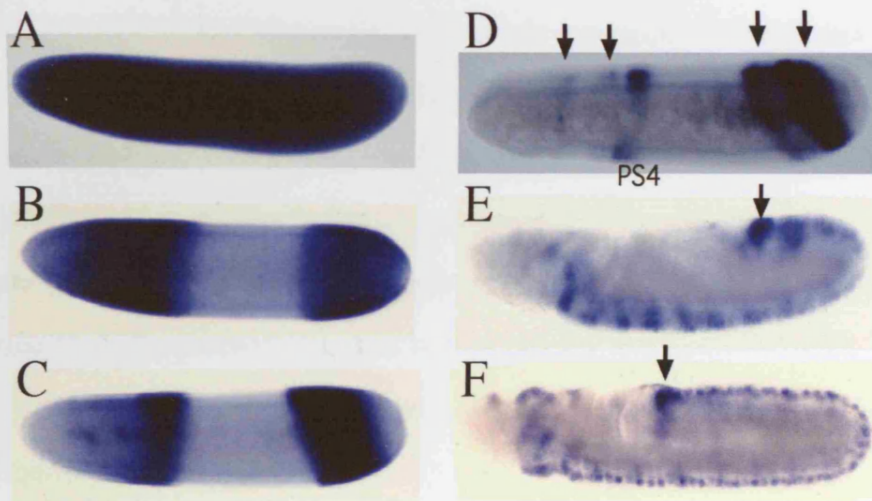


Figure 3.10A-F Expression of *hb* in *Lucilia* early embryos. All embryos are shown laterally, dorsal up and anterior to the left. Embryos in A-C are from a 4 hour collection and D-F are from an overnight collection. **A.** Maternal expression. **B.** Early zygotic expression in the anterior and posterior blastoderm. **C.** Retraction of zygotic expression from both anterior and posterior poles. **D.** Late blastoderm expression with parasegment-stripe (PS4) and two additional anterior stripes indicated by arrows. Two stripes in the posterior are also indicated by arrows. **E, F.** Germ band elongation. Expression in the developing central nervous system and in the proctodeal region (arrowed).

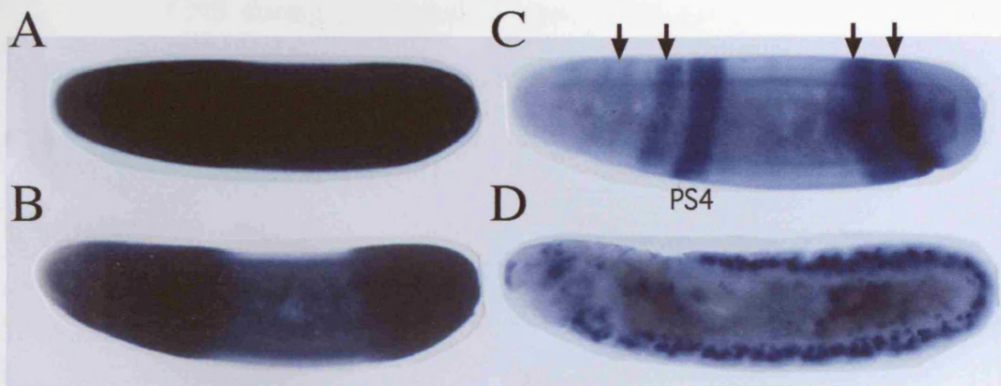


Figure 3.11A-D Expression of *hb* in *Calliphora* early embryos. All embryos are shown laterally, dorsal up and anterior to the left. All embryos are from an overnight collection. **A.** Maternal expression. **B.** Maternal expression and early zygotic expression in the anterior and posterior blastoderm. **C.** Late blastoderm expression with parasegment-stripe (PS4) and two additional anterior stripes indicated by arrows. Two stripes in the posterior are also indicated by arrows. **D.** Germ band retraction with expression in the developing central nervous system.

(Hülskamp *et al.*, 1994; Wu *et al.*, 2001). Three anterior stripes are also seen in *Drosophila* (Tautz and Pfeifle, 1989) and in *Musca* (Sommer and Tautz, 1991a); however, in these species and in *Megaselia*, there is only one stripe of *hb* expression in the posterior of the embryo and this is required for A7/A8 development in *Drosophila* (Lehmann and Nüsslein-Volhard, 1987). At germ-band elongation in *Lucilia* (figure 3.10E and 3.10F), *hb* is expressed in the developing central nervous system (CNS) and in the cephalic region. These expression patterns are similar to other species; however, the strong expression of *Lucilia hb* in the proctodeal region is more similar to expression of *hb* in the mothmidge *Clogmia albipunctata*, which is a lower Dipteran of the Nematocera (Rohr *et al.*, 1999 and see figure 1.3).

Calliphora hb mRNA expression patterns are similar to those described above for *Lucilia hb*. *Calliphora hb* has ubiquitous maternal expression (figure 3.11A) and then, as the maternal mRNAs degrade, zygotic expression is seen in the anterior and posterior domains (figure 3.11B). The later patterns of zygotic expression of *Calliphora hb* also resemble those of *Lucilia*; illustrated by at least three anterior stripes and two posterior stripes in the late blastoderm (figure 3.11C). *Calliphora* also expresses *hb* in the developing CNS during germ band elongation and germ band retraction (figure 3.11D), although *hb* expression in the proctodeal region is not as strong in this species as it is in *Lucilia*.

3.3 Discussion

3.3.1 *Calliphora* and *Lucilia* both encode Hb proteins with conserved functional domains

The *Calliphora* Hb and *Lucilia* Hb putative protein sequences are aligned with those from *Drosophila*, *Musca* and *Megaselia abdita* in figure 3.2. This illustrates that there is conservation of the Hb functional domains in these species.

It has previously been shown that the DNA binding zinc-finger domains of Hb are conserved over large phylogenetic distances (Sommer *et al.*, 1992). Mutations mapped to these DNA binding domains result in loss of thoracic segments in *Drosophila* embryos caused by mis-regulation of Hb target genes such as *Kr* and *fushi tarazu (ftz)* (Hülkamp *et al.*, 1994). The C-terminal zinc-finger domain is required for the repression of *Kr* and *kni* expression in the anterior of the *Drosophila* embryo, as well as stabilising the Hb protein (Hülkamp *et al.*, 1994). The two clusters of zinc-fingers structure of Hb are also seen in the related vertebrate gene *Ikaros* (Georgopoulos *et al.*, 1992). The C-terminal zinc-fingers of *Ikaros* are involved in protein-protein interactions resulting in dimers which can activate or repress transcription (Sun *et al.*, 1996). Indeed, dimerisation of *Kr* at high protein concentrations results in transcriptional repression (Sauer and Jäckle 1993). Dimerisation of Hb, mediated by the C-terminal zinc fingers, may result in transcriptional repression of *Kr* and *kni* depending on the concentration of Hb (Struhl *et al.*, 1992). *Ikaros* interacts with the DNA-dependent ATPase Mi-2 to mediate both transcriptional activation and repression in T-lymphocytes (Kim *et al.*, 1999; O'Neill *et al.*, 2000). Interestingly, studies of Hb in *Drosophila* have shown that the D box is involved in protein-protein interactions and binds dMi-2 (Kehle *et al.*, 1998). This is thought to result in the repression of Hox genes such as *Ubx* by histone de-acetylation and the involvement of the Polycomb group genes (Pelegri and Lehmann 1994).

The A and B boxes have a high degree of conservation among zinc-finger containing proteins such as Hb, *Kr* and HIV-1 Pol, (Tautz *et al.*, 1987). Figure 3.2 illustrates that although there is conservation of the A box between *Calliphora*, *Lucilia* and the other Dipterans, the B box has diverged beyond recognition. The significance of this is not known since specific functions have not been assigned to these domains.

The C box has also been shown to be highly conserved over large phylogenetic distances (Hülkamp *et al.*, 1994). Again this domain is conserved in *Calliphora* Hb and *Lucilia* Hb. The C box is involved in the autoregulation of *hb* expression in the PS4 domain and its amino acid sequence is also conserved in other insects such as *Tribolium* and *Schistocerca* (Wolff *et al.*, 1995; Patel *et al.*, 2001)

The E and F boxes (figure 3.2) may be involved in protein-protein interactions with transcription co-factors to mediate activation or repression. This is supported by the conservation of the F box in the *Tribolium* and grasshopper orthologs of *hb* where it has been called a 'basic box' (due to numerous basic residues such as arginine and lysine) and implicated in protein-protein interactions (Patel *et al.*, 2001).

In the more diverged regions of Hb there are numerous glutamine repeats which are extremely variable between species (figure 3.2). Glutamine repeats also vary extensively in Hb within and between *Drosophila* species (Tautz and Nigro 1998) and within *Musca* (see chapter 5). Such divergence is probably driven by a the high rate of slippage in CAG/CAA repeats (see chapter 5). The functional significance of this variation (if any) is not known, although glutamine repeats may act as species-specific *trans*-activation domains (Emili *et al.*, 1994).

3.3.2 Conserved and diverged expression patterns of *hb* in *Calliphora* and *Lucilia* suggest that some aspects of *hb* regulation in these species have changed

The functional conservation of *hb* in *Calliphora* and *Lucilia* is further demonstrated by the characterisation of *hb* mRNA expression patterns in these species (figures 3.10 and 3.11). It would appear that like *Drosophila*, *Calliphora* and *Lucilia* both have maternal and zygotic *hb* expression and it is likely that these patterns are controlled by the same factors.

In both species there is maternal expression throughout the early syncytial precellular embryo (figures 3.10A and 3.11A). The presence of NRE consensus sites in the putative *hb* 3'UTRs in these species (figure 3.8A) implies that the translation of maternal *hb* mRNAs is repressed in the posterior half of *Calliphora* and *Lucilia* embryos by Nos and Pum (Wharton and Struhl 1991; Murata and Wharton 1995; Wrenden *et al.*, 1997) as it is in other Dipterans (Curtis *et al.*, 1995). The putative NREs of *Calliphora* and *Lucilia* are similar to NREs found in the *hb* 3'UTRs of other species, as illustrated by the sequence alignment in figure 3.8B. This mechanism is likely to be ancient in origin since it probably functions in more primitive insects such as beetles (Wolff *et al.*, 1995) and grasshoppers (Patel *et al.*, 2001). In the cell division cycle (CDC), Nos and Pum

inhibit the translation of *cyclin B* mRNAs (Sonoda and Wharton 1999) and this mechanism is also a feature of translational repression of mRNAs in amphibians and in mammalian neurons (Richter and Theurkauf 2001). It is possible that the Nos/Pum mechanism has been co-opted from an ancestral role in the CDC or CNS to promote abdominal fate by inhibiting *hb* translation in the posterior half of the early embryo.

The gap role of *hb* appears to be conserved in *Calliphora* and *Lucilia* as it is expressed zygotically in the anterior half of the early blastoderm and also in a domain at the posterior (figures 3.10B and 3.11B). These expression domains and aspects of the later patterns of expression in the blastoderm are seen in other Dipterans such as *Drosophila* (Tautz *et al.*, 1987), *Musca* (Sommer and Tautz 1991a; Bonneton *et al.*, 1997), *Megaselia* (Stauber *et al.*, 2000) and *Clogmia* (Rohr *et al.*, 1999). Indeed, the gap role of *hb* is also conserved in intermediate germ band insects, such as *Tribolium* (Wolff *et al.*, 1995) and in long germ band insects such as *Schistocerca* (Patel *et al.*, 2001). However, it is possible that the posterior domain of *hb* expression appears earlier in the blowflies than in *Drosophila* and *Musca* since no embryos were observed which had anterior expression alone.

In *Drosophila* and *Musca*, the early anterior expression pattern is driven by Bcd (Schröder *et al.*, 1988; Driever and Nüsslein-Volhard 1989; Struhl *et al.*, 1989) and the posterior pattern by terminal system transcription factors (Margolis *et al.*, 1995). In *Drosophila* (Lehmann and Nüsslein-Volhard 1987) and in *Musca* (McGregor *et al.*, 2001a) these patterns are essential for the patterning of the gnathal/thoracic region and the A7/A8 abdominal regions respectively. Since *bcd* is also present in *Calliphora* and *Lucilia* (Schröder and Sander 1993; Shaw *et al.*, 2001), this implies that it is also Bcd that regulates the anterior expression of *hb* in these species. The P2 transcript, which Bcd activates the expression of, has a similar structure in *Drosophila*, *Musca*, *Calliphora* and *Lucilia* (figure 3.9). Interestingly, the arthropod transcript cap site consensus sequence is conserved in the *hb* P2 transcripts of the blowflies in addition to *Drosophila* and *Musca* (figure 3.6A). It has been suggested that this sequence may promote high levels of transcription by stabilising the interaction of TFIID and the TATA box (which is also

conserved in sequence and position in the Dipterans see figure 3.6A) and allowing enhancers to discriminate between promoters (Cherbas and Cherbas 1993). Therefore, these sequences could be important in ensuring that Bcd generates high levels of *hb* transcripts from the P2 promoter in the anterior of Dipteran embryos and especially so in the larger embryos of *Musca*, *Lucilia* and *Calliphora*. Full analysis of the *Calliphora* and *Lucilia* *hb* putative P2 promoter regions is described in chapters 4 and 5.

The later expression patterns of *hb* in *Calliphora* and *Lucilia* (figures 3.10D and 3.11C) consist of three anterior stripes, the most posterior of which corresponds to PS4. In *Drosophila* the expression of PS4 is dependent on Hb (including the C box) and Kr (Hülskamp *et al.*, 1990, 1994; Struhl *et al.*, 1992). It has recently been shown that high levels of Hb are required at PS4 to generate T2 by activating *Antp* expression and inhibiting *Kr* expression in the anterior of the embryo. This repression of *Kr* is thought to be required for head development and may explain how high levels of maternal *hb* expression can rescue head defects in *bcd* mutant embryos (Wimmer *et al.*, 2000). The conservation of *hb* PS4 expression and the C box Hb protein domain in the *Calliphora* and *Lucilia* suggests that this regulatory interaction functions in these species. In the posterior of *Drosophila* and *Musca* embryos at late blastoderm, *hb* is expressed as a single stripe. However, in both *Lucilia* and *Calliphora* at the same stage two posterior stripes form (figures 3.10D and 3.11C) and although the significance of this is not known it suggests that there are differences in the regulation of *hb* in the posterior late blastoderm between the blowflies and other Dipterans. Conserved and divergent aspects of *hb* expression at a similar stage of development have also been reported amongst *Drosophila* species (Lukowitz *et al.*, 1994; Tautz and Nigro 1998).

During germ band elongation and germ band retraction *hb* is expressed in the developing CNS of both *Calliphora* and *Lucilia* (figures 3.10E, F and 3.11D). This is again consistent with *hb* expression in other Dipterans, although the proctodeal expression appears to be stronger in *Lucilia* and *Clogmia* (Rohr *et al.*, 1999) the significance of this is not known. The proctodeal expression could derive from the posterior stripe as suggested for *Clogmia*. Is this an atavistic transition in *Lucilia* or was this expression lost in other

higher Dipterans? It has been shown that in the CNS of *Drosophila* *hb* represses the expression of *pdm-1* to regulate the determination of neural fate (Kambadur *et al.*, 1998). In more primitive organisms such as nematodes (Fay *et al.*, 1999) annelids (Savage and Shankland 1996; Iwasa *et al.*, 2000) and polychaetes (Werbrock *et al.*, 2001), *hb*-like genes are expressed in the developing CNS, but are not involved in segmentation. This suggests that the ancestral role of *hb* is in the CNS and that it has been co-opted for its gap role in the evolution of arthropods. An ancestral CNS role has also been postulated for other segmentation genes such as *en* and *eve* (Duman-Scheel and Patel 1999).

Chapter 4
Characterisation of the *hb* promoters in
Calliphora and *Lucilia*

4.1 Introduction

4.1.1 Characterisation of the putative *Calliphora* and *Lucilia hb* promoter regions

The putative *Calliphora* and *Lucilia hb* promoters were mapped as described in chapter 3 (3.2.7 and 3.2.8). To investigate whether Bcd binds to sites within these regions (a first step in defining these regions as Bcd-responsive elements) a DNaseI footprinting strategy (Lin and Shiuan and see 2.2.7) was chosen to identify sequences that were protected by Bcd binding. This had previously been used to identify Bcd-binding sites in the *Drosophila* and *Musca* promoters (Driever and Nüsslein-Volhard 1989; Bonneton *et al.*, 1997). As an alternative strategy methylation interference could have been employed, but this technique is more useful for pinpointing individual bases that are contacted by a transcription factor (Dave *et al.*, 2000) and can mask interactions detected by DNaseI (Galas and Schmitz 1978). Slight differences were observed in the sequences of the *Drosophila* and *Musca hb* promoters Bcd-binding sites, which perhaps reflects divergent sequence preferences of *Drosophila* Bcd and *Musca* Bcd (Shaw 1998). It is possible that such changes in promoter binding sites and in Bcd could have co-evolved. Therefore, it was important to characterise the Bcd-binding sites in the *Calliphora* and *Lucilia hb* promoter regions using the Bcd homeodomains from the respective species. This would also identify any species-specific differences in Bcd-binding site sequence requirements.

4.2 Results

4.2.1 Expansion of the known *bcd* sequences in *Lucilia* and *Calliphora*

To perform DNaseI footprinting on the putative *Calliphora* and *Lucilia hb* promoters the complete Bcd homeodomain sequences from these species were required. Schröder and Sander (1993) had previously sequenced the coding region corresponding to 44 amino acids from the Bcd homeodomains of both *Calliphora* and *Lucilia*. Therefore, primers (BFBCDF, LSBCDR, CVBCDR, LSBF, CVBF, LSBR and CVBR; see appendix A) were

designed based on these known sequences to perform sPCR (see 3.1.3) to extend the sequence of *bcd* both 5' and 3' in these species. The primary PCRs generated a number of bands up to approximately 1 kb, from the *ClaI*, *DraI*, *EcoRV* and *SspI* *Lucilia* libraries and from the *BglII*, *ClaI*, *DraI*, *EcoRI* and *SspI* *Calliphora* libraries. Reamplification with nested primers was performed to determine the specificity of these primary products. In the 5' direction, with respect to the known sequence, 300 bp products from the *DraI* libraries of both species and in the 3' direction 700 bp from the *ClaI* libraries of both species were sequenced. The sequence of these products overlapped with the known *bcd* homeodomain sequences from both species and extended the protein sequence by 14 amino acids towards the N-terminal in both species and 176 and 170 amino acids C-terminal in *Calliphora* and *Lucilia* respectively.

The patterns of conservation and divergence of the *Calliphora* and *Lucilia* Bcd protein amino acid sequences are illustrated in the alignment of these proteins with the Bcd proteins of other Dipterans in figure 4.1. The *Calliphora* and *Lucilia* partial Bcd proteins are 95% similar and both highly conserved with respect to the *Musca* Bcd amino acid sequence (85% and 87% respectively). Interestingly, it was found that a serine rich domain present in *Musca*, but not in *Drosophila*, is also found in the *Calliphora* and *Lucilia* Bcd proteins. Although the function of this domain is not known, it could be phosphorylated *in vivo* (see 4.3.1). A number of MAP kinase phosphorylation sites previously identified in *Drosophila* (Niessing *et al.*, 1999; Janody *et al.*, 2000) are also conserved in the Bcd proteins of *Calliphora* and *Lucilia* in addition to *Musca* Bcd. The most conserved region of the Bcd protein is the DNA binding homeodomain and this is compared in greater detail in figure 4.2. In the homeodomain, an additional difference was found 3' of the previously published sequence at position 60 between *Calliphora* (alanine) and the other species (serine) (figure 4.2). The DNA recognition helix including the specifying lysine at position 50 (Hanes and Brent 1989) is conserved in all species, but *Megaselia* is different at residues 42 and 56 (Stauber *et al.*, 1999 and see figure 4.2).

Parts of the *bcd* gene from *Calliphora* and *Lucilia*, between the codons for amino acids 5 (aspartic acid) to 86 (proline), were then amplified by PCR (see figure 4.1) and

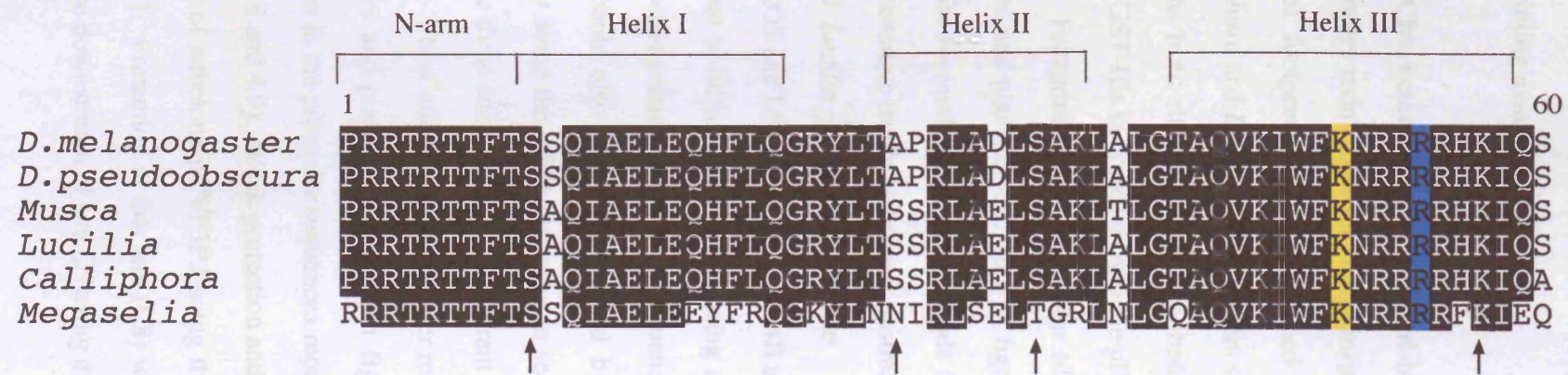


Figure 4.2 Alignment of the Bcd homeodomains from *D. melanogaster*, *D. pseudoobscura*, *Musca*, *Lucilia*, *Calliphora* and *Megaselia*. Numbering refers to the amino acid positions in the homeodomain. See figure 4.1 for the relative positions of the homeodomains within the proteins. Amino acids in the N-terminal arm (N-arm), helices I, II and III (the recognition helix) are indicated. Conserved amino acids are highlighted in black. In the recognition helix, lysine at position 50 is highlighted in yellow and arginine at position 54 in blue (see text). Amino acids involved in cooperative Bcd binding in *D. melanogaster* (at positions 10, 28, 35 and 57) are arrowed (Burz and Hanes 2001).

cloned in frame into the protein expression vector pET42b (+) (pECBCD and pELBCD, see table 2.1). P. Shaw subsequently used these plasmids to synthesise GST-tagged homeodomain proteins for either species (see 2.2.7) and these were used in the DNaseI footprinting assays below.

4.2.2 Characterisation of the Bcd-binding sites in the *Lucilia* and *Calliphora hb* P2 promoters using DNaseI footprinting

DNaseI footprinting was performed (see 2.2.7) using the plasmid subclones of the *Calliphora* and *Lucilia hb* promoters (figure 4.3) and the Bcd homeodomain-GST fusion proteins from either species described above. Control protein used in these experiments was a GST-His tag protein from the pET42B vector.

Footprinting of the *Lucilia hb* promoter was performed using the end-labelled primers and plasmid subclones in figure 4.3A. This revealed seven protected regions on both the forward and reverse strands (figures 4.4, 4.5 and 4.6). No protected regions or hypersensitive bands were seen upstream of L1 (see figure 4.4A) as far as the end of the cloned *Lucilia* promoter sequence. Double band patterns were seen using primers LRPRO5 and LPROF3 (figures 4.4B and 4.5A respectively). It is thought that this effect was due to difficulties in sequencing across a motif of thirteen adenines in this region of the *Lucilia* promoter, since the sequence was normal on either side of this adenine repeat. This could also have been caused by the primers mis-annealing to similar sequences nearby since the effect was seen in sequencing with T7 polymerase and PCR with *Taq*, despite these enzymes having different processivities.

The *Calliphora hb* promoter region was DNaseI footprinted using the end-labelled primers and plasmid subclones in figure 4.3B. This initially revealed eight protected regions in the promoter sequences represented in subclones pCVF3P2 and pCP1 (figures 4.7, 4.8 and 4.9). Weak protection and hypersensitive bands were also seen on the reverse strand of subclone pCVF3P2 using the end-labelled primer CVPRO7 (site C8 in figure 4.10B). Protection at this site (C8) was confirmed using subclone pCVF6P7 to further analyse downstream of site C5 using the universal primer T7 (figure 4.8A). No additional

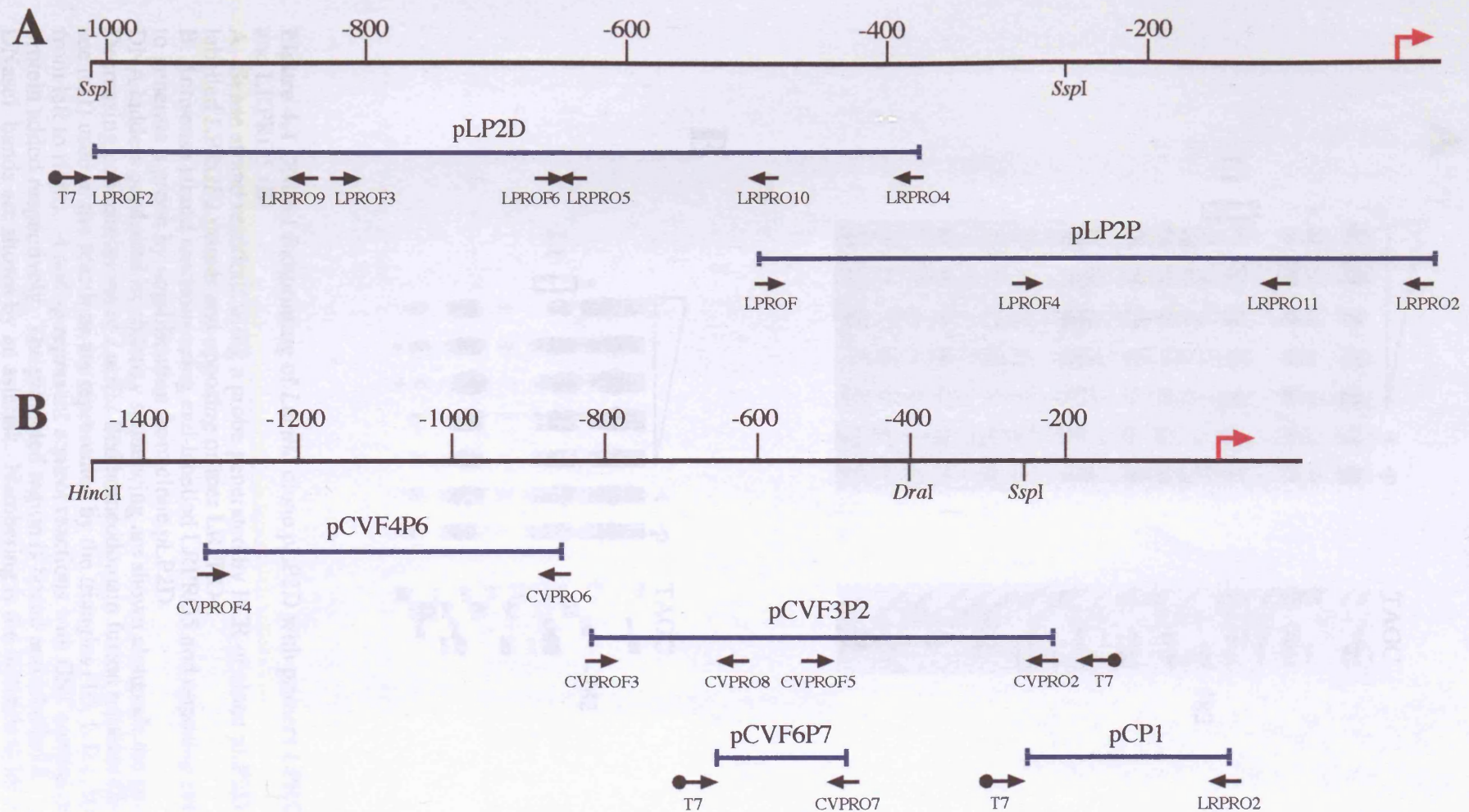


Figure 4.3 Primers and plasmid clones used in *DNaseI* footprinting of the *Lucilia* (A) and the *Calliphora* (B) *hb* P2 promoter regions. The red arrows represent the positions of the *hb* transcription start sites in both species. The scale is in bp with respect to the transcription start site (sPCR restriction sites are shown underneath, see figure 3.5). Arrows show the positions of end-labelled primers used in the *DNaseI* footprinting and dideoxy sequencing reactions (see text). Primers with a rounded end were vector based universal primers. The labelled blue lines represent plasmid clones of the promoter region from either species.

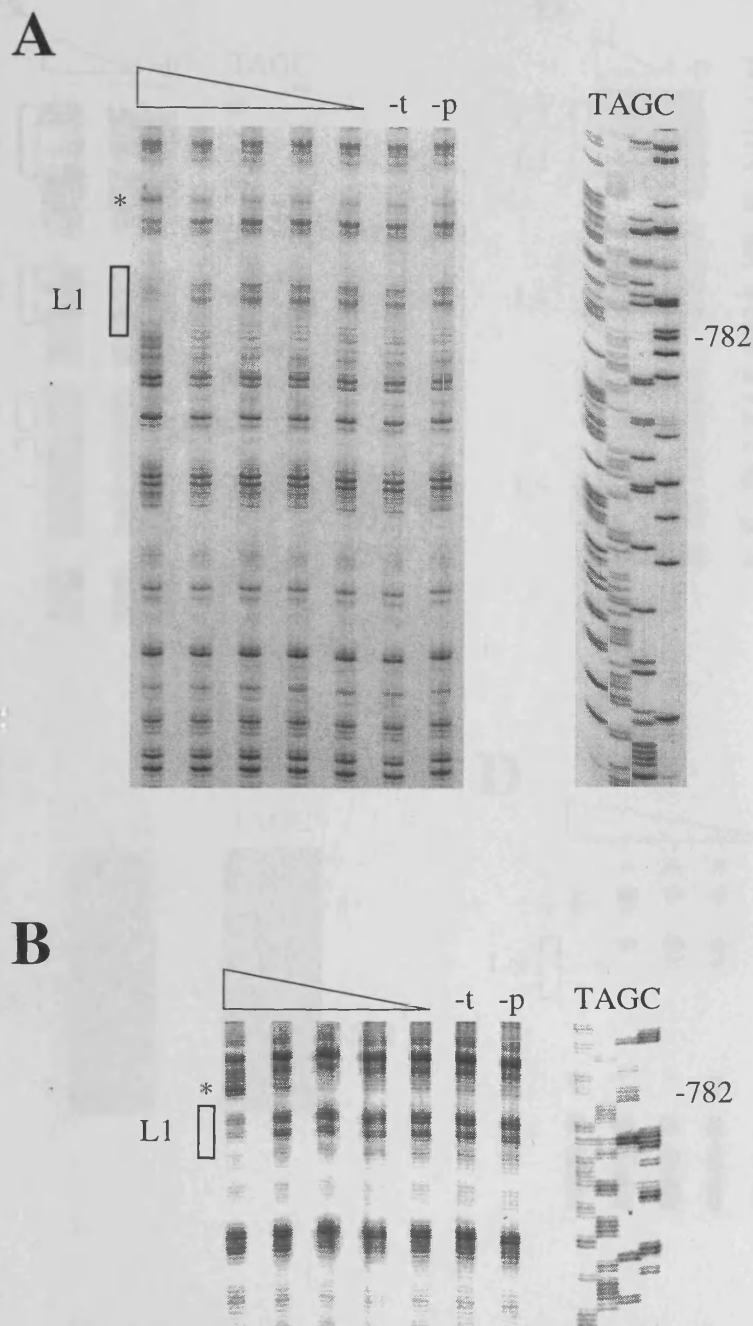


Figure 4.4 DNaseI footprinting of *Lucilia* clone pLP2D with primers LPROF2 (A) and LRPRO5 (B)

A. Sense strand reactions using a probe generated by PCR of clone pLP2D using end-labelled LPROF2 primer and opposing primer LRPRO4.

B. Antisense strand reactions using end-labelled LRPRO5 and opposing primer LPROF2 to generate a probe by amplification from clone pLP2D.

DNA ladders generated by dideoxy sequencing are shown alongside the protection patterns. Decreasing concentrations of *Lucilia* Bcd homeodomain fusion proteins (from vector pELBCD see text) used in the reactions are represented by the triangles (10, 1, 0.1, 0.01 and 0.001 nm from left to right). -t and -p represent control reactions with GST control protein added or no protein added respectively. The protected region is boxed and labelled L1. Hypersensitive DNaseI bands are shown by an asterisk. Numbering is the distance in bp 5' from the transcription start site.

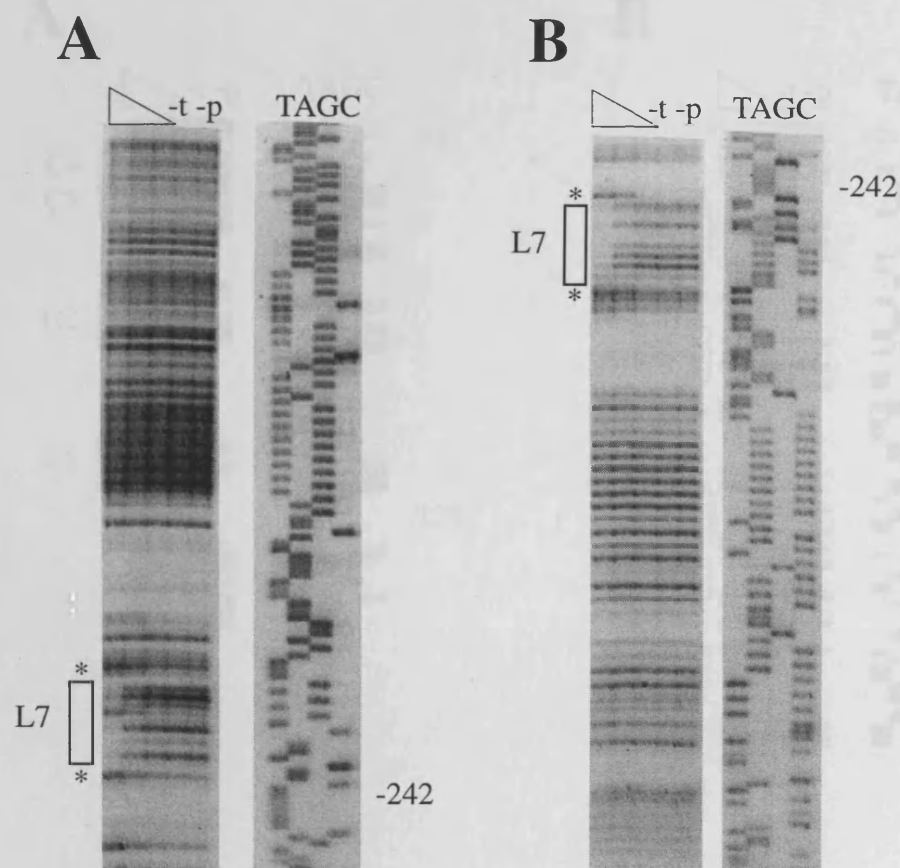


Figure 4.6 DNaseI footprinting of *Lucilia* clone pLP2P with primers LPROF4 (A) and LRPRO11 (B)

A. Sense strand reactions using a probe generated by PCR of clone pLP2P using end-labelled LPROF4 primer and opposing primer LRPRO11.

B. Antisense strand reactions using end-labelled LRPRO11 and opposing primer LPROF4 to generate a probe by amplification from clone pLP2P.

DNA ladders generated by dideoxy sequencing are shown alongside the protection patterns. Decreasing concentrations of *Lucilia* Bcd homeodomain fusion proteins (from vector pELBCD see text) used in the reactions are represented by the triangles (10, 1 and 0.1 nM from left to right). -t and -p represent control reactions with GST control protein added or no protein added respectively. The protected region is boxed and labelled L7. Hypersensitive DNaseI bands are shown by an asterisk. Numbering is the distance in bp 5' from the transcription start site.

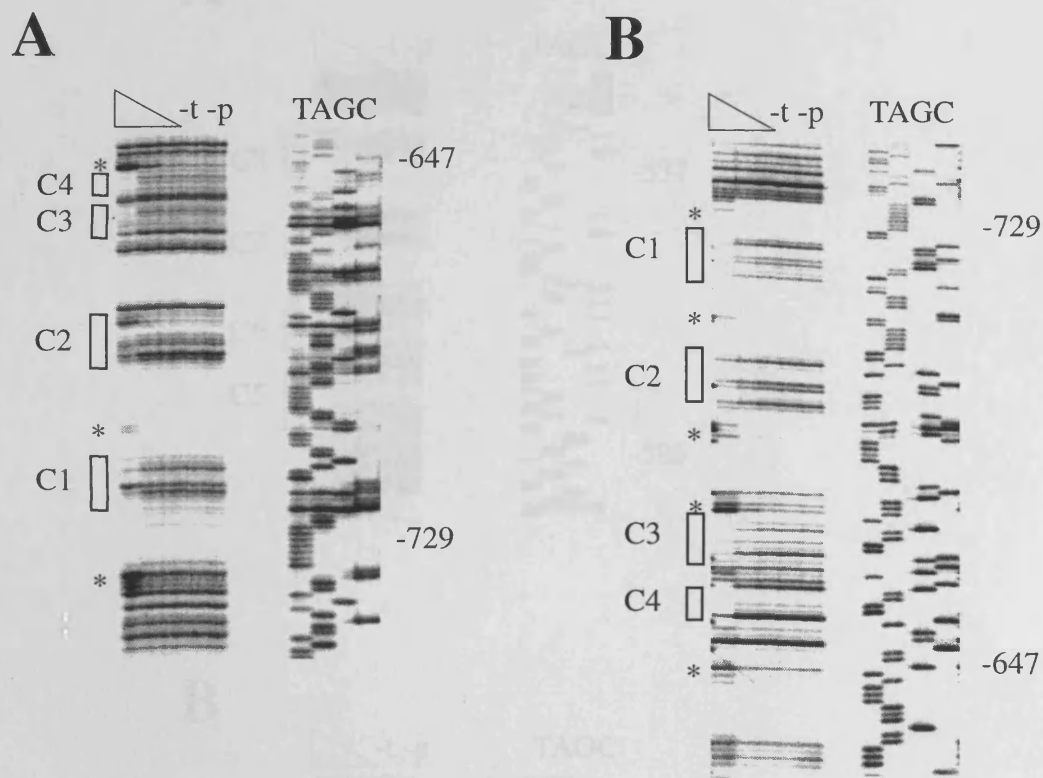


Figure 4.7 DNaseI footprinting of *Calliphora* clone pCVF3P2 with primers CVPROF3 (A) and CVPRO8 (B).

A. Sense strand reactions using a probe generated by PCR of clone pCVF3P2 using end-labelled CVPROF3 primer and opposing primer CVPRO7.

B. Antisense strand reactions using end-labelled CVPRO8 and opposing primer CVPROF3 to generate a probe by amplification from clone pCVF3P2.

DNA ladders generated by dideoxy sequencing are shown alongside the protection patterns. Decreasing concentrations of *Calliphora* Bcd homeodomain fusion proteins (from vector pECBCD see text) used in the reactions are represented by the triangles (10, 1 and 0.1, nm from left to right). -t and -p represent control reactions with GST control protein added or no protein added respectively. The protected regions are boxed and labelled C1, C2, C3 and C4. Hypersensitive DNaseI bands are shown by an asterisk. Numbering is the distance in bp 5' from the transcription start site.

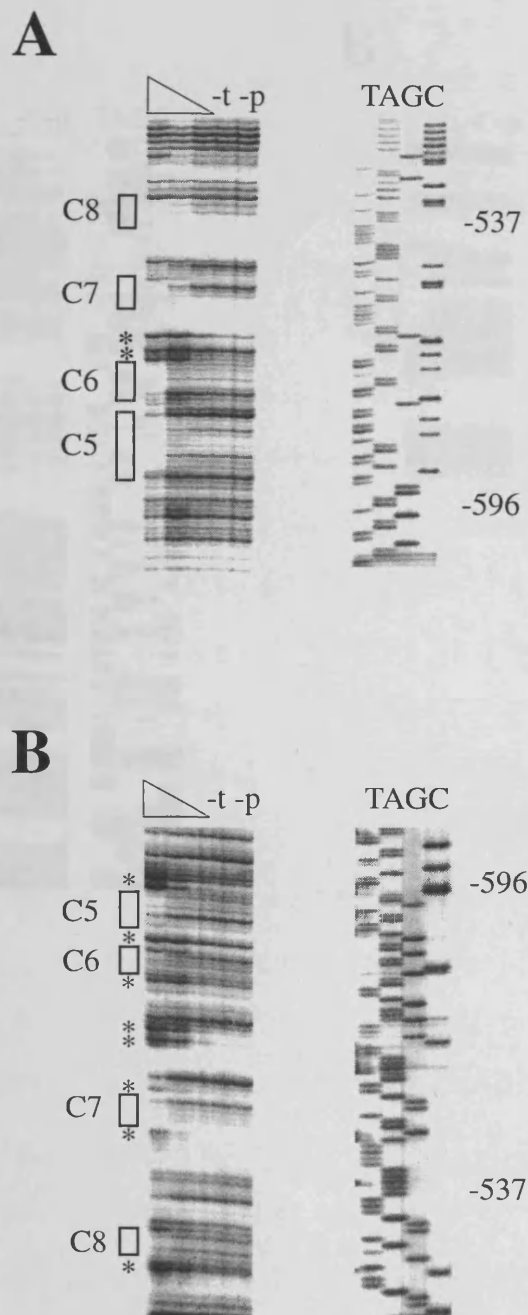


Figure 4.8 DNaseI footprinting of *Calliphora* clone pCVF6P7 with T7 primer (A) and clone pCVF3P2 with primer CVPRO7 (B).

A. Sense strand reactions using a probe generated by PCR of clone pCVF6P7 using end-labelled T7 primer and opposing primer CVPRO7.

B. Antisense strand reactions using end-labelled CVPRO7 and opposing primer CVPROF3 to generate a probe by amplification from clone pCVF3P2.

DNA ladders generated by dideoxy sequencing are shown alongside the protection patterns. Decreasing concentrations of *Calliphora* Bcd homeodomain fusion proteins (from vector pECBCD see text) used in the reactions are represented by the triangles (10, 1 and 0.1, nm from left to right). -t and -p represent control reactions with GST control protein added or no protein added respectively. The protected regions are boxed and labelled C5, C6, C7 and C8. Hypersensitive DNaseI bands are shown by an asterisk. Numbering is the distance in bp 5' from the transcription start site.

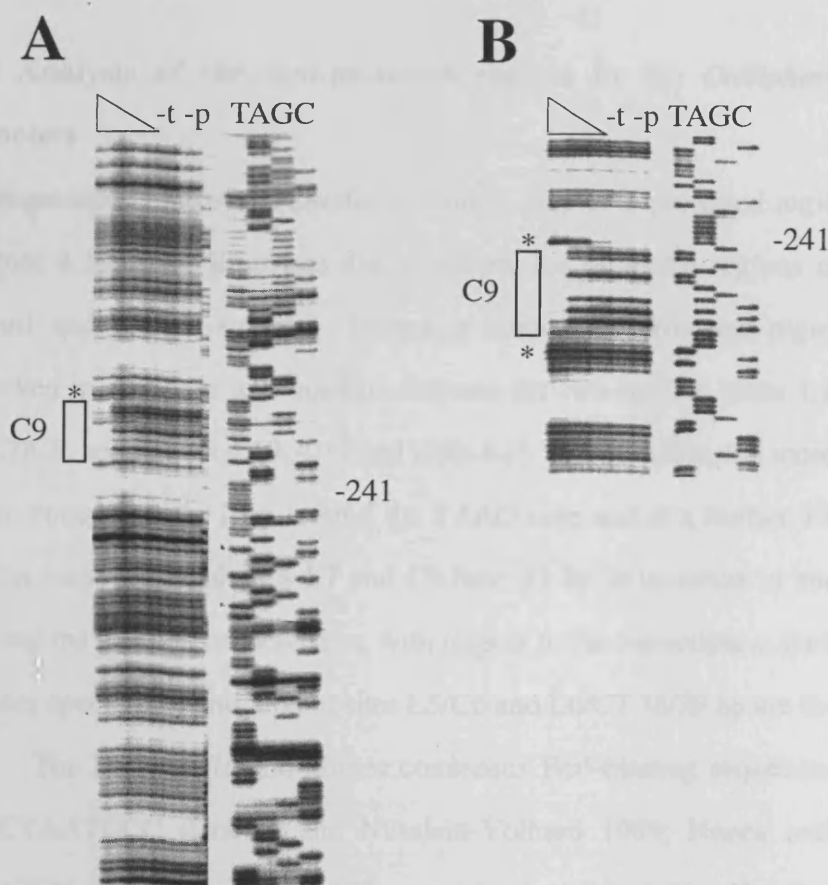


Figure 4.9 DNaseI footprinting of *Calliphora* clone pCP1 with T7 primer (A), and clone pCVF3P2 with T7 primer (B).

A. Sense strand reactions using a probe generated by PCR of clone pCP1 using end-labelled T7 primer and opposing primer LRPRO2.

B. Antisense strand reactions using end-labelled T7 and opposing primer CVPROF5 to generate a probe by amplification from clone pCVF3P2.

DNA ladders generated by dideoxy sequencing are shown alongside the protection patterns. Decreasing concentrations of *Calliphora* Bcd homeodomain fusion proteins (from vector pECBCD see text) used in the reactions are represented by the triangles (10, 1 and 0.1, nm from left to right). -t and -p represent control reactions with GST control protein added or no protein added respectively. The protected region is boxed and labelled C9. Hypersensitive DNaseI bands are shown by an asterisk. Numbering is the distance in bp 5' from the transcription start site.

protected regions were seen upstream of site C1 in either subclones pCVF4P6 using end-labelled CVPRO6 or in subclone pCVF3P2.

4.2.3 Analysis of the Bcd-protected regions in the *Calliphora* and *Lucilia hb* promoters

The sequences of the seven *Lucilia* and nine *Calliphora* protected regions are summarised in figure 4.10. This illustrates that in general the protected regions overlap for both the forward and reverse strands. Indeed, a number of protected regions appeared to be conserved in sequence and position between the two species (sites L1/C1, L5/C6, L6/C7 and L7/C9, see figures 4.10, 4.11 and table 4.1). For example, the most distal sites L1 and C1 are conserved at 11 bp around the TAAG core and at a further 13/16 bp immediately 3'. The most proximal sites L7 and C9 have 31 bp in common in and around the TAAT core and the position of these sites, with respect to the transcription start site, is very similar in either species. In and around sites L5/C6 and L6/C7 38/39 bp are the same.

The *Drosophila* and *Musca* consensus Bcd-binding sequences have been defined as TCTAATCCC (Driever and Nüsslein-Volhard 1989; Hanes and Brent 1991) and TTAATCY (Bonneton *et al.*, 1997) respectively (table 4.1). The protection patterns for sites L5, L6, L7, C2, C5, C6, C7 and C9 overlap at sequences that resemble the *Musca* and *Drosophila* Bcd-binding core consensus sequences (figure 4.10 and table 4.1). However, the sequences of the remaining protected regions in both species contain changes to these core consensus sequences. For example, sites L1, L2, L3, L4, C1, C3 and C4 show protection patterns at a core Bcd binding sequence of TAAGY (figure 4.10 and table 4.1), which is seen in *Drosophila* sites x1 and x2 (Driever and Nüsslein-Volhard 1989). *Calliphora* site C8 appears to have the core sequence AAATC, which is not found in any of the *hb* Bcd-binding sites in other species, but is present in *Drosophila kni* Bcd-dependent promoter (Rivera-Pomar *et al.*, 1995) and the *Kr* CD1 element (Hoch *et al.*, 1991).

Interestingly, sites L2, L4, L5, C3, C4, C5 and C6 (figure 4.10) appear to have Bcd target sequences on both the forward and reverse strands. Site L2 has a core sequence of

TTAAGT 5'-3' on both strands and therefore the orientation of this Bcd-binding site is unclear. Site L4 has protected sequence of CTTAAGGA on the sense strand and CTTAAGTC on the anti-sense strand. Therefore, the orientation of this site is likely to be 5'-3' with respect to the antisense strand due to the TC at positions 7 and 8 on the latter sequence being more similar to the *Drosophila* and *Musca* consensus Bcd-binding sites. The orientation of sites L5 and C6 are probably 5'-3' with respect to the top strand as TTAATC may give higher affinity binding than TTAAGC. The double protection patterns seen at *Calliphora* sites C3 and C4 also have possible Bcd-binding site sequences on both strands (figure 4.10). Therefore, different arrangements of these sites are possible and these could affect Bcd-binding at this region of the promoter (see 4.3.4).

<i>D. melanogaster</i>	<i>Musca</i>	<i>Calliphora</i>	<i>Lucilia</i>
A1 GTAATCC	A TTAATGG	C1 TTAAGCC	L1 TTAAGCC
A2 CTAATCC	B TTGATCC	C2 TTAATCA	L2 TTAAGTA
A3 CTAATCC	C TTAATCC	C3f TTAAGCG	L3 TTAAGTC
x1 CTAAGCT	D TTAACCT	C3r TTAAGTA	L4f TTAAGGA
x2 CTAAGCT	E TTAACGG	C4f TTAAGTC	L4r TTAAGTC
x3o ATCATCC	Ff TAAATCG	C4r TTAAGCT	L5f TTAATCT
x3s ATGATCC	Fr CTAATCT	C5f TTAATCT	L5r TTAAGCA
x3t CAAATCC	G TTAATCC	C5r TTAATGA	L6 TTAATCC
x4 TCAATCC	H TTGATCC	C6f TTAATCT	L7 CTAATCT
	I CTAATCT	C6r TTAAGCA	
	J CTAATCT	C7 TTAATCC	
		C8 TAAATCC	
		C9 CTAATCT	
CTAATCC	TTAATCY	TTAATCY	TTAAGCN

Table 4.1 *Drosophila*, *Musca*, *Lucilia* and *Calliphora* Bcd-binding site sequences (*D. melanogaster*: Driever and Nüsslein-Volhard 1989; Ma *et al.*, 1996; Yuan *et al.*, 1999. *Musca*: Shaw 1998; Bonneton *et al.*, 1997. *Calliphora* and *Lucilia*: McGregor *et al.*, 2001a). Consensus sequences for each species are shown in the last row in bold. For *Calliphora* and *Lucilia* the consensus sequences were calculated as described in 2.2.10. Where two cores are present in a protected region, f is the core on the forward strand and r is the core on the reverse strand.

It was not possible to quantify the affinities of Bcd for each protected region from these experiments because in most cases protection was only evident at the highest

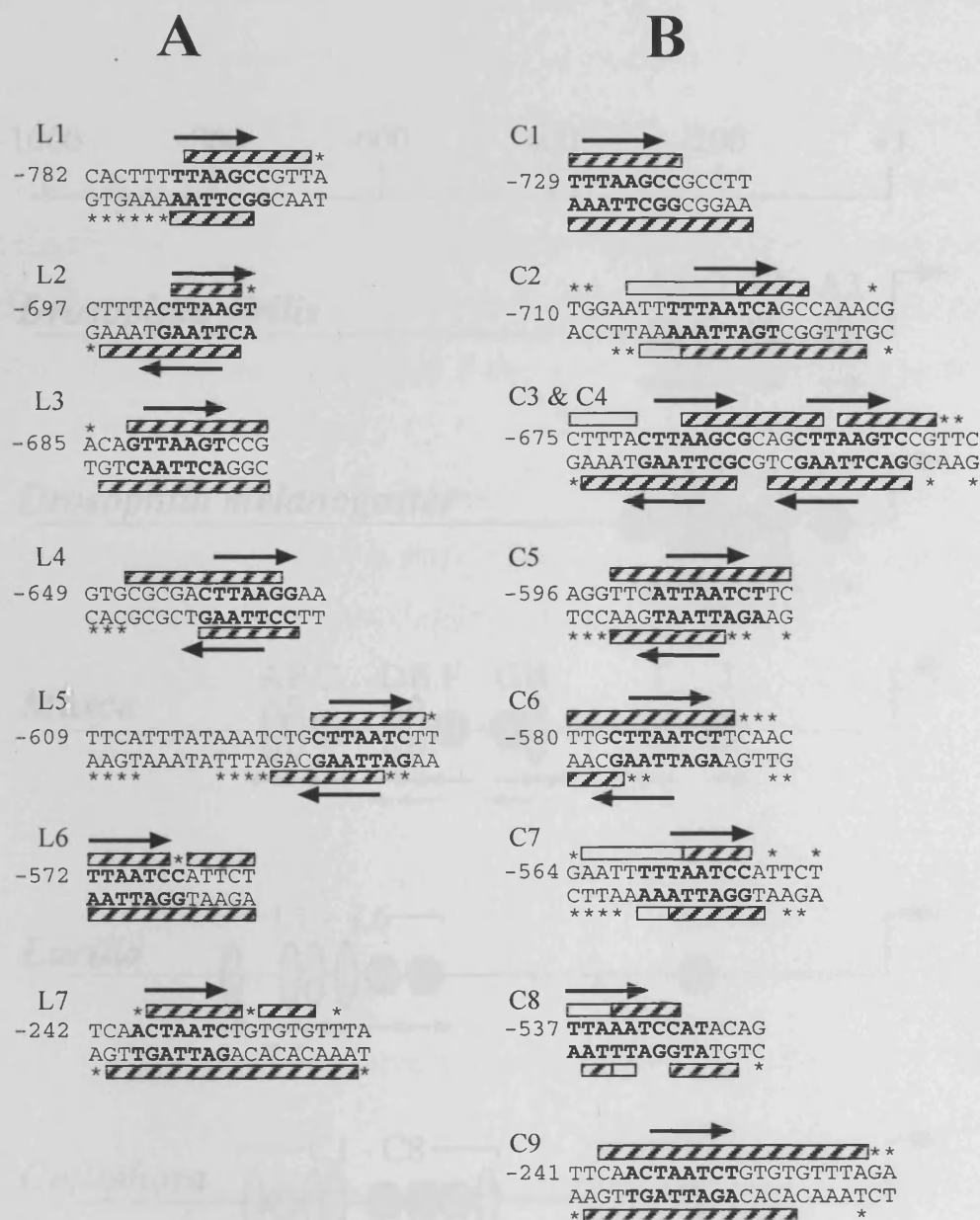


Figure 4.10 Summary of DNaseI footprinting Bcd-protected regions in the *Lucilia* (A) and *Calliphora* (B) *hb* P2 promoters regions. Sequences are shown for both strands of the protected regions and the numbering refers to the positions of these regions 5' of the transcription start site in either species. Hatched boxes indicate protected regions and asterisks mark hypersensitive sites. Arrows indicate the orientations of Bcd-binding site core sequences, which are highlighted in bold. White boxes represent regions where the protection could not be determined because the bands were weak in control lanes.

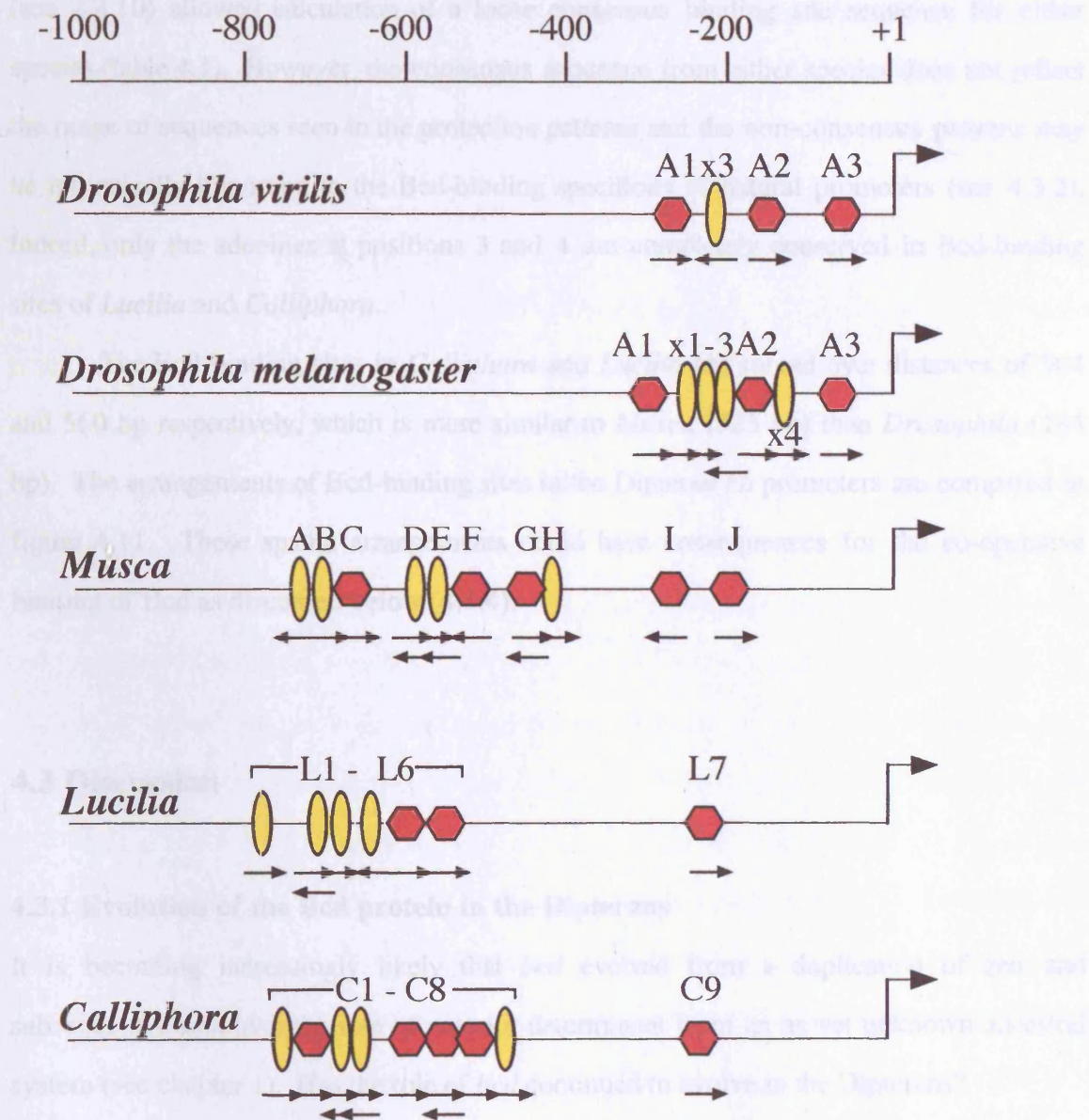


Figure 4.11 Bcd-dependent promoters of *hunchback* in higher Dipterans. The large arrow is the transcription start site. The numbered bar represents the distance in bp 5' from the transcription start site. Hexagons represent the positions of DNaseI footprinted (except for *D. virilis*) Bcd-binding sites with a canonical core sequence (TAAT), while the ovals represent sites with a non-canonical core sequence (TAAG, AAAT, CAAT, TCAT and TGAT). Binding sites are labelled according to previously published footprinting data (*Drosophila* and *Musca*). Smaller arrows represent the orientation of sites.

concentration of Bcd homeodomain. Table 4.1 summarises the core sequences of characterised *hb* Bcd-binding sites from *Calliphora*, *Lucilia*, *Musca* and *Drosophila*. Alignment of the protected region sequences of the *Calliphora* and *Lucilia hb* promoters (see 2.2.10) allowed calculation of a loose consensus binding site sequence for either species (table 4.1). However, the consensus sequence from either species does not reflect the range of sequences seen in the protection patterns and the non-consensus patterns may be intrinsically important to the Bcd-binding specificity of natural promoters (see 4.3.2). Indeed, only the adenines at positions 3 and 4 are completely conserved in Bcd-binding sites of *Lucilia* and *Calliphora*.

The Bcd-binding sites in *Calliphora* and *Lucilia* are spread over distances of 504 and 560 bp respectively, which is more similar to *Musca* (525 bp) than *Drosophila* (184 bp). The arrangements of Bcd-binding sites in the Dipteran *hb* promoters are compared in figure 4.11. These spatial arrangements could have consequences for the co-operative binding of Bcd as discussed below (4.3.4).

4.3 Discussion

4.3.1 Evolution of the Bcd protein in the Dipterans

It is becoming increasingly likely that *bcd* evolved from a duplication of *zen* and subsequently took over the role of anterior determinant from an as yet unknown ancestral system (see chapter 1). Has the role of *bcd* continued to evolve in the Dipterans?

The homeodomain and PEST domain of Bcd are generally conserved between the Dipterans with a few interesting differences (see 4.3.2). However, there is increased divergence in the less well characterised regions (figure 4.1) whose function is not known or is redundant (Schaeffer *et al.*, 1999; Janody *et al.*, 2001).

A number of studies (Driever *et al.*, 1989b; Struhl *et al.*, 1989; Schaeffer *et al.*, 1999; Janody *et al.*, 2000, 2001) have shown that the transcriptional activation function of Bcd is performed by the glutamine rich region located towards the C-terminal end of the

protein and by the S/T rich PEST domain (see figure 4.1). MAP kinase target residues in the PEST activation domain can be phosphorylated to modulate Bcd function. One consequence of this can be an increase in the strength of the morphogenetic Bcd gradient (Janody *et al.*, 2000). Many of the serine and threonine residues that can be phosphorylated *in vivo* are conserved in *Musca*, *Calliphora*, *Lucilia* and even in *Megaselia* (figure 4.1). However, the Calypttratae species also have a serine-rich domain adjacent to this activation domain (figure 4.1). Phosphorylation of these additional serines could result in Calypttratae specific modulation of Bcd activity. Furthermore, the glutamine-rich activation domain, which has been shown to interact with TAF_{II}110 (Sauer *et al.*, 1995a), is reduced in size in Calypttratae species and the PRD (Histidine-Proline-rich domain) is shorter in *Musca* (figure 4.1). However, the interaction between the Q-rich domain and TAF_{II}110 may be redundant for Bcd function in *Drosophila* (Schaeffer *et al.*, 1999). In addition the A-rich domain is also reduced in size in *Musca* (figure 4.1). In *Drosophila*, the A-rich domain may be involved in both transcriptional activation and repression mediated by interactions with TAF_{II}60 (Sauer *et al.*, 1995a) and dSAP18 (part of the Sin3/Rpd3 deacetylation complex, Zhu and Hanes 2000; Janody *et al.*, 2001) respectively.

Although the significance of these differences in Bcd between the Dipterans is not known, it is possible that such changes could result in divergent roles for the protein amongst these species. This could happen by direct modification of Bcd such as phosphorylation of species-specific residues. Alternatively, Bcd function could change in a species-specific manner through the co-evolution of Bcd and Bcd co-factors such as Chip (Torigoi *et al.*, 1999) or SAP18 (Zhu *et al.*, 2001). Indeed, it is possible that the role of Bcd has become more restricted to head development in species such as *Calliphora* during the course of Dipteran evolution (see 1.13 and figure 1.6).

4.3.2 Bcd binding to consensus and non-consensus sites

The recognition helix and N-terminal arm of the Bcd homeodomain and in particular residues arginine-3, arginine-5, isoleucine-47, lysine-50 and asparagine-51 (see 1.7 and figure 1.5) are conserved amongst the Dipterans (figure 4.2). This suggests that the

binding site sequence preferences of these proteins are also similar. Indeed, a number of the Bcd-binding sites characterised in *Calliphora* and *Lucilia* have sequences that closely resemble the consensus sites in both *Drosophila* and *Musca*. In particular sites L5, L6, L7, C2, C5, C6, C7 and C9 have TAATC core sequences (see table 4.1) followed by a C or a T at position 6.

However, 9 out of 16 of the Bcd-binding sites characterised in *Calliphora* and *Lucilia* vary from the *Drosophila* and *Musca* consensus sequences. Indeed, four *Drosophila* sites and four *Musca* sites do not match the consensus sites sequences either (table 4.1). Non-consensus Bcd-binding sites have been shown to bind the *Drosophila* Bcd homeodomain, albeit at lower affinities (Ades and Sauer 1995). It has been demonstrated that changes at positions 1 and 2 (TAATCC) lower the affinity of sites for the En (QK50) homeodomain up to 6 fold and 28 fold respectively (Ades and Sauer 1995). *Lucilia* site L8 and *Drosophila* site x4 have core sequences of AAATCC and CAATCC respectively; however, the other positions match nucleotides preferred by conserved residues in lysine-50 class homeodomains. Therefore, this change would only lower the affinity of these sites up to 6 fold.

Interestingly, a number of sites in *Calliphora* and *Lucilia* (L1, L2, L3, L4, C1, C3 and C4) resemble the non-optimal *Drosophila* sites x1 and x2 which have the sequence TAAGCT (table 4.1). These sites have a guanine at position 4 instead of a thymine, which is normally contacted by homeodomain residue isoleucine-47, and so this change would be expected to weaken the affinity of these sites for the Bcd-homeodomain (Ades and Sauer 1995). In particular, site L2 (TAAGTA) would be predicted to bind to Bcd with low affinity since it also lacks a cytosine at either bp 5 or bp 6. However two recent investigations (Dave *et al.*, 2000; Zhao *et al.*, 2000) have determined how the Bcd homeodomain binds to non-consensus sites and the importance of this binding in terms of transcriptional activation from natural Bcd-dependent promoters. These authors have demonstrated that when Bcd binds to non-consensus binding sites the arginine at residue 54 makes specific contact with the guanine of TAAG sites, while the other specific contacts in the major and minor grooves are maintained (see 1.7). Indeed, binding of the Bcd

homeodomain to x1 sites gave extended methylation interference patterns outside the basic recognition sequence, in comparison to A1 sites. This demonstrates that sites x1 and A1 are recognised in mechanistically different ways. Bcd is the only lysine-50 class homeodomain protein to have an arginine in its homeodomain at position 54 (Dave *et al.*, 2000) and this residue is conserved in all the Dipterans (figure 4.2). The altered specificity Ftz Q50K protein which does not have an arginine at position 54 can recognise optimal TAATCC Bcd-binding sites, but cannot bind TAAGCT sites and in addition fails to activate transcription from the natural *Drosophila hb* Bcd-dependent enhancer. Indeed, the complete structure of the Bcd homeodomain may be required to make specific contacts with both consensus (Tucker-Kellogg *et al.*, 1997) and non-consensus sites, because an Ftz QK50-R54 homeodomain still did not recognise x1 type sites (Zhao *et al.*, 2000; Dave *et al.*, 2000).

Therefore, it is possible that novel contacts between homeodomain variants and mixtures of multiple sites, composed of a range of sequences (a promoter signature), may be used by different proteins to generate specificity in their interactions with promoters. This mechanism in collaboration with co-operative Bcd binding (see 4.3.3) could be vital for the establishment of Bcd-dependent gene expression in the posterior regions of the embryo where the concentration of Bcd is lower. Furthermore, subtle differences between the Dipteran Bcd homeodomains in residues that do not directly contact DNA could have resulted in these homeodomains preferentially binding to different promoter signatures.

Interestingly, there are four differences between the homeodomains of the Calyptratae and *Drosophila* and single additional differences in *Musca* and *Calliphora* (figure 4.2). Three of these differences (A28S, P29S and A39T) result in a polar amino acid in the Calyptratae and a non-polar amino acid in *Drosophila*. Furthermore, in *Drosophila*, the amino acids at residues 11 and 60 are polar, but in the Calyptratae and *Calliphora* respectively, they are non-polar (figure 4.2). It is possible that the different signatures of the Dipteran *hb* promoters (figure 4.11) could have co-evolved with these differences in the Bcd homeodomains. This is discussed further in chapter 8.

4.3.3 Co-operative Bcd-binding

An important aspect of Bcd-binding to promoters containing binding sites of different sequences is co-operative binding mediated by Bcd protein-protein interactions. For example, both *in vitro* studies and *in vivo* studies (yeast) have shown that Bcd binding to an optimal binding site (A1) can induce binding of a second Bcd molecule to a non-optimal site (x1) (Ma *et al.*, 1996; Burz *et al.*, 1998). In general, it seems that Bcd monomers bound to one site induce the binding of a second Bcd at a second binding site in a pairwise DNA-dependent manner (Burz *et al.*, 1998; Burz and Hanes 2001; Ma *et al.*, 1996; Yuan *et al.*, 1996).

The N-terminal half of the Bcd protein including the homeodomain has been implicated in protein-protein interactions for the co-operative binding of Bcd to the *Drosophila hb* promoter (Yuan *et al.*, 1996). Zhao and co-workers (2001) demonstrated this further by using homeodomain swaps between Bcd and Ftz (Q50K), which determined that Bcd but not Ftz contained as yet undefined sequences outside of the homeodomain that are necessary for co-operative Bcd binding to the *Drosophila hb* promoter. These studies suggested that the homeodomain was required for co-operativity but could not interact with other Bcd molecules alone (Yuan *et al.*, 1996). However, a recent study of the Bcd homeodomain has demonstrated that a number of residues are involved in co-operative binding particularly at low Bcd concentrations (Burz and Hanes 2001 see figure 4.2). Interestingly, one of these residues (A28) is not conserved in the non-*Drosophila* species and residue T35 in *Megaselia* Bcd has the same amino acid as construct DB43 in Burz and Hanes 2001, which disrupted co-operative binding by the *Drosophila* Bcd homeodomain. The significance of these changes is not known and compensatory changes elsewhere in the Bcd protein cannot be discounted.

The requirement for these protein-protein interactions were even more apparent for the *kni* Bcd-dependent promoter which drives expression at the posterior limit of Bcd expression (Zhao *et al.*, 2000). The Bcd-dependent *kni* enhancer (*kni64*) contains six Bcd-binding sites within a 64 bp sequence (figure 4.12A) and surprisingly none of these six Bcd-binding sites match the optimal Bcd-binding sequence of TAATCC (Rivera-Pomar *et*

al., 1995). However, this element is able to drive expression of a reporter gene over the entire length of the *Drosophila* embryo (Rivera-Pomar *et al.*, 1995; Burz *et al.*, 1998) despite having no greater affinity for Bcd than the *Drosophila hb* enhancer in yeast transcriptional assays (Burz *et al.*, 1998 and see 6.4).

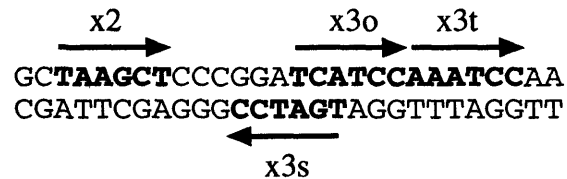
4.3.4 Spacing of Bcd-binding sites

It has been suggested that the arrangement of Bcd-binding sites in the *kni64* enhancer allows greater co-operative binding of Bcd and therefore is more sensitive to lower Bcd concentrations enabling it to drive *kni* expression in the posterior of the *Drosophila* embryo (Burz *et al.*, 1998; Yuan *et al.*, 1999). The six sites in the *kni64* enhancer are arranged in head to tail symmetrical pairs (Rivera-Pomar *et al.*, 1995 and see figure 4.12A). When this arrangement is manipulated to change the orientation of binding sites reporter gene expression at lower Bcd concentrations is adversely affected (Burz *et al.*, 1998). Using *in vitro* selection assays it has been shown that co-operative binding by Bcd has strict preferences for the spacing and orientations of binding sites (Yuan *et al.*, 1999). Bcd binds preferentially to sites arranged tail-tail separated by 7-15 bp and to sites arranged head to head separated by 3 bp. In the *Drosophila hb* enhancer only sites x2 and x3 are optimally arranged head to head separated by 3 bp (figure 4.12B). However, when either of these two binding sites are mutated transcription falls by up to 10 fold, which is a larger effect than obtained when any of the consensus binding sites in this enhancer are mutated (Ma *et al.*, 1999). A deletion in the *D. virilis hb* promoter compared to *D. melanogaster* has removed site x2 in this species (Lukowitz *et al.*, 1994 and figures 4.11 and 4.12C); however, this positions site x3 closer to A1 so that they are separated by 18 bp head to head. Furthermore, the sequence immediately downstream of A1 in *D. virilis* has diverged from that of *D. melanogaster* and this has created two non-consensus Bcd-binding sites at a distance of 4 bp head to head with A1, 9 bp head to head with x3s and 16 bp tail to tail with x3o (figure 4.12C). These sites in *D. virilis* could allow co-operative Bcd-binding and compensate for the loss of x2.

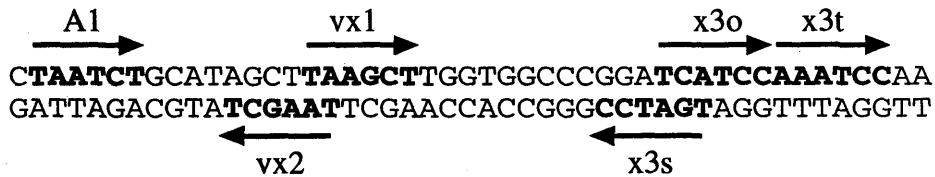
A. *Drosophila melanogaster kni*



B. *Drosophila melanogaster hb*



C. *Drosophila virilis hb*



D. *Musca domestica hb*

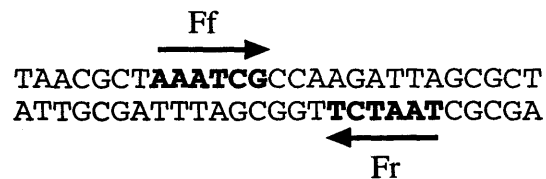


Figure 4.12 Optimally spaced Bcd-binding sites in Dipteran promoters. Shown are the six *kni* sites (A) and *hb* sites x2 and x3 (B) in *D. melanogaster*. A deletion in *D. virilis* with respect to *D. melanogaster* has removed x2, bringing x3 closer to A1 and two additional putative Bcd-binding core sequences (vx1 and vx2 in C). Within *Musca* site F there are two characterised cores separated 3 bp head to head (D).

Are any Bcd-binding sites in the *Calliphora* and *Lucilia* promoters optimally spaced? *Lucilia* sites L2 and L3 are arranged tail to tail separated by 9 bp and so are optimally spaced according to the above criteria. Sites L5 and L6 could be similarly arranged and separated by 18 bp. In *Calliphora* sites C3 and C4 are possibly arranged tail to tail separated by 9bp, although there are various possible arrangements of these sites due to Bcd-binding sites on both strands (figure 4.10) in a similar pattern to *Drosophila* x3 (see below). *Calliphora* sites C5 and C6 could also be optimally arranged for co-operative Bcd binding as they contain core Bcd-binding sequences possibly arranged 3 bp apart head to head. In addition sites C6 and C7 are arranged tail to tail 16 bp apart. In *Musca* the sites are generally further apart but within site F there are two Bcd-binding sites optimally arranged head to head 3 bp apart (figure 4.12D). However, it has not yet been determined if the spacing of sites in *Musca*, *Calliphora* and *Lucilia* are critical for transcription.

The coliphage HK022 repressor is also bound co-operatively and it has been shown that this is dependent on the spacing of the two operator sites. Indeed, different spacing of these sites induces different conformations of bound protein-DNA complexes (Mao *et al.*, 1994). This may be an important consideration in the specific DNA and protein contacts made by Bcd molecules co-operatively binding to sites of variable sequence and spacing in different Dipteran species (see 4.3.1).

Studies of Ubx binding properties (Beachy *et al.*, 1988, 1993; Ekker *et al.*, 1991) have shown that the optimal Ubx binding site sequence is TTAATGG. However, in the *Ubx* and *Antp* promoters Ubx binds to overlapping TAA tandem repeats. It is possible that these repeats act as promoter localisation points for Ubx proteins and thus facilitate co-operative Ubx binding to nearby binding sites. This mechanism may also be intrinsic to Bcd-binding sites such as x3 in *Drosophila* and sites C3 and C4 in *Calliphora* (figures 4.10B and 4.12B), which contain multiple Bcd-binding site cores on both strands and so these may be mechanistically akin to the repeats in Ubx-dependent promoters.

It has also been demonstrated that longer spaces between Bcd-binding sites allows more efficient co-operative Bcd-binding and there are less constraints on the specific

spacing and orientation of sites at larger distances (Yuan *et al.*, 1999). Given that in general the *hb* promoters of the Calyptratae have increased numbers of sites spread over a larger distance are these promoters more sensitive to lower concentrations of Bcd? Indeed, are these promoter configurations a reflection of the larger embryo sizes of the Calyptratae (figure 1.6 and see chapter 8)?

4.3.5 Do other transcription factors bind to the *Calliphora* and *Lucilia hb* promoters?

It has been suggested that the expression of genes in the anterior of the *Drosophila* embryo requires synergy between Bcd and Hb (Simpson-Brose *et al.*, 1994; Sauer *et al.*, 1995a, 1995b). Searches for possible putative Hb-binding sites in the *Calliphora* and *Lucilia hb* promoters are not straightforward because the consensus sequence for Hb-binding sites is very loose (ACNAAAAAANTA see Treisman and Desplan 1989). There is a number of thymine and adenine repeats in the *Calliphora* and *Lucilia* promoters but these would require footprinting by Hb to determine their nature. There is also evidence that *Kr* plays a minor role in setting the posterior boundary of *hb* PS4 expression (Wu *et al.*, 2001), but again footprinting would be required to define *Kr* binding sites which have a loose consensus sequence of AAGGGGTAA (Treisman and Desplan 1989).

The GAGA sequence binding transcription factor (GAF) is involved in the expression of a number of segmentation genes in *Drosophila* (Read *et al.*, 1990; Wilkins and Lis 1997). Interestingly, there are a number of putative GAF binding sites in the *Musca*, *Calliphora* and *Lucilia hb* promoters and possibly fewer in the *Drosophila* promoter (see 6.4.3). As it is known that GAF also interacts with the Bcd co-factor dSAP18 (Espinosa *et al.*, 2000), perhaps *hb* expression is up-regulated in the larger Calyptratae embryos using a mechanism involving synergy between GAF and Bcd?

Chapter 5
Analysis of intra-specific and inter-specific
variation in *hb*

5.1 Introduction

At first glance the sequences of the *Lucilia* and *Calliphora hb* promoters are unalignable with the sequence of the *Musca hb* promoter. Characterisation of Bcd-binding regions in the *Lucilia* and *Calliphora* promoters (chapter 4) revealed that, while they contain Bcd-binding sites similar to those found in the *Musca* promoter, the configurations of sites have changed between these species, in terms of number, sequence, orientation and spacing. How can the mutational mechanisms that have given rise to these different promoter configurations be investigated? Since variation between species initially arises as variation within a species, investigation of the intra-specific variation in *hb* within *Musca domestica* may reveal the patterns of polymorphisms that have given rise to the differences in *hb* that are observed between species.

What analytical tools can be used to relate the differences within a species to differences between species? Classically, intra-specific variation has been compared with inter-specific variation to suggest the adaptive evolution of the alcohol dehydrogenase protein in *Drosophila* (McDonald and Kreitman 1991). However, this analysis of amino acid replacements cannot be used to compare the promoter polymorphisms within a species to the differences in unalignable promoters between different species.

Sequence alignments are the primary tool used to compare regions of genes between species when investigating patterns of conservation and divergence. Conserved regions are usually associated with function (such as transcription factor binding sites in promoters), and non-conserved regions (such as the sequences between transcription factor binding sites) are usually described as non-functional and exposed to the whims of drift. Hence, in the absence of selection for a particular function mutations may accumulate in sequences within a population. However, changes in sequence can result in functional divergence, for example, in the *Hoxc8* early *cis*-regulatory element (see 1.2).

When sequences such as the *hb* promoters described here have diverged so far as to be unalignable, it is possible to use other methods of analysis to identify shared patterns amongst them. Features that are shared between sequences such as nucleotide bias or repetitive motifs can be illustrated using dotplots (2.2.10). However, dotplots do not allow

the precise sequences of shared motifs that cause similar patterns between sequences to be identified. Small repetitive sequence motifs can be tandemly repeated (pure) simple sequences or they can be interspersed with other motifs and termed as cryptically simple sequences (Tautz *et al.*, 1986). For example, the trimers highlighted in the sequences below illustrate the two classes of simplicity. In the purely simple sequence CAG is tandemly repeated, but in the cryptically simple sequence different motifs are interspersed with each other:

'pure' simplicity = **CAG**CAGCAGCAGCAGCAGCAGCAG

'cryptic' simplicity = GG**CAG**TAA**CAG**GCTAA**CAG**TAA

It is thought that such repetitive motifs are generated by replication slippage (Levinson and Gutman 1987; Hancock 1996) and that their evolution is subject to complex mechanisms involving slippage and other features of genomic turnover (such as point mutations, gene conversion and unequal crossing over, see Tautz *et al.*, 1986 and Dover 1993). Slippage-like mechanisms have been proposed to explain microsatellite composition and distribution in yeast and in *Drosophila* (Kruglyak *et al.*, 2000; Schug *et al.*, 1997, 1998). The sequences and distributions of simple repetitive motifs that are responsible for the shared patterns in dotplots can be analysed using the SIMPLE 34 program (Hancock and Armstrong 1994 and see below).

It has previously been proposed that slippage-like mutations in clusters of repetitive motifs have been responsible for the divergence of regions of the *Drosophila* and *Musca hb* genes and of the P2 promoters in particular (Hancock *et al.*, 1999). Are the structures of the *Musca*, *Calliphora* and *Lucilia hb* promoters the result of this mutational mechanism, balanced by compensatory selection for binding site configurations that maintain the transcriptional function of these promoters? To address this possibility, three regions of the *hb* gene were sequenced in six strains of *Musca* and these sequences were analysed using the SIMPLE 34 to investigate the distribution of polymorphisms found in each region. In addition, the distribution of simple sequence motifs was analysed in the

Lucilia and *Calliphora hb* promoters using the SIMPLE 34 program and compared with the distribution of repeats in the *Musca* promoter.

5.2 Methods

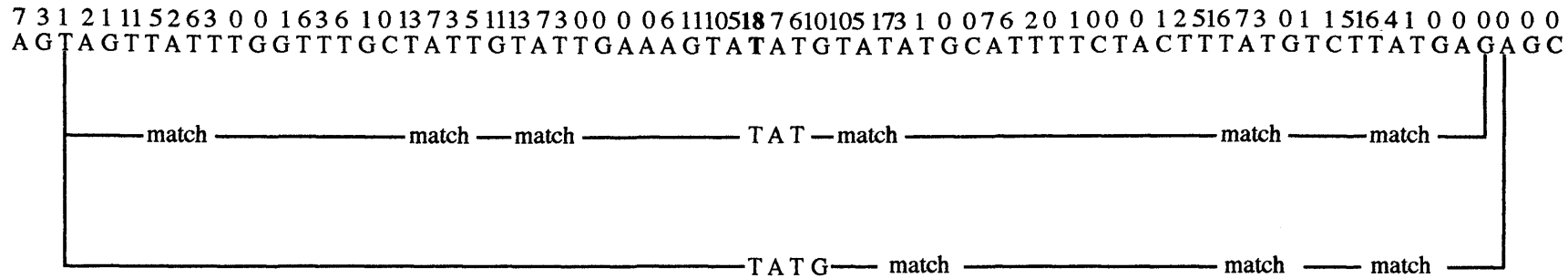
5.2.1 The analysis of simplicity

For this chapter, simple sequence profiling was carried out using the SIMPLE 34 program (Hancock and Armstrong 1994) with the help of Dr J. M. Hancock from the Department of Computer Science, Royal Holloway, University of London.

The SIMPLE 34 program can be used to investigate the tri and tetranucleotide repeat frequencies (repeat clustering) in a sequence. Thus, this program can be used firstly, to determine the positions of polymorphisms in a particular sequence within a species in relation to the positions of repetitive motifs and secondly, to investigate the occurrence of shared motifs between unalignable sequences, such as the Dipteran *hb* promoters.

A score is calculated for each nucleotide in a sequence based upon the sum of repeats of the particular tri and tetranucleotide motifs that begin at each nucleotide in the sequence. For example, in a sliding 64 bp window (32 bp upstream and 32 bp downstream of the each nucleotide), a score of 1 is given each time the trinucleotide is repeated and a score of 4 is given each time the tetranucleotide is repeated (figure 5.1). Therefore, the score of a given nucleotide position reflects the repetition of the motif starting at that nucleotide in the 64 bp window. A simplicity profile of a sequence can be generated by plotting the score for each nucleotide in the sequence against the nucleotide positions (for example, see figure 5.3).

The Overall Simplicity Factor (OSF) of a sequence is the sum of all the scores in that sequence divided by the number of nucleotides in the sequence. The Relative Simplicity Factor (RSF) of a sequence can then be calculated by dividing the OSF of that sequence by the mean OSF of ten random sequences with the same length, base



No of tetranucleotide matches: 3 (x4)= 12
 +
 No of tri nucleotide matches: 6 6
 Score = 18

Figure 5.1 How the SIMPLE 34 program calculates the score for a given nucleotide in a test sequence

Each nucleotide is given a score (the numbers directly above the sequence) based on the number of trinucleotide and tetranucleotide motifs beginning at that nucleotide that are repeated in a 64 bp window. This window extents 32 bp upstream and 32 bp downstream of the nucleotide in question. For example, the T highlighted in red is the first nucleotide in the trimer TAT and the tetramer TATG. In the 64 bp window around the T there are 6 TAT matches and 3 TATG matches and therefore it has a score of 18. The window then slides to the next nucleotide and repeats the process.

The trinucleotides are given a score of 1 and the tetranucleotide number is multiplied by four because in a random sequence composed of 25% of each nucleotide, 1 match with a trinucleotide and 1/4 match with a tetranucleotide would be expected in a 64 bp window.

The window size and trinucleotide and tetranucleotide motifs were originally chosen as a balance between reducing background noise and optimising signal match. The program has a correction system that excludes overlapping matches such as the TAT beginning 6 nucleotides to the right of the example used above.

Source: Tautz *et al.*, 1986.

composition and base doublet composition. Sequences with little or no simplicity would be expected to have RSF close to 1 and a higher RSF if they contained simple regions. The mean and variance of OSFs derived from the 10 randomised sequences allows the statistical significance of the RSF of a natural sequence to be assessed. Thus, the RSF can be used to compare the amount of simplicity between sequences.

How can the significance of the particular motifs that cause high peaks in simplicity profiles be measured to allow such motifs to be compared between sequences? The significance of a given score for a particular motif beginning at a given nucleotide in a sequence can be depicted by the Significance-value (S-value). This is calculated by dividing the highest scoring nucleotide at which that motif begins in the averaged randomised sequences by the score for the first nucleotide of the motif in the sequence of interest and then subtracting the answer from 1. Hence, the S-value never exceeds 1. For example, if motif AAAA begins at a nucleotide with a score of 16 in a sequence and the score of the motif is 5 in the averaged randomised sequences, then the S-value of AAAA at that particular position is 0.688. Negative S-values mean that the motif in question is underrepresented in the test sequence compared with the randomised sequences.

For the study of small indels in the *Musca hb* gene, which are likely to be representative of slippage events, each region analysed was divided into sequence domains of 50 bp divisions in the promoter, 5' UTR and coding regions. The average simplicity score for each domain and the average simplicity score for each region as a whole (OSF) were calculated from the previous analysis of the *Musca* Cooper strain sequence (Hancock *et al.*, 1999). The average simplicity score for each sequence domain is compared with the average simplicity score for the sequence region as a whole (region averages: promoter 3.36, 5' UTR 4.72, coding 2.72). High and low simplicity domains have scores above and below the region average respectively (see table 5.1).

5.2.2 Sequencing of three regions of *hb* from six strains of *Musca*

To investigate intra-specific variation in the *hb* gene within *Musca*, three regions of the gene were analysed in 5 different strains of *Musca* (Cardiff, Edinburgh, Rutgers, White

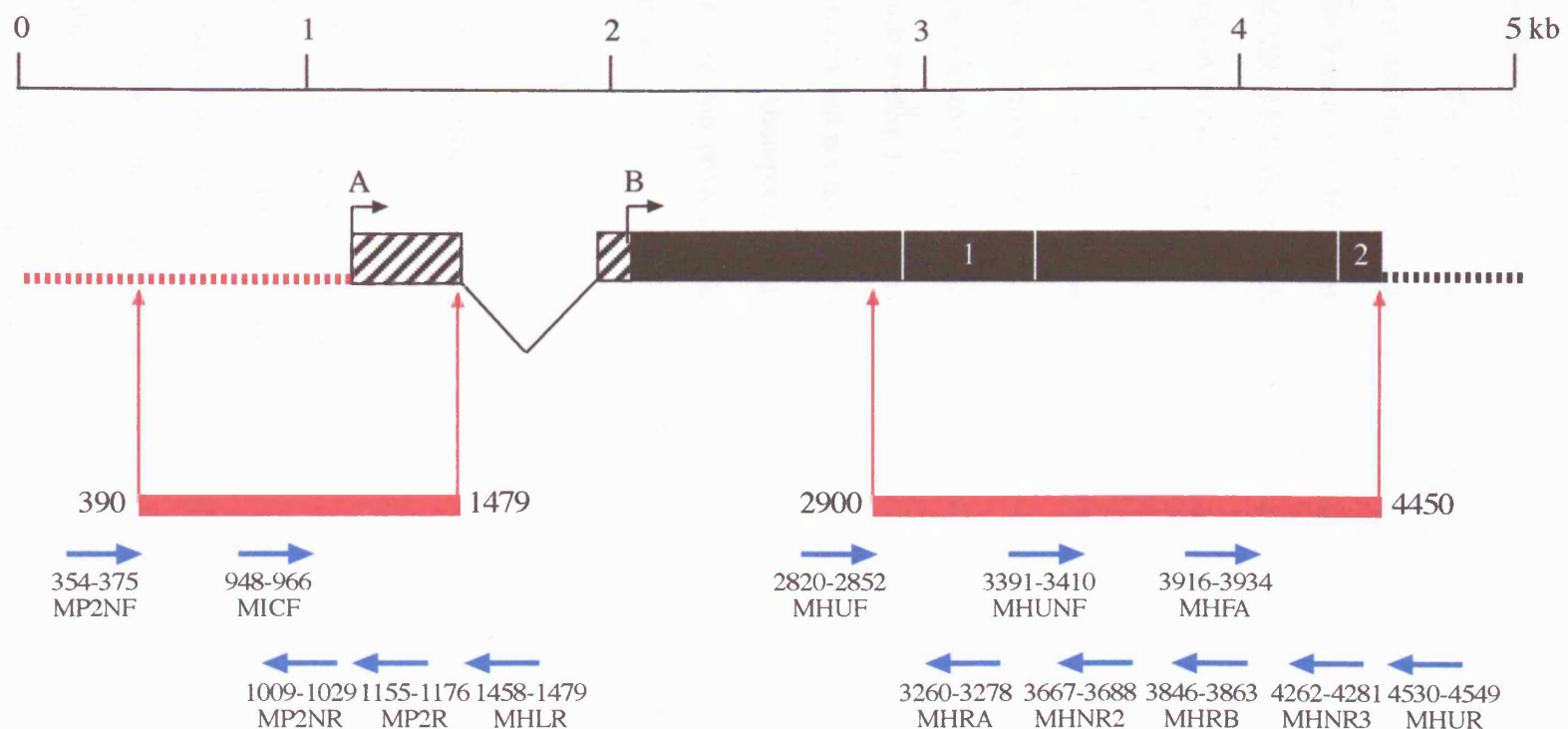


Figure 5.2 Regions of the *Musca hb* gene sequenced in six different strains

The thick black lines represent the coding region and the two zinc finger domains are indicated (1 and 2). The dashed black line is the 3' UTR. The dashed red line represents the promoter region and the 5' UTR is represented by the striped boxes separated by the diagonal lines, which represents the intron. The transcription and translation start sites are indicated by arrows labelled A and B respectively. The thick red lines show the regions sequenced in all six strains between the bases indicated by the numbers. Numbering is from the *Musca Cooper* strain sequence accession number Y13050. The numbered blue arrows represent the primer positions (see 5.2.2).

and Zurich see chapter 2), in addition to the strain in which *hb* was originally sequenced (cooper, see Bonneton *et al.*, 1997). These *Musca* strains were continuous laboratory strains originally obtained from isolated wild type cultures in which the isogenic status was not known or controlled.

The *hb* P2 promoter between bp 390 and 1154 (from -764 to the transcription start site) and the 5' UTR between bp 1155 and 1479 were amplified using genomic DNA from the 5 strains of *Musca* as templates. These PCRs were performed using primers MP2NF, MP2R, MP2NR, MHLR and MICF (figure 5.2 and appendix A). Similarly, the coding region between bp 2900 and 4450 was amplified from the 5 *Musca* strains using primers MHUF, MHUNF, MHUR, MHN2, MHN3, MHFA, MHRA and MHRB (figure 5.2 and appendix A). These regions of *hb* were then sequenced from each strain using the above primers and universal primers, either directly from the PCR products, or after cloning the products first. Sequences were obtained from both strands of at least two independent PCR products to verify that any sequence differences observed between the strains had not been introduced artifactually by PCR.

Attempts were also made to sequence the *hb* intron in each *Musca* strain; however, this region proved difficult to amplify using PCR and therefore was not used in this analysis.

5.3 Results

5.3.1 Intra-specific polymorphisms in *Musca domestica hb*

The sequences for the three regions of *hb* from each *Musca* strain (including Cooper), obtained as described in 5.2.2, were aligned using the Clustal W program (Thompson *et al.*, 1994). The polymorphisms for each *hb* region between the strains are summarised in table 5.1 and the alignments are shown in full in appendix B.

In the coding region, six nonsynonymous base differences were discovered, none of which were found in the functionally important domains of the Hb protein, such as the

zinc-fingers or the C and D boxes (Hülkamp *et al.*, 1994 and see 3.3.1). Four indels were found in the coding region; interestingly, two were found amongst CAG/CAA repeats coding for glutamine and one was found in a CAT repeat coding for histidine (see appendix B and figure 5.3).

Both base substitutions and indels were also found in the promoter region and the 5' UTR. It is possible that some of these apparent base substitutions could have resulted from slippage and subsequent mismatch repair (Schlötterer and Tautz 1992) since they occur at the ends of monopolymeric repeats, particularly in the 5' UTR. In the promoter region the sequences of all ten characterised Bcd-binding sites were the same, despite extensive indel polymorphisms in other regions of the P2 promoter between the six strains, although a single nucleotide indel was found immediately 5' to Bcd-binding site G (see appendix B).

There appears to be either an excess of base substitutions in the silent sites of the coding region or a deficit of base substitutions in the promoter and 5' UTR (table 5.1). The possible explanations for this phenomenon are discussed below (5.4.1).

Region of <i>hb</i> (size in bp)	Base substitutions	Non-synonymous	Indels in high simplicity sequence	Indels in low simplicity sequence	P
P2 (764)	24	-	5	8	NS
5'UTR (321)	8	-	5	1	<0.05
Coding (1593)	79	6	4	0	<0.01

Table 5.1 Base substitutions and indels in the *Musca hb* gene between six strains. P was calculated from χ^2 with one degree of freedom and represents the probability of indels and high simplicity coinciding by chance (see 5.2.1 for definitions of low and high simplicity).

Akashi (1994) has suggested that selection would oppose silent changes in DNA binding domains to promote the translational fidelity of functionally important domains. However, in this analysis of *hb*, the non-synonymous base substitutions in the coding

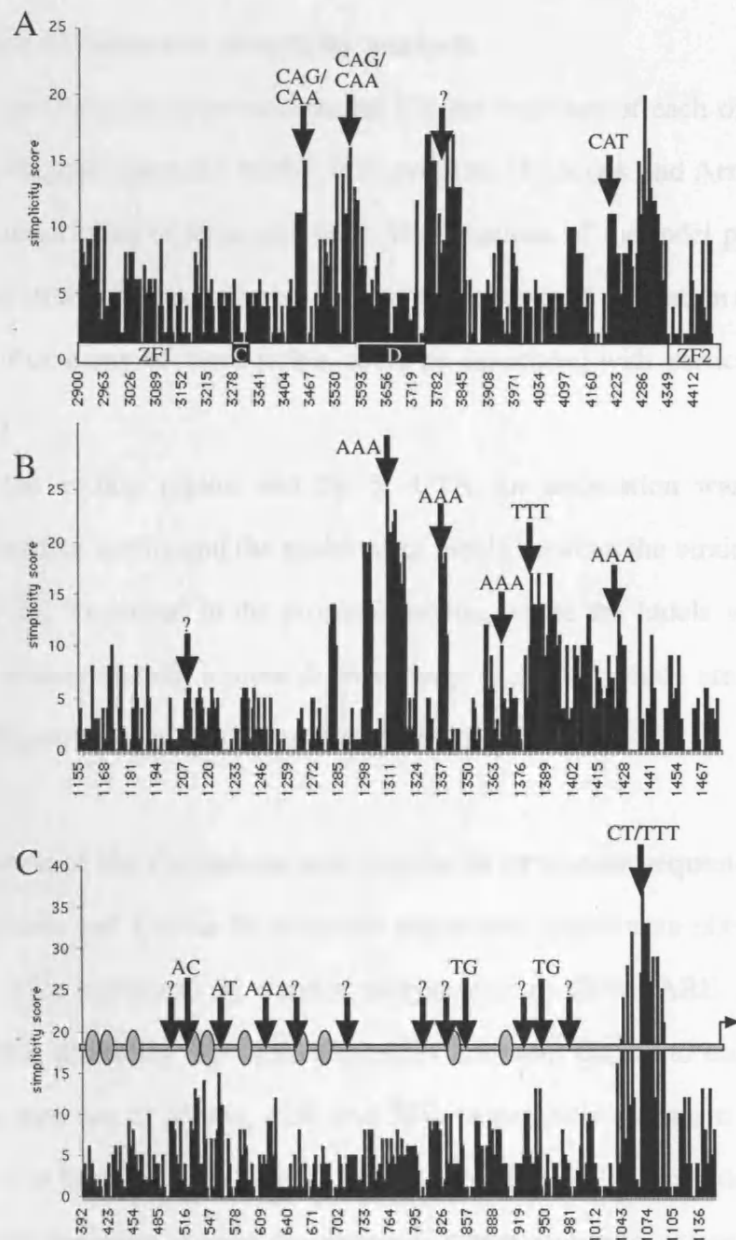


Figure 5.3 Simplicity profiles for the three regions of the *hb* gene in *Musca*: (A), coding region, (B), 5' untranslated leader (C), promoter. Simplicity scores were calculated previously using the SIMPLE34 program (see 5.2.1). Numbering of the sequences is from the *Musca* Cooper *hb* sequence, accession number Y13050. The large downward arrows represent the approximate positions of the indel differences found between the *Musca* strains with respect to the Cooper strain sequence. The motif sequence at each indel is shown where it could be identified. Question marks indicate indels where no particular motif could be identified. The grey ovals are the positions of the Bicoid binding sites in the promoter. The smaller arrow represents the transcription start site. ZF1 and ZF2 represent the two zinc finger encoding regions, while the black boxes labelled C and D encode domains which are phylogenetically conserved and likely to be involved in aspects of *hb* function (Hülskamp *et al.*, 1994). Indel positions were calculated with respect to the Cooper strain sequence from the alignments in appendix B.

region are found equally in sequences that code for functional domains (such as the zinc-fingers), and in sequences that do not encode any characterised function (data not shown).

5.3.2 *Musca hb* sequence simplicity analysis

Simplicity profiles were generated for the Cooper sequence of each of the three regions of the *Musca hb* gene using the SIMPLE 34 program (Hancock and Armstrong 1994 and see 5.2.1 for a description of these profiles). The locations of the indel polymorphisms found between the strains for each *hb* region were then examined in relation to these profiles, thus illustrating that many of these indels could be associated with particular sequence motifs (figure 5.3).

In the coding region and the 5' UTR, an association was found between the clusters of simple motifs and the positions of indels between the strains (figure 5.3A, 5.3B and table 5.1). However, in the promoter region, where the indels were more numerous, they were associated with a more diverse range of motifs, which are present in both low and high frequency clusters (figure 5.3C and table 5.1).

5.3.3 Analysis of the *Calliphora* and *Lucilia hb* promoter sequences

The *Calliphora* and *Lucilia hb* promoter sequences, which were obtained as described in chapter 3, were compared by dotplot analysis (using COMPARE and DOTPLOT, see 2.2.10). This illustrates that these sequences are more similar to each other, 61% (figure 5.4A), than they are to *Musca*, 42% and 38% respectively (compare the horizontal line in figure 5.4A to the pattern in figure 5.4B). Interestingly, in both dotplot comparisons a similar cross-matching pattern is observed, which suggests extensive sharing of short repetitive motifs in all three sequences (see 5.1). The presence of these motifs throughout the promoter sequences is further illustrated by the dotplot in figure 5.5, which shows the crossmatching patterns observed when the *Calliphora* sequence is compared to itself.

To investigate the sequence and distribution of these short repetitive motifs further, SIMPLE 34 analysis was performed on the *Calliphora* and *Lucilia hb* promoter sequences. It was found that despite a lower overall simplicity compared with the *Musca*

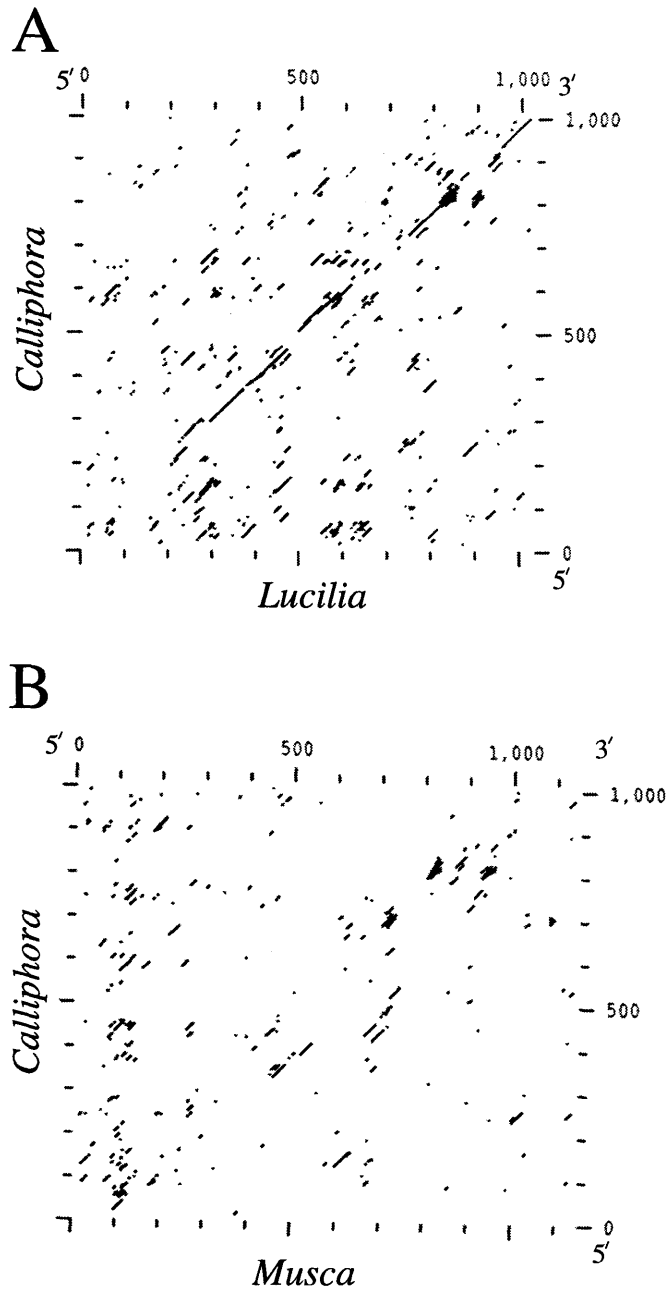


Figure 5.4 Dot-plots of inter-specific sequence comparisons of *hb* P2 promoters. (A) *Calliphora* and *Lucilia* (B) *Calliphora* and *Musca*. Stringency of 19 base perfect match in a window of 35 bp in the COMPARE algorithm was used for each dotplot. The numbering is from the start of the promoter sequence to the transcription start site in each species. A similar dotplot to B was observed when *Musca* and *Lucilia* were compared (not shown).

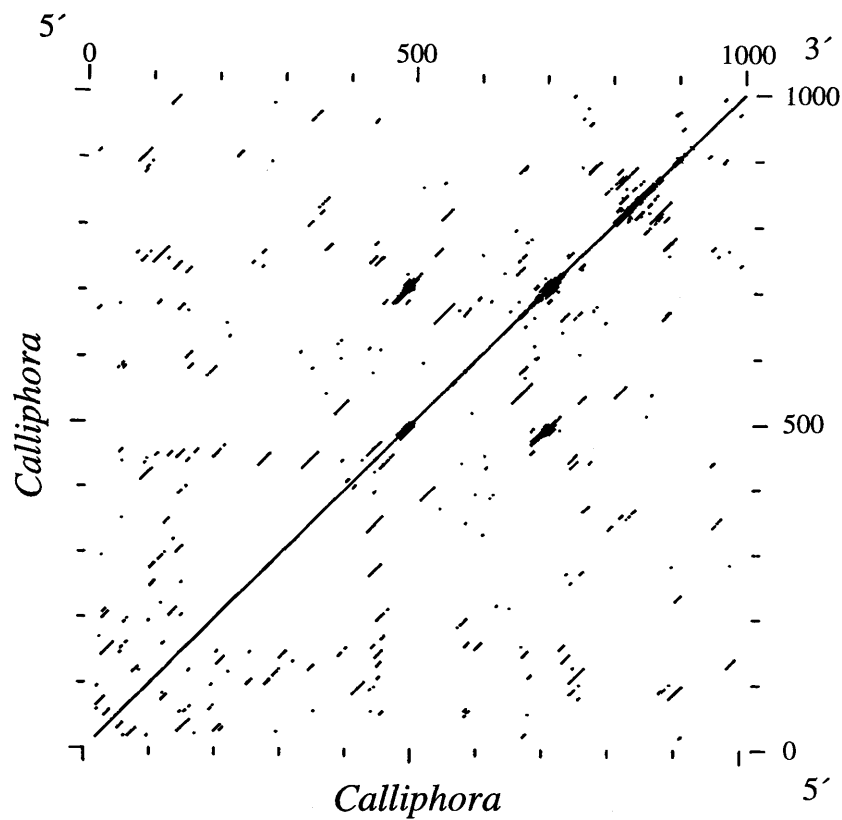
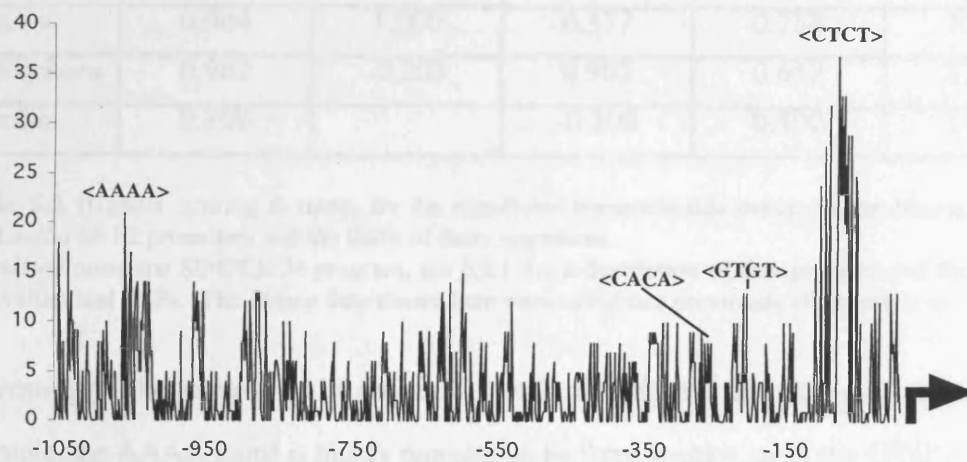
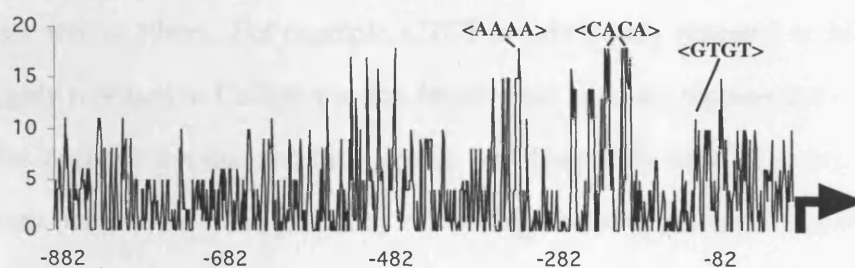


Figure 5.5 Dot-plot of intra-specific sequence comparison of the *Calliphora hb* P2 promoter. Stringency of 19 base perfect match in a window of 35 bp in the COMPARE algorithm was used for this dotplot. The numbering is from the start of the promoter sequence to the transcription start site.

Musca domestica



Calliphora vicina



Lucilia sericata

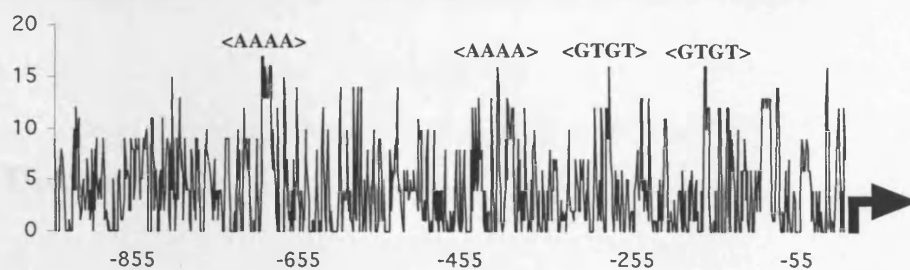


Figure 5.6 Simplicity profiles of the *hb* P2 promoter sequences from *Musca*, *Calliphora* and *Lucilia*. These were generated as described in 5.2.1. The Y-axis is the simplicity score for each nucleotide position and the X-axis the position in the sequence. Numbering of the sequences is upstream from the transcription start site in each species. Arrows indicate the transcription start sites. The positions of the highest scoring tetranucleotide repeats analysed in table 5.2 are indicated.

promoter (represented by the RSFs in table 5.2), this analysis showed that the blowfly promoters contained some highly repeated sequence motifs (table 5.2 and figure 5.6).

	AAAA	CTCT	CACA	GTGT	RSFs
<i>Musca</i>	0.964	1.000	0.337	0.768	1.628
<i>Calliphora</i>	0.962	-0.203	0.962	0.652	1.492
<i>Lucilia</i>	0.890	-	-0.200	0.890	1.427

Table 5.2 Highest scoring S-values for the significant tetranucleotide motifs in the *Musca*, *Calliphora* and *Lucilia hb* P2 promoters and the RSFs of these sequences.

Calculated using the SIMPLE 34 program, see 5.2.1 for a description of this program and the definitions of S-values and RSFs. The *Musca* data shown here were calculated previously (Hancock *et al.*, 1999).

Interestingly, the same highly repeated motifs are shared by different species. For example, the AAAA motif is highly repeated in all three species, as is the GTGT motif to a slightly lesser extent. On the other hand, some motifs are highly repeated in some species, but are rare in others. For example, CTCT is only highly repeated in *Musca*, while CACA is highly repeated in *Calliphora* and *Musca*, but is under represented in *Lucilia* as shown by the negative S-value and thus occurs less frequently than expected by chance in this species (table 5.2). The positions of the high scoring tetra-nucleotide motifs shown in table 5.2 were plotted against the simplicity profiles of each promoter sequence (generated as described in 5.2.1) and their distribution in the promoter sequences is shown in figure 5.6.

5.4 Discussion

5.4.1 Patterns of *hb* polymorphisms within *Musca domestica*

Variations were found between six strains of *Musca* in three regions of the *hb* gene (table 5.1 and appendix B). Interestingly, this variation found in *Musca hb* was high compared with intra-specific variation in *D. melanogaster hb* (Tautz and Nigro 1998). For example these authors found only two base substitutions in the P2 promoter and 12 polymorphic sites in the 5' UTR (3 of which were single base indels in mononucleotide repeats) between 12 strains of *D. melanogaster*. This may reflect a different demographic history

in *Musca* compared with *D. melanogaster* where selective sweeps have been proposed to have removed variation in this lineage (Tautz and Nigro 1998). Alternatively, low slippage rates in *Drosophila* could account for the lower level of polymorphisms (Schug *et al.*, 1997).

A high degree of association between the positions of indel polymorphisms and clusters of simple motifs was found in the 5' UTR and the coding region (table 5.1). This association is in accordance with the hypothesis that short repeated sequence motifs in *hb* sequences are prone to slippage-like mutations (Hancock *et al.*, 1999; Hancock and Vogler 2000).

While there were higher numbers of indels in the promoter and 5' UTR than in the coding region, it appeared that there were less base substitutions in these regions than at the silent sites in the coding region (table 5.1). A possible explanation for this could be that a high rate of slippage in the promoter and 5' UTR removes point mutations particularly in repetitive motifs. This is supported from studies of microsatellites in *Drosophila*, which have shown that slippage acts to prevent microsatellite repeat decay by removing point mutations (Harr *et al.*, 2000; Santibanez-Koref *et al.*, 2001).

The lower frequency of indels in the *hb* coding region is probably due to selective constraints to maintain the reading frame and protein function. Consequently, indels in this region are restricted to glutamine and histidine repeats. Polymorphic CAG repeats coding for glutamines have also been found between *hb* genes in *Drosophila* species (Treier *et al.*, 1989; Tautz and Nigro 1998). Indeed, the Hb glutamine repeats exhibit length variations between the Dipterans including *Calliphora*, *Lucilia* and *Megaselia* (see figure 3.2). Glutamine repeats can act as transcriptional activation domains (Emili *et al.*, 1994), but the significance of the slippage generated length differences within and between species is not known. However, glutamine repeat instability has been reported in many genes and species often with deleterious consequences, as in human disease (Orr 2001; Hancock *et al.*, 2001). Slippage has probably also contributed to the evolution of repetitive codons in other genes, such as *mastermind* (Newfeld *et al.*, 1994) and *period* (Peixoto *et al.*, 1992), which may be associated with protein functions. Indeed, a recent comparison of

orthologous *Drosophila* and *Tribolium* genes (Schmid and Tautz 1999) demonstrated that the *Drosophila* genes generally encode longer proteins with longer amino acid repeats, which these authors suggest have been generated by slippage. In addition to indels, six amino acid differences were found in the *hb* coding region between the strains (appendix B), although none of these were in phylogenetically conserved domains of the Hb protein (Hülkamp *et al.*, 1994 and see 3.3.1).

The percentage GC content of the *Musca hb* coding region is similar to that of the promoter (52% and 42% respectively), while the 5' UTR is AT rich (26% GC). This is reflected in the sequence and clustering of simple motifs affected by indels in each region of *hb*, such as in the CAG/CAA repeats in the coding region. In the 5' UTR, indels are mainly seen in the motifs AAAA and TTTT, which are present in high frequency clusters (figure 5.3B). These indels are probably tolerated by low selective constraints. Indeed, analysis of slippage has demonstrated that AT rich motifs are subject to slippage-like mutations at a greater rate than GC rich motifs *in vitro* (Schlötterer and Tautz 1992). Unlike the *Musca hb* 5' UTR that of *D. melanogaster* has no AT rich motifs and therefore this could explain the size difference of this region between these species (Shaw 1998). It has been suggested that slippage-like mutations in clusters of simple motifs have contributed to genome sizes in eukaryotes (Hancock 1995) since much of the size differences between genomes occurs in the non-coding regions (Cavalier-Smith 1985). Interestingly, it has been estimated that the genome of *Musca* is 5 times larger than the *Drosophila* genome (Davidson 1986). Therefore, it is possible that slippage-like mutations in simple sequence clusters have contributed to this difference and that this is reflected in the expansion of the 5' UTR and promoter regions of *hb* in *Musca* and in the blowflies (see figure 3.9).

5.4.2 Intra-specific and inter-specific variation in the *hb* promoter

The indel variation in the *Musca hb* P2 promoter demonstrates that this sequence is subject to slippage-like mutations, as was previously predicted (Hancock *et al.*, 1999). These mutations occur between Bcd-binding regions in clusters of motifs, at both low and high

frequencies. The absence of base substitutions in Bcd-binding sites might reflect the action of selection to maintain binding specificity and affinity, compared to greater tolerance of base substitutions and indels between binding sites.

The higher GC content in the *hb* coding region and promoter means that a wider range of sequence of motifs are present than in the 5' UTR (see above). In the promoter, where there are presumably less selective constraints than in the coding region, there are more indel polymorphisms due to turnover of a variety of motifs present at low and high repeat frequencies. Therefore, the lack of association between indels and high frequency motif clusters in the promoter (figure 5.3C and table 5.1) is probably due to the high turnover of several motifs present at different repeat densities. This mechanism called 'motif scrambling' was previously proposed to explain the absence of motifs with high S-values in *Tribolium hb* despite evidence of extensive repeat distribution (Hancock *et al.*, 1999).

Interestingly, despite the differences between the *Musca*, *Calliphora* and *Lucilia hb* promoters, cross-matching dotplot patterns and the sharing of repeated motifs reveal patterns of sequence conservation between the *hb* promoters in these species (see figures 5.4, 5.6 and table 5.2). This paradox of shared motifs and restructured promoters could be explained by the species-specific 'scrambling' of a relatively small number of motifs giving rise to different promoter configurations. In addition selection may promote compensatory mutations lying between binding sites in order to maintain correct binding site spacing important for co-operative interactions between bound Bcd proteins and co-activators (see 4.3.4). This is supported by analysis of another promoter in *Drosophila* species. Studies of the *eve* stripe II enhancer have shown that indels and point mutations, in some cases affecting transcription factor binding sites, have resulted in the divergence of this element between *Drosophila* species (Ludwig and Kreitman 1995). However, analysis of the expression patterns driven by inter-specific chimeric *eve* promoters has shown that the species-specific changes have probably evolved in a compensatory manner to preserve promoter function (see 1.4; Ludwig *et al.*, 2000). Hancock and Dover (1990) proposed that compensatory slippage allowed rRNA secondary structures to be maintained while the

sequence diverged. This also appears to have happened in the *hb* 3' UTRs between *Drosophila* and *Musca* (Shaw 1998; Shaw *et al.*, 2001).

It has recently been suggested that new transcription factor binding sites could evolve very quickly by base substitutions; for example, in less than 75 years for a *de novo* Hb binding site in the *eve* stripe II enhancer. Given that the rate of slippage generated mutations is approximately 100 times greater than point mutations (Schug *et al.*, 1997, 1998, Schlötterer *et al.*, 1998), a combination of both processes could mean that promoter structures can evolve very quickly and the different structures of the Dipteran *hb* promoters are examples of this.

The possible consequences of the continual restructuring of *hb* promoters by slippage to the evolution of genetic regulatory networks are discussed in chapter 8.

Chapter 6

Functional analysis of the Bcd-*hb* interaction

6.1 Introduction

6.1.1 Have *bcd* and the *hb* P2 promoters co-evolved in *Drosophila* and *Musca*?

It has been suggested that the differences between *bcd* and the *hb* promoters of *Drosophila* and *Musca* have co-evolved, resulting in species-specific interactions (Bonneton *et al.*, 1997). Binding affinity studies have shown that *Musca* Bcd preferentially binds to sites flanked by a thymine rather than a cytosine immediately 5' of the TAAT core (TTAATCC), whereas *Drosophila* Bcd prefers a cytosine (Shaw 1998; Wilson *et al.*, 1996). Furthermore, it has been suggested that a threonine rather than an alanine at homeodomain position 39 gives *Musca* Bcd greater tolerance for a thymine at position 6 (TAATCT) than *Drosophila* Bcd (Shaw 1998). However, these affinity studies have been contradicted by more recent experiments in which the binding preferences of *Musca* and *Drosophila* Bcd were not clear (Shaw *et al.*, submitted).

Given other critical features of the Bcd-*hb* promoter interaction, such as co-operative binding, site spacing and orientation (see 4.3), binding site affinities alone are only one feature of the transcriptional potential of natural Bcd-dependent enhancers. Indeed, these affinity studies (using band shift assays) employed only the homeodomain of Bcd and promoter fragments rather than the whole promoter (Shaw *et al.*, submitted). Indeed, such experiments do not consider the consequences of properties intrinsic to the whole protein and full-length promoters to the transcriptional output. Therefore, to compare the interactions between whole Bcd proteins and full length *hb* promoters in and between different species, a system is required in which the transcriptional output of these interactions can be measured.

6.1.2 A yeast system to investigate Bcd-dependent transcription

A number of previous studies of *bcd* and *hb* have employed yeast based systems to investigate features such as transcriptional activation, DNA binding properties and the effects of protein phosphorylation (Driever *et al.*, 1989b; Struhl *et al.*, 1989; Hanes and

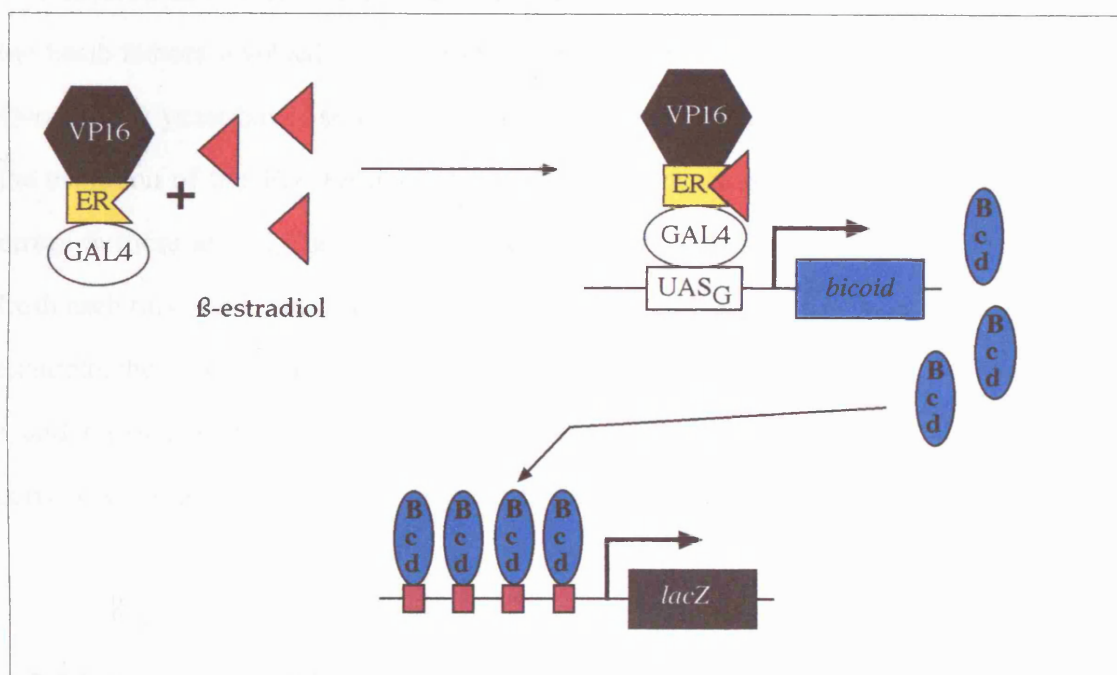
Brent 1989, 1991; Hanes *et al.*, 1994; Ma *et al.*, 1996, 1999; Yuan *et al.*, 1999; Zhao *et al.*, 2000; Burz *et al.*, 1998; Burz and Hanes 2001; Zhu *et al.*, 2001).

The system used by Burz and co-workers (1998) allows the concentration of Bcd in yeast cells to be varied over nearly three orders of magnitude. This is comparable to the two to three orders of magnitude change in Bcd concentration in the anterior of *Drosophila* embryos (Driever and Nüsslein-Volhard 1988a). Thus, this system mimics the natural Bcd concentration gradient (see figure 6.1A). However, the total cellular concentration of Bcd produced by this system in yeast can be approximately 6 orders of magnitude higher than the physiological concentration of transcription factors in embryos (Burz *et al.*, 1998, table 1; Krause *et al.*, 1988). Burz and co-workers tested the transcriptional output of a range of synthetic and natural enhancers containing Bcd-binding sites placed upstream of *lacZ*, at different concentrations of Bcd. This strategy demonstrated co-operative Bcd-binding preferences to different configurations of binding sites (see 4.3.4).

For this thesis, the above yeast system was employed to investigate and compare transcription when using the *Drosophila*, *Musca* and *Megaselia* Bcd proteins and *Drosophila* and *Musca* *hb* promoters in homogenous and heterogeneous combinations. Consequently, any co-evolutionary consequences of the differences in *bcd* and the *hb* promoters between these species were tested.

The use of this yeast strategy to investigate the Bcd-*hb* interaction had a number of advantages and disadvantages. As described above, this system had previously been used to investigate Bcd-dependent transcription from the *Drosophila* *hb* promoter and generate reproducible results. The use of a yeast-based system means that the effect of species-specific transcription co-factors can be mitigated unless they are absolutely essential for transcription. However, it was also possible that endogenous yeast factors would be responsible for any apparent differences between the fly species elements being tested. For example, it has been shown that the yeast transcriptional machinery preferentially activates transcription from closely spaced binding sites (11 bp) than from sites spaced further apart (25 bp) (Hanes *et al.*, 1994). It is possible that this effect is due to

A



B

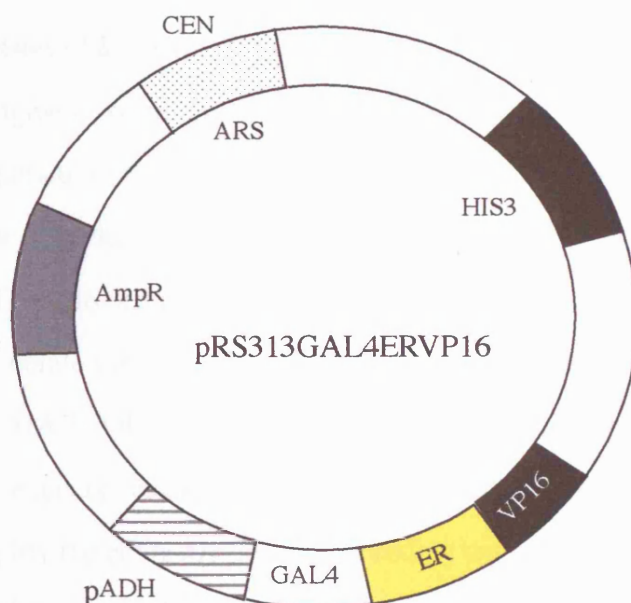


Figure 6.1 A. Yeast based strategy to study the Bcd-*hb* interaction

The GAL4 (DNA binding domain), ER (ligand binding domain), VP16 (activation domain) fusion protein is activated when the ER domain binds to its ligand, β -estradiol (red triangles). Activated fusion protein can then initiate expression of *bcd* from the *GAL1* promoter upstream of *bcd* (inserted into pBC103, see figure 6.2A and 6.2.1). Bcd can then activate *lacZ* expression by binding to Bcd-binding sites (small red squares) upstream of this reporter gene (in pLR1 Δ 1 derivatives, see figure 6.2B and 6.2.2). Therefore, the concentration of Bcd is dependent on the concentration of β -estradiol added to the yeast cultures (Burz *et al.*, 1998 and see 6.1.2) and the output of different promoter configurations at various Bcd concentrations can be measured.

B. pRS313GAL4ERVP16 shuttle vector, which expresses the GAL4ERVP16 fusion protein from the ADH promoter (Louvion *et al.*, 1993).

differences in the size of transcription factor complexes between yeast and flies (Goodrich and Tjian 1994). Yeast orthologs of factors implicated in Bcd-dependent regulation of transcription have been found such as Sin3p (Kasten *et al.*, 1997) and this suggests that the basic factors involved in transcriptional regulation are similar between yeast and flies. Overall, this yeast-based strategy represented a straightforward and quick method to test the evolution of the Bcd-*hb* interaction. It should be stressed that given the systematic errors in these assays, such as in the concentration of hormone since it had to be made fresh each time, only those assays carried out at the same time are truly comparable. For example, the independent results that the Hanes group obtained using this system (in rows 1 and 1A of table 6.1) highlights the differences that can accrue when these assays are carried out at different times (Burz *et al.*, 1998; Burz and Hanes 2001).

6.2 Materials and Methods

6.2.1 Construction of *bcd* expression vectors

A *Musca bcd* fragment containing the entire coding region (including the stop codon) was generated by restriction of pBCDR1 (see table 2.1) with *Eco*RI and *Hind*III. Vector pBC103 was also cut with these enzymes to remove the stuffer fragment and ADH region (figure 6.2A). This allowed the insertion of *Musca bcd* in frame with the HA-tag 5' to 3' in pBC103 to generate vector pBCMBCD.

Primers MABCDF and MABCDR (see appendix A) were designed based on the *Megaselia bcd* sequence and PCR was performed, using pMASB (see table 2.1) as the template, to amplify the entire *Megaselia bcd* coding region from the start codon up to and including the TAA stop codon. The resulting product was cloned and 3 clones were sequenced to verify that errors had not been introduced in the sequence artefactually by PCR. When primers MABCDF and MABCDR were designed, *Eco*RI and *Hind*III restriction sites sequences were included respectively in each. Thus, restriction of *Megaselia bcd* clones with *Eco*RI and *Hind*III generated *bcd* fragments that were

subsequently inserted in frame into pBC103 (as described above) to generate the *Megaselia bcd* expression vector pMABCD.

6.2.2 Construction of *hb* promoter *lacZ* reporter vectors

Primers MYPF and MYPR (appendix A) were designed based on the *Musca hb* P2 promoter sequence and each included *SalI* restriction sites. PCR was performed employing these primers and using pD1 (table 2.1) as the template. A 785 bp fragment of the *Musca hb* promoter from -38 to -823, which included all 10 Bcd-binding sites was amplified (figure 6.3). This fragment was then cloned to verify the sequence and these clones were subsequently digested with *SalI* to release the fragment with over-hanging ends. This allowed the *Musca hb* promoter to be inserted into the *XhoI* site of pLR1Δ1 (figure 6.2B), thereby destroying this restriction site. The orientation of the *Musca hb* promoter with respect to the reporter gene was determined by PCR between the promoter and *lacZ* using promoter specific primers and primer lacZ148 (appendix A). pMhbp2+ has the *Musca hb* promoter orientated 5' to 3' with respect to *lacZ*.

Fragments of the putative *Lucilia* and *Calliphora hb* P2 promoters were amplified by PCR using primers LYHF, CYHF and BFYR, which included *SalI* restriction site sequences (appendix A). These fragments included the *Calliphora* promoter region from +21 to -786 and the *Lucilia* promoter region from +21 to -876, which contain all of the characterised Bcd-binding sites in these promoters (figure 6.3). In addition, the *Calliphora* promoter region from -473 to -786 was amplified using primers CYHF and CV7SAL (appendix A), which contains 8 of the characterised Bcd-binding sites (figure 6.3). The sequences of these fragments were verified as described above for the *Musca* promoter. The *Calliphora* and *Lucilia* promoter fragments were then inserted into pLR1Δ1 and the orientations determined as described above (vectors pLSP+, pLSP-, pCVP+, pCVP-, pCVP5+ and pCVP5- in table 2.1).

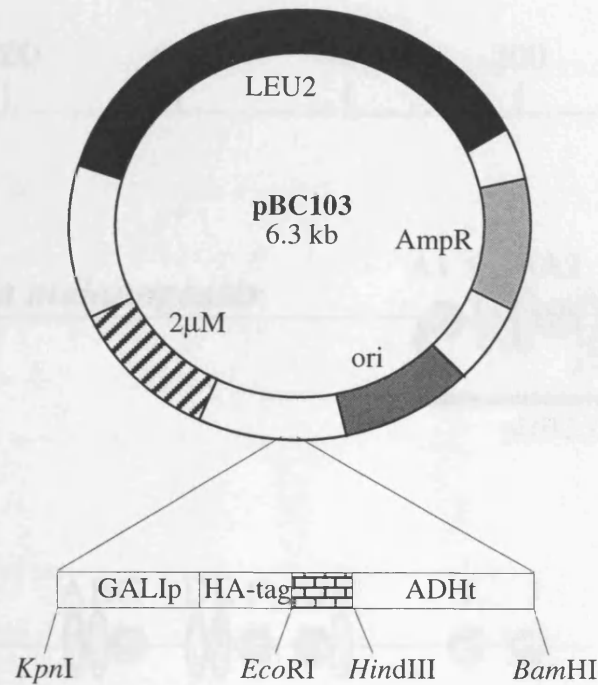
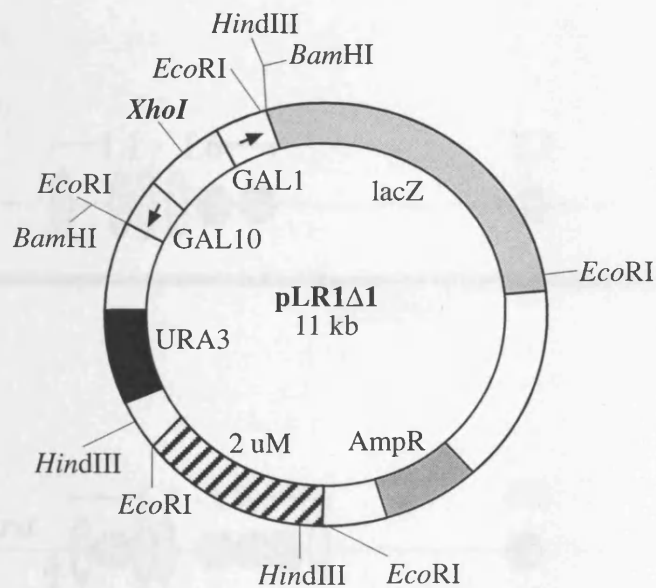
A**B**

Figure 6.2 A. pBC103 (Burz *et al.*, 1998) shuttle vector. Removal of the stuffer fragment (bricked region) and ADH-terminator (ADHt) allows the insertion of a *bcd* cDNA in frame with the HA-1 nonapeptide encoding sequence. Expression of *bcd* can then be controlled using the *GAL1* promoter (see 6.1.2 and 6.2.1).

B. pLR1Δ1 (West *et al.*, 1984) shuttle vector. Expression of the reporter gene *lacZ* can be placed under the control of Bcd by inserting DNA sequences containing Bcd-binding sites into the *XhoI* site shown in bold font (Burz *et al.*, 1998 and see 6.2.2).

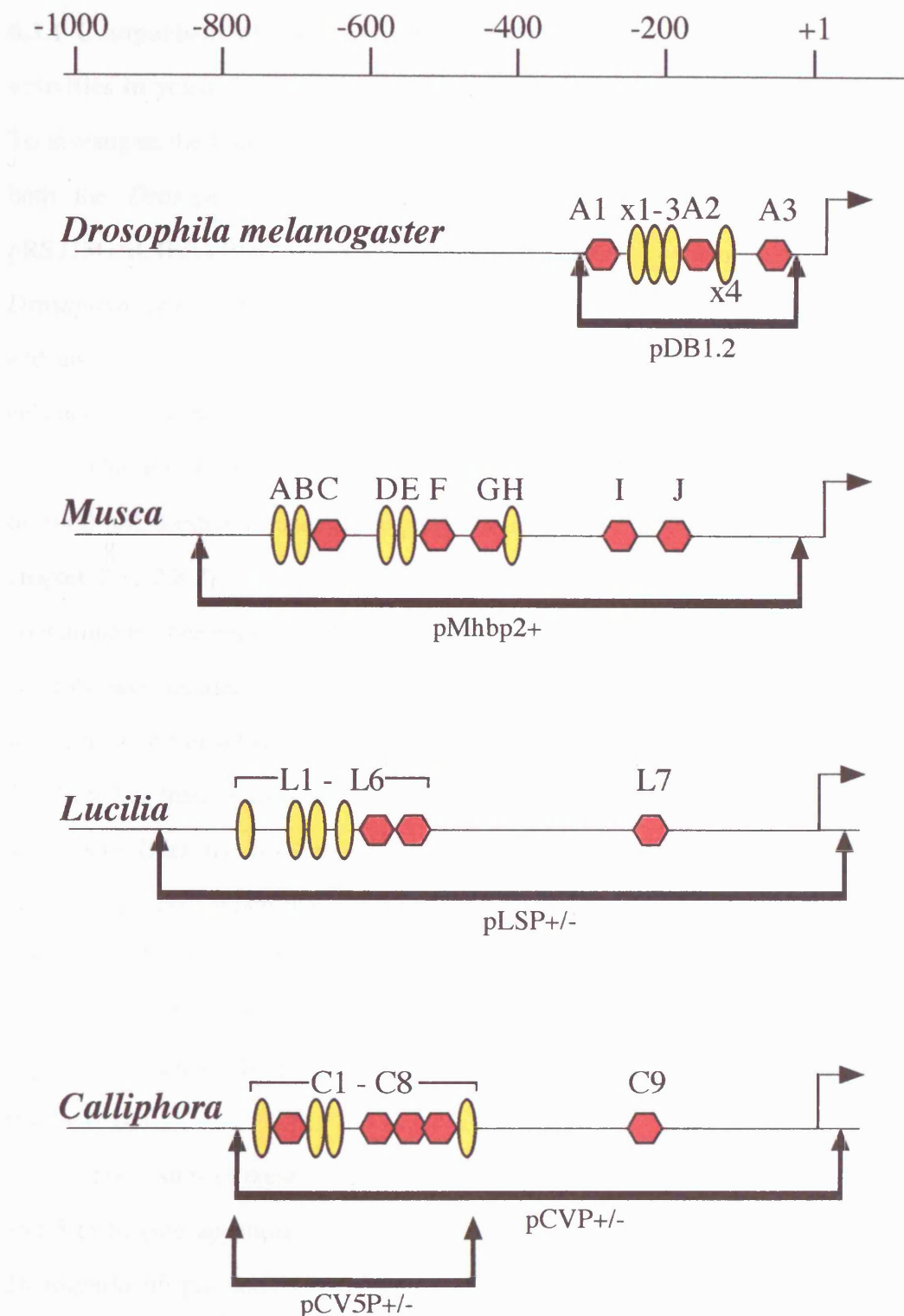


Figure 6.3 Regions of Bcd-dependent (P2) *hb* promoters used in yeast transcription studies. The large arrow is the transcription start site. The numbered bar represents the distance in bp 5' from the transcription start site. Hexagons represent the positions of DNaseI footprinted Bcd-binding sites with a canonical core sequence (TAAT), while the ovals represent sites with a non-canonical core sequence (see figure 4.11). Regions of the promoters between the vertical arrows were inserted upstream of *lacZ* in pLR1Δ1 (see 6.2.2) to generate the plasmids indicated (see table 2.1).

6.3 Results

6.3.1 Comparison of the *Drosophila*, *Musca* and *Megaselia* Bcd transcriptional activities in yeast

To investigate the transcription of *Drosophila bcd*, *Musca bcd* and *Megaselia bcd* from both the *Drosophila* and *Musca hb* promoters, yeast were co-transformed with pRS313GAL4ERVP16 (fig 6.1B) and either pDB1.2 (which expresses HA-tagged *Drosophila* Bcd, see 6.1.2 and table 2.1), pBCMBCD, pMABCD (see 6.2.1) or pBC103 and also either pDBhb.19 (which contains the 230 bp *Drosophila hb* Bcd-dependent enhancer upstream of *lacZ*, see 6.1.2, table 2.1 and figure 6.3) or pMhbp2+ (see 6.2.2).

Cultures were grown in triplicate in media containing either 0 nM, 2.5 nM, 10 nM or 1000 nM β -estradiol and then β -galactosidase assays were performed as described in chapter 2 (2.2.8.2). Yeast containing the reporter constructs with pBC103 and those containing the *bcd* constructs with pLR1 Δ 1 were used as negative controls. These negative controls gave reporter gene expression between 0 and 10 units of β -galactosidase activity, which here and in subsequent experiments was defined as background activity (appendix C). Note that there is expression of *bcd* at 0 nM β -estradiol as quantified using Western analysis by Burz and co-workers (see 6.1.2) and this has been shown to activate reporter gene expression. When there is no *bcd* inserted into pBC103 there is no reporter gene expression (Burz *et al.*, 1998). As had previously been reported (Burz *et al.*, 1998), the growth of yeast cultures was comparatively slower at 1000 nM β -estradiol and this may explain the lower levels of *lacZ* activity seen at this concentration for all three Bcds tested (figure 6.4A).

The results of these assays are summarised in figure 6.4 and table 6.1 (rows 2 to 4 and 5 to 6) (see appendix C for the results of individual assays). In assays using the *Drosophila hb* promoter (figure 6.4A), *Drosophila* Bcd activated higher levels of *lacZ* activity than did either *Musca* Bcd or *Megaselia* Bcd. This was evident at all hormone concentrations except 0 nM β -estradiol at which *Drosophila* Bcd and *Musca* Bcd gave a

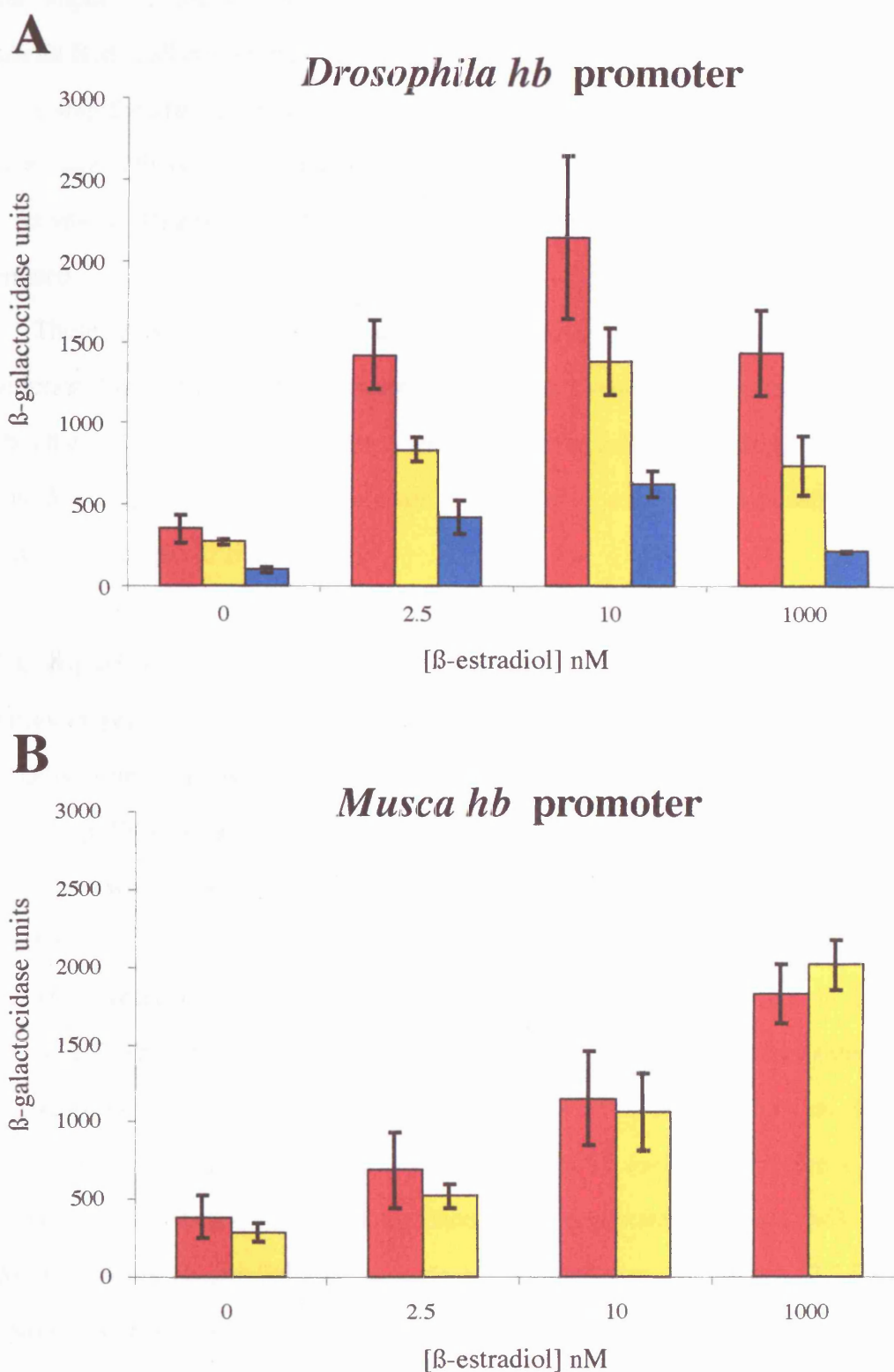


Figure 6.4. Transcriptional activation from the *Drosophila hb* promoter (**A**) and from the *Musca hb* promoter (**B**) using *Drosophila* Bcd (red columns), *Musca* Bcd (yellow columns) and *Megaselia* Bcd (blue columns). Assays For **A** and **B** were carried out at different times. β -galactocidase units are defined as the amount which hydrolyses 1 μ mol of ONPG to o-nitrophenol and D-galactose per min per cell. The hormone β -estradiol was used to vary the concentration of Bcd (Burz *et al.*, 1998 and see materials and methods). All reactions in **A** and **B** were carried out at the same time and the average was taken from a minimum of three cultures at each hormone concentration. The error bars represent the standard deviation at each hormone concentration. See appendix C for the results of individual assays.

similar output. In addition, *Musca* Bcd yielded higher levels of *lacZ* activity than did *Megaselia* Bcd at all concentrations.

Using the *Musca hb* promoter both *Drosophila* Bcd and *Musca* Bcd gave similar levels of *lacZ* activity at all hormone concentrations (figure 6.4B). Note that the reporter gene activity of *Megaselia* Bcd in combination with the *Musca hb* promoter was not determined.

These experiments show that with the *Drosophila hb* promoter, a higher transcriptional output is generated in combination with *Drosophila* Bcd than with *Musca* Bcd, but there is little difference between *Drosophila* Bcd and *Musca* Bcd in combination with the *Musca* promoter. Are these results supported by experiments in which the two promoters are compared side by side?

6.3.2 Comparison of the *Drosophila* and *Musca hb* P2 promoters transcriptional activities in yeast

Yeast cells were co-transformed (see 2.2.8.1) with pRS313GAL4ERVP16 and either pDBhb.19, pMhbP2+ (see above) or pLR1Δ1. Each of the reporter plasmid containing yeast cultures were subsequently transformed with either pDB1.2, pBCMBCD (see 6.2.1) or pBC103.

These yeast cultures were used to compare the *lacZ* reporter gene output driven by the *Drosophila hb* and *Musca hb* promoters when activated by different concentrations of either *Drosophila* Bcd or *Musca* Bcd. Negative controls were performed as described above (6.3.1) and again these gave background levels of *lacZ* activity (see 6.3.1 and appendix C). Cultures were grown in triplicate in media containing either 0 nM, 2.5 nM, 10 nM or 1000 nM β -estradiol and β -galactosidase assays were performed (2.2.8.2). All the assays described here were carried out at the same time.

The results for the output of each promoter when activated by either *Drosophila* or *Musca* Bcd are summarised in figure 6.5 and table 6.1 (rows 7 to 10) (see appendix C for the results of each individual assay). *lacZ* activity at 2.5 nM hormone was lower than for the cultures containing the same plasmids as those in 6.3.1 (see 6.1.2). When the

promoters are compared, in general the *Drosophila hb* promoter drove slightly higher average *lacZ* activity than did the *Musca hb* promoter in interactions with both *Drosophila* Bcd (figure 6.5A) and *Musca* Bcd (figure 6.5B). When the transcription factors are compared again, it appeared that while *Drosophila* Bcd activated higher *lacZ* activity from the *Drosophila* promoter than *Musca* Bcd did, the two Bcd proteins performances were more similar with the *Musca* promoter (table 6.1: compare rows 7 with 8, and 9 with 10), which supports the results described above (6.3.1).

Therefore, a difference was observed again when *Musca* and *Drosophila* Bcd were compared side by side on each promoter. However, no major difference was evident between the promoters themselves when they were compared side by side and this could be a reflection of transcriptional preferences intrinsic to yeast (see 6.4.1).

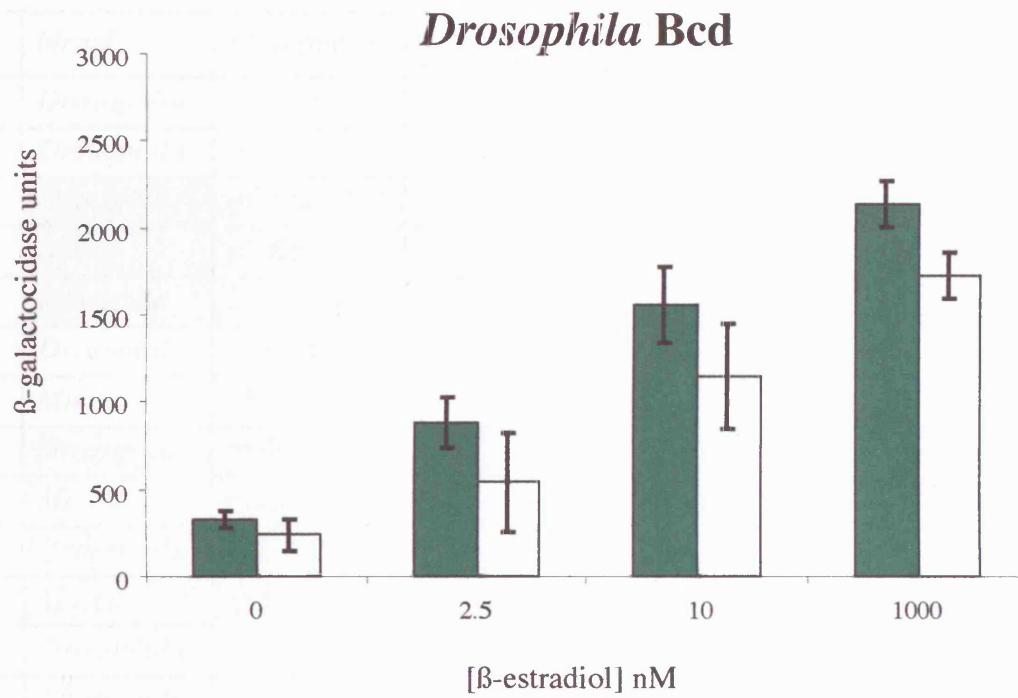
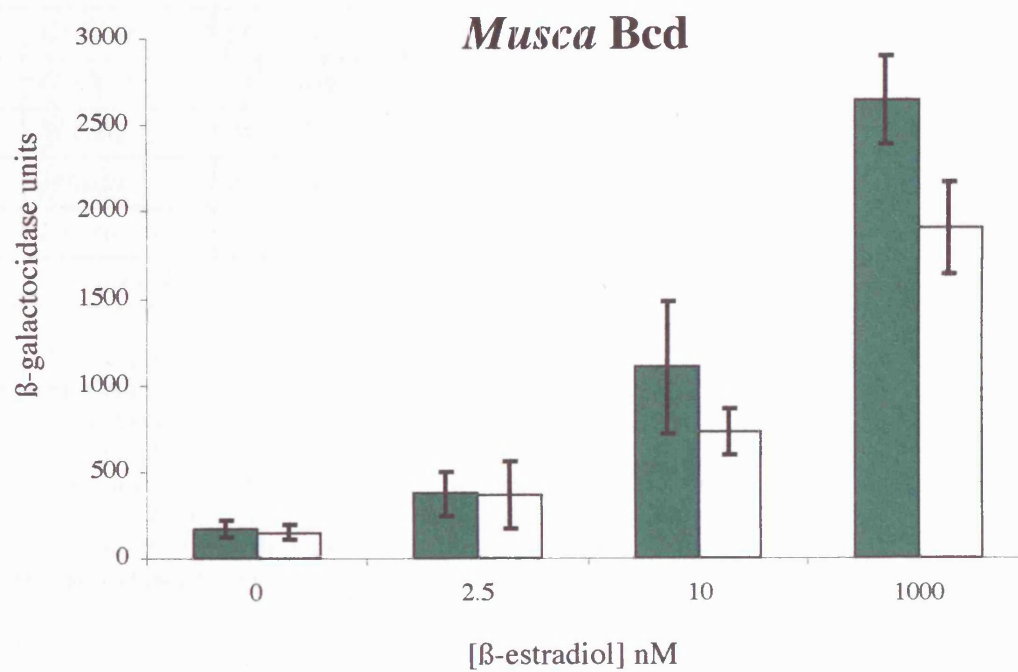
A**B**

Figure 6.5 Comparison of transcription from the *Drosophila hb* promoter (green columns) and the *Musca hb* promoter (white columns) with *Drosophila* Bcd (**A**) and *Musca* Bcd (**B**). These assays were all carried out at the same time. See the legend of figure 6.4.

	<i>bicoid</i>	<i>hb</i> reporter	0 nM β -estradiol	2.5 nM β -estradiol	1000 nM β -estradiol
1	<i>Drosophila</i>	pDBhb.19	322	1172	2499
1A	<i>Drosophila</i>	pDBhb.19	119	621	2499
2	<i>Drosophila</i>	pDBhb.19	351	1419	1451*
3	<i>Musca</i>	pDBhb.19	270	835	743*
4	<i>Megaselia</i>	pDBhb.19	101	423	216*
5	<i>Drosophila</i>	pMhbP2+	388	689	1831
6	<i>Musca</i>	pMhbP2+	288	520	2017
7	<i>Drosophila</i>	pDBhb.19	326	883	2140
8	<i>Musca</i>	pDBhb.19	170	371	2648
9	<i>Drosophila</i>	pMhbP2+	236	538	1727
10	<i>Musca</i>	pMhbP2+	152	365	1908
11	<i>Drosophila</i>	pLShb+	1	9	16
12	<i>Drosophila</i>	pLShb-	3	15	2
13	<i>Musca</i>	pLShb+	1	2	4
14	<i>Musca</i>	pLShb-	4	13	9
15	<i>Drosophila</i>	pCVhb-	10	21	8
16	<i>Drosophila</i>	pCVhb+	13	24	6
17	<i>Musca</i>	pCVhb+	6	31	9
18	<i>Musca</i>	pCVhb-	6	30	10
19	<i>Drosophila</i>	pCVhb5+	322	1505	1723
20	<i>Drosophila</i>	pCVhb5-	348	1403	1475

Table 6.1. Bcd-dependent reporter gene activities determined by β -galactosidase assays.

The figures given are averages from three cultures, see appendix C for the results of individual assays and the standard deviations. β -galactosidase units are defined as the amount which hydrolyses 1 μ mol of ONPG (to o-nitrophenol and D-galactose) per min per cell. Assays in adjacent rows of the same shading were carried out at the same time. Only the results of cultures grown in 0 nM, 2.5 nM and 1000 nM are shown. The results in 1 and 1A were obtained by other researchers at different times using the same system (Burz *et al.*, 1998; Burz and Hanes 2001). In assays marked with an asterisk the yeast cultures grew very slowly (see 6.3.1).

6.3.3 Analysis of the *Calliphora* and *Lucilia hb* promoter regions transcriptional activity in yeast

To investigate the transcriptional activity of the *Calliphora* and *Lucilia hb* promoter regions, yeast were co-transformed with either pCVP+, pCVP-, pLSP+ or pLSP- (see

6.2.2) and either pDB1.2, pBCMBCD or pBC103, in addition to pRS313GAL4ERVP16. Cultures were grown in triplicate in media containing 0 nM, 2.5 nM or 1000 nM β -estradiol and β -galactosidase assays were then performed. The results are summarised in table 6.1 rows 11 to 18 (see appendix C for results of individual assays). Surprisingly, neither the *Calliphora* nor the *Lucilia* promoters were activated by either *Drosophila* Bcd or *Musca* Bcd to much above background levels. The highest activation was observed at 2.5 nM using the *Calliphora* promoter in either orientation and using the *Lucilia* promoter orientated 3'-5'.

To further analyse transcription from the *Calliphora hb* promoter β -galactosidase assays were performed on yeast co-transformed with either pCVP5+ or pCVP5- (see 6.2.2), in addition to pRS313GAL4ERVP16 and either pDB1.2 or pBC103. The cultures had again been grown in media containing 0 nM, 2.5 nM or 1000 nM β -estradiol. *Drosophila* Bcd-dependent transcription from this *Calliphora* promoter fragment was able to activate *lacZ* activity to give a similar output to the *Drosophila hb* promoter fragment in yeast, except at the highest hormone concentration used (for example in table 6.1 compare rows 19 and 20 with rows 2 and 7). Furthermore, it did not appear that the orientation of the cluster of 8 *Calliphora* Bcd-binding sites had any affect on transcription (compare rows 19 and 20 in table 6.1).

6.3.4 Western analysis of Bcd expression in yeast

Are the yeast expressing Bcd from each species to different extents? This could explain the variation seen between these transcription factors in the above β -galactosidase assays. To investigate this possibility, yeast cultures containing pRS313GAL4ERVP16, pDBhb.19 and either pDB1.2, pBCMBCD, pMABCD or pBC103 were grown in media containing the medium concentration of hormone (2.5 nM), in parallel with and as described for yeast cultures used in β -galactosidase assays (see 6.3.1). Protein extraction from the yeast cultures was performed (see 2.2.8.3) and equal volumes of extract were run on an SDS-PAGE. Bcd (from *Drosophila*, *Musca* and *Megaselia*) is expressed from pBC103 with an N-terminal HA-tag. This facilitated universal immunodetection of Bcd from each species.

Thus, Western transfer was performed and the filter subsequently probed with anti-HA (mouse) primary antibody and HRP-conjugated anti-mouse (sheep) secondary antibody (see 2.2.8.4). Specific 70 to 80 kDa products were observed for extracts from *Musca* and *Drosophila* Bcd containing yeast cultures and these are presumably *Drosophila* and *Musca* HA-Bcds (figure 6.6A). The predicted molecular weights of *Drosophila* HA-Bcd and *Musca* HA-Bcd are 56 kDa and 53 kDa respectively (see 2.2.10). However, *Drosophila* Bcd produced in yeast cultures has previously been observed to migrate at similar sizes to the products described here (Ma *et al.*, 1996; Yuan *et al.*, 1996). A product of approximately 60 kDa was also seen, which is presumably *Megaselia* HA-Bcd (figure 6.6A), although it has a predicted molecular weight of 39 kDa. *Megaselia bcd* encodes a protein of 338 amino acids, which is smaller than the *Drosophila* and *Musca* Bcds. Interestingly, the *Musca* product appeared as a doublet, which may be due to differential phosphorylation of *Musca* Bcd in yeast as described previously for *Drosophila* Bcd (Driever and Nüsslein-Volhard 1988b). Indeed, post-translational modifications could explain the higher than expected molecular weights of the Bcd proteins produced in yeast. However, this analysis cannot exclude the possibility that one of the *Musca* bands is an artefact and this is most likely to be the larger band since *Musca* Bcd (468 amino acids) is slightly smaller than *Drosophila* Bcd (489 amino acids).

The above yeast protein extracts were again run on an SDS-PAGE, but this time the quantities loaded were equalised with respect to the cell densities at harvest. Western analysis was again performed as described above. The specific products observed above were also observed in this analysis and *Musca* HA-Bcd again appeared as a possible doublet (figure 6.6B). If the lower band of the *Musca* doublet is Bcd (arrowed in figure 6.6B) then it is expressed in comparable amounts to *Drosophila* HA-Bcd. However, if both *Musca* bands are differentially phosphorylated forms of Bcd then it is being expressed at higher concentrations than *Drosophila* HA-Bcd. There is no ambiguity with respect to *Megaselia* HA-Bcd as both figure 6.6A and 6.6B illustrate that it is being expressed at higher concentrations than both *Drosophila* and *Musca* HA-Bcds. The pixel densities of the putative Bcd bands in figure 6.6B and 6.6C (see below) were

measured (see 2.2.10) and are recorded in table 6.2. These support the notion that *Drosophila* HA-Bcd and *Musca* HA-Bcd are expressed in approximately equal amounts when the density of the non-specific (control) bands on the gel are considered (indicated in figure 6.6B). However, *Megaselia* HA-Bcd is being expressed in higher quantities at this particular concentration of hormone.

To determine whether an increase in hormone concentration did indeed result in an increase in Bcd expression and thus higher levels of β -galactosidase reporter gene activity. Yeast cultures containing pRS313GAL4ERV16, pDBhb.19 and either pBC103, pDB1.2 or pBCMBCD were grown as in media containing 0 nM, 2.5 nM or 1000 nM β -estradiol. Protein was extracted from these cultures as described above. Samples were then loaded on an SDS-PAGE in volumes equalised for the cell density of each culture at harvest. Western analysis was then performed as described above. In cultures containing pDB1.2 or pBCMBCD, it is clear that a product of approximately 70 to 80 kDa is produced in higher quantities when the yeast are grown in the presence of increasing concentrations of β -estradiol (figure 6.6C). These products were not seen in the control yeast extracts and are therefore presumably *Drosophila* HA-Bcd and *Musca* HA-Bcd expressed from pDB1.2 and pBCMBCD respectively. *Musca* HA-Bcd is detected at the low concentration of hormone, but *Drosophila* HA-Bcd is not (figure 6.6C). At the medium hormone concentration it appears that *Drosophila* and *Musca* HA-Bcds are expressed equally supporting the analysis carried out above (table 6.2), although at the highest concentration *Drosophila* HA-Bcd expression appears to be higher (figure 6.6C). The double band seen previously in the above *Musca* HA-Bcd lanes was not observed in this experiment, although this gel was not run as far as those described above.

In an attempt to resolve whether the HA-specific products seen in the above experiments were indeed Bcd, Western analysis was repeated as described above using an anti-*Musca* Bcd polyclonal antibody as the primary antibody. This antibody was known to cross react with Bcd from all three species (P. Shaw personal communication). However, no specific bands were observed on any of the filters used above under various conditions. Therefore, any speculations made regarding the relative expression of each *bcd* gene in

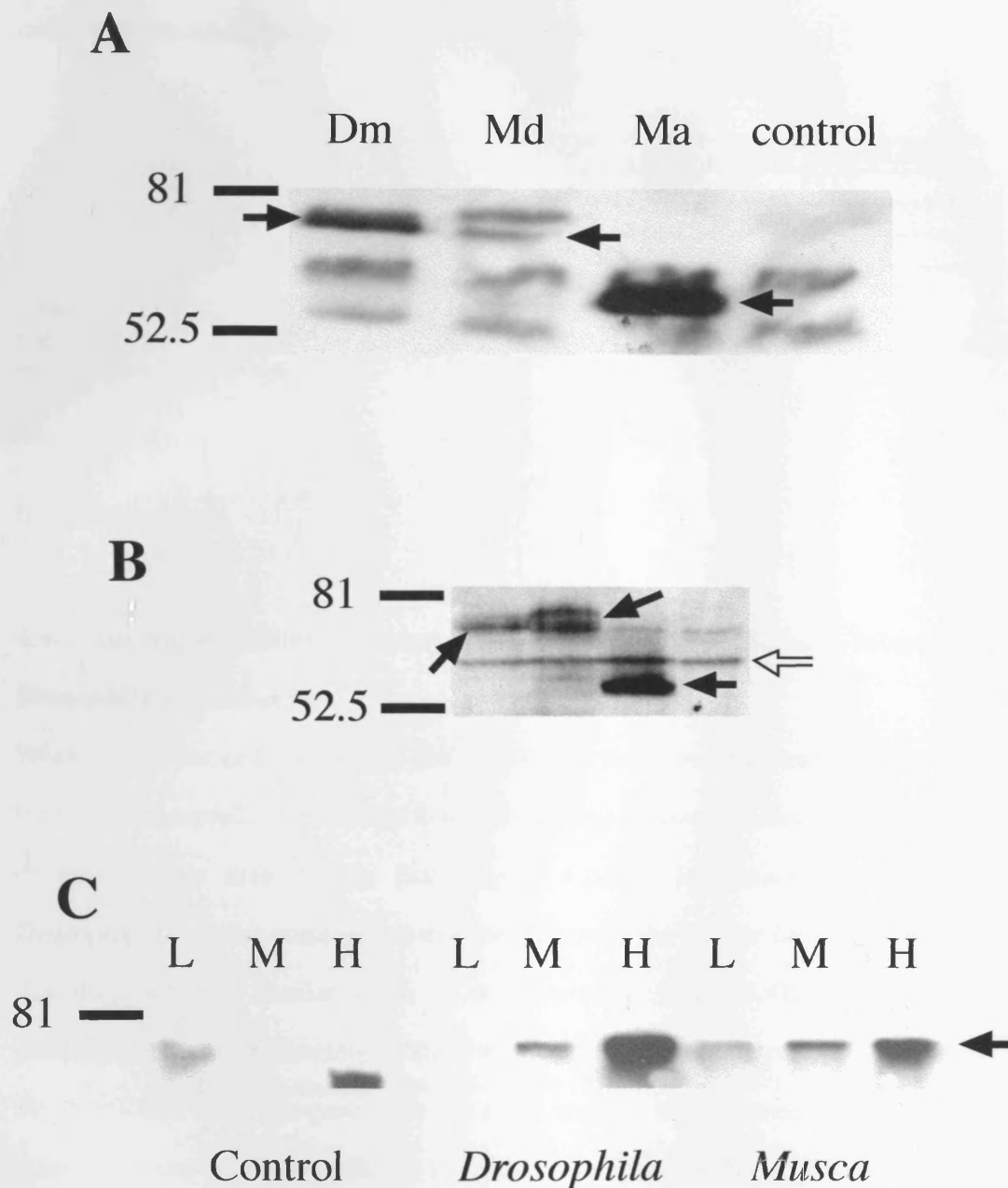


Figure 6.6 Western analysis of *bicoid* expression in yeast cultures

A. Yeast protein extracts from cultures grown in the presence of 2.5 nM β -estradiol. Control lane yeast cells were transformed with vector pBC103. Samples Dm (*Drosophila*), Md (*Musca*) and Ma (*Megaselia*) had *bcd* from each species inserted in frame into vector pBC103 (pDB1.2, pBCMBCD and pMABCD respectively). Arrows indicate specific, putatively HA-Bcd bands in the Dm, Md and Ma lanes (see 6.3.4).

B. Same samples as described for A, but the gel loading was equalised with respect to the cell density of the yeast cultures at harvest (see 6.3.4). The white arrow indicates the control bands described in table 6.2.

C. Protein extracts from yeast cultures containing plasmids pBC103 (control), pDB1.2 (*Drosophila*) and pBCMBCD (*Musca*) grown in the presence of **Low** (0 nM), **Medium** (2.5 nM) and **High** (1000 nM) concentrations of β -estradiol. The arrow indicates the size of the specific HA-Bcd bands.

The size markers are given in kDa. Volumes of protein extract loaded for each sample onto SDS-page gels were equalised for yeast culture cell density at harvest (except for **A**). For Westerns the primary antibody used was 1/1600 diluted anti-HA (mouse) and the secondary was 1/1000 diluted HRP conjugated sheep anti-mouse polyclonal antibody (see 2.2.8.4).

yeast are based on the assumption that the bands bound by the anti-HA antibody correspond to the HA-Bcds.

HA-Bcd	6.6B	6.6C	control band (6.6B)
<i>Drosophila</i>	3452	3838	1335
<i>Musca</i>	5553	4974	1694
<i>Megaselia</i>	9705	-	1834

Table 6.2. Pixel densities of putative HA-Bcd bands in figure 6.6. Units are given in square pixels (see 2.2.10). The control band represents the density of a non-specific band in figure 6.6B for comparison.

6.4 Discussion

6.4.1 Incompatibilities between components of the Bcd-*hb* interaction from *Drosophila* and *Musca*

When the *Musca* and *Drosophila bcd* genes were compared by their transcriptional output from the *Drosophila hb* promoter it was found that *Drosophila* Bcd activated higher levels of transcription than *Musca* Bcd (figure 6.4A). Interestingly, when *Musca* and *Drosophila* Bcd were compared in combination with the *Musca hb* promoter it was found that they activated similar levels of *lacZ* activity (figure 6.4B). Therefore the two promoters behaved differently in these transcription factor comparisons. However, when the promoters were compared side by side, slightly higher levels of *lacZ* activity were generated from the *Drosophila hb* promoter than from the *Musca hb* promoter with both *Drosophila* Bcd and *Musca* Bcd (figure 6.5).

It has been shown that yeast preferentially initiate transcription from closely spaced binding sites rather than widely spaced binding sites (Hanes *et al.*, 1994, see 6.1.2) and the average spacing of binding sites in the *Drosophila* and *Musca hb* promoters is 30 bp and 54 bp respectively. This could account for the average output of the *Drosophila* promoter being slightly higher than the *Musca* promoter at higher hormone concentrations when the promoters were compared side by side (figure 6.5). Indeed, the effect could also mask the different performances of these two promoters observed when *Musca* Bcd and *Drosophila*

Bcd were directly compared (figure 6.4). Therefore, the inferences made below concern only the results obtained when the transcription factors were compared side by side on either promoter.

How can the differences between the promoters seen when *Musca* Bcd and *Drosophila* Bcd were compared (figure 6.4) be rationalised? It has been demonstrated that *Drosophila* Bcd binds with a higher affinity than does *Musca* Bcd to *hb* promoter fragments from both species (Shaw *et al.*, submitted). This may in part explain the higher level of transcription of *Drosophila* Bcd than *Musca* Bcd from the *Drosophila hb* promoter (figure 6.4A). However, it is possible that the *Musca hb* promoter has properties that are not present in the *Drosophila hb* promoter, which *Musca* Bcd, but not *Drosophila* Bcd, requires for transcription. This could explain why *Drosophila* Bcd and *Musca* Bcd drive similar levels of transcription from the *Musca hb* promoter (figure 6.4B). These incompatibilities between the *Drosophila* and *Musca* components could be evidence for the co-evolution of Bcd and the *hb* promoter in these species and this is discussed in relation to other evidence from these species in chapter 8. If species-specific differences in *bcd* are accumulating over time then do Bcd proteins from other species recognise the *Drosophila hb* promoter?

6.4.2 *Megaselia* Bcd function?

Megaselia Bcd is very different in amino acid sequence to the other Bcd orthologs (see chapter 4). However, the critical residues in the homeodomain for binding site sequence recognition at positions 47, 50, 51 and 54 in the recognition helix and at positions 3 and 5 in the N-terminal arm are all conserved (see figures 4.1 and 4.2). Interestingly, three putative Bcd-binding sites have been found in the region upstream of the *Megaselia hb* zygotic transcription start site (P. Shaw personal communication).

To test how well a Bcd protein from a species more distantly related to *Drosophila* than *Musca* (see figure 1.3) activated transcription from the *Drosophila hb* promoter, *Megaselia* Bcd was used. It was found that *Megaselia* Bcd activated reporter gene

expression poorly from the *Drosophila* promoter in comparison to both *Drosophila* Bcd and *Musca* Bcd (figure 6.4A).

While it has not yet been determined which sequences *Megaselia* Bcd preferentially binds to, this result could mean that they are different to those bound by *Drosophila* Bcd. The Otd homeodomain, which also has a lysine at position 50, also preferentially binds to the sequence TAATCC (Wilson *et al.*, 1996), despite differences at residues 47 and 54 (reviewed in Laughon 1991). However, Otd recognises 'non-consensus' Bcd-binding sites poorly (Dave *et al.*, 2000). Therefore, given the importance of these sites for transcription (see 4.3.3 and 4.3.4), if *Megaselia* Bcd did not recognise them, then this would explain its performance in the yeast transcription results described above. Indeed, *Megaselia bcd* gave no signs of rescue in transgenic *Drosophila bcd* mutant embryos (P. Shaw personal communication). It is possible that *Megaselia* Bcd is a poor activator of transcription without specific co-factors, particularly in yeast, but unfortunately no *Megaselia* Bcd-dependent promoters were available to test. However, these results may be indicative of the functional divergence of *bcd* within the Diptera.

6.4.3 Bcd-dependent transcription from the *Calliphora* and *Lucilia hb* promoters

The putative *hb* promoters from *Calliphora* and *Lucilia* were unable to initiate transcription to much above background levels in combination with either *Drosophila* Bcd or *Musca* Bcd (table 6.1 rows 11 to 18). However, when only the distal half of the *Calliphora* promoter, which contains 8 Bcd-binding sites, was used (see figure 6.3) in either orientation, transcription driven by *Drosophila* Bcd was comparable to the levels driven by the *Drosophila hb* promoter (table 6.1 rows 19 and 20). This suggests that there are sequences in the proximal part of the *Calliphora* and *Lucilia* promoters that cause transcriptional repression in yeast. Studies of the *Drosophila hb* promoter have revealed that the region between -50 and -94, with respect to the transcription start site, contains sequences that repress transcription in yeast (Ma *et al.*, 1999). Interestingly, this sequence contains CT rich sequences (for example CCTCTGCCC), which could be bound by the GAGA transcription factor (GAF) (Wilkins and Lis 1998). GAF interacts with SAP18,

which is part of the Sin3-HDAC complex. When this complex is recruited to a promoter transcription is repressed through histone deacetylation (Kasten *et al.*, 1997; Espinas *et al.*, 2000). The proximal region of both the *Calliphora* and *Lucilia hb* promoters contain approximately 10 and 9 putative GAF binding sites respectively. Perhaps these sites are causing transcriptional repression in yeast by a similar mechanism to that used by GAF, since a *Sin3* gene has been found in yeast (see 6.1.2). However, the *Musca hb* promoter used in the above yeast assays contains approximately 6 putative GAF binding sites and it was able to initiate transcription. GAF may also have a positive role in activating *hb* expression in flies (see 4.3.5).

It is also possible that *Musca Bcd* and *Drosophila Bcd* do not recognise the *Calliphora* and *Lucilia* promoters as well as their own promoters as a result of co-evolution, but that the cluster of *Calliphora* binding sites alone can overcome this due to favourable conditions in yeast (see 6.1.2). This could be tested to a certain extent by examining the expression patterns driven by the *Calliphora* and *Lucilia* promoters in transgenic *Drosophila*.

6.4.4 Measuring promoter sensitivity in yeast

In the yeast transcription assays described here unfortunately the concentration of Bcd is probably too high to discern a direct difference in sensitivity between the promoters at Bcd concentrations representative of the posterior limits of the gradient in embryos (see 6.1.2). Indeed, the *Drosophila* and *Musca* promoters were indistinguishable at the lowest Bcd concentrations used in these assays and the results obtained from using the *Calliphora* and *Lucilia* promoters did not shed any light on this matter either. However, it remains possible that more sensitive promoters are part of the mechanism used by species with larger embryos to read the Bcd gradient.

Interestingly, while the *kni64* element can drive reporter gene expression almost throughout the *Drosophila* embryo, its performance was the same as the *Drosophila hb* promoter in yeast transcription assays similar to those performed here (Rivera-Pomar *et al.*, 1995; Burz *et al.*, 1998).

Chapter 7

Characterisation of *orthodenticle* in *Musca*

7.1 Introduction

The differences in the Bcd-*hb* interaction between *Drosophila* and *Musca* must be considered in the context of the whole network of genes that are regulated by Bcd. Any changes in the DNA binding properties of Bcd will affect the expression patterns of all genes regulated by this transcription factor and so if Bcd and *hb* have co-evolved then Bcd and all its target promoters must have co-evolved. Therefore, have other Bcd-regulated genes and their promoters in *Musca* evolved in a similar manner to that described for *hb*?

To begin answering this question, this chapter describes the characterisation of the structure and expression patterns of the head gap gene *otd* in *Musca*. In *Drosophila*, *otd* encodes a *prd* class homeodomain containing transcription factor that has similar binding site preferences to Bcd, mediated by a lysine at homeodomain position 50. *otd* is required for correct development of the head and CNS (Finkelstein *et al.*, 1990). *otd* mutants are embryonic lethal and exhibit defects in the development of the labral, intercalary and gnathal segments as a result of disrupted head involution (Wieschaus *et al.*, 1984).

otd was chosen because it has been shown that it is directly activated by Bcd to give an anterior domain of expression in the early *Drosophila* embryo (Finkelstein and Perrimon 1990). The *Drosophila otd* Bcd-dependent promoter is composed of only 3 Bcd binding sites spread over 180 bp of DNA and is located approximately 5 kb upstream of the transcription start site (Gao and Finkelstein 1998). Interestingly, Hb may also contribute to the expression of *otd* in the anterior of the embryo through a single binding site in the above 180 bp *cis*-regulatory element. Therefore, *Drosophila otd* provides a simple Bcd-dependent promoter and expression pattern to compare with *Musca* and since the homeodomain of *otd* is highly conserved over large evolutionary distances (Simeone *et al.*, 1992) the cloning of *otd* genes has proven to be relatively straightforward in other species (Li *et al.*, 1996).

7.2 Results and Discussion

7.2.1 Amplification of the *Musca otd* homeodomain using degenerate PCR

Degenerate primers NGF and GOR (see appendix A) were designed to amplify the codons of the first 40 amino acids of the *Musca otd* homeodomain, based on the *Drosophila otd* sequence (Finkelstein *et al.*, 1990 and see figure 7.1A for the structure of the *Drosophila otd* transcript). These primers were employed in PCR to amplify a 119 bp product from *Musca* genomic DNA. A BLAST search revealed that the sequence of this fragment had the highest similarity to the *Drosophila otd* homeodomain sequence and therefore that the *Musca otd* homeodomain sequence has been successfully cloned.

Comparison of the *Drosophila* and *Musca otd* homeodomain sequences revealed that there were 23 synonymous differences and 1 non-synonymous difference, which resulted in an alanine and a serine at homeodomain position 18 respectively in each species.

7.2.2 5' RACE PCR to map the *Musca otd* transcription start site

First strand cDNA synthesis was performed on approximately 240 ng of *Musca* mRNA, (see 2.2.5) using RT and a *Musca otd* homeodomain primer (RTOTD, appendix A). Primary PCR was then carried out using AAP and the gene specific primer OALR2. Reamplification of the primary reaction with AAP and OALR3 generated a product of approximately 1.1 kb (figure 7.2A). Sequencing of this RACE product revealed that it overlapped with the *Musca otd* homeodomain sequence and allowed the *Musca otd* translation start site to be identified by its similarity to that of *Drosophila otd*. The 5' end of the RACE product was identified by its similarity to the arthropod consensus transcription start site sequence (see figure 3.6A).

7.2.3 3' RACE PCR to amplify the 3' region of the *Musca otd* transcript

Primer OTD3R1, based on the *Musca otd* homeodomain sequence, was employed in a primary PCR reaction on a cDNA template synthesised from 120 ng of *Musca* mRNA

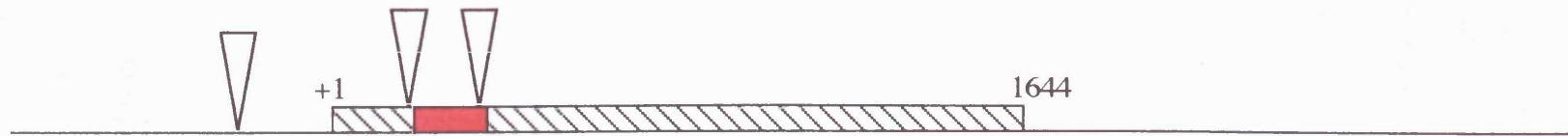
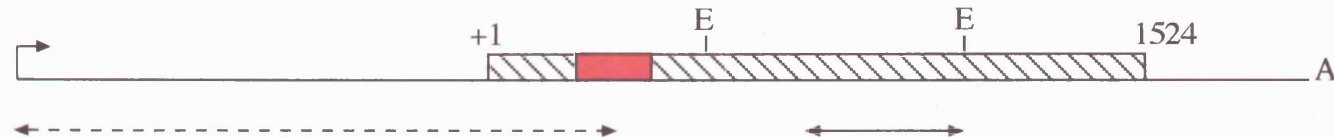
A**B**

Figure 7.1 Transcript structure of *otd* in *Drosophila* (**A**) and *Musca* (**B**)

Dashed boxes represent exon sequences and triangles illustrate the positions of introns in *Drosophila*. The red boxes represent the positions of the homeodomain coding sequences. The black lines represent the 5' and 3' UTRs of *otd* in each species. The putative transcription start site (black arrow) and polyadenylation signal represented by A are shown for *Musca*. The double headed black arrow illustrates the region used as a probe in Southern analysis (figure 7.3) and the double headed dashed arrow represents the region used as a probe for *in situ* hybridisations (figure 7.5). The positions of two *Eco*RI (E) sites are indicated. The *Drosophila otd* transcript structure was obtained from Gadfly using accession number CG12154.

(see 2.2.5 and 2.2.6). Reamplification with the nested primer OTD3R2 resulted in a major product of approximately 1.7 kb and several other smaller products (figure 7.2B). Sequencing of the 1.7 kb product revealed that it overlapped with the *Musca otd* homeodomain sequence and encoded a similar protein to that of *Drosophila otd*. A putative polyadenylation signal (not shown) was found approximately 400 bp downstream of the first in frame stop codon (figure 7.1B). Therefore, it appears that the *Musca otd* 3' UTR is approximately 900 bp shorter than that of *Drosophila*. Discrepancies between RACE product clones were answered by sequencing such regions from genomic DNA with primers ORR1, ORR2, ORR3, ORF1 and ORF2 (see appendix A).

7.2.4 Southern analysis of *Musca otd*

To verify that the above PCR sequences were representative of *Musca* genomic sequences and to determine the copy number of *otd* in *Musca*, Southern analysis was carried out on *Musca* genomic DNA restricted with *Bgl*III, *Eco*RI, *Dra*I and *Hind*III using a 381 bp *Musca otd* probe (see figure 7.1B). This probe hybridised to an *Eco*RI fragment of approximately 600 bp (figure 7.3) and so confirmed the positions of the two *Eco*RI sites shown in the transcript map (figure 7.1B). The other hybridising bands that can be observed in this digest are likely to be the result of incomplete digestion. A single hybridising band of approximately 3 kb was observed in the *Hind*III digestion (figure 7.3), which suggests that there is only a single *otd* gene in *Musca*. However, this does not exclude the possibility that other *otd* genes are present in *Musca*, which have diverged in the sequence corresponding to the probe used here. Unfortunately, it appears that too much DNA was used in both the *Bgl*III and *Dra*I digests for the probe hybridisation patterns to be observed clearly.

7.2.5 *Musca otd* transcript and protein structure

Assembly of the above PCR product sequences reveals that the putative *Musca otd* transcript is approximately 3 kb long (figure 7.1B) and encodes a protein of 508 amino acids, which is 79% similar to *Drosophila* Otd (figure 7.4).

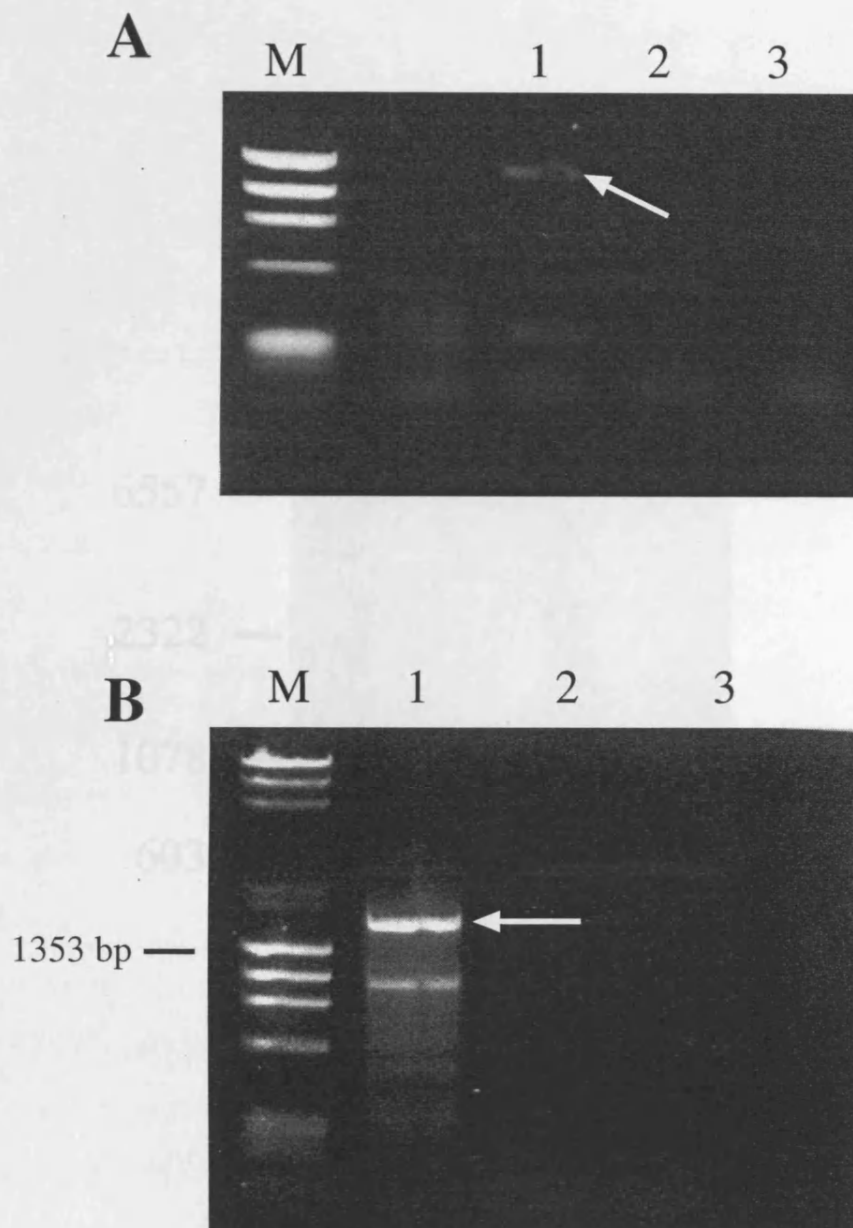


Figure 7.2 Mapping of the *Musca otd* transcript using RACE PCR

A. Results of reamplification of 5' RACE reactions using primers OALR3 and AAP. Lane 1: Reamplification of primary reaction template. Resulting band indicated by the white arrow. Lane 2: Untailed cDNA template. Lane 3: No template control. Marker (M) sizes were 1353, 1078, 872, 603 and 310 bp from top to bottom.

B. Results of reamplification of 3' RACE reactions using primers OTD3R2 and AUAP. Lane 1: Result of reamplification of the primary reaction (cDNA template). An arrow indicates the specific product amplified (see text). Lane 2: Reamplification of mRNA template primary reaction. Lane 3: No template reamplification.

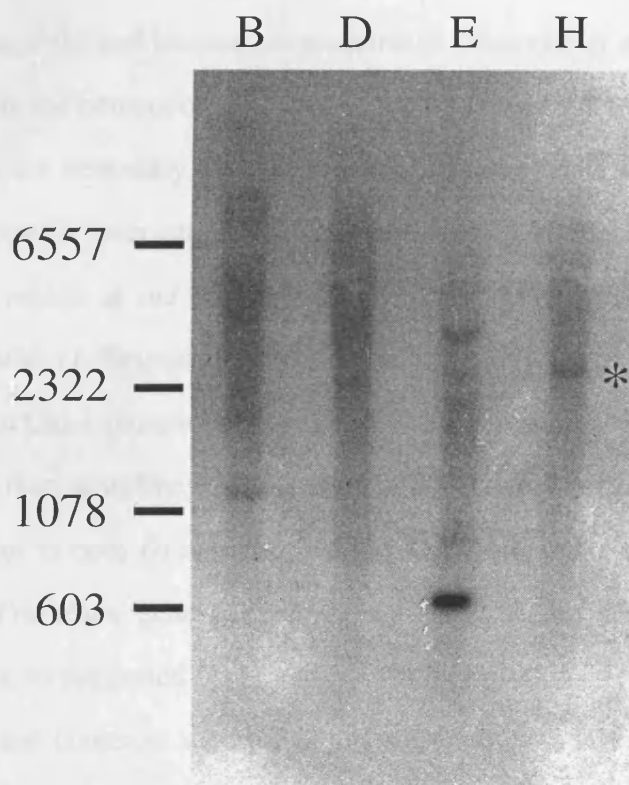


Figure 7.3 Southern analysis of *Musca* genomic DNA digested with *Bg/II* (B), *DraI* (D), *EcoRI* (E) and *HindIII* (H). Hybridised with a 381 bp *Musca otd* coding sequence probe as illustrated in figure 7.1B and washed to a stringency of 0.1X SSC. The size markers are given in bp. The single band in the *HindIII* lane is marked by an asterisk.

The patterns of amino acid conservation between the *Musca* and *Drosophila* Otd sequences suggest that the two proteins are functionally equivalent. Indeed, only a single difference is found between the *Musca* and *Drosophila* Otd homeodomains (see above). This is not surprising since Otd/Otx homeodomains are highly conserved across large phylogenetic distances; for example, there are only three amino acid differences between the *Drosophila* and murine homeodomains (Simeone *et al.*, 1992). The amino acids N-terminal to the homeodomain are also highly conserved between these two species (figure 7.4) and are necessary for *otd* function in *Drosophila* where they may be involved in protein-protein interactions (J. Reischl personal communication). In addition, the C-terminal region of *otd* has been characterised as a transcriptional activation domain in *Drosophila* (J. Reischl personal communication) and is also similar in *Musca* and the *Tribolium* Otd-1 protein (Li *et al.*, 1996). Furthermore, the Southern data presented above suggests that, as in *Drosophila*, there is only a single *otd* gene in *Musca*. The similarity of *Musca otd* to both *Drosophila otd* and *Tribolium otd-1* supports the hypothesis that the second *Tribolium* gene (*Tc otd-2*) was lost in the arthropod branch leading to the Dipterans, as suggested by Li and co-workers (1996). These authors have also proposed that the last common ancestor of the arthropods and vertebrates had a single *otd* gene, which shared structural features of both *Tribolium otd* genes.

The two *Tribolium otd* genes encode proteins of 371 and 301 amino acids and these are comparatively shorter than those of *Drosophila* and *Musca* (548 and 508 amino acids respectively). A comparison of *Drosophila* and *Tribolium* genes carried out by Schmid and Tautz (1999) found that *Drosophila* genes, including *otd*, encoded proteins that are on average 30% longer than those of *Tribolium* are. It appears that this phenomenon is a consequence of long runs of repeated codons in the *Drosophila* genes, which may have been generated by slippage (Schmid and Tautz 1999). Indeed, the RSFs (see chapter 5) of the *Drosophila* genes were all higher than those of the orthologous *Tribolium* genes studied. Interestingly, the alignment of the *Musca* and *Drosophila otd* amino acid sequences (figure 7.4) demonstrates that there are numerous amino acid repeats, of varying composition, in the *otd* genes of both these species. This suggests that

slippage may be responsible for the divergence of *otd* sequences between *Drosophila* and *Musca* as suggested for *hb* in chapter 5. Analysis of the *cis*-regulatory regions of *otd* in *Musca* would determine whether this affect is restricted to the less constrained domains of the coding region or whether slippage has also driven the restructuring of *otd* promoters between *Drosophila* and *Musca*.

7.2.6 Analysis of *otd* expression patterns in *Musca*

Sense and anti-sense ribo-probes of approximately 1.1 kb were synthesised (see 2.2.9.1) corresponding to the region of *Musca otd* indicated in figure 7.1B. *In situ* hybridisations were then performed on *Musca* embryos (up to 6 hours old and overnight collections) using these probes (see 2.2.9). No staining was observed in any embryos when the sense probe was used. The anti-sense probe revealed that in *Musca otd* is expressed in similar patterns to that of the *Drosophila* ortholog (Finkelstein *et al.*, 1990; Gao and Finkelstein 1998).

In *Musca*, *otd* expression is first observed in the syncytial blastoderm as a stripe between 90% and 70% egg length (figure 7.5A). Such a stripe of *otd* expression is also observed in *Drosophila*, where it is dependent on activation by Bcd (Finkelstein and Perrimon 1990). Therefore, this suggests that *otd* expression in *Musca* is also regulated by Bcd. Interestingly, in *Drosophila* the *otd* stripe is formed by the retraction of expression from the anterior cap of the embryo and this is possibly a result of the terminal system modulating Bcd activity in this region of the embryo (Janody *et al.*, 2000). Indeed, *tor* is required for the retraction of *otd* expression from the anterior pole of the *Drosophila* embryo (Finkelstein and Perrimon 1990). In the *in situ* hybridisation experiments described here no *otd* expression was observed in the anterior cap of any *Musca* embryos. This difference in anterior cap staining between *Drosophila* and *Musca* has also been observed for *tll* expression in these species (N. Wratten personal communication) and suggests that there may be a difference in the timing of Bcd modification by the terminal system between these two species. However, this may not apply to all Bcd target genes

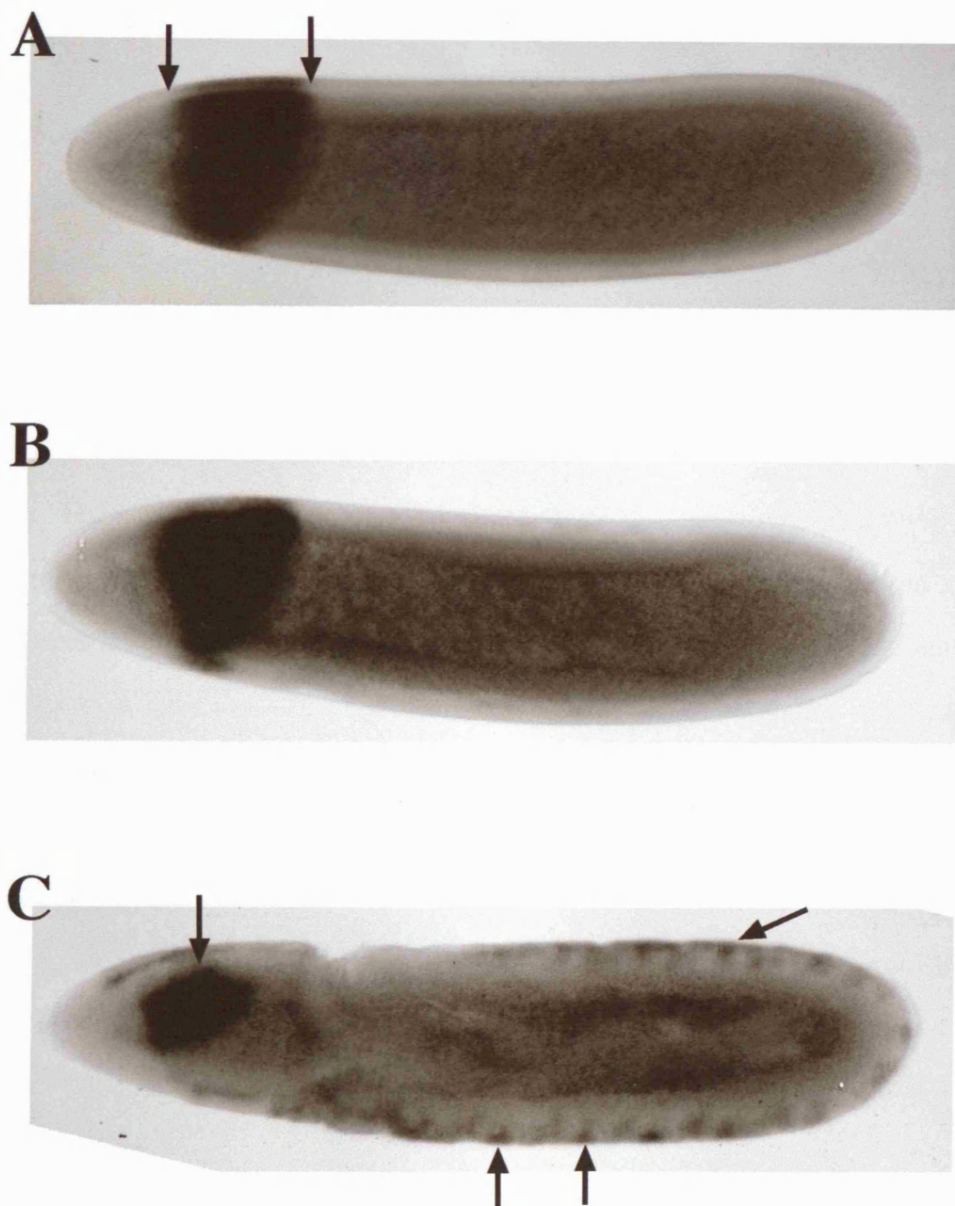


Figure 7.5 Expression patterns of *otd* in *Musca* embryos

A. Blastoderm stage embryo with expression in a broad anterior stripe from approximately 90% to 70% egg length (indicated by the arrows).

B. Later blastoderm embryo with expression retracted from the ventral side of the embryo.

C. Embryo at germ band elongation exhibiting expression in the CNS (for example, as indicated by the arrows).

All embryos were obtained from an overnight collection and were fixed and stained as described in chapter 2. All embryos are shown dorsal side up and anterior to the left.

since *Musca hb* is expressed in the anterior of the embryo and subsequently retracts in a similar manner to *Drosophila hb*.

As the *Musca* embryo becomes cellularised the stripe of *otd* expression retracts from the *ventral* side of *Musca* embryos (figure 7.5B). This is also observed in *Drosophila* and is caused by the repression of *otd* expression by Dorsal (Gao and Finkelstein 1998) and again this suggests that a similar regulatory mechanism is employed in *Musca*. During germ band elongation in *Musca* embryos, *otd* is expressed in the developing brain and in the developing nervous system in cells along the ventral midline (figure 7.5C) and these expression patterns are also observed for *otd* in *Drosophila* embryos (Finkelstein *et al.*, 1990).

The *otd/Otx* genes are expressed in the developing brains of both arthropods and vertebrates and indeed the *otd/Otx* genes appear to play highly conserved roles in brain development. It has been demonstrated that *Drosophila otd* can functionally replace the murine Otx-1 gene and the human Otx-1 and Otx-2 genes can rescue *otd* mutant defects in *Drosophila* (Acampora *et al.*, 1998; Leuzinger *et al.*, 1998). This suggests that the differences seen between vertebrate and arthropod *otd/Otx* genes outside of the homeodomain can be overcome and that the conserved homeodomain in particular is vital for the function of these genes. Furthermore, these experiments reveal that the genetic regulatory networks containing the *otd/Otx* genes are also conserved and are used to control brain development in both vertebrates and arthropods (reviewed by Hirth and Reichert 1999).

7.2.7 Conclusions and future work

In this chapter the structure and expression patterns of *otd* in *Musca* have been determined and shown to be similar to those described for *Drosophila otd*. Importantly, this suggests that *otd* is also regulated by Bcd in *Musca*. Therefore, it would now be possible to characterise the Bcd-dependent promoter of *otd* in *Musca* and compare this *cis*-regulatory sequence to the *Drosophila otd* promoter. This would determine whether the *otd* promoters of these two species are subject to the slippage generated restructuring of the *hb*

promoters. Indeed, the variation in codon repeats between the *Drosophila* and *Musca otd* genes indicate that this is highly probable. Characterisation and subsequent functional analysis of the *Musca otd* Bcd-dependent promoter could also be used to investigate the possible co-evolution of Bcd and its target promoters between *Drosophila* and *Musca* further, since a pattern may become more visible through the analysis of multiple Bcd-dependent promoters including *hb*, *tll* and *otd*.

The *Drosophila otd* Bcd-dependent promoter (see 7.1) is positioned approximately 5 kb upstream from the transcription start site. This distance is likely to be further in the larger genome of *Musca* and therefore screening of a *Musca* genomic library would probably be a more profitable approach to clone the *Musca otd* promoter than using sPCR.

Chapter 8

General discussion

8.1 Results summary

In this thesis I have investigated the evolution of genetic regulatory interactions in higher Dipteran species. These evolutionary studies have focused on the well characterised transcription factor encoded by *bcd* and two of the genes whose expression is regulated by Bcd, namely *hb* and *otd*. The results of this work are summarised below.

- *hb* genes were isolated from *Calliphora* and *Lucilia* and the *Lucilia hb* P2 transcription start site was mapped using 5' RACE PCR. The amino acid sequence of the *Calliphora* and *Lucilia hb* coding regions contain a number of domains (some of which have been functionally characterised in *Drosophila*) that are conserved in *hb* genes from other species such as *Drosophila*, *Musca* and *Tribolium*. These results also suggest that the *hb* transcripts have a similar structure in *Calliphora* and *Lucilia* as they do in other Dipterans, based on the conservation of putative splice site sequences and the transcription start site sequence. In addition, consensus NRE sequences were found in the putative *hb* 3' UTRs of *Calliphora* and *Lucilia*, suggesting that *hb* is also regulated by Nos and Pum in these species. From 3' RACE experiments in *Musca* it appears that this species, like *Drosophila*, employs only the more distal of two possible polyadenylation signals.
- The known sequence of *bcd* in *Calliphora* and *Lucilia* was extended in both 5' and 3' directions using sPCR. This confirmed a number of interesting amino acid differences in the Bcd homeodomain between the Calypttratae species and *Drosophila*. Interestingly, this analysis also revealed that a number of residues outside the homeodomain, which have been functionally characterised in *Drosophila*, were also conserved in *Calliphora* and *Lucilia* (as well as *Musca*). However, it appears that the Calypttratae species all have an additional serine rich domain that is not found in *Drosophila* species, although the function of this domain is not known.
- The putative *hb* P2 promoter regions from *Calliphora* and *Lucilia* were also cloned using sPCR. These regions were then footprinted to characterise Bcd-binding sites, using the homogeneous Bcd homeodomains from each species. This demonstrated that

the *hb* promoters of these two species differed from each other and from the *Drosophila* and *Musca hb* promoters in terms of the number, sequence, orientation and spacing of Bcd-binding sites that they contain.

- Three regions of *hb* (P2 promoter, 5' UTR and coding region) were sequenced in six strains of *M. domestica*. This revealed that there were extensive polymorphisms in all three regions at a higher frequency than was found during a similar study of *Drosophila* (Tautz and Nigro 1998). Many of these polymorphisms were indels and these associated with high frequency repeats of simple motifs in the coding region and 5' UTR and with both high and low frequency simple motifs in the promoter. This suggests that all three regions of *hb* studied are subject to slippage generated turnover of simple motifs and that extent of this turnover is dependent upon region specific constraints. Furthermore, analysis of the *Calliphora* and *Lucilia hb* promoter regions using the SIMPLE 34 program revealed that while a number of simple motifs were shared between these promoters and with the *Musca* promoter, the frequency of some motifs varied between the promoters of all three species. This implicates mechanisms of turnover in the restructuring of these *cis*-regulatory sequences.
- To investigate any functional consequences of the observed sequence differences in *bcd* and *hb* between *Drosophila* and *Musca*, transcription assays were carried out using homogeneous and heterogeneous combinations of these two components in yeast. The results of these assays reveal that *Drosophila* Bcd can activate transcription similarly from both the *Drosophila* and *Musca* promoters. In comparison, while *Musca* Bcd gave a lower transcriptional output than *Drosophila* Bcd did with the *Drosophila* promoter, *Musca* Bcd drove a similar output to *Drosophila* Bcd with the *Musca* promoter. This may suggest that one or more properties of the *Musca hb* promoter (i.e. the spacing, sequence, orientation and number of binding sites) are necessary for *Musca* Bcd, but not *Drosophila* Bcd, mediated transcription.
- In contrast to both *Drosophila* Bcd and *Musca* Bcd, the transcriptional output of *Megaselia* Bcd was comparably poor in combination with the *Drosophila* promoter. One explanation for this result is that it is the outcome of different binding properties

defined by the large number of differences between the *Drosophila* and *Megaselia* Bcd proteins.

- To investigate the evolution of another Bcd-regulated gene in the higher Diptera, the *otd* gene was isolated from *Musca*. This analysis revealed that the regions of *otd* that have been functionally characterised in *Drosophila*, such as the homeodomain, are also conserved in *Musca*, while other regions of *otd* have diverged in sequence between these species. Interestingly, there is only a single difference between the *Drosophila* and *Musca* Otd homeodomains, in comparison to the 5 differences between the Bcd homeodomains of these two species.

Analysis of the *otd* expression patterns in *Musca* revealed that while these were also similar to those observed in *Drosophila*, there was one interesting difference. In contrast to *Drosophila* embryos, *otd* expression was not observed in the anterior termini of any *Musca* embryos. This suggests that *otd* is not expressed in this region of the *Musca* embryo and may reveal differences in the regulation of *otd* between these species.

8.2 Implications of these data

8.2.1 The evolution of *bcd* and Bcd-dependent regulation

The determination of anterior-posterior polarity by Bcd appears to be a derived developmental mechanism, which is limited to Cyclorrhaphan flies (figure 1.3). A recent study by Stauber and co-workers (2002) discovered that the non-Cyclorrhaphan flies *C. albipunctata*, *Haematopota pluvialis* and *Empis livida* only have a single *Hox3* gene and in each species this gene is more similar to *zen* than to *bcd*. Interestingly, these *Hox3* genes are expressed both maternally and zygotically in the nurse cells and early embryo respectively of each species. Therefore, these single *Hox3* genes have expression patterns characteristic of both *bcd* and *zen* and this study suggests that the duplication of such a *Hox3* gene at the stem of the Cyclorrhaphan lineage resulted in *bcd* and *zen* (Stauber *et al.*, 2002 and see 1.6).

After this duplication event it can be envisaged that divergence in the function of the two proteins would have occurred; for example, glutamine changing to lysine at homeodomain position 50 and evolving the ability to bind the *cad* mRNA in one paralog. In addition, divergent *cis*-regulatory evolution between the two paralogs would then explain the loss of maternal *zen* expression and the loss of zygotic *bcd* expression (figure 8.1). Yet, how did *bcd* assume the role of anterior determinant and more specifically where did all the Bcd-binding sites in the promoters of Cyclorrhaphan genes come from? In due course further studies of the non-Cyclorrhaphans will no doubt characterise the role of the single *Hox3* gene present and perhaps suggest the genetic mechanism by which anterior-posterior polarity is established in these species. This may reveal how the developmental role of *bcd* evolved from more ancestral modes of development, such as that exhibited by *Tribolium*, in which it seems that no *bcd* gene is present (Brown *et al.*, 2001).

This thesis has shown that the configurations of Bcd-binding sites in the *hb* promoters of higher Dipterans can evolve rapidly; for example, in less than approximately 20 to 40 million years in the case of the Calyptratae (see figure 1.3). Indeed, this *cis*-regulatory evolution is probably the result of turnover mechanisms (such as slippage, gene conversion, unequal crossing over and transposition) acting on these *hb* promoter sequences as well as the promoters of other genes such as *tlx* (N. Wratten personal communication). The core sequences of Bcd-binding sites such as TAAT and TAAG are likely to have been present in the upstream regions of genes at a range of frequencies before the evolution of *bcd*. Therefore, as *bcd* acquired the role of anterior determinant the spread of Bcd-binding sequences, by mechanisms of turnover, may have been selected for resulting in the control of gene expression by Bcd. If maternal *hb* was the ancestral anterior determinant (as might be revealed by further investigations of non-Cyclorrhaphan flies) then this may explain why both Bcd and Hb regulate many genes.

Therefore, the evolution of *cis*-regulatory regions may have facilitated the rewiring of the networks that control early development in insects and this has resulted in the different strategies used by the Cyclorrhaphans and other insects.

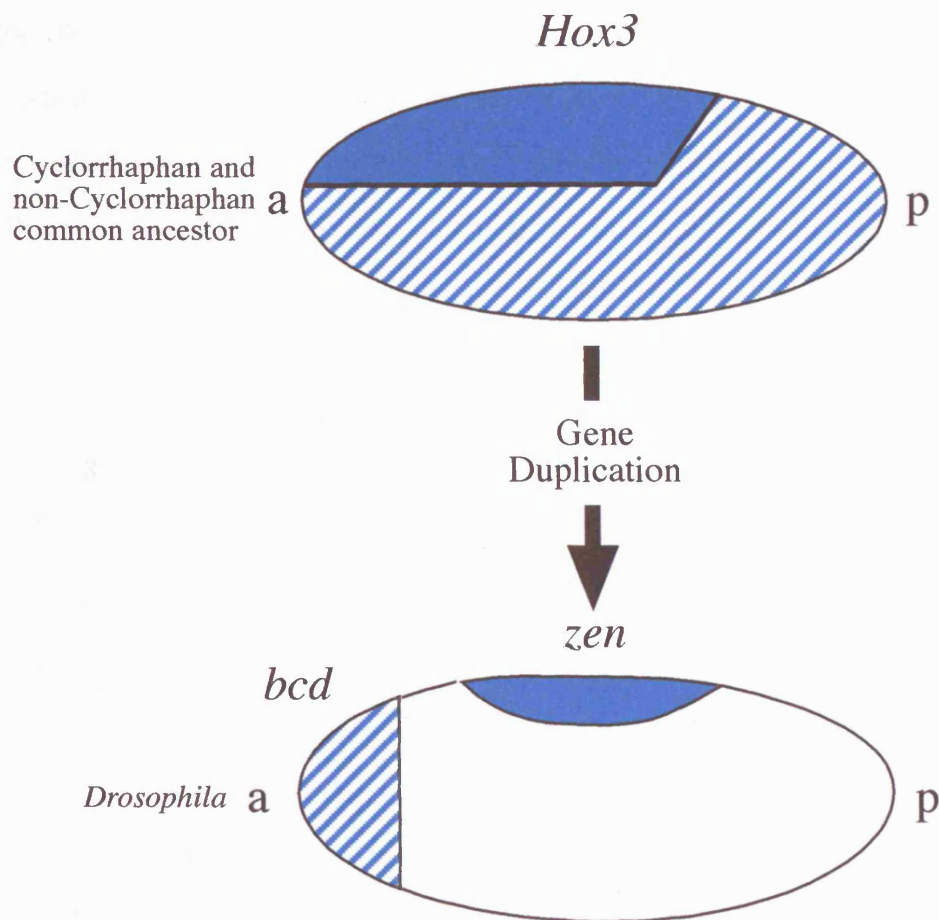


Figure 8.1 Expression patterns of the *Hox3* progenitor and of *bcd* and *zen*. Embryos are illustrated with the anterior (a) to the left and posterior (p) to the right, dorsal up. The expression patterns of the *Hox3* progenitor are represented in the top embryo by both intact blue shading and hatched shading. The expression patterns of *bcd* (hatched) and *zen* (intact blue) are shown in the bottom embryo. The gene duplication event took place approximately 140 MYA. Source: Stauber *et al.*, 2002, figure 5.

It can be envisaged that when Bcd-dependent regulation had been wired into the network of developmental interactions, mechanisms of turnover would continue to restructure promoters resulting in the different configurations seen in the *hb* and *tll* promoters of *D. melanogaster*, *D. virilis*, *Musca*, *Calliphora* and *Lucilia* (Driever and Nüsslein-Volhard 1989; Treier *et al.*, 1989; Bonneton *et al.*, 1997; McGregor *et al.*, 2001a; Shaw *et al.*, submitted). It has been suggested that such turnover in the regulatory regions of genes may co-evolve with point mutations in the DNA binding domains of transcription factors (with the latter tracking the former) to maintain the regulatory interaction (Bonneton *et al.*, 1997; Hancock *et al.*, 1999). What are the predicted consequences of changing the properties of factors that contribute to the Bcd-*hb* interaction?

8.2.2 Modelling the Bcd-*hb* interaction: predicting the consequences of change

The flow diagram in figure 8.2 summarises the inputs that control the Bcd-*hb* interaction and the properties of *hb* expression that result based on a 'fractional occupancy' model proposed by Gibson (1996). Basically, the properties of the *trans*-acting factor Bcd (concentration, binding affinity and protein-protein co-operativity) combine with properties intrinsic to the promoter signature, in addition to the contributions of any co-factors (figure 8.2A), to generate a *hb* transcriptional response along the anterior-posterior axis of the embryo (figure 8.2B).

Gibson used the above model to predict the affect on *hb* expression of changing aspects of the input parameters. Intriguingly, Gibson suggested that the contribution of the configuration of binding sites was central to the output of the interaction (i.e. to the values of the output parameters shown in figure 8.2). Specifically, this model predicts that increasing the number of binding sites would result in an increase in transcriptional response, a posterior shift in threshold position and a narrowing of the threshold width. Furthermore, while changes in the Bcd-binding affinity and co-operative properties of the interaction would be predicted to have little affect on the threshold width, increases in values of these two inputs would result in a posterior shift in the threshold position.

Interestingly, Gibson suggests that the same expression response can be generated

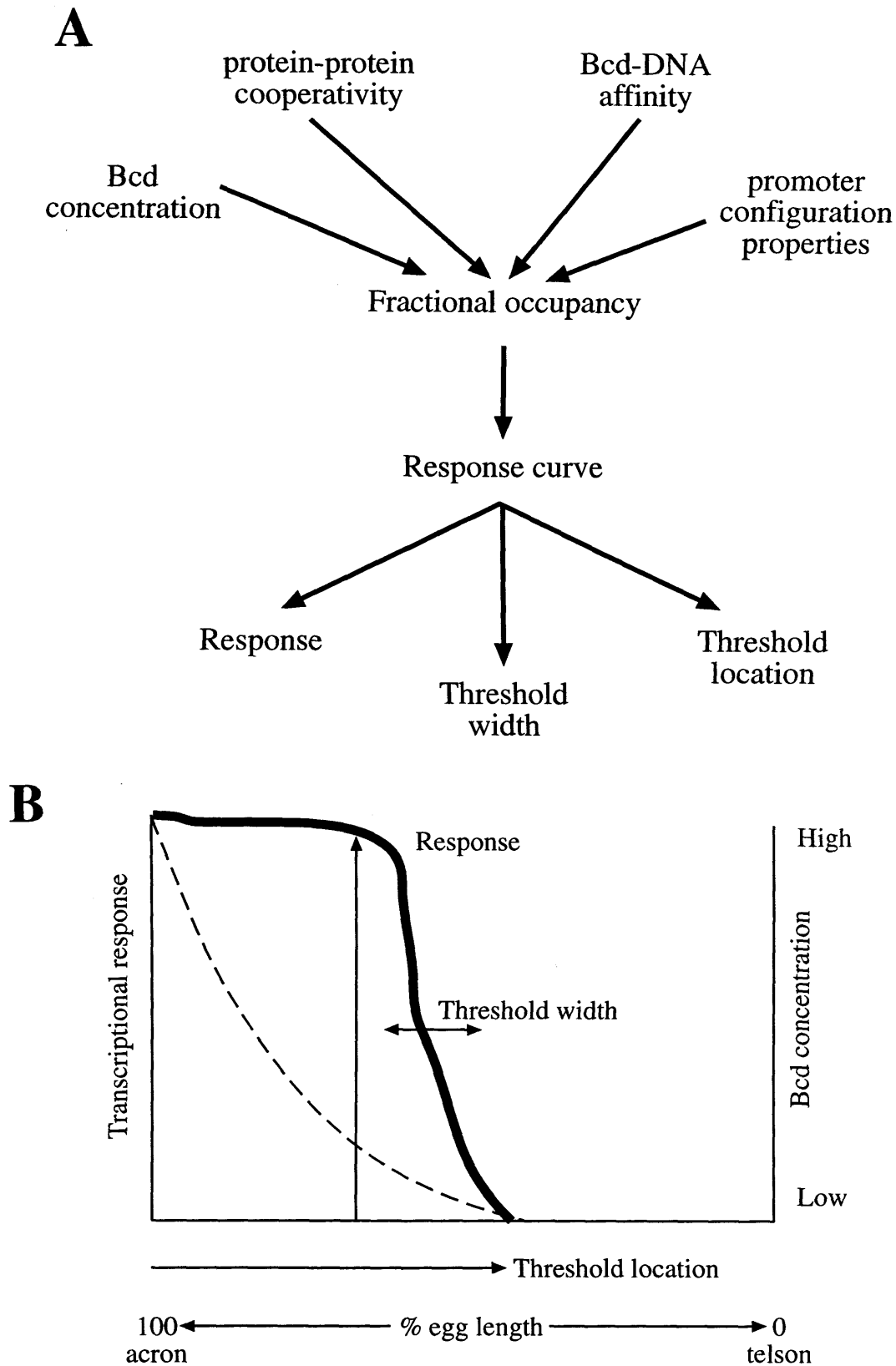


Figure 8.2 Parameters of the Bcd-*hb* interaction (**A**) and response curve (**B**)
A. The fractional occupancy of the *hb* promoter is determined by the Bcd concentration, co-operativity and binding affinity, as well as the properties of the configuration of binding sites (sequence, spacing, number and orientation). **B.** The fractional occupancy allows the determination of the transcriptional response, threshold width and threshold position along the anterior-posterior axis of the embryo. The response curve is represented by the thick line and the Bcd concentration by the dashed line.
 Source, Gibson 1996.

from different combinations of input values. In general terms, the response of the Bcd-*hb* interaction is a balance between the binding affinity of Bcd and the sensitivity of the configuration of binding sites in the promoter. Therefore, the same response can result from Bcd in one species binding with a low affinity to a promoter with many optimally spaced binding sites (a sensitive promoter) and in another species Bcd binding with high affinity to fewer sites of variable spacing (a less sensitive promoter). Given that there are 5 differences between the *Musca* and *Drosophila* Bcd homeodomains (figure 4.2) and the configurations of binding sites in the *hb* promoter have diverged, what are the functional differences in these components between *Musca* and *Drosophila*?

8.2.3 Co-evolution of *bcd* and Bcd-dependent promoters?

The yeast transcription results, presented in chapter 6 of this thesis, reveal that there are incompatibilities in the components of the Bcd-*hb* interaction between *Drosophila* and *Musca*. In comparisons between *Drosophila* Bcd and *Musca* Bcd, the latter transcription factor requires properties specific to the *Musca hb* promoter (for example, additional binding sites) for it to give the same transcriptional output as *Drosophila* Bcd (see 6.4.1). Have other experiments revealed functional differences between the components of the Bcd-*hb* interaction in *Drosophila* and *Musca*?

In transgenic experiments the *Musca hb* promoter is recognised by *Drosophila* Bcd and in *Drosophila* embryos it can rescue *hb* mutant defects in anterior structures (Bonneton *et al.*, 1997). However, *Musca bcd* does not appear to rescue *Drosophila bcd* mutants as well as when *Drosophila bcd* constructs are used (Shaw *et al.*, submitted; Berleth *et al.*, 1988). Indeed, *Drosophila* Bcd binds with a higher affinity than does *Musca* Bcd, to a range of binding sites in both the *Drosophila* and *Musca hb* promoters (Shaw *et al.*, submitted). Furthermore, these experiments revealed that *Drosophila* and *Musca* Bcd preferentially bound to their cognate *hb* promoters.

Are these differences between *Drosophila* and *Musca* evident in other promoters? Characterisation of the Bcd-dependent *tll* promoter in *Musca* has revealed that, as with the *hb* promoters, it is also restructured in terms of the sequence, number, orientation and

spacing of Bcd-binding sites when compared with the *Drosophila tll* promoter (figure 8.3; Shaw *et al.*, submitted). Intriguingly, this analysis of another Bcd-dependent promoter between *Drosophila* and *Musca* showed that *Drosophila* Bcd again bound with a higher affinity than *Musca* Bcd. However, Bcd from either species preferentially bound to the *Musca tll* promoter rather than to the *Drosophila tll* promoter, which may also suggest that *Musca* promoters are more sensitive to Bcd-binding than *Drosophila* promoters.

Therefore, evidence from a range of experiments reveals that there are incompatibilities between Bcd and its target promoters in *Musca* and *Drosophila*. Indeed, further incompatibilities are exhibited by the divergent protein encoded by *Megaselia Bcd*, which does not recognise the *Drosophila hb* promoter in yeast (6.3.1) or in transgenic *Drosophila* (P. Shaw personal communication).

Given the functional and molecular differences between the *Drosophila* and *Musca* components, it is possible the Bcd-*hb* interactions in each species gives a similar output as a consequence of co-evolution between the Bcd-binding affinity (high in *Drosophila* and low in *Musca*) and the promoter signature (more sensitive in *Musca*, but less so in *Drosophila*). This is in agreement with Gibson's model based on the output of the Bcd-*hb* interaction (the extent of *hb* expression along the anterior-posterior axis of the embryo) being a balance between the Bcd binding affinity and the configuration of Bcd-binding sites in the *hb* promoter. However, this could be an overly simplistic explanation since the effects of other species-specific co-factors such as Chip and SAP18 are as yet unknown (Torigoi *et al.*, 1999; Zhu *et al.*, 2001).

8.2.4 The Bcd gradient and egg size

The larger sizes of the *Musca*, *Calliphora* and *Lucilia* embryos compared with those of *Drosophila* raises the question of how the Bcd gradient functions along the anterior-posterior axis of larger embryos (see figure 1.6). Gibson's model predicts that the threshold position of *hb* expression can be moved further towards the posterior by increasing the number of Bcd-binding sites in the *hb* promoter and/or by generating a more sensitive configuration of these binding sites. Alternatively, increasing the Bcd

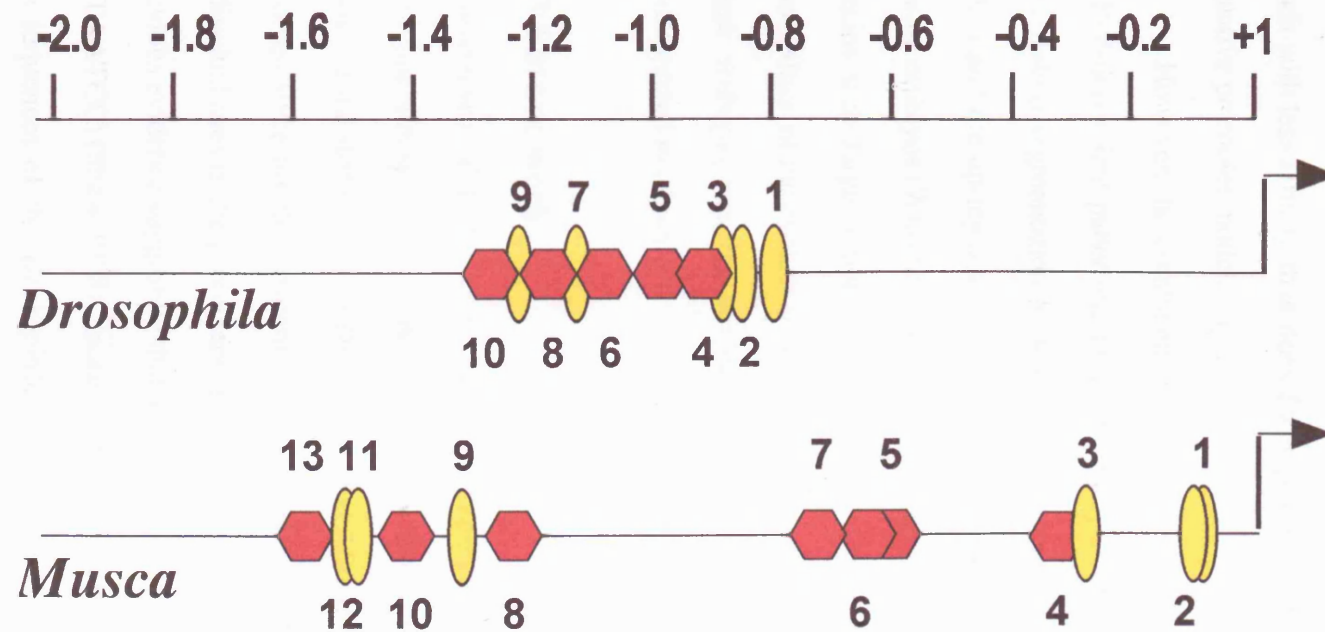


Figure 8.3 Configuration of Bcd-binding sites in the *Drosophila* and *Musca tll* promoters. The transcription start site is represented by the arrows and the scale is in kb upstream of the transcription start site. The positions of Bcd-binding sites are illustrated by red hexagons (TAAT core sites) and yellow ovals (non-TAAT sites), which are numbered from 3' to 5'.

binding affinity would also result in a posterior shift in the threshold position.

Interestingly, although *Lucilia* has the same number of sites as *Drosophila*, both *Musca* and *Calliphora* have additional sites and the results described above for both the *Musca hb* and *tll* promoters suggest that these could be more sensitive to Bcd than the equivalent *Drosophila* promoters. In addition, since it has been found that *Musca* Bcd binds with less affinity than does *Drosophila* Bcd the functional evidence favours a more sensitive promoter model.

However, the combined effects of a number of other factors could also contribute to Bcd-dependent patterning in larger embryos. For example, it has been shown that the *Drosophila* segmentation protein Chip can potentiate Bcd activity *in vivo* (Torigoi *et al.*, 1999) and the up-regulation of maternal *hb* can rescue Bcd-dependent structures in *bcd* mutant embryos (Wimmer *et al.*, 2000). Indeed, *bcd* could encode a more stable protein in species with larger embryos or its expression could be up-regulated. Therefore, the contribution of the evolution of the Bcd-*hb* interaction to the reading of the Bcd gradient in larger embryos remains unclear, especially since the Bcd gradient has not yet been characterised in *Musca*.

8.3 Future work

Co-evolution of Bcd and its target promoters may have resulted in the inter-specific incompatibilities that are observed between these components. However, the molecular basis for the differences in *Drosophila* and *Musca* Bcd binding affinities remains to be resolved since no clear pattern was observed in the binding preferences of either Bcd to individual sites in the promoters of either species (Shaw *et al.*, submitted). This contradicts previous evidence suggesting that *Musca* Bcd preferentially bound to sites flanked by a T (TTAATCC) (Shaw 1998). Indeed, there is no clear species-specific relationship between the sequences of the *Drosophila* and *Musca* Bcd-binding sites in the *hb* and *tll* (N. Wratten personal communication) promoters of these species. This could be investigated by testing the effect of individually and combinatorially replacing residues in the

Drosophila Bcd homeodomain with the *Musca* specific residues. These tests could be carried out using band shift assays and transgenic *Drosophila*.

To investigate the co-evolutionary hypothesis further it would be necessary to perform the reciprocal *in vivo* tests to those carried out to date by studying the performance of *Drosophila bcd* and *hb* in *Musca* embryos. However, the development of suitable vectors and markers to do this is still in its infancy (see O'Brochta *et al.*, 1996) despite the recent generation of transgenic *Musca* using the *piggyBac* transposable element (Hediger *et al.*, 2001).

Another possible method to compare the *hb* promoters would be *in vitro* transcription using nuclear extracts (for example as used by Read *et al.*, 1990) from *Drosophila*, *Musca* and even *Calliphora* and *Lucilia* embryos. This would allow tests of promoter sensitivity to Bcd with the inclusion of as yet unknown species-specific co-factors. Indeed, using *in vitro* transcription comparisons, the configurations of binding sites in the *Musca*, *Calliphora* and *Lucilia* promoters could be manipulated to test the importance of sequence, spacing, number and orientation and the species-specific Bcd preferences for all these promoter properties. This would require the *Lucilia* and *Calliphora* Bcd genes to be sequenced in full and could pinpoint the changes between these species which result in a difference in their abilities to rescue anterior structures in *Drosophila* embryos (Schröder and Sander 1993 and see 1.13).

It appears that the key to understanding how the Bcd gradient patterns the anterior of larger embryos, in comparison to those of *Drosophila*, is to characterise the Bcd gradient in *Musca* embryos. This could be carried out using the same method as was used to characterise the *Drosophila* Bcd gradient (Driever and Nüsslein-Volhard 1988a) and an antibody that recognises *Musca* Bcd.

References

- Acampora, D., Avantaggiato, V., Tuorto, F., Barone, P., Reichert, H., Finkelstein, R. and Simeone, A. (1998). Murine Otx1 and *Drosophila otd* genes share conserved genetic functions required in invertebrate and vertebrate brain development. *Development* **125**, 1691-1702.
- Ades, S. E. and Sauer, R. T. (1994). Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. *Biochemistry* **33**, 9187-9194.
- Ades, S. E. and Sauer, R. T. (1995). Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry* **34**, 14601-14608.
- Akam, M. (1989). Hox and HOM: homologous gene clusters in insects and vertebrates. *Cell* **57**, 347-349.
- Akam, M. (1998). Hox genes, homeosis and the evolution of segment identity: no need for hopeless monsters. *International Journal of Developmental Biology* **42**, 445-451, SI 1998.
- Akam, M. (2000). Arthropods: developmental diversity within a (super) phylum. *PNAS USA* **97**, 4438-4441.
- Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster* - natural selection and translational accuracy. *Genetics* **136**, 927-935.
- Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851-1864.
- Averof, M. and Akam, M. (1993). HOM/Hox genes of *Artemia*: implications for the origin of insect and crustacean body plans. *Current Biology* **3**, 73-78.
- Beachy, P. A., Krasnow, M. A., Gavis, E. R. and Hogness, D. S. (1988). An Ultrabithorax protein binds sequences near its own and the *Antennapedia* P1 promoters. *Cell* **55**, 1069-1081.
- Beachy, P. A., Varkey, J., Young, K. E., von Kessler, D. P., Sun, B. I. and Ekker, S. C. (1993). Cooperative binding of an Ultrabithorax homeodomain protein to nearby and distant DNA sites. *Mol Cell Biol* **13**, 6941-6956.

- Belting, H., Shashikant, C. S. and Ruddle, F. H. (1998). Modification of expression and *cis*-regulation of *Hoxc8* in the evolution of diverged axial morphology. *PNAS USA* **95**, 2355-2360.
- Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. and Nüsslein-Volhard, C. (1988). The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J* **7**, 1749-1756.
- Beverley, S. M. and Wilson, A. C. (1984). Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J Mol Evol* **21**, 1-13.
- Bonneton, F., Theodore, L., Silar, P., Maroni, G. and Wegnez, M. (1996). Response of *Drosophila* metallothionein promoters to metallic, heat shock and oxidative stresses. *FEBS Lett* **380**, 33-38.
- Bonneton, F., Shaw, P. J., Fazakerley, C., Shi, M. and Dover, G. A. (1997). Comparison of *bicoid*-dependent regulation of *hunchback* between *Musca domestica* and *Drosophila melanogaster*. *Mech Dev* **66**, 143-156.
- Brown, S., Fellers, J., Shippy, T., Denell, R., Stauber, M. and Schmidt-Ott, U. (2001). A strategy for mapping *bicoid* on the phylogenetic tree. *Curr Biol* **11**, R43-44.
- Burke, A. C., Nelson, C. E., Morgan, B. A. and Tabin, C. (1995). Hox genes and the evolution of vertebrate axial morphology. *Development* **121**, 333-346.
- Burz, D. S., Rivera-Pomar, R., Jäckle, H. and Hanes, S. D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold- dependent gene activation in the *Drosophila* embryo. *EMBO J* **17**, 5998-6009.
- Burz, D. S. and Hanes, S. D. (2001). Isolation of mutations that disrupt cooperative DNA binding by the *Drosophila bicoid* protein. *J Mol Biol* **305**, 219-230.
- Carroll, S. B. (1994). Developmental regulatory mechanisms in the evolution of insect diversity. *Development supplement*, 217-223.
- Carroll, S. B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature* **376**, 479-485.

- Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577-580.
- Carroll, S. B., Grenier, J. K. and Weatherbee, S. D. (2001). *From DNA to Diversity, Molecular Genetics and the Evolution of Animal Design*. Malden: Blackwell Science.
- Cavalier-Smith, T. (1985). *Eukaryote gene numbers, non-coding DNA and genome size*. In *The evolution of genome size*, (ed. T. Cavalier-Smith), pp. 69-103. London: John Wiley and Sons Ltd.
- Cherbas, L. and Cherbas, P. (1993). The arthropod initiator: the capsite consensus plays an important role in transcription. *Insect Biochem Mol Biol* **23**, 81-90.
- Church, G. M. and Gilbert, W. (1984). Genomic sequencing. *PNAS USA* **81**, 1991-1995.
- Curtis, D., Apfeld, J. and Lehmann, R. (1995). *nanos* is an evolutionarily conserved organizer of anterior-posterior polarity. *Development* **121**, 1899-1910.
- Dave, V., Zhao, C., Yang, F., Tung, C. S. and Ma, J. (2000). Reprogrammable recognition codes in *bicoid* homeodomain-DNA interaction. *Mol Cell Biol* **20**, 7673-84.
- Davidson, E. H. (1986). *Gene Activity in Early Development*. Orlando: Academic Press.
- Davidson, E. H. (2001). *Genomic regulatory systems : development and evolution*. San Diego: Academic Press.
- Dearden, P. and Akam, M. (1999). Developmental evolution: Axial patterning in insects. *Curr Biol* **9**, R591-594.
- Dearden, P. K. and Akam, M. (2001). Early embryo patterning in the grasshopper, *Schistocerca gregaria*: *wingless*, *decapentaplegic* and *caudal* expression. *Development* **128**, 3435-3444.
- Devon, R. S., Porteous, D. J. and Brookes, A. J. (1995). Splinkerettes improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* **23**, 1644-1645.
- Dover, G. A. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111-117.

- Dover, G. A. and Flavell, R. B. (1984). Molecular coevolution: DNA divergence and the maintenance of function. *Cell* **38**, 622-623.
- Dover, G. A. (1993). Evolution of genetic redundancy for advanced players. *Curr Opin Genet Dev* **3**, 902-910.
- Dover, G. (2000). How genomic and developmental dynamics affect evolutionary processes. *Bioessays* **22**, 1153-1159.
- Driever, W. and Nüsslein-Volhard, C. (1988a). A gradient of *bicoid* protein in *Drosophila* embryos. *Cell* **54**, 83-93.
- Driever, W. and Nüsslein-Volhard, C. (1988b). The *bicoid* protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell* **54**, 95-104.
- Driever, W., Thoma, G. and Nüsslein-Volhard, C. (1989a). Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the *bicoid* morphogen. *Nature* **340**, 363-367.
- Driever, W., Ma, J., Nüsslein-Volhard, C. and Ptashne, M. (1989b). Rescue of *bicoid* mutant *Drosophila* embryos by *bicoid* fusion proteins containing heterologous activating sequences. *Nature* **342**, 149-154.
- Driever, W. and Nüsslein-Volhard, C. (1989). The *bicoid* protein is a positive regulator of *hunchback* transcription in the early *Drosophila* embryo. *Nature* **337**, 138-143.
- Driever, W. (1993). *Maternal control of anterior development in the Drosophila embryo*. In *The development of Drosophila melanogaster*, vol. 1 (ed. M. Bate and A. Martinez Arias), pp. 301-324. New York: Cold Spring Harbor Press.
- Dubnau, J. and Struhl, G. (1996). RNA recognition and translational regulation by a homeodomain protein. *Nature* **379**, 694-699.
- Duboule, D. and Wilkins, A. S. (1998). The evolution of 'bricolage'. *Trends Genet* **14**, 54-59.
- Duman-Scheel, M. and Patel, N. H. (1999). Analysis of molecular marker expression reveals neuronal homology in distantly related arthropods. *Development* **126**, 2327-2334.

- Ekker, S. C., Young, K. E., von Kessler, D. P. and Beachy, P. A. (1991). Optimal DNA sequence recognition by the Ultrabithorax homeodomain of *Drosophila*. *EMBO J* **10**, 1179-1186.
- Emili, A., Greenblatt, J. and Ingles, C. J. (1994). Species-specific interaction of the glutamine-rich activation domains of SP1 with the TATA box-binding protein. *Mol Cell Biol* **14**, 1582-1593.
- Espinas, M. L., Canudas, S., Fanti, L., Pimpinelli, S., Casanova, J. and Azorin, F. (2000). The GAGA factor of *Drosophila* interacts with SAP18, a Sin3-associated polypeptide. *EMBO Reports* **1**, 253-259.
- Fay, D. S., Stanley, H. M., Han, M. and Wood, W. B. (1999). A *Caenorhabditis elegans* homologue of *hunchback* is required for late stages of development but not early embryonic patterning. *Dev Biol* **205**, 240-253.
- Feinberg, A. P. and Vogelstein, B. (1984). "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. *Anal Biochem* **137**, 266-267.
- Finkelstein, R. and Perrimon, N. (1990). The *orthodenticle* gene is regulated by *bicoid* and *torso* and specifies *Drosophila* head development. *Nature* **346**, 485-488.
- Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, A. C. and Perrimon, N. (1990). The *orthodenticle* gene encodes a novel homeodomain protein involved in the development of the *Drosophila* nervous system and ocellar visual structures. *Genes Dev* **4**, 1516-1527.
- Frohnhofer, H. G. and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the *Drosophila* embryo by the maternal gene *bicoid*. *Nature* **324**, 120-125.
- Galas, D. J. and Schmitz, A. (1978). DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157-3170.
- Gao, Q. and Finkelstein, R. (1998). Targeting gene expression to the head: the *Drosophila orthodenticle* gene is a direct target of the Bicoid morphogen. *Development* **125**, 4185-4193.
- Gehring, W. J., Affolter, M. and Burglin, T. (1994b). Homeodomain proteins. *Annu Rev Biochem* **63**, 487-526.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G. and Wuthrich, K. (1994a). Homeodomain-DNA recognition. *Cell* **78**, 211-223.

Georgopoulos, K., Moore, D. D. and Derfler, B. (1992). Ikaros, an early lymphoid-specific transcription factor and a putative mediator for T-cell commitment. *Science* **258**, 808-812.

Gibson, G. (1996). Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor Popul Biol* **49**, 58-89.

Gibson, G. (2000). Evolution: Hox genes and the cellared wine principle. *Curr Biol* **10**, R452-455.

Goodrich, J. A. and Tjian, R. (1994). TBP-TAF complexes: selectivity factors for eukaryotic transcription. *Curr Opin Cell Biol* **6**, 403-409.

Grbic, M. and Strand, M. R. (1998). Shifts in the life history of parasitic wasps correlate with pronounced alterations in early development. *PNAS USA* **95**, 1097-1101.

Gyuris, J., Golemis, E., Chertkov, H. and Brent, R. (1993). CD11, a human G1-phase and S-phase protein phosphatase that associates with CDK2. *Cell* **75**, 791-803.

Hamilton, B. A., Palazzolo, M. J., Chang, J. H., VijayRaghavan, K., Mayeda, C. A., Whitney, M. A. and Meyerowitz, E. M. (1991). Large scale screen for transposon insertions into cloned genes. *PNAS USA* **88**, 2731-2735.

Hancock, J. M. and Dover, G. A. (1990). 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res* **18**, 5949-5954.

Hancock, J. M. and Armstrong, J. S. (1994). SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* **10**, 67-70.

Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. *J Mol Evol* **41**, 1038-1047.

- Hancock, J. M. (1996). Simple sequences and the expanding genome. *Bioessays* **18**, 421-425.
- Hancock, J. M., Shaw, P. J., Bonneton, F. and Dover, G. A. (1999). High sequence turnover in the regulatory regions of the developmental gene *hunchback* in insects. *Mol Biol Evol* **16**, 253-265.
- Hancock, J. M. and Vogler, A. P. (2000). How slippage-derived sequences are incorporated into rRNA variable- region secondary structure: implications for phylogeny reconstruction. *Mol Phylogenet Evol* **14**, 366-374.
- Hancock, J. M., Worthey, E. A. and Santibanez-Koref, M. F. (2001). A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol Biol Evol* **18**, 1014-1023.
- Hanes, S. D. and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell* **57**, 1275-1283.
- Hanes, S. D. and Brent, R. (1991). A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **251**, 426-430.
- Hanes, S. D., Riddihough, G., Ish-Horowicz, D. and Brent, R. (1994). Specific DNA recognition and intersite spacing are critical for action of the *bicoid* morphogen. *Mol Cell Biol* **14**, 3364-3375.
- Harding, K. and Levine, M. (1988). Gap genes define the limits of *Antennapedia* and *Bithorax* gene-expression during early *Drosophila* development. *EMBO J* **7**, 205-214.
- Harlow, E. and Lane, D. (1988). *Antibodies a laboratory Manual*. New York: Cold Spring Harbor Laboratory.
- Harr, B., Zangerl, B. and Schlötterer, C. (2000). Removal of microsatellite intrusions by DNA replication slippage: phylogenetic evidence from *Drosophila*. *Mol Biol Evol* **17**, 1001-1009.
- Hediger, M., Niessen, M., Wimmer, E. A., Dübendorfer, A. and Bopp, D. (2001). Genetic transformation of the housefly *Musca domestica* with the lepidopteran derived transposon *piggyBac*. *Insect Molecular Biology* **10**, 113-119.

- Hirsch, J. A. and Aggarwal, A. K. (1995). Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *EMBO J* **14**, 6280-6291.
- Hirth, F. and Reichert, H. (1999). Conserved genetic programs in insect and mammalian brain development. *Bioessays* **21**, 677-684.
- Hoch, M., Seifert, E. and Jäckle, H. (1991). Gene expression mediated by *cis*-acting sequences of the *Krüppel* gene in response to the *Drosophila* morphogens *bicoid* and *hunchback*. *EMBO J* **10**, 2267-2278.
- Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A. and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development* supplement, 125-133.
- Holland, P. W. H. (1999). The future of evolutionary developmental biology. *Nature* **402**, C41-C44.
- Hülskamp, M., Schröder, C., Pfeifle, C., Jäckle, H. and Tautz, D. (1989). Posterior segmentation of the *Drosophila* embryo in the absence of a maternal posterior organizer gene. *Nature* **338**, 629-632.
- Hülskamp, M., Pfeifle, C. and Tautz, D. (1990). A morphogenetic gradient of *hunchback* protein organizes the expression of the gap genes *Krüppel* and *knirps* in the early *Drosophila* embryo. *Nature* **346**, 577-580.
- Hülskamp, M., Lukowitz, W., Beermann, A., Glaser, G. and Tautz, D. (1994). Differential regulation of target genes by different alleles of the segmentation gene *hunchback* in *Drosophila*. *Genetics* **138**, 125-134.
- Irish, V., Lehmann, R. and Akam, M. (1989). The *Drosophila* posterior-group gene *nanos* functions by repressing *hunchback* activity. *Nature* **338**, 646-648.
- Ito, H., Fukuda, Y., Murata, K. and Kimura, A. (1983). Transformation of intact yeast cells treated with alkali cations. *J Bacteriology* **153**, 163-168.
- Iwasa, J. H., Suver, D. W. and Savage, R. M. (2000). The leech *hunchback* protein is expressed in the epithelium and CNS but not in the segmental precursor lineages. *Dev Genes Evol* **210**, 277-288.

- Jacob, F. (1977). Evolution and tinkering. *Science* **196**, 1161-1166.
- Janody, F., Sturny, R., Catala, F., Desplan, C. and Dostatni, N. (2000). Phosphorylation of *bicoid* on MAP-kinase sites: contribution to its interaction with the *torso* pathway. *Development* **127**, 279-289.
- Janody, F., Sturny, R., Schaeffer, V., Azou, Y. and Dostatni, N. (2001). Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway. *Development* **128**, 2281-2290.
- Jeffreys, A. J., Neumann, R. and Wilson, V. (1990). Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60**, 473-485.
- Kambadur, R., Koizumi, K., Stivers, C., Nagle, J., Poole, S. J. and Odenwald, W. F. (1998). Regulation of POU genes by *castor* and *hunchback* establishes layered compartments in the *Drosophila* CNS. *Genes Dev* **12**, 246-260.
- Kasten, M. M., Dorland, S. and Stillman, D. J. (1997). Large protein complex containing the yeast Sin3p and Rpd3p transcriptional regulators. *Mol Cell Biol* **17**, 4852-4858.
- Kehle, J., Beuchle, D., Treuheit, S., Christen, B., Kennison, J. A., Bienz, M. and Muller, J. (1998). dMi-2, a *hunchback*-interacting protein that functions in polycomb repression. *Science* **282**, 1897-1900.
- Kim, J., Sif, S., Jones, B., Jackson, A., Koipally, J., Heller, E., Winandy, S., Viel, A., Sawyer, A., Ikeda, T. et al. (1999). Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. *Immunity* **10**, 345-355.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge ; New York: Cambridge University Press.
- Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. and Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell* **63**, 579-590.
- Krause, H. M., Klemenz, R. and Gehring, W. J. (1988). Expression, modification and localisation of the Fushi-Tarazu protein in *Drosophila* embryos. *Gene Dev* **2**, 1021-1036.

- Kruglyak, S., Durrett, R., Schug, M. D. and Aquadro, C. F. (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol Biol Evol* **17**, 1210-1219.
- Laughon, A. (1991). DNA binding specificity of homeodomains. *Biochemistry* **30**, 11357-11367.
- Lawrence, P. A. (1992). *The making of a fly: the genetics of animal design*. Oxford [England] ; Cambridge, Mass., USA: Blackwell Science.
- Lehmann, R. and Nüsslein-Volhard, C. (1987). *hunchback*, a gene required for segmentation of an anterior and posterior region of the *Drosophila* embryo. *Dev Biol* **119**, 402-417.
- Leuzinger, S., Hirth, F., Gerlich, D., Acampora, D., Simeone, A., Gehring, W. J., Finkelstein, R., Furukubo-Tokunaga, K. and Reichert, H. (1998). Equivalence of the fly *orthodenticle* gene and the human OTX genes in embryonic brain development of *Drosophila*. *Development* **125**, 1703-1710.
- Levinson, G. and Gutman, G. A. (1987). Slipped-strand mispairing - a major mechanism for DNA-sequence evolution. *Mol Biol Evol* **4**, 203-221.
- Li, X. and Noll, M. (1994). Evolution of distinct developmental functions of three *Drosophila* genes by acquisition of different *cis*-regulatory regions. *Nature* **367**, 83-87.
- Li, Y., Brown, S. J., Hausdorf, B., Tautz, D., Denell, R. E. and Finkelstein, R. (1996). Two *orthodenticle*-related genes in the short-germ beetle *Tribolium castaneum*. *Dev Genes Evol* **206**, 35-45.
- Lin, K. C. and Shiuan, D. (1995). A simple method for DNaseI footprint analysis. *J Biochem Biophys Methods* **30**, 85-89.
- Louvion, J. F., Havaux-Copf, B. and Picard, D. (1993). Fusion of GAL4-VP16 to a steroid-binding domain provides a tool for gratuitous induction of galactose-responsive genes in yeast. *Gene* **131**, 129-134.
- Ludwig, M. Z. and Kreitman, M. (1995). Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Mol Biol Evol* **12**, 1002-1011.

- Ludwig, M. Z., Patel, N. H. and Kreitman, M. (1998). Functional analysis of *eve* stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* **125**, 949-958.
- Ludwig, M. Z., Bergman, C., Patel, N. H. and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564-567.
- Luk, S. K. S., Kilpatrick, M., Kerr, K. and Macdonald, P. M. (1994). Components acting in localisation of *bicoid* messenger RNA are conserved among *Drosophila* species. *Genetics* **137**, 521-530.
- Lukowitz, W., Schröder, C., Glaser, G., Hülskamp, M. and Tautz, D. (1994). Regulatory and coding regions of the segmentation gene *hunchback* are functionally conserved between *Drosophila virilis* and *Drosophila melanogaster*. *Mech Dev* **45**, 105-115.
- Ma, X., Yuan, D., Diepold, K., Scarborough, T. and Ma, J. (1996). The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. *Development* **122**, 1195-1206.
- Ma, X., Yuan, D., Scarborough, T. and Ma, J. (1999). Contributions to gene activation by multiple functions of Bicoid. *Biochem J* **338**, 447-455.
- Macdonald, P. M. and Struhl, G. (1988). *cis*-acting sequences responsible for anterior localization of *bicoid* mRNA in *Drosophila* embryos. *Nature* **336**, 595-598.
- Macdonald, P. M. (1990). *bicoid* messenger RNA localization signal - phylogenetic conservation of function and RNA secondary structure. *Development* **110**, 161-171.
- Mao, C., Carlson, N. G. and Little, J. W. (1994). Cooperative DNA-protein interactions Effects of changing the spacing between adjacent binding sites. *J Mol Biol* **235**, 532-544.
- Margolis, J. S., Borowsky, M. L., Steingrimsson, E., Shim, C. W., Lengyel, J. A. and Posakony, J. W. (1995). Posterior stripe expression of *hunchback* is driven from two promoters by a common enhancer element. *Development* **121**, 3067-3077.
- McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature* **351**, 652-654.

- McGregor, A. P., Shaw, P. J. and Dover, G. A. (2001b). Sequence and expression of the *hunchback* gene in *Lucilia sericata*: a comparison with other Dipterans. *Dev Genes Evol* **211**, 315-318.
- McGregor, A. P., Shaw, P. J., Hancock, J. M., Bopp, D., Hediger, M., Wratten, N. S. and Dover, G. A. (2001a). Rapid restructuring of bicoid-dependent *hunchback* promoters within and between Dipteran species: implications for molecular co-evolution. *Evol Dev* **3**, 397-407.
- Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. and Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res* **20**, 4255-4262.
- Murata, Y. and Wharton, R. P. (1995). Binding of pumilio to maternal *hunchback* mRNA is required for posterior patterning in *Drosophila* embryos. *Cell* **80**, 747-756.
- Nagy, L. M. (1994). Insect segmentation. A glance posterior. *Curr Biol* **4**, 811-814.
- Namba, R., Pazdera, T. M., Cerrone, R. L. and Minden, J. S. (1997). *Drosophila* embryonic pattern repair: how embryos respond to *bicoid* dosage alteration. *Development* **124**, 1393-1403.
- Newfeld, S. J., Tachida, H. and Yedvobnick, B. (1994). Drive-selection equilibrium - homopolymer evolution in the *Drosophila* gene *mastermind*. *J Mol Evol* **38**, 637-641.
- Niessing, D., Dostatni, N., Jackle, H. and Rivera-Pomar, R. (1999). Sequence interval within the PEST motif of Bicoid is important for translational repression of *caudal* mRNA in the anterior region of the *Drosophila* embryo. *EMBO J* **18**, 1966-1973.
- O'Brochta, D. A., Warren, W. D., Saville, K. J. and Atkinson, P. W. (1996). Hermes, a functional non-drosophilid insect gene vector from *Musca domestica*. *Genetics* **142**, 907-914.
- O'Neill, D. W., Schoetz, S. S., Lopez, R. A., Castle, M., Rabinowitz, L., Shor, E., Krawchuk, D., Goll, M. G., Renz, M., Seelig, H. P. et al. (2000). An Ikaros-containing chromatin-remodeling complex in adult-type erythroid cells. *Mol Cell Biol* **20**, 7572-7582.
- Ohta, T. and Dover, G. A. (1984). The cohesive population genetics of molecular drive. *Genetics* **108**, 501-521.

- Orr, H. T. (2001). Beyond the Qs in the polyglutamine diseases. *Genes Dev* **15**, 925-932.
- Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., Gehring, W. J. and Wuthrich, K. (1990). Protein DNA contacts in the structure of a homeodomain complex determined by nuclear magnetic resonance spectroscopy in solution. *EMBO J* **9**, 3085-3092.
- Padegimas, L. S. and Reichert, N. A. (1998). Adaptor ligation-based polymerase chain reaction-mediated walking. *Anal Biochem* **260**, 149-153.
- Palopoli, M. F. and Patel, N. H. (1996). Neo-Darwinian developmental evolution: can we bridge the gap between pattern and process? *Curr Opin Genet Dev* **6**, 502-508.
- Patel, N. H. (1994). Developmental evolution: insights from studies of insect segmentation. *Science* **266**, 581-590.
- Patel, N. H., Hayward, D. C., Lall, S., Pirkel, N. R., DiPietro, D. and E., B. E. (2001). Grasshopper *hunchback* expression reveals conserved and novel aspects of axis formation and segmentation. *Development* **128**, 3459-3472.
- Peixoto, A. A., Costa, R., Wheeler, D. A., Hall, J. C. and Kyriacou, C. P. (1992). Evolution of the threonine-glycine repeat region of the *period* gene in the *melanogaster* species subgroup of *Drosophila*. *J Mol Evol* **35**, 411-419.
- Pelegri, F. and Lehmann, R. (1994). A role of polycomb group genes in the regulation of gap gene expression in *Drosophila*. *Genetics* **136**, 1341-1353.
- Percival-Smith, A., Muller, M., Affolter, M. and Gehring, W. J. (1990). The interaction with DNA of wild-type and mutant fushi-tarazu homeodomains. *EMBO J* **9**, 3967-3974.
- Pignoni, F., Steingrimsson, E. and Lengyel, J. A. (1992). *bicoid* and the terminal system activate *tailless* expression in the early *Drosophila* embryo. *Development* **115**, 239-251.
- Qian, S., Capovilla, M. and Pirrota, V. (1993). Molecular mechanisms of pattern formation by the BRE enhancer of the *Ubx* gene. *EMBO J* **12**, 3865-3877.
- Raff, R. A. (1996). *The shape of life : genes, development, and the evolution of animal form*. Chicago: University of Chicago Press.

- Read, D., Nishigaki, T. and Manley, J. L. (1990). The *Drosophila even-skipped* promoter is transcribed in a stage-specific manner *in vitro* and contains multiple overlapping factor-binding sites. *Mol Cell Biol* **10**, 4334-4344.
- Richards, O. W. and Davies, R. G. (1977). *Imms' general textbook of entomology*. London: Chapman and Hall.
- Richter, J. D. and Theurkauf, W. E. (2001). Development - The message is in the translation. *Science* **293**, 60-62.
- Rivera-Pomar, R., Lu, X. G., Perrimon, N., Taubert, H. and Jackle, H. (1995). Activation of posterior gap gene expression in the *Drosophila* blastoderm. *Nature* **376**, 253-256.
- Rivera-Pomar, R. and Jäckle, H. (1996). From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. *Trends Genet* **12**, 478-483.
- Rohr, K. B., Tautz, D. and Sander, K. (1999). Segmentation gene expression in the mothmidge *Clogmia albipunctata* (Diptera, psychodidae) and other primitive dipterans. *Dev Genes Evol* **209**, 145-154.
- Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.
- Santibanez-Koref, M. F., Gangeswaran, R. and Hancock, J. M. (2001). A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol Biol Evol* **18**, 2119-2123.
- Sauer, F. and Jackle, H. (1993). Dimerisation and the control of transcription by Krüppel. *Nature* **364**, 454-457.
- Sauer, F., Hansen, S. K. and Tjian, R. (1995a). DNA template and activator-coactivator requirements for transcriptional synergism by *Drosophila bicoid*. *Science* **270**, 1825-1828.
- Sauer, F., Hansen, S. K. and Tjian, R. (1995b). Multiple TAFIIs directing synergistic activation of transcription. *Science* **270**, 1783-1788.
- Sauer, F., Rivera-Pomar, R., Hoch, M. and Jäckle, H. (1996). Gene regulation in the *Drosophila* embryo. *Philos Trans R Soc Lond B Biol Sci* **351**, 579-587.

Savage, R. M. and Shankland, M. (1996). Identification and characterization of a *hunchback* orthologue, *Lzf2*, and its expression during leech embryogenesis. *Dev Biol* **175**, 205-217.

Schaeffer, V., Janody, F., Loss, C., Desplan, C. and Wimmer, E. A. (1999). Bicoid functions without its TATA-binding protein-associated factor interaction domains. *PNAS USA* **96**, 4461-4466.

Schlötterer, C. and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20**, 211-215.

Schlötterer, C., Ritter, R., Harr, B. and Brem, G. (1998). High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Mol Biol Evol* **15**, 1269-1274.

Schmid, K. J. and Tautz, D. (1999). A comparison of homologous developmental genes from *Drosophila* and *Tribolium* reveals major differences in length and trinucleotide repeat content. *J Mol Evol* **49**, 558-566.

Schröder, C., Tautz, D., Seifert, E. and Jäckle, H. (1988). Differential regulation of the two transcripts from the *Drosophila* gap segmentation gene *hunchback*. *EMBO J* **7**, 2881-2887.

Schröder, R. and Sander, K. (1993). A comparison of transplantable Bicoid activity and partial Bicoid homeobox sequences in several *Drosophila* and blowfly species (Calliphoridae). *Roux's Archives of Developmental Biology* **203**, 34-43.

Schug, M. D., Mackay, T. F. and Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nat Genet* **15**, 99-102.

Schug, M. D., Hutter, C. M., Wetterstrand, K. A., Gaudette, M. S., MacKay, T. F. C. and Aquadro, C. F. (1998). The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol Biol Evol* **15**, 1751-1760.

Schulte, P. M., Glémet, H. C., Fiebig, A. A. and Powers, D. A. (2000). Adaptive variation in lactate dehydrogenase-B gene expression: Role of a stress-responsive regulatory element. *PNAS USA* **97**, 6597-6602.

- Seeger, M. A. and Kaufman, T. C. (1990). Molecular analysis of the *bicoid* gene from *Drosophila pseudoobscura* - identification of conserved domains within coding and noncoding regions of the *bicoid* messenger RNA. *EMBO J* **9**, 2977-2987.
- Shashikant, C. S., Kim, C. B., Borbély, M. A., Wang, W. C. H. and Ruddle, F. H. (1998). Comparative studies on mammalian *Hoxc8* early enhancer sequence reveal a baleen whale specific deletion of a *cis*-acting element. *PNAS USA* **95**, 15446-15451.
- Shaw, P. J., Wratten, N. S., McGregor, A. P. and Dover, G. A. Co-evolution in *bicoid*-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. *Evol Dev* (submitted).
- Shaw, P. J. (1998). PhD Thesis. *Molecular characterisation of the interaction between the bicoid and hunchback genes in Musca domestica: insights into the evolution of a regulatory interaction*. In Department of Genetics. pp. 143. Leicester: University of Leicester.
- Shaw, P. J., Salameh, A., McGregor, A. P., Bala, S. and Dover, G. A. (2001). Divergent structure and function of the *bicoid* gene in Muscoidea fly species. *Evol Dev* **3**, 251-262.
- Shubin, N., Tabin, C. and Carroll, S. (1997). Fossils, genes and the evolution of animal limbs. *Nature* **388**, 639-648.
- Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995). An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* **23**, 1087-1088.
- Simeone, A., Acampora, D., Gulisano, M., Stornaiuolo, A. and Boncinelli, E. (1992). Nested expression domains of 4 homeobox genes in the developing rostral brain. *Nature* **358**, 687-690.
- Simpson-Brose, M., Treisman, J. and Desplan, C. (1994). Synergy between the *hunchback* and *bicoid* morphogens is required for anterior patterning in *Drosophila*. *Cell* **78**, 855-865.
- Small, S., Blair, A. and Levine, M. (1992). Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *EMBO J* **11**, 4047-4057.

- Small, S., Blair, A. and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev Biol* **175**, 314-324.
- Sommer, R. and Tautz, D. (1991a). Segmentation gene expression in the housefly *Musca domestica*. *Development* **113**, 419-430.
- Sommer, R. and Tautz, D. (1991b). Asynchronous mitotic domains during blastoderm formation in *Musca domestica* (Diptera). *Roux's Archives of Developmental Biology* **199**, 373-376.
- Sommer, R. J., Retzlaff, M., Goerlich, K., Sander, K. and Tautz, D. (1992). Evolutionary conservation pattern of zinc-finger domains of *Drosophila* segmentation genes. *PNAS USA* **89**, 10782-10786.
- Sonoda, J. and Wharton, R. P. (1999). Recruitment of Nanos to *hunchback* mRNA by Pumilio. *Gene Dev* **13**, 2704-2712.
- Stauber, M., Jackle, H. and SchmidtOtt, U. (1999). The anterior determinant *bicoid* of *Drosophila* is a derived Hox class 3 gene. *PNAS USA* **96**, 3786-3789.
- Stauber, M., Taubert, H. and Schmidt-Ott, U. (2000). Function of *bicoid* and *hunchback* homologs in the basal cyclorrhaphan fly *Megaselia* (Phoridae). *PNAS USA* **97**, 10844-10849.
- Stauber, M., Prell, A. and Schmidt-Ott, U. (2002). A single *Hox3* gene with composite *bicoid* and *zerknüllt* expression characteristics in non-Cyclorrhaphan flies. *PNAS USA* in press.
- Stern, D. L. (1998). A role of *Ultrabithorax* in morphological differences between *Drosophila* species. *Nature* **396**, 463-466.
- Struhl, G., Struhl, K. and Macdonald, P. M. (1989). The gradient morphogen *bicoid* is a concentration-dependent transcriptional activator. *Cell* **57**, 1259-1273.
- Struhl, G., Johnston, P. and Lawrence, P. A. (1992). Control of *Drosophila* body pattern by the *hunchback* morphogen gradient. *Cell* **69**, 237-249.
- Subramaniam, V., Jovin, T. M. and Rivera-Pomar, R. V. (2001). Aromatic amino acids are critical for stability of the *bicoid* homeodomain. *J Biological Chemistry* **276**, 21506-21511.

Sun, L., Liu, A. P. and Georgopoulos, K. (1996). Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. *EMBO J* **15**, 5358-5369.

Tautz, D., Trick, M. and Dover, G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**, 652-656.

Tautz, D., Lehmann, R., Schnürch, H., Schuh, R., Seifert, E., Kienlin, A., Jones, K. and Jäckle, H. (1987). Finger protein of novel structure encoded by *hunchback*, a 2nd member of the gap class of *Drosophila* segmentation genes. *Nature* **327**, 383-389.

Tautz, D. (1988). Regulation of the *Drosophila* segmentation gene *hunchback* by two maternal morphogenetic centres. *Nature* **332**, 281-284.

Tautz, D. and Pfeifle, C. (1989). A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*. *Chromosoma* **98**, 81-85.

Tautz, D. and Nigro, L. (1998). Microevolutionary divergence pattern of the segmentation gene *hunchback* in *Drosophila*. *Mol Biol Evol.* **15**, 1403-1411.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680.

Torigoi, E., Bennani-Baiti, I. M., Rosen, C., Gonzalez, K., Morcillo, P., Ptashne, M. and Dorsett, D. (2000). Chip interacts with diverse homeodomain proteins and potentiates *bicoid* activity *in vivo*. *PNAS USA* **97**, 2686-2691.

Treier, M., Pfeifle, C. and Tautz, D. (1989). Comparison of the gap segmentation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. *EMBO J* **8**, 1517-1525.

Treisman, J. and Desplan, C. (1989). The products of the *Drosophila* gap genes *hunchback* and *Krüppel* bind to the *hunchback* promoters. *Nature* **341**, 335-337.

- Treisman, J., Gonczy, P., Vashishtha, M., Harris, E. and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**, 553-562.
- Tucker-Kellogg, L., Rould, M. A., Chambers, K. A., Ades, S. E., Sauer, R. T. and Pabo, C. O. (1997). Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. *Structure* **5**, 1047-1054.
- Werbrock, A. H., Meiklejohn, D. A., Sainz, A., Iwasa, J. H. and Savage, R. M. (2001). A polychaete hunchback ortholog. *Dev Biol* **235**, 476-488.
- West, R. W., Yokum, R. R. and Ptashne, M. (1984). *Saccharomyces cerevisiae* GAL1-GAL10 divergent promoter region: Location and function of the upstream activating sequence UASG. *Mol Cell Biol* **4**, 2467-2478.
- Wharton, R. P. and Struhl, G. (1991). RNA regulatory elements mediate control of *Drosophila* body pattern by the posterior morphogen *nanos*. *Cell* **67**, 955-967.
- Wieschaus, E., Nüsslein-Volhard, C. and Jurgens, G. (1984). Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. 3. Zygotic loci on the X-chromosome and 4th chromosome. *Roux's Archives of Developmental Biology* **193**, 296-307.
- Wilkins, R. C. and Lis, J. T. (1997). Dynamics of potentiation and activation: GAGA factor and its role in heat shock gene regulation. *Nucleic Acids Res* **25**, 3963-3968.
- Wilkins, R. C. and Lis, J. T. (1998). GAGA factor binding to DNA via a single trinucleotide sequence element. *Nucleic Acids Res* **26**, 2672-2678.
- Wilson, A. C., Maxson, L. R. and Sarich, V. M. (1974). Two types of molecular evolution: evidence from studies of intraspecific hybridization. *PNAS USA* **71**, 2843-2847.
- Wilson, D. S., Sheng, G., Jun, S. and Desplan, C. (1996). Conservation and diversification in homeodomain-DNA interactions: a comparative genetic analysis. *PNAS USA* **93**, 6886-6891.
- Wimmer, E. A., Carleton, A., Harjes, P., Turner, T. and Desplan, C. (2000). Bicoid-independent formation of thoracic segments in *Drosophila*. *Science* **287**, 2476-2479.

- Wolberger, C., Pabo, C. O., Vershon, A. K. and Johnson, A. D. (1991). Crystallization and preliminary x-ray diffraction studies of a mat- α -2-DNA complex. *J Mol Biol* **217**, 11-13.
- Wolff, C., Sommer, R., Schröder, R., Glaser, G. and Tautz, D. (1995). Conserved and divergent expression aspects of the *Drosophila* segmentation gene *hunchback* in the short germ band embryo of the flour beetle *Tribolium*. *Development* **121**, 4227-4236.
- Wolff, C., Schroder, R., Schulz, C., Tautz, D. and Klingler, M. (1998). Regulation of the *Tribolium* homologues of *caudal* and *hunchback* in *Drosophila*: evidence for maternal gradient systems in a short germ embryo. *Development* **125**, 3645-3654.
- Wreden, C., Verrotti, A. C., Schisa, J. A., Lieberfarb, M. E. and Strickland, S. (1997). Nanos and pumilio establish embryonic polarity in *Drosophila* by promoting posterior deadenylation of *hunchback* mRNA. *Development* **124**, 3015-3023.
- Wu, X., Vasisht, V., Kosman, D., Reinitz, J. and Small, S. (2001). Thoracic patterning by the *Drosophila* gap gene *hunchback*. *Dev Biol* **237**, 79-92.
- Xue, L. and Noll, M. (1996). The functional conservation of proteins in evolutionary alleles and the dominant role of enhancers in evolution. *EMBO J* **15**, 3722-3731.
- Yuan, D., Ma, X. and Ma, J. (1996). Sequences outside the homeodomain of bicoid are required for protein- protein interaction. *J Biol Chem* **271**, 21660-21665.
- Yuan, D., Ma, X. and Ma, J. (1999). Recognition of multiple patterns of DNA sites by *Drosophila* homeodomain protein Bicoid. *J Biochem (Tokyo)* **125**, 809-817.
- Zhang, C. C., Muller, J., Hoch, M., Jackle, H. and Beinz, M. (1991). Target sequences for *hunchback* in a control region conferring *Ultrabithorax* expression boundaries. *Development* **113**, 1171-1181.
- Zhao, C., Dave, V., Yang, F., Scarborough, T. and Ma, J. (2000). Target selectivity of bicoid is dependent on nonconsensus site recognition and protein-protein interaction. *Mol Cell Biol* **20**, 8112-8123.
- Zhu, W. and Hanes, S. D. (2000). Identification of *Drosophila* Bicoid-interacting proteins using a custom two-hybrid selection. *Gene* **245**, 329-339.

Zhu, W. C., Foehr, M., Jaynes, J. B. and Hanes, S. D. (2001). *Drosophila* SAP18, a member of the Sin3/Rpd3 histone deacetylase complex, interacts with Bicoid and inhibits its activity. *Dev Genes Evol* **211**, 109-117.

Appendix A PCR primer sequences

Name	Primer sequence (5'-3')	TM °C
MP2F	AACATGGATTGTGAAGCTCTG	55
MP2R	TCGATGCTGAAATGCAACTGAT	55
MP2NF	TCTCGTCGTGTAAATTATTAGCG	55
MP2NR	GATGGGTGATTGTTGTTCTAAGCA	57
MHUF	CAGTGAAGATCTCAAATACATTG	54
MHUR	TGTGAGTCAAGTCGGGTGTA	60
MHUNF	CACTTCTCAGTTGTCGGCTG	59
MHNR2	TTGCAGGTTCAAGTTCCAGTAG	55
MHNR3	CTGTTGGAGCTGTTGCCTGA	55
MHRA	TACCCTCCTCATCCAATAC	45
MHRB	TCCATGGCGGAATCCACA	55
MHFA	TGAATTGGCCATGAACTTG	50
MICF	TAGATGCGTAAATGAGGCT	60
MHLR	GTGGAATCCAATATTTTCGCGGT	64
CALHB	AGAAGTGAGACATGGCGGGTA	64
LUCHB	GCCTGTTGGAAATGTTGCTGAT	65
LSRSF	GCAGAATTGGGACACCATGCA	65
CZF	AAAAGAATTCCGCAAACACAAGAACTTG	58
CZR	AAAAGGATCCCAATTTGAATGAATGACAGT	58
ABF	GCNAAAYATHAARCARGARCC	-
LSZF	AAAAAAGCTTTTCCAGTGYGACAAATGCTC	50
LSZR	AAAAGGATCCTTGGTGGCATARTCACAGTC	50
OCZ	TGAATGACAGTATTTGGTGGCATAATCACA	68
NCZ	CAGCACATCTGTATTGATACACCGAGGA	68
AOL995	CCCTTTACGCGTTTTGTGCGACGAATTCTTTCCC	72
o1992	TTTACGCTTTTGTGCGACGAATTCTTTCCCTCTAGATCT AGACTC	
o1993	GAGTCTAGATCTAGA-NH ₂	
ABLS	GCNAAAYATYAAACAGGARCC	-
LSHZR	CAGCACATCTATATTGATAG	49
BFHB1	CATATACGCAAACACAAGAACTTG	62
BFHB2	TGTGTCAATAAGTCCATGTTGAAT	61
BFHB3	AGGCGGCCGCACAAAATCAGGT	75
BFHB4	CCCATCTACAATTCWCATGG	-

BFHB5	CAGGATTTGCACTTGTARTTC	-
CALHF	GTCACCTTCTCAATTGTCAGCTGCT	65
CALHR	TTCAGGACCACTAACAGCCTGA	65
CALFA	CTACTGGCTCCAGCAGTGCAGCT	73
CALFB	CAGCCAACTCCAGTCCTTCGGCTA	72
LHBCR	TGAGGTGATGTTGGAGCTGA	64
LHBCF1	GAGTTACCAGCCTTGACACT	58
LHBCF2	CTGCCTCTAGCAGTGCATCA	64
LHBCF3	TCAACTACAGCCAACTCAAGTC	60
NLHR	CTGCAAAGACATATTAAGTGTC	55
NLHF	GCCATGAATCTGAAATTAGC	58
NLHRA	CAATATTTGCACTCGTACATG	58
CCF1	TACTGTCATTTCATCAAAATTGCA	61
CCR1	TGCGTGGCGAGAGTTGAGCCAAGA	75
CCF2	CATTAACCTTAAACTAGTTGA	50
CCR2	TTACATACATTTAGTTGAGAC	48
BLOSR	TACCAGTTATTGTGTTCCACATAATTAGC	64
LSHBRT	AGTCATATTGGTATCCATG	51
LSHBRA	TGGCGGCAAACATATTGTTATACCAG	69
JALR	TTGCATGGTGTCCCAATTCTGCAT	71
LRPRO1	ATTATATCCACAGTGATGTCTTGT	58
LRPRO2	TGATGCTAGAATTGCAACTGAT	61
LRPRO3	TCTTTGAGCATTAACTGAGTATTTG	57
LRPRO4	AGCATTAACTGAGTATTTGTATCA	52
LRPRO5	TTCCTTAAGTCGCGCACTAT	50
LRPRO6	AAGTAAAGTGTAATTTTCAGCAT	52
LRPRO9	ACTCTTTATTCGCATCCCCCTCTT	56
LRPRO10	ACACAGGTAAGGTCAGGTAT	55
LRPRO11	CCCAATATGCACTAAAGCA	58
LPROF	ATACCTGACCTTACCTGTGTGCGA	57
LPROF2	ATTAGTATTTGTTAGAGAAATCAT	55
LPROF3	AAGGCATAAAGAATTTCTTGGA	59
LPROF4	ACATACACGCTTACAGTTG	53
LPROF5	ACACTTTTTAAGCCGTTAGA	55
LPROF6	AATTTCTTAAAGAAATAGTGC	51
CVPRO1	CCTCTAAACACACAGATTAGTTGA	55
CVPRO2	GAAATTCCTCTAAACACACAGATTA	57

CVPROF2	TATTATATCCTGTAAAGATCA	49
CVPROF3	AGAGATTTTCATCATACCTTAA	52
CVPROF4	CCAAATATTGAATACCACATGCCA	66
CVPROF5	TAAATCCATACAGAATCCGCA	62
CVPROF6	CGCAGCTTAAGTCCGTTCAT	64
CVPRO4	AAGGTTACATGTTATAACCAAC	54
CVPRO6	AAATCTGAAGGGTTGAAAACCTGT	62
CVPRO7	TAAATTAAATGTCGCCTCT	54
CVPRO8	TCGTTGAAGATTAAGCAAGA	58
BFHB7	ATCAGTTGCATTCTAGCATCA	56
M3R1	TAGGTCACCTTAAGTGAATCG	52
M3R2	AAGTGAATCGTTGTCATGAATTGT	57
MYPF	AAAAGTCGACGACATTTGTCACTTGGCCTCT	56
MYPR	AAAAGTCGACCAGAGAAGGCCCGTTGTGA	58
BFYR	AAAAGTCGACTGATGCTAGAATGCAACTGATG	63
LYHF	AAAAGTCGACGGAATGCGAATAAAGAGTTA	56
CYHF	AAAAGTCGACATTCAAGGCAAATAAACAG	54
CVF5SAL	AAAAGTCGACTAAATCCATACAGAATCCGCA	62
CV7SAL	AAAAGTCGACTAAATTAAATGTCGCCTCT	54
MABCDF	AAAAGAATTCATGGCGCAACCTCCGCCACCTCTG	68
MABCDR	AGATAAGCTTTTAATTGAAACAGTAGGCGAATTGA	64
BFBCDF	ACGAGAACAACATTTACTAGTGCT	56
LSBCDR	AATTTTCACTTGAGCTGTGCC	50
CVBCDR	ACTCTTGGCATTAGAAATTCCTC	56
LSBF	GCTGAATTGGAACAACATTTTTTACAA	59
CVBF	GCAGAACTGGAACAGCATTTTTTACAA	59
LSBR	CCAAGTGCTAGTTTTGTCTGAGAGTT	56
CVBR	CCAAGAGCAAGTTTAGCTGAGAGTT	56
NGF	GNGARMGNACNACATTTAC	-
GOR	GTCKGGRTTDATTTT	-
RTOTD	TCGCGCATAAAGATGT	46
OALR2	GCCAAATAGTGATTCCAATACGTC	60
OALR3	TAGTGATTCCAATACGTCCAG	62
OTD3R1	CAGCTGGACGTATTGGAATCACTA	65
OTD3R2	CACGTTATCCTGACATCTTTATGC	63
ORR1	CATATTTACATACTTGTCTGT	57
ORF1	ACTCGGCGAACTCGAACAAT	65

ORR2	CAAATCTAAGCATAATCCGAG T	58
ORR3	GGTAATGCTGGAATTCGAATC	62
ORF2	CAACTCGCATTTGCATCA	54
lacZ148	GTTGGGAAGGGCGATCGGTG	70

Appendix B Clustal W sequence alignments of *hb* from six *Musca* strains

Numbering refers to the cooper sequence (accession number Y13050).

B.1 Alignment of *hb* coding region amino acid sequences

Amino acids in zinc finger domains are italicised and indels are highlighted. Asterisks indicate conserved amino acids and dots indicate similar amino acid types.

```
cooper_      HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN 341
edinburgh_   HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN
rentokil_    HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN
rutgers_     HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN
white_       HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN
zurich_      HGKMKNHKCKSCGMVAITKMAFWEHARTHMKPEKILQCPKCPFVTELKHHLEYHIRKHKN
*****

cooper_      LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM 401
edinburgh_   LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM
rentokil_    LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM
rutgers_     LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM
white_       LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM
zurich_      LKPFQCDKCSYSCVNKSMNLNSHRKSHSSVYQYRCADCDYATKYCHSFKLHLRKYEHKPGM
*****

cooper_      VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPLQQQ 461
edinburgh_   VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPLQQQ
rentokil_    VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPLQQQ
rutgers_     VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPIQQQ
white_       VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPLQQQ
zurich_      VLDEEGIPNPSVVIDVYGTRRGPKNKSAANAALKKACSDLKIPPTSQLSAALQGFPLQQQ
*****

cooper_      QQPQPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP 521
edinburgh_   QQPQPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP
rentokil_    QQQ-QPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP
rutgers_     QQQQPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP
white_       QQ-QPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP
zurich_      QQHQPASPAPAKSSSSVASELPALTNLMSLQQNLAQQQQQQQSPGAQSHSSQQQINNLLP
**      *****

cooper_      PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGE- 580
edinburgh_   PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGEN
rentokil_    PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGEN
rutgers_     PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGEN
white_       PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGEN
zurich_      PLASLLQQNRNMAFFPYWNLNLQMLAAQQQAAVLAQLSPRMREQLQQQQQNKQANENGEN
*****

cooper_      ---EDEEDNDEVDEDEEEFDGKSVDSAMDLSQGTPTKEEQQTPELAMNKLKSEEHGETPL 637
edinburgh_   HGEDEEDNDEIDEDEEEFDGKSVDSAMDLSQGTPTKEDQQTPELAMNKLKSEEHGETPL
rentokil_    HGEDEEDNDEVDEDEEEFDGKSVDSAMDLSQGTPTKEEQQTPELAMNKLKSEEHGETPL
rutgers_     HGEDEEDNDEVDEDEEEFDGKSVDSAMDLSQGTPTKEEQQTPELAMNKLKSEEHGETPL
white_       HGEDEEDNDEVDEDEEEFDGKSVDSAMDLSQGTPTKEEQQTPELAMNKLKSEEHGETPL
zurich_      HGEDEEDNDEVDEDEEEFDGKSVDSAMDLSQGTPTKEEQQTPELAMNKLKSEEHGETPL
*****
```

```

cooper_      FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV 697
edinburgh_   FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV
rentokil_    FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV
rutgers_     FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV
white_       FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV
zurich_      FSSSAAARRKGRVLKLDQEKTAGHLQIASAPTSPOHHLHHNNEMPPTTSSPIHPSQVNGV
*****

```

```

cooper_      AAGAADHSSADESMETG--HHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS 755
edinburgh_   AAGAADHSSADESMETG--HHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS
rentokil_    AAGAGDHSSADESMETG--HHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS
rutgers_     AAGAADHSSADESMETGHHHHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS
white_       TAGAADHSSADESMETG--HHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS
zurich_      AAGAADHSSADESMETG--HHHHHHNPTTANTSASSTASSSGNSSNSSSTSTSSNSNSSS
.*** *****

```

```

cooper_      AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCGEKCDGPVGLFVHMARN 815
edinburgh_   AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCGEKCDGPVGLFVHMARN
rentokil_    AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCGEKCDGPVGLFVHMARN
rutgers_     AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCGEKCDGPVGLFVHMARN
white_       AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCGEKCDGPVGLFVHMARN
zurich_      AGNSPNTTMYECKYCDIFFKDAVLYTIHMGYHSCDDVFKCNMCG
*****

```

```

cooper_      AHS 818
edinburgh_   AH
rentokil_    AHS
rutgers_     AHS
white_       AH

```

B.2 Alignment of *hb* promoter and 5' UTR

Asterisks indicate conserved nucleotide position and indels are highlighted. Bcd-binding sites are underlined and the transcription start site is shown in bold font.

```

cooper_      TTTTTCAGCTTAATGGCAATATTAGGCTAAATCTCGGCGCATTTGATCCCTTTTTTTTAC 441
edinburgh_   TTTTTCAGCTTAATGGCAATATTATGCTAAATCTCGTTCGCATTTGATCCCTTTTTTTTAC
rentokil_    TTTTTCAGCTTAATGGCAATATTAGGCTAAATCTCGTTCGCATTTGATCCCTTTTTTTTAC
rutgers_     CCAGCTTAATGGCAATATTATGCTAAATCTCGGCGCATTTGATCCCTTTTTTTTAC
white_       TTTTTCAGCTTAATGGCAATATTATGCTAAATCTCGGCGCATTTGATCCCTTTTTTTTAC
zurich_      CCAGCTTAATGGCAATATTATGCTAAATCTCGCAGCATTTGATCCCTTTTTTTTAC
*****

```

```

cooper_      AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAC-A 500
edinburgh_   AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAC-A
rentokil_    AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAGGA
rutgers_     AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAC-A
white_       AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAGGA
zurich_      AAAGTCATTTAATCCATTTCTTAATTCCGTTTCATAAAATCCCCGAGGCGAGTGTGTAC-A
*****

```



```

cooper_      AAAAAATATTGAAAAAAAAAAAAATTAACCGGATTATCAAAAAAA-TATAACTTCCAAGAAG 1357
edinburgh_   AAAAA-TATTGAAAAA---AAATTAACCGGATTATCAAAAAAA-TATAACTTCCAAGAAG
rentokil_    AAAAAATATTGAAAAAAGAAAAATTTTACCGGATTATCAAAAAAA-TATAACTTCCAAGAAG
rutgers_     AAAAA-----ATTAACCGGATTATCAAAAAAA-TATAACTTCCAAGAAG
white_       AAAAAATATTGAAAAAGAAAAATTTAA-CGGATTATCAAAAAAA-TATAACTTCCAAGAAG
zurich_      AAAAAATATTGAAAAAGAAAAATTTAA-CGGATTATCAAAAAAA-TATAACTTCCAAGAAG
*****

```

```

cooper_      TTTTATTAAAAAGAAAAAAACAATTTTTTTTATTTTTTTATCAATCAATTATTTTTTACAA 1417
edinburgh_   TTTTATTAAAAAAATA---CAATTTTTTA-----A-AAATTTATTTTTTACAA
rentokil_    TTTTATTAAAAAAACAATTTTTTA-----A-AAATTTATTTTTTACAA
rutgers_     TTTTATTAAAAAAACA-----TTTTTTA-----A-AAATTTATTTTTTACAA
white_       TTTTATTTTAAA-AAAAAAACAATTTTTTA-----AAGCAATTATTTTTTATAA
zurich_      TTTTATTAAAAAAACA-----CAATTTTTTA-----A-AAATTTATTTTTTACAA
*****

```

```

cooper_      AAAA---TTAACAAAAAGTGACAACGAAATTGTTTAAACAAAAACACCGCGAAATATTGGA 1473
edinburgh_   AAAAAAATTAAACAAAAAGTGACAACGAAATTGTTTAAACAAAAACACCGCGAAATATTGGA
rentokil_    AAAAAA-TTAACAAAAAGTGACAACGAAATTGTTTAAACAAAAACACCGCGAAATATTGGA
rutgers_     AAAAAA-TTAACAAAAAGTGACAACGAAATTGTTTAAACAAAAACACCGCGAAATATTGGA
white_       AAAAAA-TTAACAAAAAGTGACAACGAAATTGTTTAAAGAAAAACACCGCGAAATATTGGA
zurich_      AAAAAAATTGACAAAAAGTGACAACGAAATTATTTAAACAAAAACACCGCGAAATATTGGA
****

```

```

cooper_      TTCCAC 1479
edinburgh_   TTCCAC
rentokil_    TTCCAC
rutgers_     TTCCAC
white_       TTCCAC
zurich_      TTCCAC
*****

```

	1	2	3	4
1	348.2	417.4	343.6	335.5
2	30.4	70.3	33.2	181.8
3	309.6	334.6	338.1	341.5
4	306.2	421.1	348.4	216.4
5	113.1	321.5	139.4	219.5
6	82.8	306.3	301.5	221.2
7	15.8	102.4	77.0	5.4
8	10.0	42.0	33.2	218.0

Table C1 Raw data for rows 2 to 4 of table 6.1. See figure 6.4A. The values marked with an asterisk did not grow.

Appendix C β -galactosidase assay results

The results in each table are of assays that were carried out at the same time. β -galactosidase units are defined as the amount which hydrolyses 1 μ M of ONPG per min per cell. The yeast cultures assayed at each concentration are numbered and the standard deviation (stdev) and the average β -galactosidase activity are given. The results of negative controls (-ve) for assays where no Bcd was present are also given.

			β-estradiol concentration (nM)			
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	10	1000
pDB1.2	pDBhb.19	1	290.2	1267.8	1791.8	1213.5
		2	412.1	1570.2	2498.9	1689.3
		3*	-	-	-	-
		stdev	86.2	213.8	500.0	336.4
		average	351.2	1419.0	2145.4	1451.4
		-ve	3.2	2.3	0	1.7
pBCMBCD		1	288.9	813.4	1436.6	876.5
		2	271.8	773	1550.1	815.4
		3	248.2	917.4	1153.6	535.5
		stdev	20.4	74.5	204.2	181.8
		average	269.6	834.6	1380.1	742.5
pMABCD		1	106.2	421.1	548.4	216.4
		2	113.1	321.5	637.4	210.5
		3	82.9	526.3	701.8	221.2
		stdev	15.8	102.4	77.0	5.4
		average	100.7	423.0	629.2	216.0

Table C1 Raw data for rows 2 to 4 of table 6.1. See figure 6.4A. The cultures marked with an asterisk did not grow.

			β-estradiol concentration (nM)			
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	10	1000
pDB1.2	pMhbP2+	1	500.5	933.5	852.4	1980
		2	235.4	450.9	1143.3	1617.8
		3	427.6	681.7	1465	1896.3
		stdev	137.0	241.4	306.6	189.6
		average	387.8	688.7	1150.6	1831.4
pBC103		-ve	0	1.8	1.6	3.6
pBCMBCD		1	220.7	458.4	1135.3	1961.3
		2	277.7	609.3	971.0	1892.1
		3	364.1	492.4	962.2	2198.3
		stdev	72.2	79.2	97.5	160.6
		average	287.5	520.0	1022.8	2017.2

Table C2 Raw data for rows 5 and 6 of table 6.1. See figure 6.4B

			β -estradiol concentration (nM)			
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	10	1000
pDB1.2	pDBhb.19	1	297.6	843.6	1397.8	2259.0
		2	382.6	1045.0	1802.2	2171.5
		3	297.1	759.9	1480.6	1990.9
		stdev	49.2	146.5	213.6	136.7
		average	325.8	882.8	1560.2	2140.4
pBC103		-ve	0	3.2	2	4.1
pDB1.2	pMhbP2+	1	137.3	223.1	852.4	1867.9
		2	327.2	759.8	1134.3	1610.2
		3	243.8	632.4	1456.0	1703.5
		stdev	95.2	280.4	302.0	130.5
		average	236.1	538.4	1147.6	1727.2
pBC103		-ve	5.4	5.5	25.5	8.6
pBCMBCD	pDBhb.19	1	157.4	183.4	720.1	1626.5
		2	106	340.9	605.5	1949.2
		3	191.8	571.4	866.9	2148
		stdev	46.5	125.5	382.5	256.4
		average	169.6	370.6	1102.9	2647.6
	pMhbP2+	1	126.2	278.5	755.4	2591.0
		2	164.0	319.8	1040.6	2424.2
		3	218.7	513.5	1512.7	2927.5
		stdev	43.2	195.1	131.0	263.2
		average	151.7	365.2	730.8	1907.9

Table C3 Raw data for rows 7 to 10 of table 6.1. See figure 6.5.

		β -estradiol concentration (nM)			
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	1000
pDB1.2	pLShb+	1	1.1	2.3	2.7
		2	0.6	2.2	3.7
		3	2.4	22.6	40.7
		stdev	0.9	11.7	21.7
		average	1.4	9.0	15.7
pBC103		-ve	0.6	0.8	3.5
pDB1.2	pLShb-	1	2.9	13.4	1.1
		2	3.3	15.3	1.5
		3	3.3	15.6	3.9
		stdev	0.2	1.2	1.5
		average	3.2	14.8	2.2
pBC103		-ve	0.9	0.7	2.9
pBCMBCD	pLShb+	1	2.1	1.4	4.2
		2	1.0	2.5	5.0
		3	0	1.5	3.0
		stdev	1.1	0.6	1.0
		average	1.0	1.8	4.1
	pLShb-	1	2.4	12.9	7.5
		2	6.4	11.5	11.2
		3	2.3	13.5	9.2
		stdev	2.3	1.0	1.9
		average	3.7	12.6	9.3

Table C4 Raw data for rows 11 to 14 of table 6.1

		β -estradiol concentration (nM)			
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	1000
pDB1.2	pCVhb+	1	17.0	31.6	6.3
		2	10.5	20.1	5.0
		3	11.6	19.9	6.6
		stdev	3.5	6.7	0.9
		average	13.0	23.9	6.0
pBC103		-ve	0	3.4	0
pDB1.2	pCVhb-	1	15.5	24.0	10.2
		2	5.1	18.7	5.8
		3	8.3	21.4	6.4
		stdev	5.3	2.7	2.4
		average	9.6	21.3	7.5
pBC103		-ve	0	18.1	0
pBCMBCD	pCVhb+	1	7.0	30.0	7.2
		2	5.3	25.1	9.2
		3	6.4	36.5	11.2
		stdev	0.9	5.7	2
		average	6.2	30.5	9.2
	pCVhb-	1	8.2	26.0	10.0
		2	5.7	33.4	10.0
		3	3.2	31.6	10.9
		stdev	2.5	3.9	0.5
		average	5.7	30.3	10.3

Table C5 Raw data for rows 15 to 18 of table 6.1

			β -estradiol concentration (nM)		
<i>bcd</i> plasmid	<i>hb</i> plasmid		0	2.5	1000
pDB1.2	pCVhb5+	1	298.5	1488.4	1687.5
		2	310.6	1300.0	1475.0
		3	355.7	1725.6	2007.6
		stdev	30.1	213.3	268.1
		average	321.6	1504.7	1723.4
pBC103		-ve	0	0	1.1
pDB1.2	pCVhb5-	1	340.0	1256.4	1441.0
		2	371.2	1451.0	1539.3
		3	333.3	1501.0	1444.9
		stdev	20.2	129.2	55.7
		average	348.2	1402.8	1475.1
pBC103		-ve	0	0.5	0

Table C6 Raw data for rows 19 and 20 of table 6.1