Evolution and development of insect wings: A comparative analysis of the genome wide targets of the Hox protein Ultrabithorax in *Bombyx mori*, *Apis mellifera* and *Drosophila melanogaster*.

A Thesis

submitted in partial fulfillment of the requirements

of the degree of

Doctor of Philosophy

by

SHREEHARSHA T T

20083007

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH, PUNE

2014

to my father

...who nurtured the curiosity in me...

CERTIFICATE

Certified that the work incorporated in the thesis entitled **"Evolution and development of insect wings: A comparative analysis of the genome wide targets of the Hox protein Ultrabithorax in *Bombyx mori*, *Apis mellifera* and *Drosophila melanogaster*."** Submitted by Mr. Shreeharsha T T was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.

Prof. L S Shashidhara
Supervisor

Date: 11th July 2014

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Shreeharsha T T

20083007

Date: 11th July 2014

ACKNOWLEDGEMENTS

I would like to thank my advisor Shashi (Prof L S Shashidhara) for the untiring support and guidance he extended to me all through my PhD and for believing in me even in times when I myself could not. He has been an inspiring guide throughout the walk of my PhD. I thank the CSIR for fellowship, IISER and CCMB for the excellent facility and ambience which made learning science fun and work very smooth.

I would like to thank my Research Assessment Committee members especially Suren (Prof.Surendra Ghaskadbi) for being critical of my work and at the same time being a very encouraging mentor. I thank Girish (Dr. Ratnaparkhi) for his critical comments and help whenever I was stuck with Protein or fly work. I also extend my sincere gratitude to Sanjeev (Prof Galande) and Nagraj (Dr. Balasubramaniam) for their helpful discussions and insights. I thank Farhat (Dr.Habib) for his constant guidance and support when I forayed into computational analysis, without his help this work would have not been possible. I would like to thanks Aurnab (Dr.Ghose) and Girish (Dr.Deshpande) for discussions and guidance which helped complete my PhD and move ahead.

I would like to thank all my labmates at IISER, Pune, with whom I started this journey and saw the lab take shape as we started our path into PhD. I would specially mention Abhishek, Ameya, Aniruddh, Anurag, Mithila, Payal, Reshma, Rini, Senthil, Shraddha and Sneha for the refreshing ambience they created for both science and life in the lab. I would like to specially thank Payal for the support, discussion and cheer while I wrote this thesis. I am grateful to Ranveer, Sourabh and Mouli from Galande lab for discussions while doing the ChIP standardizations. I thank Kanika and Vimal for all the awesome discussions on science in general and in particular, for helping me learn new fascinating aspects of physics. I have fond memories of my life in hostel due to all my friends who stayed with me and kept me in good spirits. I thank all of them including Shomu, Amar and Murthy for the radiance they always brought in the hostel life.

I thank Arun N and Boominathan for helping me with computational aspects with their intriguing discussions and helping write little codes to sort out nagging issues. Without their help it would have been a difficult and time consuming endeavor to do all the computational analysis.

I would like to also thank all my CCMB batchmates for their extensive support and discussions. Particularly the group of Abhishek, Hardik, Pandit and Manish from whom I have learnt a lot while having loads of fun. I am grateful to Virender and Poornima from Dr.Madhusudan Rao s lab for all their help and encouragement. I thank all the scientists who taught me during my coursework

at CCMB, it was a indeed a wonderful learning experience. I thank the students and the staff (especially Dr Sundaram) at the proteomics facility at CCMB which helped me immensely in initial days of my proteomics work. I also thank the lab management, fine chemicals, the animal facility staff, Fly media staff and the support staff both at CCMB and IISER whose support helped move work smoothly. I thank Nanaji and Kelkar sir from admin who took personal care of us for paper work when we were starting at a brand new institute. I am thankful to Tushar and Mahesh Rote for the help they extended for the administrative work at various times.

I am grateful to my lab first at CCMB, Particularly Ramesh Y for the mentoring during the initial times of my PhD and his encouragement there on. I like to thank my seniors Prashant, Usha, Sudha and Pavan for their help and guidance, by teaching in the lab and later for their advice. I extend my gratitude to the seniors Mohit, Ruchi, Guru and Pallavi who returned at various points with encouraging words. I thank my labmates in IISER, Naveen, Savita and JP who helped manage the lab, supported me with my work and made working fun. I thank the Postdocs in our lab Arkaja, Anu and Shital for their guidance in times of need. I thank all the summer interns who made the place lively and pepped up with intriguing questions particularly Purba, Dhanashree and Navaneet.

I thank Fujiwara sensei (Prof. Haruhiko Fujiwara) at the University of Tokyo for his encouragement, support and advice during my stay in their laboratory. I thank all his lab members for making my first venture outside India a wonderful and a memorable one. I specially thank Ando san, Yamaguchi san, Fujii san and Miyagi san for their patient support and help during my stay. I also thank Shashank san, and all the Kashiwa International lodge members, for the bonhomie during my stay. I thank the APDBN, JSDB and the Development fellowships for their funding for two of my very resourceful visits to Japan. I thank DBT for partially supporting my travel to Lepidoptera and Butterfly meetings in Europe. I thank Mita sensei (Prof Kazuei Mita) and Suetsugu sensei from NIAS, Tsukuba for their extensive help in understanding the Silkworm genome and their incredible support whenever I requested.

In Mysore I would like to thank the Department of Zoology, Shyamala madam for her constant support, words of encouragement and advice. I thank the then head of the department Dr Ramesh for his support in tough times and encouragement. Venkatesh, for helping me do my work in the lab and for the hospitality. Other students at the department for making me feel at home, Maithri, Shruti, Ranjita, Srinivas, Chaitra and Vineet. I specially thank Joseph for his support during all my stays and for the intriguing discussions. I would also thank the Sericulture Department for their initial support, the then Head Dr. Seetaramaiah, Dr.Jagadish and all the support staff who taught me the traditional methods of maintaining silkworms.

I also thank the CSR&TI, Mysore which helped me maintain silkworms and where all the people welcomed me always with a warm smile and they helped me in all possible ways in this PhD. I thank Dr Sharmila for her untiring support and help without which this PhD would have been impossible. I also thank Dr Kalpana, Dr Ashwath and Dr Nirmal Kumar for their help and support.

I thank all the people who helped and guided me through initial silkworm resource hunting days, including the Central Silk Board, SBRL, Bangalore, Silkworm germplasm repository in Hosur, Silkboard at Hindupur. I received a lot of support from the Lab of Dr Nagaraju at CDFD for their guidance and direction when I first ventured into an unfamiliar world of silkworms.

I thank Genotypic and Abexome staff for their scientific discussions while they extended their services in my work.

I thank Dr Upendra Nongthomba lab and his lab members at MRDG, IISc for the support they extended for my dissections. I specially thank Mohan and Herojeet for their unceasing help.

My acknowledgements would remain incomplete if I did not thank the KVPY fellowship which supported me during my undergraduate days and inspired me through the exposure it gave me to pursue and cherish understanding science. I also received a continued support from many of my summer internship mentors in my PhD days. I am deeply indebted to Dr Gayatri Ramakrishna, CDFD, Dr DP Kasbekar, CCMB and Dr Rene Borges, CES, IISc, for their encouragement and guidance whenever I crossed their paths, sometimes intentionally.

In the end I thank the most important pillar which supported me during trying times in research, and always with a smile, my Family. I am really grateful to my mother for encouraging me to pursue my work with utmost sincerity and never expecting me to call or be home for any necessities. It helped me keep relaxed without which I may have tumbled in troubled times. I also thank my brother for the encouragement, support and discussions. I thank my sister-in-law and my little nephew for bringing cheer in my life.

Shreeharsha

February 2015

Contents

Chapter 1

Introduction

1.1: Body plan formation/ segmentation and patterning 37

1.2: Hox genes and pattern formation 39

1.3: Body form diversity and evolution 41

1.4: Evolution of diverse body plans vis-à-vis Hox genes 43

1.5: Origin and evolution of wings in Insects 44

1.6: Hox gene Ubx and the specification of the third thoracic segment 46

1.7: Mechanism of Ubx function in *Drosophila* 47

1.8: Wing development in *Drosophila* 48

1.9: Development of wing in Lepidoptera 50

1.10: Expression pattern of the Hox protein Ultrabithorax 51

1.11: Silkworm as a Lepidopteran model 53

Objectives 55

Chapter 2

Identification of the targets of *Bombyx* Ubx by Chromatin immunoprecipitation & sequencing

Introduction

2.1: Chromatin Immunoprecipitation 70

2.2: Array hybridization to identify ChIP enriched regions 70

2.3: High throughput sequencing methods to identify ChIP enriched fragments 72

2.4: Sequencing with Illumina Genome Analyzer 73

2.5: Requirements for ChIP sequencing 74

Materials & Methods

2.6: Silkworm race used for this study 77

2.7: Comparison of sequences from silkworm races to that of Genome database 78

2.8: Identifying silkworm wing buds and appropriate larval stage for ChIP 79

2.9: Generation and validation of Antibodies 80

2.10: Chomatin Immunoprecipitation 87

Results & Discussion

2.11: Silkworm race used for the study 93

2.12: Identifying silkworm wing buds and appropriate larval stage for ChIP 94

2.13: Generation of Antibodies 94

2.14: Chromatin Immuno-precipitation 97

Summary 100

Chapter 3

Analysis of DNA sequences enriched in ChIP

Introduction

3.1: Alignment to the genome 117

3.2: Identification of Peaks 118

3.3: Tools used in ChIP-seq analysis workflow 119

3.4: Silkworm Genome databases 122

Materials & Methods

3.5: Quality control analysis of ChIP-seq reads 125

3.6: Creating index of the genome with Bowtie 127

3.7: Alignment of the ChIP-seq reads to the genome 127

3.8: Conversion of file format and indexing 128

3.9: Visualization with IGV viewer 129

3.10: Peak calling using MACS 129

3.11: Identification of genes associated with the peaks 130

3.12: Annotation of genes associated with the peaks 133

Results & Discussion

3.13: FastQC quality control analysis of reads 134

3.14: Creating index of the genome with Bowtie 135

3.15: Alignment of the ChIP-seq reads to the genome 136

3.16: Visualization with IGV viewer 136

3.17: Peak calling with MACS 136

3.18: Identification of genes associated with the peaks 138

3.19: Annotation of genes and identification of homologs 138

Summary 140

Chapter 4

Comparison of targets of Ubx from *Bombyx*, *Drosophila* and *Apis*

Introduction

4.1: Comparative genomics 155

4.2: Direct targets of Ubx from *Drosophila* and *Apis* 155

4.3: Gene Ontology databases 156

4.4: Visualization through BioVenn and Circos 158

Materials & Methods

4.5: Comparative analysis of targets of Ubx in insects 159

4.6: Gene Ontology analysis of target sets 160

4.7: Comparative Gene Ontology analysis 160

Results & Discussion

4.8: Comparative analysis of targets of Ubx in different insects 162

4.9: Gene Ontology analysis of target sets 164

4.10: Comparative Gene Ontology analysis 164

Summary 166

Chapter 5

Trascriptome analysis of *Bombyx* wing buds

Introduction

5.1: RNA-Sequencing 198

5.2: RNA-Seq Analysis 199

5.3: Gene expression studies in haltere of *Drosophila* 201

Materials & Methods

5.4: Isolation of wing buds for RNA preparation 203

5.5: Analysis of the RNA-Seq reads 203

5.6: Correlating ChIP targets of Ubx and gene expression in *Bombyx* HW buds 207

Results & Discussion

5.8: Isolation of wing buds for RNA preparation 208

5.9: Analysis of the RNA-Seq reads 208

5.10: Comparison of ChIP-Seq data and transcriptome data 211

5.11: Differential expression of targets of Ubx in *Bombyx* and *Drosophila* 211

5.12: Future direction: Analysis of Ubx binding regions 212

5.13: Comparative analysis of targets of Ubx and differentially expressed genes between fore and hind wing appendages. 213

5.14: Mechanims of Hox regulation: A discussion on an evo-devo perspective of Ubx regulation in Insects. 217

Summary 220

Future directions 239

Appendices

Appendix Chapter 2 243

Appendix Chapter 3 246

Bibliography 253

LIST OF FIGURES AND TABLES

A LIST OF FIGURES

Fig	Chapter 1	Page
1.1	Homeobox genes and body patterning	57
1.2	Divergence of Insects	58
1.3	Comparison of Hox gene clusters in different insects	59
1.4	Variation in the wing appendages across different insect orders	60
1.5	Role of Ultrabithorax in the specification of segmental identity	61
1.6	Wing development in *Drosophila*	62
1.7	Morphogen gradients in wing development	63
1.8	Expression patterns of Ubx in Embryos of different insects	64
1.9	Expression patterns of Ubx in wing primordia of different insects	66
1.10	Life cycle of the silkmoth *Bombyx mori*.	67
	Chapter 2	
2.1	An overview of the Chromatin Immunoprecipitation (ChIP) method.	102
2.2	An overview of ChIP-chip methodology	103
2.3	Illumina® high throughput sequencing chemistry	104
2.4	Maintenance of silkworm	106
2.5	Comparison of sequences from locally available races against genome database	107
2.6	Identification of wing bud and larval stages in *Bombyx*	108
2.7	Expression and purification of *Bombyx* N terminal Ubx Protein	109
2.8	Western blot to show specificity of antibodies to *Bombyx* N terminal Ubx	110
2.9	Detection of purified and endogenous *Bombyx* Ubx protein byWestern blot	111
2.10	Immuno Histo-chemistry (IHC) with *Bombyx* wing buds	112

2.11	Comparison of Ubx expression in the hind and fore wing appendages of insects	113
2.12	Standardizations for ChIP in *Bombyx* wing buds	114
2.13	Sonication standardizations for ChIP	115
	Chapter 3	
3.1	MACS flow chart explaining the method it uses to identify peaks	143
3.2	Silkworm genome databases.	144
3.3	Visualization with IGV viewer and assignment of genes to peaks	145
3.4	Number of Peaks in replicate 1 and 2 of hind wing	146
3.5	Peaks common between two IgG filtered replicates	147
3.6	FastQC quality report of Hind wing Input dataset 1	148
3.7	FastQC quality report of Hind wing Bm Ubx ChIP datset 1	149
3.8	FastQC quality report of Hind wing IgG negative control dataset 1	150
3.9	FastQC quality report of Hind wing Input dataset 2	151
3.10	FastQC quality report of Hind wing Bm Ubx ChIP datset 2	152
3.11	FastQC quality report of Hind wing IgG negative control dataset 2	153
	Chapter 4	
4.1	Venn diagram:extent of overlap between the Ubx targets in FW and HW	169
4.2	Gene Ontology analyses of the Ubx targets of the hind wings in *Bombyx.*	170
4.3	A comparative graph of the GO analysis between Ubx targets of insects	171
4.4	Venn diagram comparison of Ubx targets between *Bombyx* and *Drosophila* (R)	172
4.5	Comparative GO analysis *Bombyx* and *Drosophila* (R) : Wing development terms	173
4.6	Comparative GO analysis *Bombyx* and *Drosophila* (R) : Imaginal disc specific terms	174
4.7	Venn diagram comparison of Ubx targets between *Bombyx* and *Drosophila* (P)	175
4.8	Comparative GO analysis *Bombyx* and *Drosophila* (P) : Wing development terms	176
4.9	Comparative GO analysis *Bombyx* and *Drosophila* (P) : Imaginal disc specific terms	176

4.10	Venn diagram comparison of Ubx targets between *Bombyx* and *Apis*	177
4.11	Comparative GO analysis *Bombyx* and *Apis* : Wing development terms	178
4.12	Comparative GO analysis *Bombyx* and *Apis* : Imaginal disc specific terms	178
4.13	Comparative wing development GO term for common targets	179
4.14	Venn diagram comparison of Ubx targets between *Apis* and *Drosophila* (R)	180
4.15	Venn diagram comparison of Ubx targets between the three insects.	181
4.16	Circos plot comparison of Ubx targets between the three insects.	182
4.17	A comparative GO analysis of pathways of Ubx targets the three insects.	183
4.18	A comparative GO analysis of wing development pathways between Bm and Dm.	184
4.19	A comparative GO analysis of wing development pathways between Bm and Am.	185
4.20	The phylogenetic tree of evolutionary relationships of insect wings.	186
	Chapter 5	
5.1	RNA sequencing methodology	224
5.2	Overview of the programs in the Tuxedo suite	225
5.3	RNA yield QC from *Bombyx* fore- (FW) and hind (HW) wing buds	226
5.4	Quality control analysis of the hindwing bud RNA-Seq reads using FastQC.	227
5.5	Quality control analysis of the forewing bud RNA-Seq reads using FastQC.	228
5.6	Visualization of the transcriptome data.	230
5.7	Comparison of number of genes expressed in FW and HW in *Bombyx*	233
5.8	Comparison of genes from ChIP seq and RNA seq data	233
5.9	Comparison of genes from ChIP seq and *Drosophila* microarray data	235

B	LIST OF TABLES	
Table	**Chapter 3**	
3.1	Alignment statistics of the ChIP-seq reads to the *Bombyx* genome	141
	Chapter 4	
4.1	Targets common between *Bombyx* hindwing and *Drosophila* (R)	187
4.2	Targets common between *Bombyx* hindwing and *Drosophila* (P)	190
4.3	Targets common between *Bombyx* hindwing and *Apis* hindwing	192
4.4	Targets common between *Bombyx* hindwing, *Drosophila* (R) and *Apis*	195
	Chapter 5	
5.1	Alignment statistics forRNA seq reads from FW and HW datasets	229
5.2	Differentially expressed genes between fore- and hindwings of *Bombyx*	231
5.3	Direct targets of Ubx that are differential expressed between FW and HW	234
5.4	Direct targets of Ubx in Bm HW and differentially expressed in *Drosophila* (M)	236
5.5	Direct targets of Ubx in Bm HW and differentially expressed in *Drosophila* (A)	237

ABBREVIATIONS USED

APF	-	After Puparium Formation
ach	-	achaete
Amp	-	Ampicillin
ap	-	apterous
A-P	-	Anterior- Posterior
bcd	-	bicoid
BGI	-	Beijing Genomics Institute
BLAST	-	Basic Local Alignment Search Tool
Bm	-	*Bombyx mori*
bp	-	base-pairs
BSA	-	Bovine Serum Albumin
CBB	-	Coomassie Brilliant Blue dye
CCD	-	Charge Coupled Device camera
ChIP	-	Chromatin Immunoprecipitation
CPU	-	Central Processing Unit (of a computer)
CRE	-	Cis-regulatory elements
ci	-	cubitus interruptus
ct	-	Cut
DAVID	-	Database for Annotation, Visualization, and Integrated Discovery
DGRC	-	Drosophila Genomics Resource Center
Dl	-	Delta
Dll	-	Distalless
dpp	-	decapentaplegic
D-V	-	Dorsal- ventral
EDTA	-	Ethylene Diamine Tetra acetic Acid
EGFR	-	Epidermal growth factor receptor
en	-	engrailed

Exd	-	Extradenticle
FPKM	-	Fragments Per Kilobase of exons per million fragments Mapped
FTP	-	File Transfer Protocol
FW	-	Fore Wing
Gb	-	Giga bases
hh	-	Hedgehog
hth	-	homothorax
Hox	-	Homeobox
HW	-	Hind Wing
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
kn	-	knot
Kr	-	Kruppel
MQ	-	Milli Q
NCBI	-	National Center for Biotechnology Information, USA
NMWL	-	Nominal Molecular Weight Limit
nub	-	nubbin
omb	-	optiomoter blind
PAGE	-	Poly Acrylamide Gel Electrophoresis
PBS	-	Phosphate Buffered Saline
PBST	-	Phosphate Buffered Saline (with Tween- 20)
PBTx	-	Phosphate Buffered saline with Triton X-100
PBTx	-	Phosphate Buffered Saline (with Triton X-100)
ptc	-	Patched
PVDF	-	Poly Vinyl Di Fluoride
RIPA	-	Radio Immuno Precipitation Assay
RT	-	Room Temperature
RT-qPCR	-	Real Time- quantitative Polymerase Chain Reaction
sal	-	spalt
SDS	-	Sodium Dodecyl Sulphate
Ser	-	Serrate
SilkDB	-	Silkworm DataBase
TAE	-	Tris acetate EDTA

TBS	-	Tris Buffered Saline
TBST	-	Tris Buffered Saline with Tween20
tsv	-	Tab Separated Value file
Ubx	-	Ultrabithorax
Utx	-	Ultrathorax
vg	-	vestigial
vn	-	vein
vol	-	Volume
wg	-	wingless

Synopsis

Title, Evolution and development of insect wings, A comparative analysis of the genome wide targets of the Hox protein Ultrabithorax in *Bombyx mori, Apis mellifera* and *Drosophila melanogaster.*

Name of the Student,	**SHREEHARSHA T T**
Roll number,	20083007
Name of the thesis advisor,	Prof L S Shashidhara
Date of Registration,	11th July 2007

Indian Institute of Science Education and Research (IISER), Pune, India

Introduction

Insects are the first animals to have acquired flight during evolution. Amongst all the animals, they belong to the order with the largest number and diversity of species (Mora et al, 2011). A part of this plethora of body forms is also evident in their flight appendages. Most insects have four wings, while beetles and flies have only one pair of wings. In beetles, the forewings are modified as thick protective cover called elytra and only hind-wings perform the flight function. In contrast, only the forewings perform the flight function in flies, while the hind-wings are modified to a small club-shaped balancing organ called haltere. In addition, except in few early insect groups, they all show differences between forewing and hindwing morphology.

The insect body is divided into segments in which the development and fate of different organs is mainly controlled by a set of master control genes of the Hox complex. The Hox genes are highly conserved across the animal kingdom and are the main players in generating morphological diversity along with body axis within an organism. Hox genes are homeodomain-containing transcription factors, which function by regulating the expression of downstream target genes. Thus, the mechanism of organ specification and body plan development,

which allows a variety and range of modifications, is well conserved in the animal kingdom (Carroll and Grenier, 2005).

Suppression of wing fate and specification of haltere fate in *Drosophila melanogaster* by the Hox gene *Ultrabithorax* (*Ubx*) is a classic example for Hox regulation of serial homology, which has served as a paradigm for understanding Hox gene function (Lewis, 1978). The differential development of wing and haltere constitutes a good genetic system to study cell fate determination at different levels such as growth, cell shape, size and its biochemical and physiological properties. They also represent the evolutionary trend that has established the differences between fore and hindwings in insects. Ubx, which is required to specify haltere development in *Drosophila*, is expressed in T3 segments during development of all insects studied so far. Furthermore, the Ubx protein itself has not evolved amongst the diverse insect groups, although there are significant differences in Ubx sequences between *Drosophila* and crustacean Arthropods (Galant and Carroll, 2002; Ronshaugen et al., 2002). Interestingly, over-expression of Ubx from these organisms in T2 segment lead to wing to haltere transformations in *Drosophila* (Grenier and Carroll, 2000; Kanhale D and Shashidhara L S, unpublished results). This suggests that in the Dipteran lineage, certain wing patterning genes may have come under the regulation of Ubx.

Flight appendages in *Bombyx, Drosophila* and *Apis*

In *Bombyx mori* (Lepidoptera), the forewings are appendages of the thoracic segment T2 and the hindwing emerge from the thoracic segment T3. The morphology of both these wings is similar, with some changes in the overall shape. Both the fore- and hindwings are flat structures without much patterning differences. Whereas in *Drosophila* (Diptera), the thoracic segment T2 gives rise to a pair of wings while the T3 gives rise to a pair of globular structures called halteres. In the *Apis mellifera* (Hymenoptera), the two pairs of wings are morphologically identical, except that the hindwing is smaller than the forewing.

The wing discs/buds of *Bombyx* develop as flat bilayered epithelial buds that resemble miniature adult wings. The wing buds are small in the first four instars and in the fifth and last instar they grow rapidly. The wing venations are clearly visible from the late fourth instar onwards. This bud like mode of wing development is an ancestral mode also common to Hymenoptera (Macdonald et al. 2010). In case of *Drosophila,* the wings develop much differently than the pattern described above, in that the axis is specified in an epithelial monolayer (Fristrom1993; Pastor-Pareja et al. 2004) and during pupation this layer undergoes an eversion, folding back on itself to form a bilayer, which extends away from the body axis to form the adult wing.

The primary goal of this study is to identify the direct targets of Ubx in the wing buds of the Lepidopteran model organism *Bombyx* and then to use a comparative analysis to understand the insect appendage diversity. A comparative study is aimed at understanding developmental and molecular events downstream of Ubx that causes differences in wing morphology amongst insect orders by comparing direct targets of Ubx in *Bombyx* (silkworm, Lepidoptera), *Apis* (Honey bee, Hymenoptera) and *Drosophila* (fruitfly, Diptera). During evolution, Hymenopterans diverged from other insects more than 300 million years ago, while Lepidoptera and Diptera have diverged almost 250 million years ago. Thus, this study would be tracing the evolution of function of Ubx over the past 300 million years. Genome-wide direct targets of Ubx in *Apis* (Prasad N, 2013) and *Drosophila* (Agrawal et al. 2011) have been identified earlier in our laboratory.

Objectives

The first step in an attempt to understand the role of Ubx is to identify the direct binding regions and their target genes that are thereby regulated by Ubx in the hindwing buds of *Bombyx*. The next step involves a detailed comparative analysis of targets of Ubx in different insect orders in order to understand the kind of genes that are conserved and the ones that have diverged. The role these genes play in development across these insect orders will allow us to decipher

the mechanisms that may be involved in the specification of haltere in Diptera. Based on these objectives, following specific aims were defined for this project,

1. Identify direct binding regions of Ubx in the developing hind wing of *Bombyx*.

2. Identify genes associated with regions bound by Ubx in *Bombyx* and compare these target genes to that of haltere in *Drosophila* and hindwing of *Apis*.

3. To carry out a Gene Ontology (GO)-based functional analysis on the targets of *Bombyx* in comparison to targets of Ubx in *Drosophila* and *Apis*

4. To find out if any of the target genes of Ubx in *Bombyx* are differentially expressed between fore- and hindwings and compare them with the genes differentially expressed between wing and haltere in *Drosophila* (from published microarray data, Mohit Prasad et al. 2006 and Pavlopoulos and Akam 2011).

Results and Discussion

1. Identification of the direct binding regions of Ubx in the developing hind wing buds of *Bombyx mori*.

Expression patterns of few developmental genes that regulate wing disc development in *Drosophila* have been studied in Lepidoptera, mostly through butterfly as a model system. Expression of some of the developmental markers is also known through a preliminary study on *Bombyx* wing buds (Singh et al., 2001). Based on these studies and directly observing the morphology of wing buds in *Bombyx*, we decided to use the late fourth instar of the *Bombyx* larva, as an equivalent of late third instar larval wing imaginal disc in *Drosophila* for Chromatin Immunoprecipitation (ChIP).

As the approach used for identification of the direct targets was a sequencing based method, we relied heavily on the silkworm genome information available on public databases to assign the binding region and

locate the target genes. For this, we had to ensure that the silkworm races available in India were very close to the races in China (Dazao) and Japan (Daizo p50T), whose genome sequences are available.

We PCR-amplified and sequenced both exonic and intronic regions of Cytoplasmic Actin A4 and cubitus interruptus from two *Bombyx* races available in India, Daizo (multivoltine) and C108 (bivoltine) and compared the sequences to that of the genome databases. Both exon and the more variable intron regions were found to be highly similar to the sequences in the genome databases, with identity of at least 92% for most of the regions sequenced. This ensured that the races available here in in India are indeed suitable to carry out sequencing-based approach to identify the genomic regions in a ChIP experiment.

The DNA binding homeodomain is conserved across various Hox and non-Hox proteins within an organism. In order to raise Ubx-specific antibodies, DNA corresponding to the N-terminal region of the Ubx protein (excluding the Homeodomain, YPWM motif and the C terminal region) of *Bombyx* was cloned into an expression vector. The protein was expressed in bacterial system, purified and was used to raise polyclonal antibodies in rabbit. The antibodies were purified by using a protein-A column and validated for specificity on immunoblotting with *Bombyx* embryonal, larval and wing disc lysates. A single band at around 30KDa, which is the expected molecular weight of the full length Ubx protein in *Bombyx*, was observed. This band was observed only in the hindwing lysate and not in the forewing lysate, consistent with the observation that Ubx is not expressed in the developing forewing in Lepidoptera (Warren et al.1994).

To further validate the antibodies, immunohistochemistry was carried out on the fourth instar discs. High levels of Ubx expression was observed in the nuclei of hindwing discs, whereas its expression in forewing discs was confined only to the peripodial membrane, which is the outermost covering of the wing bud. Earlier Sean Carroll's group had reported that Ubx is expressed only in the hindwing discs in *Junonia coenia* (common buck eye butterfly) (Warren et al, 1994). This is probably because the protocol used

for butterfly immune-histochemistry involved removal of the peripodial membrane.

Thus, in spite of morphologically similar fore- and hindwings, Lepidopteran wing discs show differential Ubx expression. This resembles the pattern seen in the fly, where Ubx is expressed in the nuclei throughout the haltere disc, but is confined to the peripodial membrane in the wing disc.

Antibodies validated as above were used for ChIP experiments. Nuclei from fore- and hindwing discs were extracted, fixed, lysed and sonicated and then subjected to the ChIP with anti-Ubx antibodies. A normal Rabbit IgG (Invitrogen®) was used as a negative control; the input chromatin was used for normalization of both these experiments. Two such biological replicates of ChIP were performed and the resultant DNA of all the experiments and the control were sequenced on an Illumina® deep sequencing platform to obtain the reads.

2. Identification of targets of Ubx in wing buds of *Bombyx*

The sequencing reads were checked for their quality by analyzing with the tool FastQC. The sequences were then aligned to the Silkworm genome version 2.0 (Xia et al, 2004). The peaks (binding regions) were identified by using the program MACS (Zhang et al, 2008). Peaks were identified for each sequencing dataset corresponding to ChIP using anti-Ubx and normal IgG by normalizing to their respective input control reads. The peaks from the negative control (IgG) were considered non-specific and were deleted from the Ubx-ChIP peaks.

Post IgG filtering, 1128 peaks were mapped for the hindwing discs and 340 peaks were mapped for the forewing discs of *Bombyx*, of which only 28 peaks were found to be common between the two wing discs. This is expected as the forewing of *Bombyx* does not express Ubx, except in the peripodial membrane. The peaks observed in the forewing maybe

originating from the peripodial membrane, which does not directly contribute to the wing development.

Genes associated with the binding regions (peaks) were identified by using both the BGI SilkDB database and the Kaikobase database (Shimomura et al, 2009). The Kaikobase provides additional detail on identification of the gene region with data on EST, full-length cDNA and mRNA.

In the hindwing disc data set, 870 genes were associated to the 1128 peaks, and 245 genes were associated to the 340 peaks of forewing disc data set. Of these genes, 548 genes in hindwing and 181 genes in forewing had corresponding fly homologs, with 36 genes being common to the two discs. The identification of significantly lesser number of targets of Ubx in the fore wing compared to the hind wing was not surprising as Ubx expression in the forewing is limited only to the peripodial layer.

The *Drosophila* and *Apis* homologs of putative candidates of Ubx in *Bombyx* were found using the Ensembl metazoa database with the tool Biomart (Kasprzyk, 2011) for *Bombyx* (genome version 2). These homologs were then used for a comparative study with the targets of Ubx in haltere of *Drosophila* (from published ChIP-chip studies, Agarwal et al, 2011 and Choo et al, 2011) and targets of Ubx in the hindwing discs of *Apis* (recently completed ChIP seq study in our lab, Prasad N, 2013). The comparative analysis included the identification of developmental processes and pathways targeted by Ubx in *Bombyx* as against *Drosophila* (and *Apis*) and possible evolutionary trend in *Bombyx* lineage as against *Drosophila* lineage from the ancestral *Apis* lineage.

3. Gene ontology analysis and a comparative study of the targets of Ubx of hindwing in *Bombyx*, *Apis* and the haltere in *Drosophila*.

Targets of Ubx were directly compared with each other to identify targets that are common and specie specific between the three organisms. As the study intends to understand the targets that have come under the regulation of Ubx during Dipteran evolution, comparisons were made only for the

subset of the targets of Ubx in *Bombyx* and *Apis,* for which fly homologs are listed in databases.

When the target genes of Ubx in hindwing of *Bombyx* were compared to the two studies on *Drosophila* haltere, it was found that many genes such as *brinker, engrailed, hedgehog, vestigial* etc, known to be relevant in wing development are common targets of Ubx in both species. The haltere specific set too has genes, which were previously studied as wing development genes such as *ten-m, vein, Wingless, dpp, homothorax, notch* etc. These targets specific to haltere, may have come under the regulation of Ubx after the divergence of Diptera from Lepidoptera and may play an important role in the suppression of hindwing and the specification of haltere.

Comparatively higher percentages of targets were shared between *Drosophila* and *Bombyx* as compared to *Drosophila* and *Apis*. This is reflective of the fact that Lepidopteran and Dipteran lineages diverged much later compared to Hymenopteran lineage. However, fewer targets are common between *Bombyx* and *Apis* than *Drosophila* and *Apis*. Probably, more targets of Ubx in the ancestral (Hymenoptera) *Apis* have been selected and retained in the *Drosophila* lineage during evolution to specify a haltere.

When all the three sets namely *Bombyx, Apis* and *Drosophila* targets of Ubx were compared together, the genes common to all the three data sets were very few. Most of these genes are well known in the context of *Drosophila* wing development, suggesting an essential role for Ubx in the hindwing development/modification in all insect groups, even when the diversification of forewing-hindwing morphology is minimal.

In order to functionally categorize the target genes in the three insect orders into biological processes and pathways they belong to, Gene ontology (GO) analysis (using the database DAVID, Huang et al, 2009) was carried out. The resulting representations of percentage of genes in each of the three insect sets were then compared.

We observed that the molecular and cellular processes that are essential in shaping the wing in *Drosophila* were represented in greater proportions in all the three insect orders studied. As a general trend, genes representing each biological process are represented in similar proportions across the three insect orders. This suggests that the targets under Ubx regulation in similar biological processes have been similarly regulated across these three very diverse insect orders in evolutionary time for the past 350 million years. Interestingly, a trend was observed wherein the more ancestral *Apis* has the least percentage of genes represented each of GO category. It is followed by *Bombyx*, while *Drosophila* has the highest representation. This increase in the proportion of specific GO category amongst the total targets of Ubx in *Drosophila* is more prominent in case of processes such as cell adhesion and regulation of growth. Cell adhesion, proliferation and growth control are some of the developmental tools that may be regulated by Ubx to shape a small globular haltere from a default wing state in the T3 segment. These novel targets specific to *Drosophila* may have played a part in quantitatively increasing the differences between the wing and the haltere. *Drosophila* (Dipteran) lineage further diverged from *Bombyx* (Lepidopteran) some 250 million years ago. This suggests that evolution of Diptera is correlated (and perhaps a main driving force) with the increased number of wing development genes coming under the regulation of Ubx

Targets of Ubx, which are common between two insect sets and their respective species-specific target sets, were subjected to a separate gene ontology analysis and a pairwise comparison was carried out between insects. Here, we observed that the percentage of genes representing a biological process in the common set between *Bombyx* and *Drosophila* had a substantial enrichment over the individual species-specific sets, especially in the wing development and transcription related processes.

A similar comparison was carried out between the common and species-specific targets of Ubx in the hindwings of *Bombyx* and *Apis*. While a similar increase in enrichment of common genes over the species-specific

genes was observed, the enrichment was comparatively lower when compared to the common targets of *Bombyx* and *Drosophila*.

Drosophila appear to have retained and added more Ubx targets to specify the haltere, which is reflected in the comparison between common and species-specific targets between *Apis* and *Drosophila* (Prasad N, 2013). The common targets between *Apis* and *Drosophila* show a higher enrichment than the common targets between *Apis* and *Bombyx* for specific GO categories. This suggests that as *Bombyx*, which has a flat hindwing structure, has comparatively lesser ancestral targets retained and/or enriched as compared to *Drosophila*.

In summary, there appears to be enrichment for the wing development related genes amongst the targets of Ubx. A comparison of such genes suggests that *Bombyx* is far more diverged from *Apis* than *Drosophila*, although at the morphological level *Bombyx* and *Apis* are both 4-winged insects with near-identical fore- and hindwings. The targets of Ubx that are common to *Drosophila* and *Apis* had higher proportional representation of genes related to wing development compared to the targets of Ubx in *Bombyx* and *Apis*. This suggests that Ubx in *Drosophila* appear to have retained as well as acquired more wing-development genes as its targets. Thus, diversity in the morphology in insect wings may be at the level of evolutionary changes in the regulatory sequences, which is being investigated now.

4. Identification of genes differentially expressed between *Bombyx* fore- and hindwings and a comparison to genes differentially expressed between wing and haltere in *Drosophila*

The targets that are differentially regulated between wing and haltere discs in *Drosophila* seem to be crucial for the specification of the haltere fate. In order to see if there is such differential expression of certain targets of Ubx (identified by ChIP-seq) between the fore- and hindwings of *Bombyx,* we carried out RNA-seq of transcriptomes of fore- and hindwing buds isolated from the fourth instar larvae of *Bombyx*.

We observed that 241 genes are differentially expressed (fold change > or =2) between fore and hindwings, and amongst these only 10 are the direct targets of Ubx in hindwing of *Bombyx* identified in this study by ChIP-seq. Thus, very few genes are differentially expressed between the fore- and hindwings and amongst them even smaller number of genes are direct targets of Ubx. The absence of any gene expression differences is reflective the morphological similarities between the fore- and hindwings.

We then compared the *Drosophila* homologues of targets of Ubx (from ChIP-seq) in *Bombyx* hindwing, to two microarray studies (Mohit Prasad et al. 2006 and Pavlopoulos and Akam 2011) that identified genes differentially regulated between wing and haltere in *Drosophila*. We observed that more genes (37 and 65) are differentially expressed in wing and haltere than between fore- and hindwing discs (only 10) in *Bombyx.* Amongst the genes differentially expressed between fore- and hindwing buds are *Ubx, engrailed (en), cheerio and bent.* While in *Drosophila* too Ubx is expressed only in the haltere, *en* is not differentially expressed between wing and haltere. Regulatory regions of the genes that are differentially regulated only between wing and haltere could have evolved in *Drosophila* to respond more strongly to the presence of Ubx.

Future Directions

Sequence Analysis of ChIP-seq data

The homeodomain in Ubx is shown to bind to a TAAT core motif-based heptamer in *Drosophila* (Ekker et al, 1991). Studies from our lab have shown that such motifs cannot be recognition sites for Ubx to identify its targets on the chromatin. However, our earlier work suggests that binding sites for other transcription factors such as GAGA-associated factor are enriched in regions bound by Ubx in both *Drosophila* and *Apis* (Agrawal et al, 2011; Prasad, 2013). This suggests that Ubx recognizes its targets by recognizing a complex of transcription factors already bound to the chromatin.

The preliminary MEME-ChIP motif analysis of the Ubx-binding regions in *Bombyx* and its comparison to *Drosophila* and *Apis* reveals that in *Bombyx* too, there is no clear target recognition sequence for Ubx. This also indicates that binding of Ubx alone may not be sufficient for the regulation of targets.

There is also a possibility of recruitment of other cofactors to regulate the expression of targets of Ubx. As many genes are targeted by Ubx in both *Bombyx* and *Drosophila,* but are differentially regulated only in *Drosophila*, they may have evolved to be regulated differently in different insect orders. Understanding the way these genes are regulated and the organization of their regulatory regions will probably allow us to unravel the mechanisms of specification and evolution of haltere in Diptera against hindwing in Lepidoptera. Therefore, the next step in this work would involve a comparative analysis of the Ubx-binding regions determined by ChIP-seq. This helps us to identify what sequences changes in a given gene has allowed it to (i) become a target of Ubx in *Drosophila* (if the gene in question is not a target of Ubx in other insect species) and (ii) differentially express between wing and haltere in *Drosophila*, but not between fore- and hindwing in *Bombyx* or *Apis* (this is in cases wherein a gene in question is a target of Ubx in more than one species, but differentially expressed only in flies).

BIBLIOGRAPHY

Agrawal, P., Habib, F., Yelagandula, R. & Shashidhara, L.S. (2011). Genome-level identification of targets of Hox protein Ultrabithorax in *Drosophila* , novel mechanisms for target selection. Sci. Rep. 1, 205,1-10

Carroll SB, Grenier JK, Weatherbee SD. (2005). From DNA to Diversity, Molecular Genetics and the Evolution of Animal Design. Malden, MA, Blackwell Sci. 2nd ed.

Choo SW, White R and Russell S . (2011).Genome wide analysis of the binding of the Hox protein Ultrabithorax and the Hox co factor Homothorax in *Drosophila* . Plos One 6(4),1-14

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.

Galant R, Walsh CM, Carroll SB. (2002). Hox repression of a target gene, extradenticle -independent, additive action through multiple monomer binding sites.Development,129,3115–26

Grenier, J. K., & Carroll, S. B. (2000). Functional evolution of the Ultrabithorax protein. PNAS, 97(2), 704-709.

Huang DW, Sherman BT, Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc.4(1),44-57

Kasprzyk A, (2011).BioMart, driving a paradigm change in biological data management. doi,10.1093 bar049

Lewis EB. (1978). A gene complex controlling segmentation in *Drosophila* . Nature 276,565–70

Mohit Prasad et al. (2006). Modulation of AP and DV Signaling Pathways by the Homeotic Gene Ultrabithorax During Haltere Development in *Drosophila*. Developmental biology 291.2,356–67.

Mora C, Tittensor DP, Adl S, Simpson AG, Worm B. (2011). How many species are there on Earth and in the ocean? PLoS Biol. 9(8),e1001127

Macdonald WP, Martin A, and Reed RD. (2010). Butterfly wings shaped by a molecular cookie cutter, evolutionary radiation of Lepidopteran wing shapes associated with a derived Cut/wingless wing margin boundary system. Evo & Dev. 12-3,296–304.

Pavlopoulos A and Akam M, (2011). Hox gene Ultrabithorax regulates distinct sets of target genes at successive stages of *Drosophila* haltere morphogenesis. PNAS. 108-7,2855-2860

Prasad N. (2013). Thesis, Hox genes and evolution of arthropod body plan, A comparative analysis of targets of Ultrabithorax in *Drosophila melanogaster* and *Apis mellifera*.

Ronshaugen M, McGinnis N, McGinnis W. (2002). Hox protein mutation and macroevolution of the insect body plan. Nature 415,914–17

Shimomura M, Minami H, Suetsugu Y, Ohyanagi H, Satoh C, Antonio B, Nagamura Y, Kadono-Okuda K, Kajiwara H, Sezutsu H, Nagaraju J, Goldsmith MR, Xia Q, Yamamoto K, Mita K. (2009). KAIKObase, an integrated silkworm genome database and data mining tool. BMC Genomics. 10,486.

Singh MK, Singh A and Gopinathan KP. (2001). The wings of *Bombyx mori* develop from larval discs exhibiting an early differentiated state, a preliminary report. J. Biosci. 26-2,167-177

Warren, R. W., Nagy, L., Selegue, J., Gates, J., & Carroll, S. (1994). Evolution of homeotic gene regulation and function in flies and butterflies. Nature 372,458-461

Xia et al. (2004). Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx mori*). Science 306,1937-40

Zhang et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol vol. 9 (9), R137

Chapter 1

Introduction

Chapter 1: Introduction

How a single celled egg gives rise to a multi cellular complex organism has fascinated humanity from ancient times. The diversity and myriad variety of forms of animals that exist on this earth fill us with a sense of wonder as to how such a plethora of animal forms arose from a visibly similar single celled form. Questions, such as what instructions guide the growth direction and what mechanisms regulate it to achieve the body forms, have always intrigued researchers over the past two hundred years. Some answers that remained unsolved for the past two centuries have come about to be explained in a span of the last two decades by our growing knowledge of how genes regulate many processes to bring about the variety and form. However, there is still a huge gap in our understanding of the mechanisms of development, exploration of which need invention of newer experimental approaches.

1.1 Body plan formation/ segmentation and patterning

When a single celled embryo develops into a complex adult, the initial process of pattern formation allows the organization of spatial and temporal pattern of cellular processes to give rise to a well-ordered structure. This process involves the instructions to direct the formation of organs or structures in the right location by controlling cellular growth and movement. This orchestration is brought about by an array of cellular and molecular mechanisms at different stages of development.

Pattern formation involves laying out the body plan that starts with defining the coordinates, which provide reference for positional information. Origins of these coordinates are in the polarization of single-celled embryo itself. The first phase of polarization occurs along two axes, antero-posterior (head to tail) and dorso-ventral (back to belly), to provide a coordinate system to localization of organ progenitors and regionalized patterning (Nusslein-Volhard and Wieschaus, 1980). This follows, mostly in vertebrates, specification of the third axis, the left-right axis.

The laying of body plan is followed by processes of morphogenesis and differentiation to form various body forms with complex structures. During morphogenesis the developing embryo develops into three germ layers, which involves cell migration and growth. These layers give rise to different organs and structure in the adult. During cell differentiation, the cells gradually start to become structurally and functionally different in order to form different organs and structures. Cell differentiation also contributes to the pattern forming mechanism by facilitating the formation of different structures from similar types of cells. Once the differentiation has occurred growth brings about the increase in size and changes in shape controlled by various signaling pathways for different body locations.

The basic mechanisms of axis formation and body plan organization are well conserved across animals with bilateral symmetry and they are understood mostly from the studies based on the model insect *Drosophila*. All arthropods have the characteristic segmented body plan. Arthropod embryos pass through segmented germ band stage, which is very well conserved and is also referred to as the phylotypic stage. The segmentation genes that function at the bottom of the *Drosophila* segmentation cascade just before and after phylotypic stage seem to be well conserved among arthropods. These genes include the segment polarity genes like *engrailed (en), Wingless (Wg)* and *hedgehog (hh)* and proteins that encode parasegment boundaries. The larval body of *Drosophila* is made up of three thoracic and eight abdominal segments. Although different regions of the body of larvae are different, all segments also show similarity in certain morphological features.

In an animal body, the axes determine the allocation of cells for developing specific structures in the designated position. Morphogen gradients, which are initiated from maternal signals help in assigning the cells to their coordinates from the two reference axes. Allocated cells in Dorso-ventral (DV) axis constitute the germ layers, whereas the AP axis subdivides into segments (Lawrence and Struhl, 1996). The sequential expression of different sets of genes establishes the body plan along the antero-posterior (AP) axis. The four classes of genes acting along the AP axis are the gap genes, pair-rule genes,

segmentation genes and the homeotic (Hox) genes. Selector genes such as segment-polarity gene *engrailed* and Hox gene complexes confer identity to a region by controlling activity of their downstream genes. Selector genes give unique instructions for development to the founder cells of compartments and their descendants. These genes are activated in different combinations in different segments, which determine the fate of the founder cells in that region.

1.2 Hox genes and pattern formation

The homeotic or Hox genes were first identified and classified based on the phenotypes exhibited by mutations in which a part of the body is transformed and took the identity of another part of the same organism, which is termed as homeotic transformation. For example, loss-of-function mutations in Ultrabithorax in *Drosophila* show a transformation of third thoracic segments bearing halteres into the second thoracic segment bearing wings, while flies with dominant gain-of-function Antennapedia mutation show antennae to leg transformations. Hox genes encode homeodomain-containing transcription factors that regulate a variety of developmental processes including patterning along the antero-posterior axis, segmentation, cell cycle regulation, differentiation etc. The homologs of these genes are found in all bilaterians playing similar roles in development of body plan.

In *Drosophila* there are 8 such genes, while mammals have close to 40 Hox genes, which determine the identity of segment along the AP axis of the embryo. They are found clustered as gene complexes on the chromosome/s. In *Drosophila*, there are two major complexes of Hox genes, the Antennapedia complex (ANT-C) and the Bithorax complex (BX-C). From mutation, studies it was determined that the bithorax complex controls the identity of the third thoracic and all abdominal segments, while Antennapedia complex controls more anterior segments. The ANT-C includes *labial (lab), proboscipedia (pb), Deformed (Dfd), Sex comb reduced (Scr)* and *Antennapedia (Antp)* and the BX-C includes *Ultrabithorax (Ubx), abdominal-A (abd-A)* and *Abdominal-B (Abd-B)* (Fig1.1). The gene number in each complex and gene arrangement varies across different animals. For example, the *lab* is far off from the rest of the

ANT-C in *Bombyx* (Yasukochi et al, 2004). The gene *fushi tarazu (ftz)* does not play Hox-like role in *Drosophila*, while it is a Hox gene in basal arthropods (Gibson 2000). Nevertheless, in *Drosophila*, *ftz* is part of ANT-C on the chromosome.

Drosophila body is divided into 14 segments specified by 8 Hox genes. This difference in the number of Hox genes and the number of independent segments may be explained by the fact that some of the segments of the abdomen are very similar and do not exhibit any segmental differences and certain segments are specified by a combination of Hox proteins and not by just one Hox protein. It is interesting to note that in spite of the body form differences, animals have the same array of Hox genes, which is paradoxical. It has been proposed that Hox genes demarcate relative positions in the animal forms rather than specify one particular structure (Carroll, 1995). A given Hox gene may regulate a specific segment in different ways in two species, while within an animal different Hox genes control the morphology of different regions. The Hox proteins bind to the DNA by recognizing a core sequence of four nucleotides to directly or indirectly relay their influence on large number of downstream targets. As all Hox proteins have similar DNA-binding domain, it is not well understood how they induce morphological diversity between segments.

Hox proteins function by modifying a pre-existing developmental program. The default pattern of a trunk segmental identity in *Drosophila*, for example, is that of the second thoracic segment. Loss of function mutations in of all genes of the Bx-C transforms all abdominal segments to look like second thoracic segment (Struhl, 1982). Considering the potential of Hox proteins to cause such major morphological changes to a given segment, regulation of their expression pattern is very critical. This has lead to the evolution of elaborate molecular mechanisms in the animal kingdom to keep Hox proteins expressed in specific domains. They are subjected to both negative (to keep their expression in switched-off state in specific segments) and positive regulation (to keep their expression in switched-off state in specific segments). In most organisms, polycomb class of proteins keep the Hox genes in repressed state and members of the Trithorx family of genes keep them in activated state.

The organization of the Hox genes on the chromosome is in the same order, from the centromere towards the distal tip, as in which they confer identity to the segments along the antero-posterior axis (Sanchez-Herrero et al. 1985; Kaufman et al. 1990). For example, *Ubx*, *Abd-A* and *Abd-B* of the Bx-C are arranged in that order from centromere to the distal end of the third chromosome in *Drosophila*. As mentioned above, Ubx is expressed in segments more anterior to that of Abd-A, which in turn is expressed in segments more anterior to the Abd-B. This co-linearity in arrangement and expression of gene products is highly conserved across animal kingdom. In addition to maintaining this co-linearity, they also exhibit a specific pattern of interactions amongst each other when more than one Hox protein is expressed in the same cell. In general, the Hox genes that are distal to the centromere dominate over the Hox genes that are proximal to the centromere by suppressing the latter's function. This phenomenon is known as posterior prevalence. In summary, the Hox genes are arranged collinearly on the chromosome and they interact with each other in complex ways to define the segmental identities along the body axis (Fig 1.1).

In spite of such wealth of knowledge on Hox genes, we do not yet completely understand the downstream targets and pathways that Hox proteins control and interact to confer segmental identities.

1.3 Body form diversity and evolution

As Hox genes regulate the processes leading to the specification of segmental identities in animals, they could play a key role in bringing about the diversity of animal life during evolution. Among animals, insects are the most abundant and diverse class. There is evidence for changes in the expression of Hox genes causing the evolutionary changes in the body patterning of insects (Carroll, 1995). The differences between the body plans of insects and other arthropods also relates to the differential Hox gene expression as well as to the changes in the sequences of the Hox proteins, particularly at the C-terminus (Grenier, 1997, Hughes and Kaufman, 2002).

To appreciate the myriad body forms and how they may have been derived through evolution in insects, we need to understand the evolution of insects. The body plan of all animals is organized in similar fashion, with bilateral symmetry and having anterior and posterior ends. The bilaterians arose from an Urbilaterian common ancestor about 550 million years ago (Erwin and Davidson, 2002). Arthropods belong to the phylum arthropoda and are segmented, appendage-bearing protostomes, protected by cuticle that is shed periodically during development (Brusca and Brusca, 1990). The phylum Arthropoda, a phylum with the highest number of species is classified into the subphyla Mandibulata and Chelicerata. Mandibulate can be further divided into Myriapoda nd Pancrustacea, Pancrustacea includes the group Hexapoda to which the class Insecta belongs (Regier et al, 2010).

The order Hymenoptera (ants and bees) is derived from an early branch of holometabolous insects around 350 million years ago. Diverging from hymenoptera is the clade consisting of Coleoptera (beetles), Lepidoptera (moths and butterflies) and Diptera (flies). Diptera is the most diverged and modern form of insects that have two wings and two halters for flight (Fig 1.2). Strepsiptera, which has a haltere in the T2 segment and a wing in T3 segment was earlier thought to be a closer relative of Diptera (Whiting and Wheeler, 1994). However, recent evidence suggests that they are closer to the Colepotera (McKenna et al, 2010), thus, indicating two independent, but convergent, changes leading to the evolution of halteres.

Hox clusters were initially discovered in *Drosophila* and were later identified in many other insects. The Hox cluster in *Drosophila* is split into to two complexes (ANTP-C and BX-C), while other insects retain the single cluster of the presumed Urbilaterian ancestor. The Hox cluster in *Bombyx mori* harbors a tandem duplication of 12 Hox genes between *pb* and *zen*, which is unique to this lineage (Fig 1.3). Furthermore, *labial* is located in the opposite end of chromosome in *Bombyx mori*, compared to *Drosophila*. Thus, the clustering and duplication events of the Hox clusters are undergoing evolutionary changes and may have a role in the evolution of diverse body plans.

1.4 Evolution of diverse body plans vis-à-vis Hox genes

The first model attempting to explain the evolution body forms through the regulation of Hox genes was put forth by Ed Lewis in 1978, where he proposed that segmental diversity in insects involved the evolution of homeotic genes that were not present in the ancestral Arthropod forms. However, it was found later that acquisition of novel genes is not very evident between non-inspect arthropods and insects, although duplication of a Hox gene resulting in two similar Hox genes with different expression patterns are found amongst related species.

In the diverse insect world, in spite of having the same set of Hox proteins guiding body pattern formation, numerous changes in the number and type of appendages have occurred through evolution. The study of fossils and existing insect forms allows us to understand the way Hox genes have come to control features like larval limb or adult wing in these plethora of insect forms, which gives us an idea of the fundamental processes that have paved their origin and diversification.

There are several potential evolutionary mechanisms by which Hox gene regulation may bring about the diversity and modification in insects (Pick L and Heffer A, 2012). They are

(i) Changes in the number of Hox genes in a given lineage. Between two insect species, changes in the number of Hox proteins may confer different body plans.

(ii) Changes in the transcriptional and post-transcriptional regulation of a given Hox gene. Differences in transcriptional regulation of a given Hox gene between two species could be due to changes in the upstream regulators and/or changes in the enhancer sequences of that Hox gene. Differences in the regulation at the post-transcriptional level may be due to miRNA profile and/or changes in 5’ and 3’ untranslated regions (UTRs) of the transcripts.

(iii) Changes in coding sequences of Hox genes acquiring novel functions.

(iv) Changes in cis-regulatory regions of downstream genes thereby inducing changes in the way they respond to a given Hox protein.
(v) Changes in coding sequences of genes downstream to a given Hox protein and thereby inducing newer morphological features, even if all else (as above) is similar between two species.

1.5 Origin and evolution of wings in Insects

Insects are the first to have acquired flight in the evolutionary history. The ability to fly in living organisms appears to have independently evolved at least four times: in insects, pterosaurs, birds and bats.

There are two theories on the evolutionary origins of wings in insects, one suggests that they evolved as an outgrowth of the dorsolateral cuticle (Flower, 1964) for gliding before powered flight evolved and the other hypothesis suggests that they were modified from ancestral dorsal projections of the ventral legs of early insects (Kukalova-Peck, 1983). Fossil evidences suggest that the pterygotes (winged insects) evolved wings as derivatives of legs from an apterygote ancestor, and thus the ancestors had wings on all the segments. These may have subsequently lost by the action of Hox genes, except for the second and third thoracic wing appendages (Carroll et al, 1995). Wing and leg imaginal disc primordia are derived from shared set of precursor cells, with similar signaling processes for setting up polarity. Studies on different animal species support this theory and also suggest that insect wings, crustacean epipods, xiphosuran book gills and arachnid book lungs and spinners all share a common ancestry (Angelini and Kauffman, 2005).

Insect flight has gone through two stages of evolution, where the crucial difference between the two modes of flight has been the signaling between nervous system and wings. Butterflies and moths use a more ‘primitive’ mode of flight called the synchronous flight, where each wing beat is generated by a single nerve impulse. This is the same was as in birds and bats. Smaller insects, such as flies, have evolved asynchronous flight, where a nerve impulse is not directly correlated to a wing beat, but in turn there are secondary steering

muscles, which stimulate the primary wing muscles. This helps them to adapt to the smaller body sizes, smaller than that of a bumblebee. In these insects, the wing beat frequency required to support flight becomes unsustainable if the synchronous model is used. However the complete mechanism as to how a single impulse triggers a 'flight engine' is not yet completely understood.

Asynchronous flight modes allow aerodynamic feats like hovering and backward flight amongst other advantages, which are useful for survival (Hunter, 2007). In Diptera, such a powerful flight mode is accompanied with a hassle of instability of the insect body. They have evolved a mechano-sensory dumbbell shaped organs called halteres, which are modified hindwings. They produce anti-phase beats, which provides the inertial forces to stabilize the flight in two-winged flies (Dickinson, 1999). This helps to counter the possibility of the fly to rotate during the flight with rapid wing beat.

Studies in *Drosophila* have shown that Hox genes are not required for the development of wings in insects (Carroll et al, 1995). Wings form on the second thoracic segment, which is the domain of the *Antp* gene. However when Antp is removed from the wing primordia, wings develop normally (Carroll et al, 1995). This is in confirmation of the fact that the second thoracic segment is the ground state of trunk identity and Hox proteins act on this developmental plan to bring about newer segmental identities. In *Drosophila*, Hox gene *Ubx* specifies haltere development in the third thoracic segment by modifying wing developmental pathway (Lewis, 1998). In *Tribolium*, Ubx represses the default elytron (a protective structure modified from wing) formation to promote the wing development in the third thoracic segment (Tomoyasu et al, 2005). Hox genes are also thought to cause minor differences between fore- and hindwings in a given insect species. For example, differences in the eyespot patterns in fore- and hindwings of butterflies (Weatherbee et al, 1999). To summarize, Hox genes do not directly constitute an instructional code to specify leg or wing. They, however, are involved in suppression of legs or wings or their modifications. This may be achieved by regulating many downstream molecular pathways and biological processes.

1.6 Hox gene Ultrabithorax (Ubx) and the specification of the third thoracic segment in insects

Insects are the first animals to have acquired flight during evolution. Amongst all the animals, they belong to the order with the largest number and diversity of species (Mora et al, 2011). A part of this plethora of body forms is also evident in their flight appendages (Fig 1.4). Most modern insects have four wings, a pair each on the T2 and T3 segments. Many of them have similar fore- and hindwings like in the case of dragonflies and damselflies, which is the ancestral state of the wings in insects. In butterflies, the fore- and hindwings display differences in patterns and shape. In bees, the hindwing is slightly smaller in size than the forewing. In Diptera, which are the most recently diverged form of insects, the hindwings are reduced to small balancing organs called halteres. Halteres have no direct role in flight unlike the forewings. Small muscles beat halters back and forth in an antiphase motion to wings to provide inertial forces that stabilize the flight in small insects. Wing and haltere differ in size and morphology. Halteres are globular club shaped organs, whereas the wings are flat bilayered structures (Roch and Akam, 2000). Halteres do not have vein and intervein patterns and also lack marginal bristles of the wings The size of the haltere capitellum is reduced by fivefold and by an eightfold reduction in surface area as compared to the wing blade in *Drosophila* (Crickmore and Mann, 2006: Roch and Akam, 2000).

Ultrabithorax (Ubx) specifies the identity of the third thoracic segment in insects. It is necessary for the proper development of hindwing appendages in Lepidoptera, Coleoptera and Diptera (Tomoyasu et al, 2005; Weatherbee et al, 1998, 1999). In *Drosophila*, Ubx is expressed in the haltere imaginal disc but not in the wing disc (albeit in the peripodial membrane which does not contribute to development of wing proper). The loss of function mutations in *Ubx* cause the transformation of haltere-to-wing in the T3, conversely ectopic expression of Ubx in T2 causes wing-to-haltere transformation. These results suggest that Ubx suppresses wing development and specifies the haltere fate in the T3 segment (Lewis, 1963; 1978; Cabrera et al., 1985; White and Akam, 1985; White and Wilcox, 1985) (Fig 1.5). Ubx functions at different levels of

wing developmental pathway to direct haltere development, by suppressing genes involved in dorso-ventral specification of the disc, organ size and shape and bristle formation (Mohit et al, 2003; Weatherbee et al, 1998).

The origin of the *Ubx* gene can be traced back to times much before the advent of insects and their body plan in evolutionary time scale (Averof and Akam, 1995; Grenier et al, 1997). Ubx is expressed in other arthropods like chelicerates and crustaceans and in Onychorphora (velvet worms, a sister group of Arthropoda), where it is known to regulate limb and appendage development (Hughes and Kauffman, 2005). It has been shown that when Ubx from a non-arthropod lineage organism is ectopically expressed in *Drosophila*, it induces similar transformations of antenna-to-leg or wing-to-haltere and regulates downstream genes as Ubx from *Drosophila* (Fig 1.5). However, unlike *Drosophila* Ubx, Ubx of Onychophora is incapable of transforming the embryonic thoracic ectoderm towards abdominal identity or to repress a limp development by repressing the key target gene *Distal-less* (*Dll*). This functional divergence is mapped to regions on the Ubx protein outside of the well conserved homeodomain (Grenier and Carroll, 2000). It is possible that Hox proteins apart from binding to different targets in different organisms, might recruit different set of cofactors to achieve diverse developmental patterns.

1.7 Mechanism of Ubx function in *Drosophila*

What downstream pathways and genes does Ubx target and regulate? To answer these questions, many studies have come up in the last two decades identifying genes that could be potential targets of Ubx and some of them have also been validated.

In the initial studies to identify the targets Ubx, a few targets regulated by Ubx were identified by candidate gene approaches (Weatherbee et al, 1998; Shashidhara et al., 1999). Later studies showed that Ubx binds to cis-regulatory elements of two such targets, *spalt* and *knot* (Galant et al, 2002 and Hersh and Carroll, 2005). Subsequently, other unbiased approaches have been used to identify direct and indirect targets of Ubx.

Microarray-based studies were used to identify Ubx targets that are differentially expressed between wings and halteres (Mohit et al, 2006, Crickmore ans Mann, 2006; Weatherbee et al, 1998). More recently, studies on tiling arrays have been performed with different stages of development in *Drosophila* by Pavlopoulos and Akam (2011) to identify the targets of Ubx. However, the identified genes in these studies could be either direct or indirect targets of Ubx. Nevertheless, these studies identified key signaling pathways, which may be relevant for haltere development. Makhijani et al in 2006 further validated *thickveins (tkv)* and *dally* as direct targets based on ChIP-qPCR studies and additional genetic analysis suggested that regulation of these targets are critical for haltere development.

Hersh et al. in 2007 used whole transcriptome and custom microarrays to identify target genes of Ubx. They identified the cis regulatory region of the gene *CG13222 (Cuticular protein 47e)* to which Ubx binds and positively regulates the expression of the gene. This work showed that Ubx not only represses a genetic pathway regulating wing development, it also directly activates specific targets required for haltere specification.

There have been many attempts to identify direct taregts of Ubx at the genome-level using Chromatin immunoprecipitation (ChIP) followed by microarray approaches (Agrawal et al, 2011; Choo et al, 2011; Slattery et al, 2011). These studies have given us a comprehensive list of direct targets of Ubx during the specification of haltere in *Drosophila*. Gene ontology studies of these targets suggest that Ubx may regulate genes that are themselves transcriptional regulators and/or key components of the major signaling pathways to specify haltere development.

1.8 Wing development in *Drosophila*

Drosophila is a holometabolous insect, where the adult form is derived after a process of metamorphosis and a resting pupal stage. Virtually the entire adult ectoderm is formed from primordia called imaginal discs. Embryonic cells of

the wing primordia become morphologically distinct when they invaginate from embryonic ectoderm in late embryogenesis to form imaginal discs (Cohen, 1993). The imaginal discs are molecularly distinct from the surrounding larval tissue. The discs grow extensively during the larval phase and at the time of metamorphosis the discs evert through a process of cell rearrangement to form adult appendages.

The adult wing in *Drosophila* is a homologous structure to the leg and it develops from a larval wing imaginal disc. The imaginal disc is an epithelial monolayer which consists of undifferentiated, proliferating cells. The wing disc primordia starts at ~20 cells during embryonic development and goes up to ~75000 cells in late third instar larvae. The patterning of the wing disc for development is regulated by two major patterning centers or coordinates, which are the antero-posterior (AP) and dorso-ventral (DV) axes established at the boundaries of the DV and AP compartments.

The late third instar wing imaginal disc is a flat two layered structure with a thin peripodial membrane and a thicker disc epithelium. The disc epithelium in the distal part called the pouch gives rise to the wing blade and the proximal wing epithelium region called notum gives rise to the thoracic body wall. A region between the pouch and notum gives rise to the hinge region of the adult wing. On complete maturation at the pupal stage, the wing disc invaginates, folding upon itself to form two layers of the wing blade (Fig 1.6). The lacunae between the dorsal and ventral surface give rise to the proteins, which develop into veins (Blair, 2007).

In a wing disc, the signaling centers are set up along the AP and DV compartment boundaries. Cells at the AP boundary set up a signaling cascade that specifies pattern along the AP axis (Fig 1.7). The morphogen engrailed (en) is expressed in the posterior compartment of the disc. en activates hedgehog (hh) and represses the mediator of Hh, cubitus interruputs (ci) in the posterior compartment. Hh is released from these cells and acts on the anterior compartment cells through Ci. As Patched (Ptc), the receptor for Hh is also a target of Ci, more and more Hh is received at the AP boundary. Main target of

Ci is Decapentaplegic (Dpp), a TGF beta signaling protein, which is secreted by the compartment boundary cells and it serves as a signaling molecule for regulating growth and pattern formation in both anterior and posterior compartments along the antero-posterior axis. Dpp regulates the localized expression of Spalt-related (Sal) and Optomotor blind (Omb) through short-range and long range signaling effects. Hh also activates Knot (kn), which is required to specify vein/intervein differences.

The DV boundary forms the signaling axis for the patterning along the DV axis. apterous (ap) is expressed from the first larval instar in the dorsal compartment and it defines the dorsal state. apterous induces Serrate (Ser), a ligand of Notch (N) in dorsal cells, while restricting the expression of another ligand of N, the Delta (Dl) to ventral cells. N signaling at this boundary induces Wingless (Wg), which acts as a signaling molecule analogous to Dpp along the DV axis. Wg regulates Delta and Serrate at the boundary to maintain its own expression. Wg activates expression of various genes at specific thresholds to pattern the wing along the DV axis. *vestigial (vg)* is a pro-wing gene, which is induced by Wg to express in a broad stripe around the DV boundary.

1.9 Development of wing in Lepidoptera

The knowledge about the Lepidopteran wing development comes chiefly from some studies carried out on butterflies, particularly on the molecular mechanisms of formation eyespots on fore- and hindwings.

In Lepidoptera, forewings and hindwings are the flight appendages of the thoracic segment T2 and T3, respectively. The wing discs/buds of Lepidopteran insects develop as flat bilayered epithelial buds that resemble miniature adult wings. The wing buds are small in the first four instars and they grow rapidly in the fifth and the last instar stages. This (bud-like) mode of wing development is the ancestral mode, which is also reported in Hymenoptera (Macdonald et al. 2010). The wing buds do not invaginate or undergo massive cell rearrangements like the wing discs of *Drosophila*, but grow out laterally into a fully grown wing. The development of scales on Butterfly wings involves the expression of the

gene *Achaete-scuete (AS-C)* and is similar to the development of sensory bristles in the fly (Galant et al, 1998, Hartenstein and Posakony, 1989).

Some of the developmental markers of the developing Lepidopteran wing disc are known. For example, *ap* is expressed in butterflies in a pattern similar to that in the fly. *Dll* expression in butterfly discs to form eyespots adapts a mechanism similar to positional information along proximo-distal axis to specify leg development in *Drosophila*. *Dll* is expressed in the center of an eyespot as central focus signal to be regulated by other genes to limit its boundary (Brakefield et al, 1996; Weatherbee, 1998). The expression patterns of genes in developing larval wing (butterfly/silkmoth) are known for *Dll, ptc, ci, nubbin (nub), Wg* and *en* from the work of Carroll et al. (1994), Singh et al. (2001) and Keys et al. (1999). However detailed roles of all the players in the context of wing development in Lepidoptera are not yet studied.

The Ubx protein was found to regulate scale morphology, pigmentation and eyespot specification in *Precis coenia* (Weatherbee et al, 1998). It has been observed that several genes regulated by Ubx in *Drosophila* haltere are not repressed by Ubx during butterfly hindwing development (Weatherbee et al, 1998). This suggests that different sets of targets exist for Ubx in different insect lineages, which lead to the morphological divergence in insect wing appendages.

1.10 Expression pattern of the Hox protein Ultrabithorax

In *Drosophila* embryos, the Hox protein Ubx is expressed from the posterior thoracic region to most of the abdomen. In the thorax, the expression of Ubx is limited to the posterior of T2 and the entire T3 and in the abdominal segments it extends upto A8. The strongest Ubx expression is in the embryonic parasegment 6 that leads to the posterior of T3 and the anterior of A1 (Akam and Martinez-Arias 1985, Akam et al. 1985, Carroll et al. 1988, Martinez-Arias and White 1988) (Fig 1.8). Ubx is expressed in the peripodial membrane of the wing disc, which does not contribute to the development of wing blade proper, in the entire

haltere disc and in the second and third leg discs (Akam 1983; White and Wilcox 1984).

In most of the insects, the expression pattern of Ubx itself does not seem to vary a lot and is found to be very similar to the expression in the fly. The development of the long jumping leg in grasshopper is partially due to the strong expression of Ubx in the third thoracic limb (Kelsh et al, 1994). In *Tribolium,* the Ubx homolog Ultrathorax (Utx) expresses in the posterior of T1 to anterior of T3 (Bennett et al 1999), however the expression levels retract from anterior extending towards the posterior in midway of development. Utx expression in the third thoracic segment suppresses the development of elytra to give rise to a wing appendage. Here hindwing represents a more ancestral state of wing appendage, while the wing program recruits several elytron genes and Utx represses these to promote the formation of hindiwngs (Tomoyasu et al, 2005). Ubx is also known to regulate appendage development in crustaceans. Expression of Ubx in Hymenoptera is similar to that of *Drosophila* (Walldorf, 2000), but *Apis* shows substantial (but lower than hindwing) expression in the forewing disc unlike in *Drosophila* (Prasad, 2013) (Fig 1.9).

Studies of Ubx expression in Lepidoptera have come chiefly from the study on butterfly appendage development. In butterflies (*Precis*), Dll is expressed in the abdomen in the regions wherever Ubx/Abd-A expression is absent, and this allows the development of prolegs. But, in the Hawk moth *Manduca*, Ubx is expressed in the proleg primordia (Zheng et al, 1999). The Octopod mutations in *Ubx* of *Manduca* results in reduction in Ubx expression and results in the transformation of abdominal segments A1 and A2 to thoracic identity (Zheng et al, 1999). In butterflies, Ubx is expressed to the highest levels in the anterior A1 region, with diminishing levels in further posterior segments. It is also expressed in the lateral regions of T2 and T3. Though earlier thought to be 'wing suppressing gene', Ubx is expressed in the hindwing of butterflies as opposed to its absence in the forewings (Weatherbee et al, 1998). Thus, the difference between a butterfly and fly is not the mere presence of Ubx, but the target genes that it regulates.

In *Bombyx* embryos, highest levels of Ubx expression were found in A1 with weak expression in T3 and A9. No expression was found in the lateral regions of T3-A9 (Masumoto et al, 2009). In *Bombyx,* a deletion removes *Ubx* and *abd-A* (E^N) resulting in the development of thorax like legs on abdominal A1-8 segments (Ueno et al, 1992).

1.11 Silkworm as a Lepidopteran model

The common mulberry silkworm (*Bombyx mori*) is the only truly domesticated insect in human history, the start of domestication (sericulture) dates back to 2500 BC in China for silk production. It has been since cultured world over to produce silk for textile industry. Apart from its economic value, silkworm has developed into a valuable model for genetics and molecular developmental studies. Silkworms are easy to rear and the availability of genetically homogenous inbred lines makes them suitable for genetic analysis. Genetic manipulation tools are also available making silkworm the next best insect after *Drosophila* for genetic analysis. The genome of Silkworm was the first complete genome to be sequenced of a Lepidopteran insect in 2004, independently by Chinese and Japanese groups (Xia et al, 2004 and Mita et al, 2004). *B. mori* has 28 chromosomes and a large genome of about 530 Megabases. The quality of the genome has increased since the release to public and now silkworm database is supported with mRNA and EST based annotations making it a good system for studies at the genomic level. It has served as a platform for more such sequencing projects in Lepidoptera as the *Heliconius* and the Monarch butterfly genome sequencing.

The life cycle of Silkworm

The completely domesticated Bombyx moths are unable to fly and survive in the wild. Larvae hatch from fertile eggs in about 7 days. The newly hatched worms are ant-like with hair on the body. They grow prolifically into larger hairless, smooth, creamy white worms about six to eight centimeters long with shiny mouthparts and yellow hemolymph. *Bombyx* larvae are grown on mulberry (*Morus alba*) leaves or on artificial media. They feed on finely cut

mulberry leaves during active growth and moult into the second instar silkworm in 3-4 days depending on the temperature conditions. As the silkworm approaches moulting, the mouth parts become smaller and the larvae stop feeding on leaves. They moult into the next instar by shedding skin. The second instar moults after 2-3 days into third and the third in 3-4 days into the fourth.

The fourth to fifth transition takes 6-8 days and at this stage the worm is shiny with a translucent cuticle and starts to spin a wooly white/yellow cocoon in 2-3 days. The fifth instar silkworm develops into pupae in the cocoon and in 14-21 days, develops into an adult silk moth. The males and females are identified in the pupal stages and kept enclosed in chambers for mating. The female silk moth is larger than the male moth with smaller antennae and wings; neither of the moths can fly and can only survive for two weeks. They mate and the female lays a large number of eggs, which are white in color, which turn yellow and finally grey. These eggs hatch in 9-12 days in case of non-diapause races.

Objectives

The Hox regulation of serial homology has come to be best understood by the studies on the Ubx-mediated haltere specification in *Drosophila*. However, intense studies in the past two decades have not been able to provide sufficient insights into the mechanism by which Ubx specifies haltere. As Ubx itself has not evolved amongst various insect species, although we see much diversity in wing morphology, it has been suggested different sets genes have come under the regulation of Ubx in different insect orders. Therefore, evolutionary developmental biology (Evo-Devo) approach may help us understand the role of Ubx in the evolution of a two-winged fly from a four winged ancestor and at the same time to understand the mechanism of Ubx-mediated specification of haltere. The approach in this study is to compare the role of Ubx between *Bombyx*, *Apis* and *Drosophila*.

The first step is to identify direct targets of Ubx and through them identify developmental mechanisms that are different in these three insect groups. Direct targets of Ubx at the genome level have already been identified for *Drosophila* (Choo et al., 2012; Agrawal et al., 2012) and *Apis* (Prasad, 2013). This study, therefore, focused on identifying direct targets of Ubx in the hindwing buds of *Bombyx* and compared the same to those of *Drosophila* and *Apis*. Following are the specific aims of this project,

5. Identify direct targets of Ubx in the developing hindwing of *Bombyx* by Chromatin-immunoprecipitation followed by deep sequencing (ChIP-seq).

6. Compare the target genes to those of haltere in *Drosophila* and hindwing of *Apis* and identify genes that are being targeted by Ubx in all the lineages and targets that are species-specific.

7. To carry out a Gene Ontology (GO) based functional analysis on the targets of Ubx in *Bombyx* and compare the same to similar analyses carried out for targets of Ubx in *Drosophila* and *Apis*.

8. To find out if any of the target genes of Ubx in *Bombyx* are differentially expressed between fore- and hindwings and compare them to the genes differentially expressed between wing and haltere in *Drosophila*.

Plates

Chapter 1

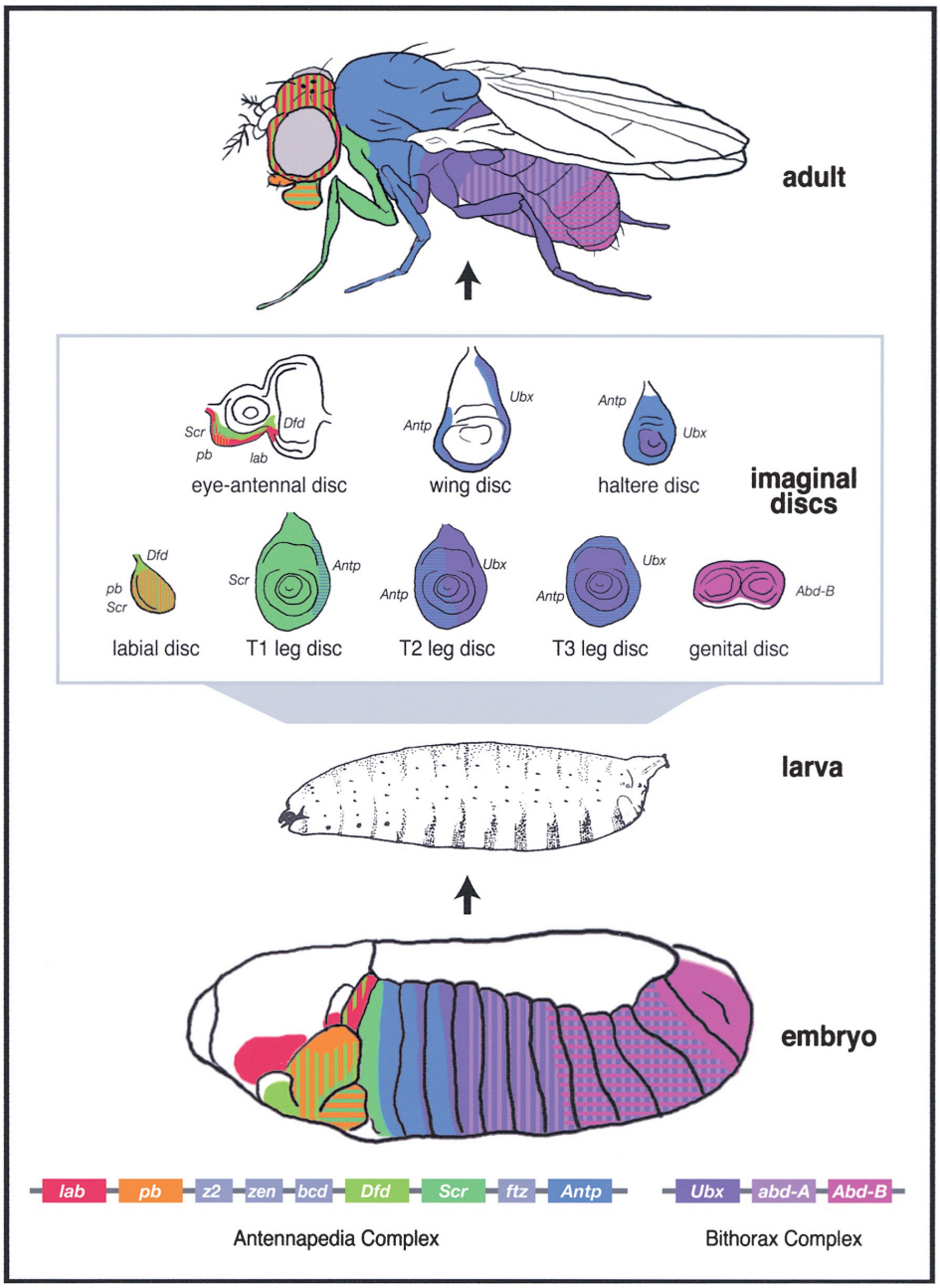


Figure 1.1 Homeobox genes and body patterning

Schematic showing the expression of Hox genes in *Drosophila* adult fly, imaginal discs and the developing embryo. Hox complex comprises of a cluster of 8-10 genes that determine the identity of segments along antero-posterior axis of the embryo and are arranged in the same order in which they lie on the chromosome (from the centromere to the distal tip) in a collinear fashion. (Image: Hughes and Kaufman, 2005).

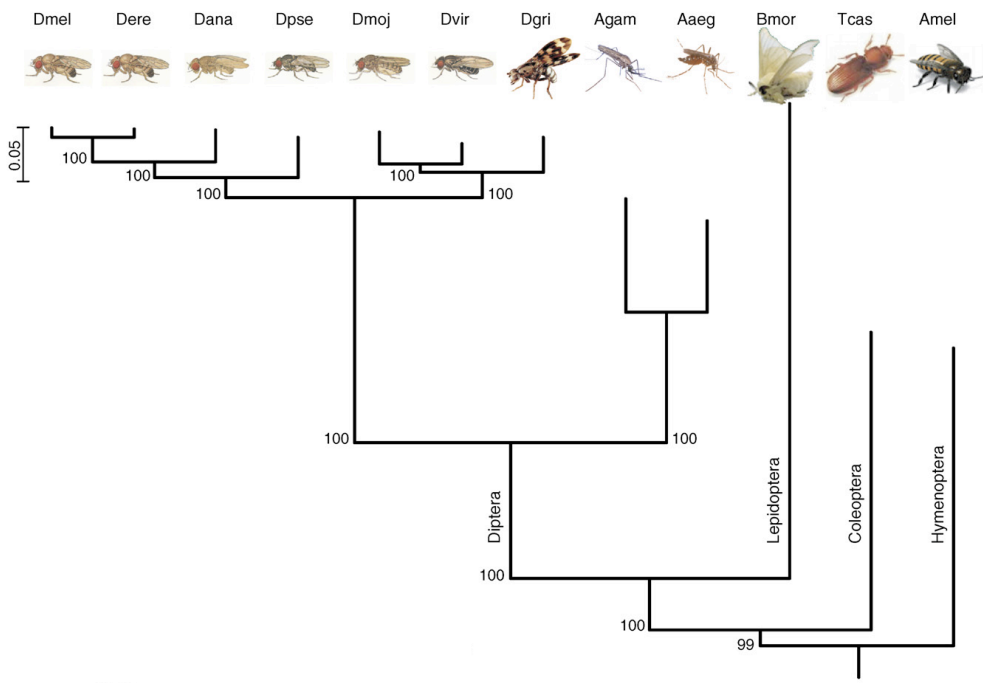


Figure 1.2 Divergence of Insects

The above phylogeny shows the pairwise divergence of sequenced insect genomes in terms of average protein identities. The tree corresponds to the well-established phylogeny of the species of insects (Zdobnov and Bork, 2006)

Abbreviations: Dmel, *Drosophila melanogaster;* Dere, *Drosophila erecta;* Dana, *Drosophila ananassae;* Dpse, *Drosophila pseudoobscura;* Dmoj, *Drosophila mojavensis;* Dvir, *Drosophila virilis;* Dgri, *Drosophila grimshawi;* Agam, *Anopheles gambiae;* Aaeg, *Aedes aegypti;* Bmor, *Bombyx mori;* Amel, *Apis mellifera;* Tcas, *Tribolium castaneum.*

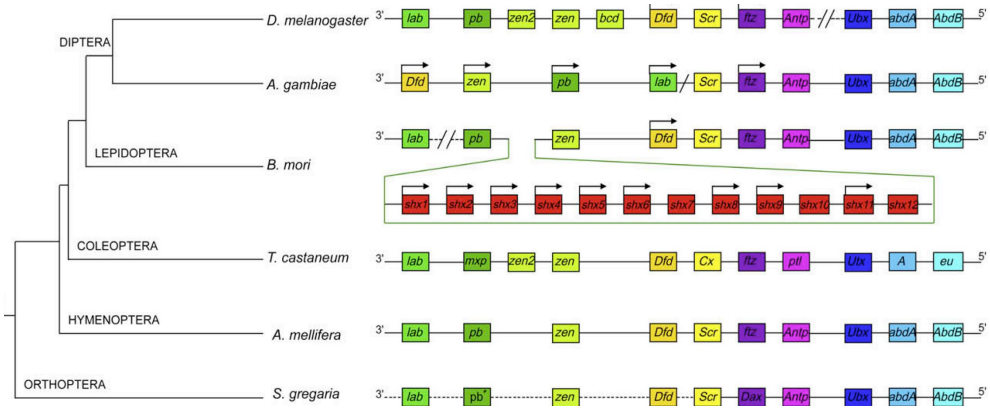


Figure 1.3 Comparison of Hox gene clusters in different insects

Hox genes are arranged in clusters on the chromosome in the same order as they are expressed along the antero-posterior axis. The arrangement of Hox clusters is well conserved across animal kingdom but variations do occur. The gene *labial* in *Bombyx* is isolated and located far away from the Hox cluster. The *Bombyx* Hox cluster also hosts a new Hox group of Bmshx genes, which are well conserved amongst themselves and are split into two sub-clusters (Image: Chai et al, 2008).

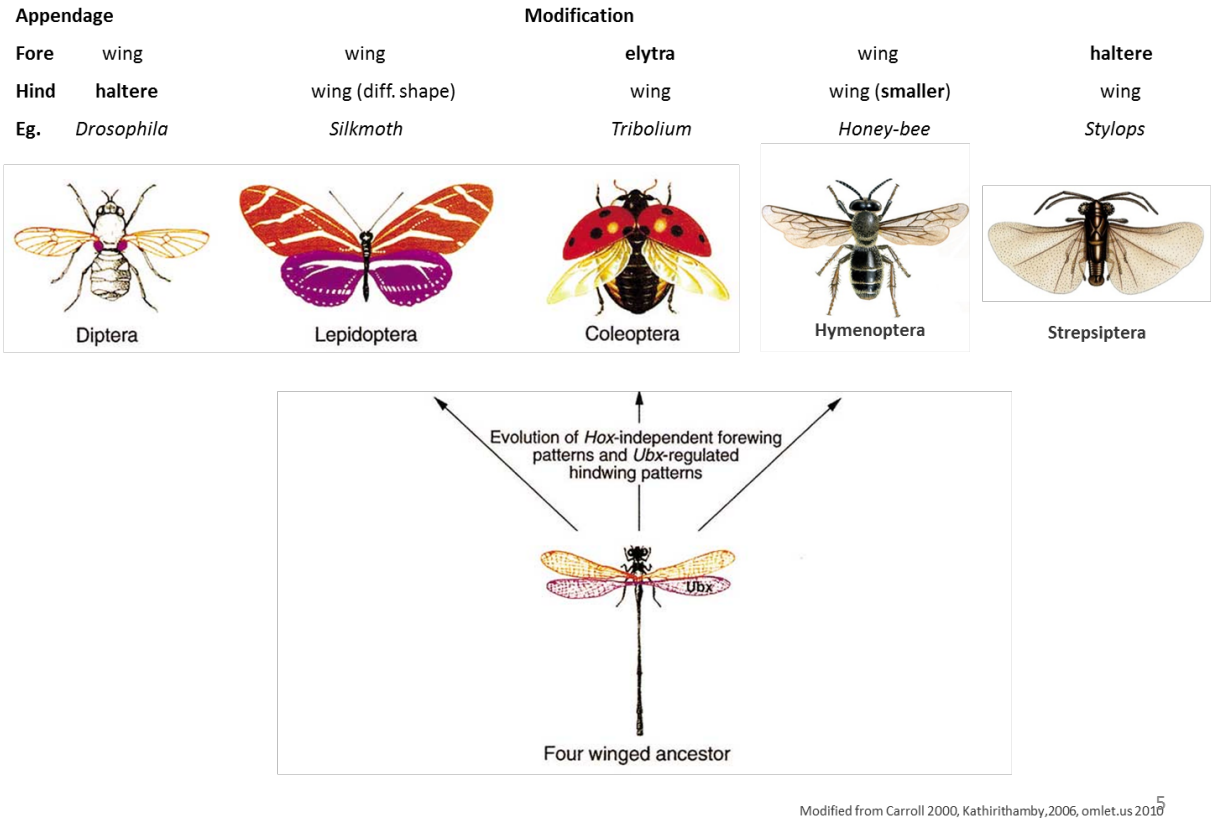
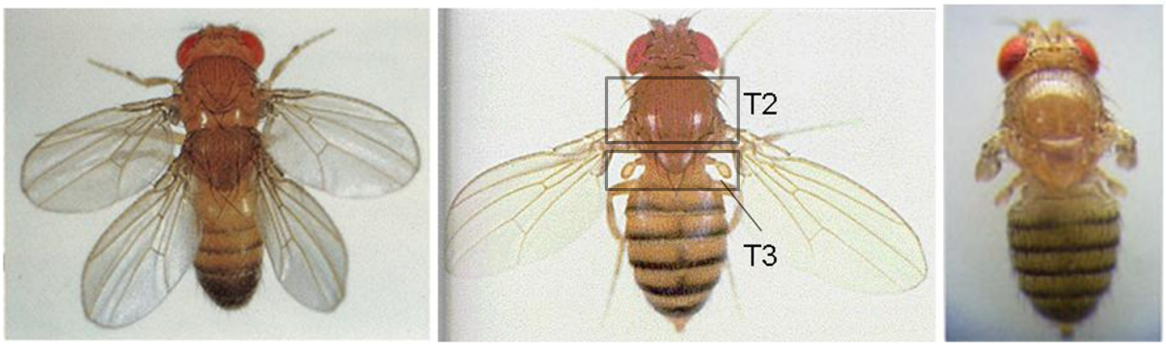


Figure 1.4 Variation in the wing appendages across different insect orders

Insects show a range of wing appendage modifications from change in wing size to pattern and even complete modification. This change is brought about by the action Hox genes through evolution.

Ubx is known to transform a wing into haltere in the third thoracic segment of *Drosophila*. In *Tribolium* however it represses the default elytron formation in the third thoracic segment promoting wing appendage.

Strepsiptera is an order with the forewing modified into a haltere, the mechanisms of this appendage modification is not explored yet, but this order is known to be closer to the Coleoptera than Diptera. Thus, the action of Hox genes on serial homology can be best understood by studying the regulation of these organ modifications across insect orders.



No *Ubx* in T3 | Wildtype | *Ubx* in both T2 and T3

Precis coenia (Lepidoptera) | *Tribolium catanum* (Coleoptera) | *Acanthakora kaputensis* (Onychophora)

Figure 1.5 Role of Ultrabithorax in the specification of segmental identity

Ubx was found to be both necessary and sufficient to cause a wing-to-haltere transformation in the thoracic segments of *Drosophila*. Loss of function mutants of Ubx in the third thoracic segment cause a haltere-to-wing transformation while ectopic expression of Ubx in second thoracic segment causes wing-to-haltere transformation. This transformation can be effected not only by Ubx obtained from *Drosophila*, but also other insect orders (Lepidoptera or Coleoptera) and even from Ubx of a wingless sister order of Arthropods, Onychophorans. This shows that functionally Ubx protein is well conserved and has not changed much through evolution (Images: Thesis Ruchi Bajpai and internet sources)

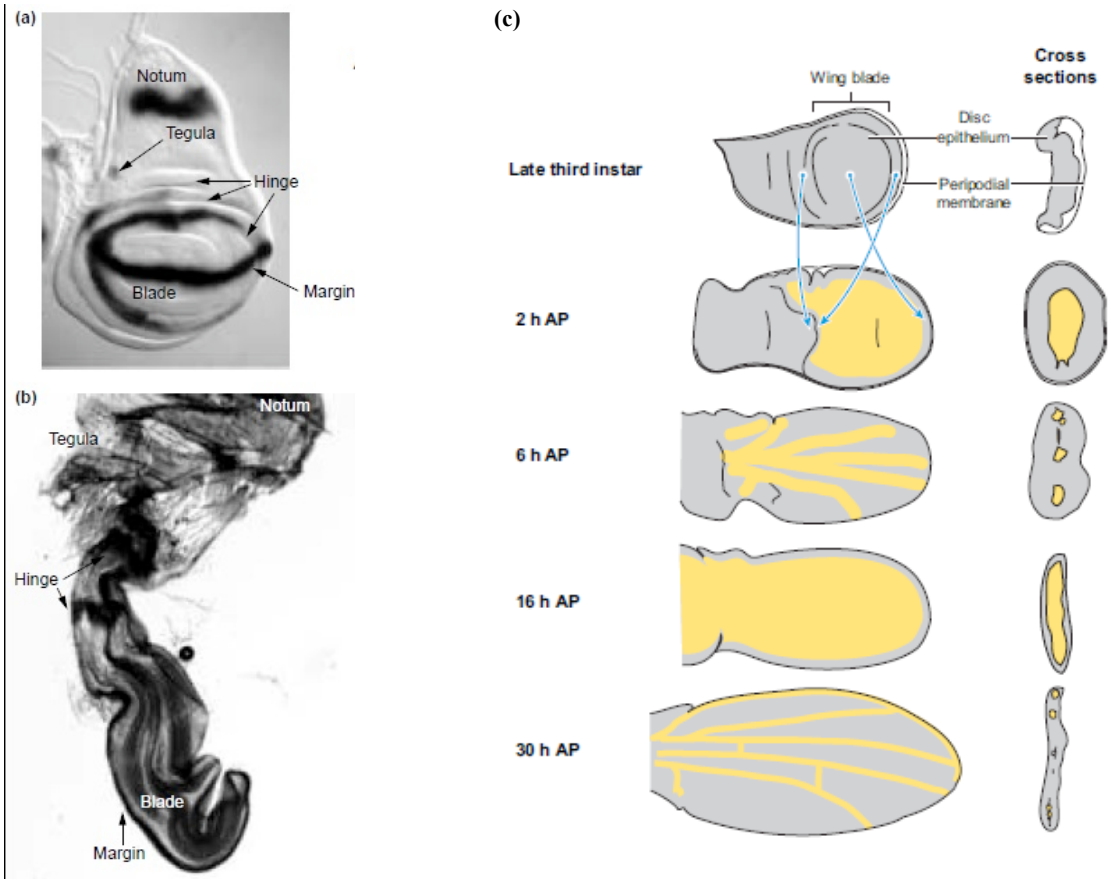


Figure 1.6 Wing development in *Drosophila*

(a) The elements of a developing late third instar wing disc anlange are described in the background of Wingless (Wg) staining. Wg is expressed in the DV boundary, which divides the pouch region into dorsal and ventral compartments, which later become two the dorsal and ventral layers of the adult wing blade. It is also expressed in the hinge region and the notum region, which give rise to adult wing hinge and the thoracic part of the body, respectively. The Wg expressed in DV boundary also gives rise to the wing margin in adult wing.

(b) An image of the developing adult wing with all the recognizable structures derived from anlage elements described in (a).

(c) Stage wise development of wings from late third instar wing disc: dorsal and transverse section views. Arrows from the late-third-instar imaginal disc to the 2 hour wing after pupariation (AP) show the event of eversion where, the prospective dorsal and ventral portions of the pouch in the imaginal disc fold to form the dorsal and ventral surfaces of the wing blade. (Blair, 2007).

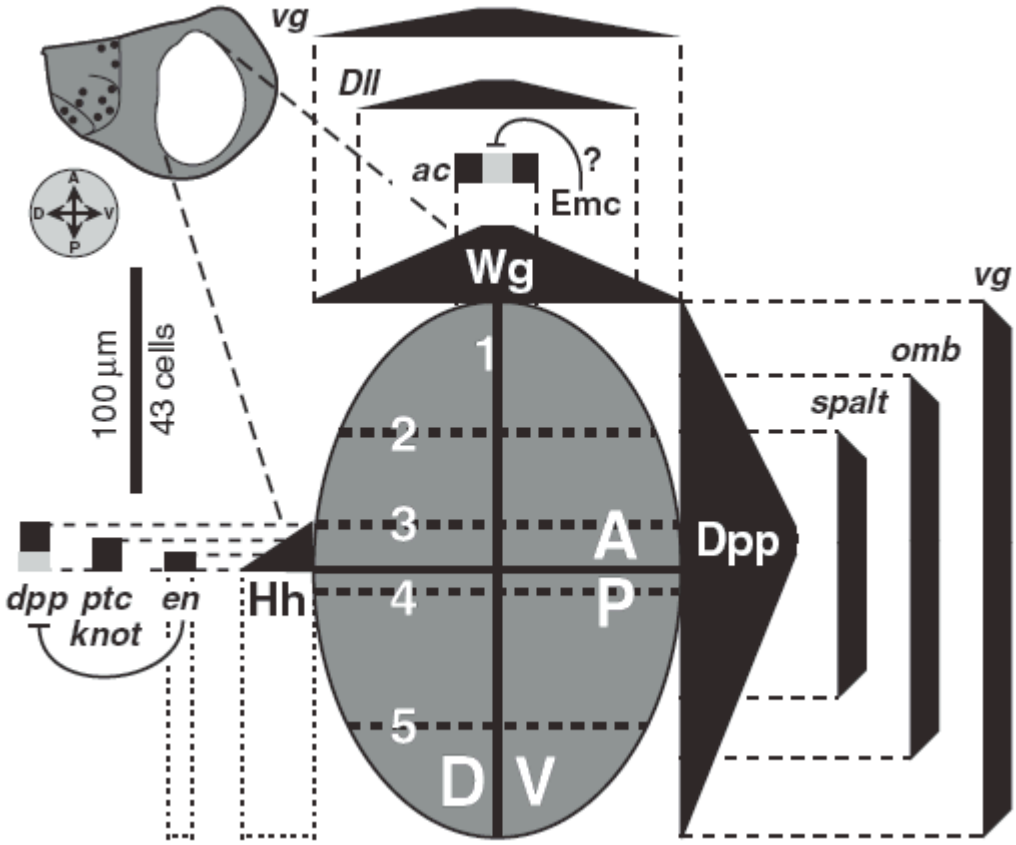


Figure 1.7 Morphogen gradients in wing development

The elliptical region shown in grey is the pouch region showing the compartmentalization by the formation of AP and DV boundaries by various morphogens to specify the coordinates of the wing imaginal disc. (From Held L, 2002)

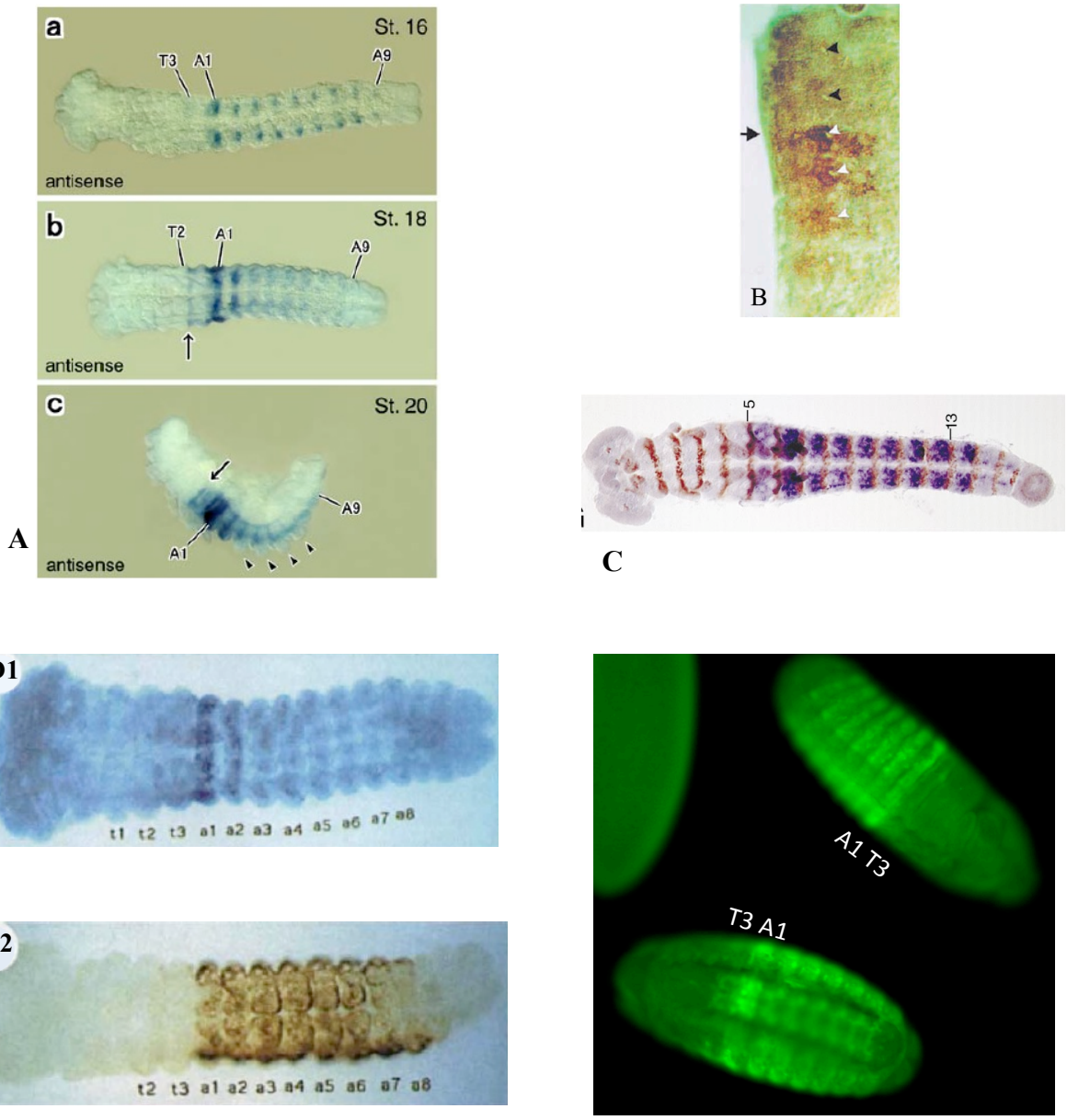


Figure 1.8 Expression patterns of Ubx in Embryos of different insects

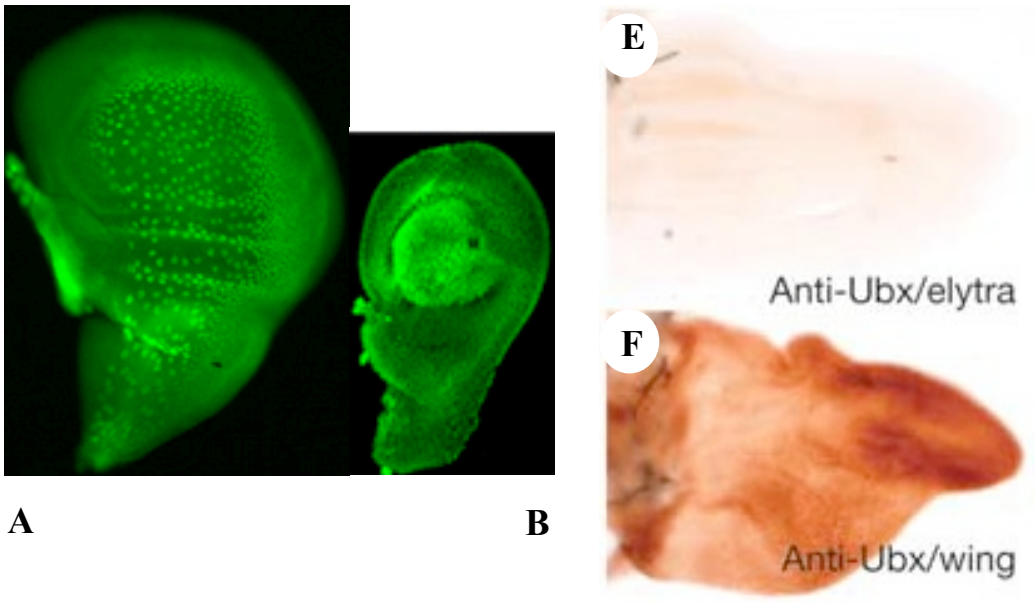
(A) Ventral view of expression on Ubx in *Bombyx mori* embryos of different stages at the transcript level. The highest levels of Ubx are found in A1 segment (Masumoto et al, 2009).

(B) Expression of Ubx/Abd-A in *Apis mellifera* embryos. Arrow shows the parasegment between thorax and abdomen, which expresses highest levels of Ubx (Walldorf et al, 2000).

(C) Expression of Ubx transcripts in *Tribolium castaneum* embryos. The expression is seen between parasegments 5-16 in the embryos (Bennett et al, 1999).

(D1-D2) Early expression pattern of Ubx in the embryos of butterfly *Precis coenia*. Ubx is expressed the highest in A1 and diminishes in the posterior segments (Warren et al, 1994).

(E) Dorsal and ventral views of *Drosophila* embryos showing Ubx expression.



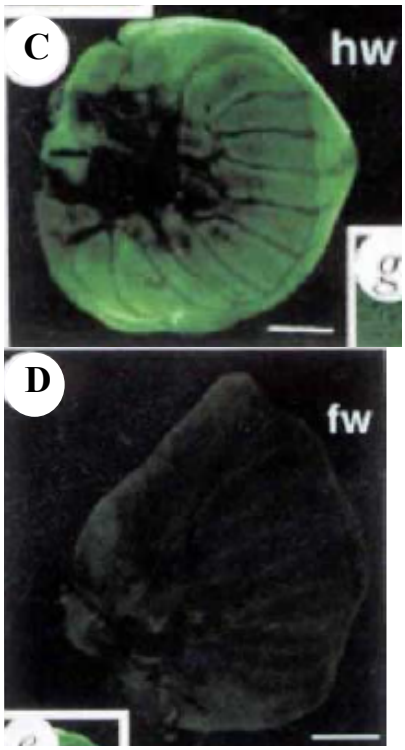


Figure 1.9 Expression patterns of Ubx in wing primordia of different insects

A-B. Ubx is expressed only in the peripodial membrane of wing disc in *Drosophila,* but is expressed in the entire haltere.

C-D. Ubx is not expressed in the forewing bud of *Precis,* but is expressed in the hindiwng bud (Warren et al, 1994).

E-F. Ubx is expressed in the hindwing appendage of *Tribolium* and absent in the forewing (Elytra) disc (Tomoyasu et al, 2005).

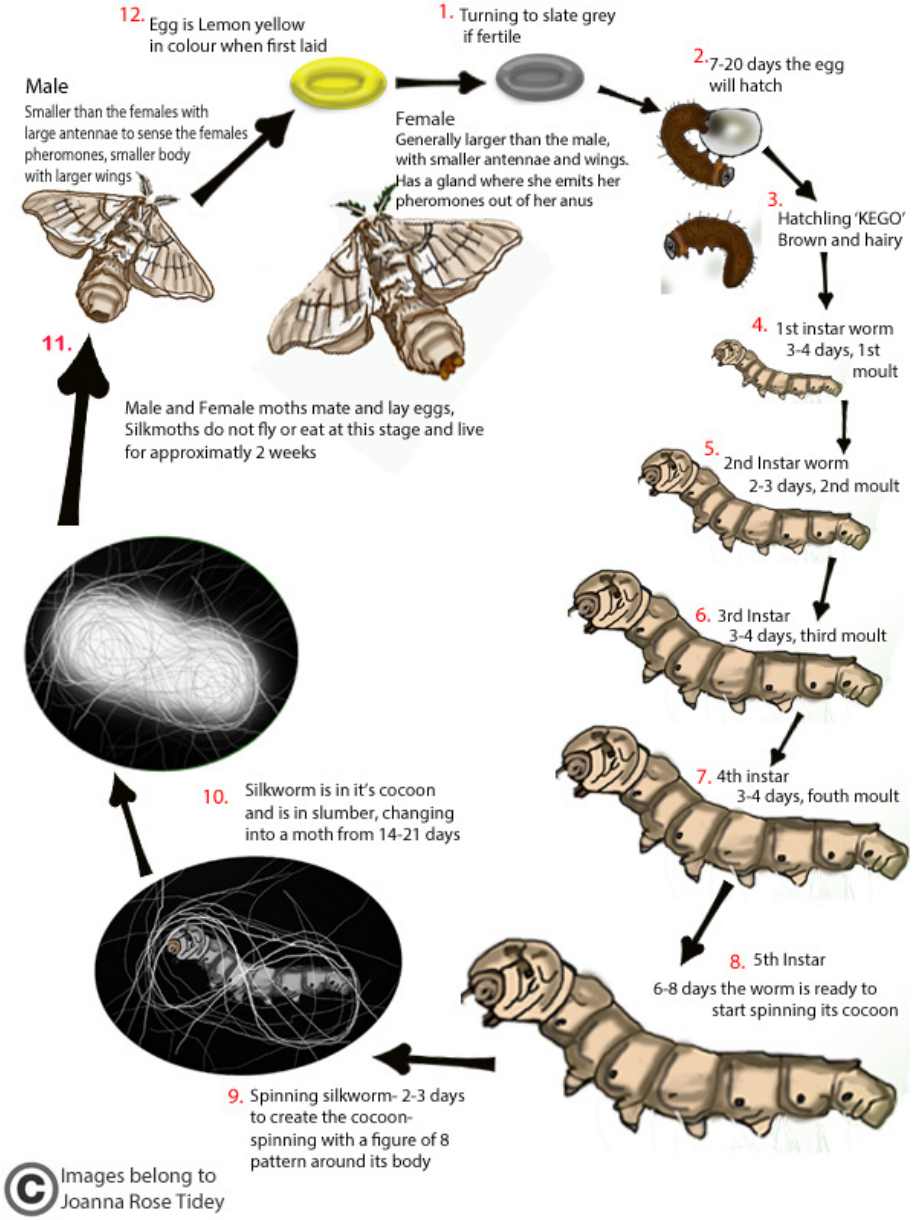


Figure 1.10 Life cycle of the silkmoth *Bombyx mori.*

The above schematic describes the life cycle of *Bombyx mori* from the egg stage through larvae and pupae to the moths. The whole cycle completes in about 2 month duration under 25 ºC conditions. The larvae hatch out of fertile eggs and eat mulberry leaves voraciously and grow rapidly in size. The life cycle consists of 5 larval instars. The late fifth instar larva spins a cocoon and pupates before emerging as a moth. The moths are short lived; they die quickly after mating and laying eggs. (Image credit: joannarosetidey.com)

Chapter 2

Identification of the targets of *Bombyx* Ubx by Chromatin Immunoprecipitation and Sequencing

Introduction

Mechanism of transcriptional regulation can be understood by studying the protein-DNA interactions on chromatin. The mapping of binding sites of transcriptional machinery is crucial to decipher the gene regulatory networks and their manifestations (Farnham et al, 2009).

To understand the regulatory mechanisms that are controlled by Ubx to bring about the diversity we see in the flight appendages of insects in general, the role of Ubx in particular and its downstream target genes to bring about the specification of haltere in diptera, we need information on genome-wide binding regions of Ubx in different insects groups. Towards this direction, in this study, we have employed Chromatin-immunoprecipitation followed by deep-sequencing (ChIP-seq) to identify the genome wide binding sites of Ubx in the developing hind and fore wing buds of *Bombyx*.

Chromatin-immunoprecipitation (ChIP) is a powerful technique that enables selective enrichment of DNA sequences bound by a particular protein in living cells (Solomon et. al., 1988). ChIP is generally carried out by crosslinking DNA-protein interactions using chemicals or radiation, and then using antibodies against the protein of interest to achieve a pull down. As the chromatin bound by the protein will co-precipitate, one could use this technique to potentially identify all the regions of the chromosomes bound by the protein of interest. Earlier the identification of the bound fragments was carried out by making tagged probes of the bound regions and hybridizing to an array (ChIP-chip) allowing genome wide view of DNA-protein interactions (Ren et al, 2000). The array based methods were noisy, low on resolution and biased to the regions represented on the fabricated array. The advent of next generation sequencing methods has enabled the direct sequencing of the DNA fragments instead of hybridizing to a tiling-array (inclusive of promoter/enhancer regions) or a microarray.

In recent times, the enriched DNA obtained in a ChIP experiment is subjected to massively parallel sequencing using platforms like Illumina® or Solid®

sequencers to follow up with computational approaches to identify the targets regulated by that protein. Such deep sequencing methods have better resolution, greater coverage, lesser bias and fewer artifacts as compared to array based methods (Park, 2009). The genome-wide, improved and accurate mapping of binding sites by ChIP seq enables accuracy in identifying and mapping targets of transcription factors, enhancers and also allows identification of binding motifs with higher precision.

2.1 Chromatin Immuno-precipitation

In a ChIP experiment, chromatin regions bound by a particular protein are enriched. It is carried out on live cells or tissue after crosslinking the DNA-protein or protein-protein interactions on the chromatin by using chemical agents like formaldehyde (Fig 2.1). Then cell or nuclear extract prepared is subjected to sonication to shear the chromatin into fragments with an average size between 100-500 base pairs. This sonicated lysate is immuno-precipitated with antibodies specific to the DNA-binding protein of interest to pulldown the regions it binds throughout the genome. Sequences that are bound by a protein factor are selectively enriched in the immuno-precipitated sample. The crosslinking can be reversed by heating to recover and purify the DNA, which is then subjected to quantitation by qPCR, probed to an array or sequenced on a genome analyzer like Illumina® or Solid®. The fold enrichment of certain bound genomic regions relative to non-binding regions provides quantitative information on levels of association of the protein with different genomic regions. It also provides information on the DNA motifs a protein uses to identify specific regions and regulate its targets.

2.2 Array hybridization to identify ChIP enriched regions.

The enriched DNA fragments obtained after a ChIP experiment can be identified at a genome wide scale by hybridization to a microarray (Ren et al, 2000). High density tiling arrays with oligonucleotide probes for the entire genome with interval separated regions include promoter and enhancer regions at a preferred resolution (Park, 2009).

After the enrichment, de-crosslinking and purification on fragments, the fragments are subjected to ligation mediated PCR amplification reaction to amplify in the presence of a labeling fluorescent dye (Cy5) (Fig 2.2). A sample of DNA which was not enriched by immune-precipitation acts as a control and is labeled with a different dye (Cy3). Both these labeled samples are then hybridized to a single microarray containing the genomic regions of the organism used as the source of tissue.

A whole genome array consists of probes of around 60 bp length synthesized and printed on a chip, mapping the entire genome with gaps at regular intervals. These arrays can be used to identify the *in vivo* genome wide direct binding sites of transcription factors by ChIP coupled with hybridization. These arrays have high reproducibility and the coverage is genome wide with multiple overlapping probes representing binding regions. A resolution of fine mapping of binding regions upto 25bp can be achieved by using the tiling arrays.

The ratio of fluorescence intensity between enriched and un-enriched experiments is used to calculate the relative binding of the protein to the sequence. Three independent biological replicates are used and a weighted average analysis is performed to obtain the relative binding values. Genome wide analysis of the binding regions under different conditions /tissues *in vivo* has proven to be an effective tool to discover and understand regulatory networks. However ChIP-chip arrays are limited to the available arrays on certain organisms and turns out to be expensive to cover the whole genome in replicates.

The *Drosophila* ChIP data used in this study for the comparative analysis of targets of Ubx in *Bombyx* (Chapter 4) were generated by such ChIP-on-chip studies (Agrawal et al, 2011 and Choo et al, 2011). Agrawal et al., (2011) had used Agilent® *Drosophila* whole-genome array, 2004 build with, 488,000 probes (each, 58 bp long and with average 233bp spacing), while Choo et al., (2011) used an Affymetrix® *Drosophila* genome-wide tiling array 2.0 with higher resolution and density.

2.3 High throughput sequencing methods to identify ChIP-enriched fragments

Sanger sequencing was one of the greatest technologies to come up, pioneering the field of Genomics in the later parts of the last century. It allowed us to understand the genome organizations to a great detail in various organisms, including that of human beings. However, as opposed to the great expectations at that time, the genome information by itself failed to explain the complex biology that is rooted in all life forms. The researchers then emphasized that regulation in the genome probably plays a very important role. Emergence of the second generation high throughput sequencing in the post-Sanger sequencing era has led to a rapid and greater understanding of the regulatory processes that operate to control various biological processes. The next generation sequencing coupled with ChIP experiments has unraveled many new aspects of gene regulation.

ChIP-seq allows the usage of any species for study with a sequenced genome, whereas ChIP-chip relies on the handful of chips available by vendors in the market or expensive custom arrays. The starting material required for ChIP-seq is very low, as low as 10 ng, which was not possible with the sensitivity levels of ChIP-chip. ChIP-seq is also cost effective method, especially for large genomes, where ChIP-chip requires many chips for the whole experiment with replicates, turning out to be very expensive.

Various platforms are available for high throughput sequencing, which can be used to identify genomic regions which include both larger fragments for *de novo*-whole genome sequencing and smaller fragments from ChIP. They differ in their chemistry and resolution and hence are suitable for different kind of experiments. Some of the commercially available platforms for next (second) generation sequencing are Illumina/Solexa, ABI SOLiD, Roche 454, Helicos Biosciences, and Pacific Biosciences. The third generation sequencing platforms involving single molecule sequencing are set to prevail in the next tier of sequencing technology.

2.4 Sequencing with Illumina (Solexa) Genome Analyzer

The Illumina/Solexa sequencing platform is based on the principle of sequencing by synthesis chemistry, with a special DNA polymerase incorporating reversible terminator nucleotides for four bases, each labeled with a different fluorescent dye (Fig 2.3).

In this system DNA fragments to be sequenced are ligated at both ends to adapters, denatured and then immobilized on a solid support. The surface, which is on a flow-cell, is coated with complementary adapters. Thus, each single stranded DNA fragment with adapters attached forms a bridge by hybridizing with its free end to complementary adapter on the surface. These bridges are then amplified to form localized clusters by using adapter specific primers, via an isothermal amplification process (Ansorge, 2009).

These clusters are then subjected to another amplification step with a reaction mixture containing primers and four reversible terminator nucleotides each labeled with a different fluorescent dye and DNA polymerase. When a terminator nucleotide is incorporated into the nucleotide strand, its position and fluorophore based identity is detected by a CCD camera. Then the terminator group is removed to continue the sequencing reaction and continue the synthesis and detection cycles (Fig 2.3a).

The sequence read length generally is of 35, 72 or 100 base pairs, and it can be single- or paired-end mode of sequencing. The sequencing of at least 40 million clusters can be generated simultaneously and sequenced in parallel resulting in a very high throughput output. Hence the second/next generation sequencing is also called massively parallel sequencing to signify this process of formation of multiple clusters and simultaneous amplification and sequencing of many such clusters on the flow cell. This system generates at least 1.5 Gb of single read data per run and 3Gb per paired end run with a runtime of 36 bp single end cycle of two days (Ansorge, 2009).

This study employed Illumina GAII automated genome analyzer for sequencing in two experiments namely the ChIP-seq study (current chapter) and the RNA-seq study on *Bombyx* wing discs (Chapter 5).

2.5 Requirements for ChIP sequencing:

ChIP sequencing involves enrichment of protein bound sequences using a specific antibody and then sequencing the enriched DNA fragments by using a high throughput sequencing methodology. Here is the detailed discussion on what components are the important pre-requisites for a ChIP-seq experiment.

1. Antibody

ChIP relies on very specific and strong antibodies to perform the pull-down of the protein cross-linked to the DNA in a chromatin state. The enrichment of the protein bound-DNA fragments maybe hampered by the unavailability of the epitopes as the protein is bound to the DNA. Therefore, to ensure successful immuno-precipitation, the preferred kind of antibodies is the polyclonal type. They bind to multiple epitopes and are efficient in high affinity pull-down experiments.

Antibody specificity and quality is governed by two factors: one is the reactivity towards the protein of interest by binding to multiple epitopes and second is the minimal cross reactivity to other proteins (Landt et al, 2012). Therefore, it is imperative that the antibodies are characterized well so that the reagent itself is not a limiting factor in the efficiency of the ChIP.

Ubx contains a well-conserved DNA-binding homeodomain, which is also present in many other proteins within the same organism. Therefore, polyclonal antibodies were raised against the N terminal region of the Ubx protein, which is specific to the Ubx and does not contain epitopes to the C terminal regions, which consist of YPWM and homeodomain motifs (Fig 2.7a).

2. Tissue: Wing buds

In *Bombyx*, the fore- and hindwings are the flight appendages of the thoracic segment T2 and T3, respectively (Fig. 2.7). The wing discs/buds of *Bombyx* develop as flat bilayered epithelial buds that resemble miniature adult wings. The wing buds are small in the first four instars and in the fifth and last instar they grow rapidly. The wing venations are clearly visible from the late fourth instar onwards. This bud like mode of wing development is an ancestral mode, which is also reported in Hymenoptera (Macdonald et al. 2010).

Expression patterns of few developmental genes that regulate wing disc development in *Drosophila* have been studied in Lepidoptera, mostly through butterfly as a model system. Expression of some of the developmental markers is also known through a preliminary study on *Bombyx* wing buds (Singh et al.2001). We also explored the development of wing buds right from the second instar of the *Bombyx* larvae till fifth instar to understand the morphological changes and feasibility of acquiring enough tissue for ChIP (at least 10^6 cells).

Based on these studies and directly observing the morphology of wing buds in *Bombyx*, we decided to use the late fourth instar of the *Bombyx* larva as an equivalent of late third instar larval wing imaginal disc in *Drosophila* for ChIP.

3. Silkworm races for ChIP

As the approach used for identification of the direct targets was a sequencing based method, we relied heavily on the silkworm genome information available on public databases to assign the binding region and locate the targets. We planned experiments in India based on the races of silkworms that were locally maintained. For this, we had to ensure that the silkworm races available in India were very close to the actual sequenced races in China (Dazao) and Japan (Daizo p50T). The Daizo available in India has been brought from Japan to the germplasm center about 50 years ago, so we had to ensure that sequences from this strain also matched to the available genome databases.

We PCR-amplified and sequenced both exonic and intronic regions of *Cytoplasmic Actin A4* and *cubitus interruptus (ci)* from the *Bombyx* races Daizo (multivoltine) and C108 (bivoltine) and compared the sequences to that of the genome databases. Both exon and the more variable intron regions were found to be highly similar to the sequences in the genome databases, with identity of at least above 92% for most of the regions sequenced. This experiment ensured that the races available here in in India were indeed suitable to carry out sequencing-based approach to identify the genomic regions in a ChIP experiment.

Using the polyclonal antibodies we carried out ChIP on nuclear lysates from wing buds of *Bombyx* (race: Daizo) as described in the section below.

Materials and Methods

2.6 Silkworm race used for the study

Silkworms were maintained in Centre for Sericulture Research and Training Institute (CSR&TI) at Mysore, India at 25ºC (Fig 2.4). They were reared on Mulberry (*Morus alba*) leaves as feed. Initially two races (Daizo and C108) were obtained from Central Sericultural Gemplasm Resources Centre (CSGRC), Hosur, India.

The eggs of silkworm can be stored at 4°C for a month before the release of the larva. The eggs from bivoltine race like C108 need to be treated with acid mixture before storing in order for them to be hatched successfully. In case of multivoltine eggs of Daizo, an exposure to light is sufficient to hatch to the 1st instar larvae. From early first instar to third instar, mulberry leaves, uniformly cut into small pieces, are used to feed the silkworms. The size of the cut leaves is important as the silkworm is entirely domesticated and the nascent larvae are unable to feed on intact leaves. However, this brings in the problem of drying of leaves and the need to constantly change the feeder beds. Therefore, once the larvae reach late first instar with well-developed mouthparts, larger leaf pieces are used. After the fifth moult, entire leaves are used to feed the larvae.

The larvae were transferred to new leaf bed twice every day. The leaves were surrounded with moistened sponge to maintain the humidity and freshness in the leaves. The larvae were not fed on the day they moulted into the next instar. This is recognized by the reduction in size of the mouth parts and change in the color of the skin (in late instars) before they actually moult to the next instar. Afterthe fifth instar, the skin of the larvae turns yellowish and the mouthparts reduce as the larva readies itself to go into the pupation by starting to weave a cocoon around it. The larvae develop into a moth from the cocoon in about 10 days. Generally the pupae are sex separated and kept in mating cups to harvest the eggs from the moths. The moths emerge out of the cocoon, they mate, lay eggs and are short lived. The eggs can be stored at 4°C until the next hatching.

The race Daizo is the closest race to the race used for sequencing the genome, both in China (Dazao) and Japan (Daizo p50T) and it is a multivoltine race which can be continually cultured without any dormancy periods. Daizo larvae have a pair of dark crescent markings on their 5th segment and a pair of star spots on the 8th segment and an eyespot on the 1st segment (Fig 2.6a) (Yamaguchi et al, 2013, Nie et al, 2014). C108, is a bivoltine strain used for silk production, it can only be cultured twice a year.

2.7 Comparison of sequences from silkworm races to that of Genome database.

Intronic and exonic sequences of these two races, namely Daizo and C108, were used to verify if they differed from the genome sequence available on the public databases. Two genes (*cubitus interrruptus, ci* and *Cytoplasmic Actin A4*) with well described gene features (exon, intron, CDS, stop etc.) in *Bombyx* were used to test the sequences of the two races by amplification and sequencing. The eggs of these two races were brought from Mysore for the genomic DNA extraction. However, after sequence comparisons to database only the Daizo larvae were used for the ChIP experiment.

2.7.1 Extraction of Genomic DNA from silkworm eggs.

The protocol for extraction of genomic DNA from eggs was modified from Nagata et al. (1996). Whole eggs (50-60) were frozen in liquid nitrogen, crushed with a mortar and pestle and the ground powder was suspended in 20mM Tris HCl (pH7.5), 100mM NaCl, 1mM EDTA, 2% SDS and 0.5% Sodium lauryl sarcosinate. The lysate was vortexed at 1500 rpm for 15' then it was centrifuged at 10,000 rpm for 10' at RT.

Supernatant was taken into a fresh microcentrifuge tube and mixed with 1:1 volume of Phenol: Chloroform mixture and mixed thoroughly. It was then centrifuged at 5000 rpm to separate layers; upper aqueous layer was aspirated and followed with three such Phenol: Chloroform extractions. The upper aqueous phase was mixed well with 1:1 volume of Chloroform and the step was repeated twice. The DNA in the upper aqueous layer was precipitated with 1/10

volume of 5M Potassium acetate with 2 volumes of chilled ethanol by keeping the mixture at -20ºC overnight.

On the next day, the precipitated DNA was pelleted at 14000 rpm at 4ºC for 15 minutes. The pellet was washed with chilled 70% ethanol twice. The pellet retains a brown color due to the pigments present in the egg coat, which was not removed and it was found not to affect the downstream processing of DNA and amplification.

The pellet obtained was dissolved in 20mM Tris-HCl (pH7.5), 100mM NaCl, 1mM EDTA, and 0.5% SDS. RNA was eliminated by treating twice with 2M LiCl for one hour at -20ºC. Pellet was obtained by centrifugation at 14000 rpm for 15 minutes at 4ºC and washed twice with chilled 70% ethanol. The pellet was finally dried at room temperature and dissolved in 20µl TE buffer (Tris-HCl (pH 8.0) and 0.1mM EDTA) and stored at -20ºC.

2.7.2 Comparison of the sequences to Genome database.

To compare the DNA sequences of the locally available races to the genome database we selected two *Bombyx* genes (*Ci* and *Actin A4*) with well described gene features. The intention was to compare sequences of the more variable regions, the introns, between the sequence from local races (Daizo and C108) and the genome database and see what identities are retained. Sequences of primers used are shown in Table 2.1 in Appendix Chapter 2. Amplified regions were sequenced in two sequencing reactions, one each for left and right primers.

The sequences obtained were subjected to BLAST analysis against the BGI SilkDB database (Xia et al, 2004).

2.8 Identifying silkworm wing buds and the appropriate larval stage for ChIP

The method to isolate larval wing buds in *Bombyx* and their location in the larval body was derived from studies on segmental transplantation of wing buds

in fifth instar from a Japanese group (Fig 2.6 b, Hojyo and Fujiwara, 1997). Therefore, initially wing bud isolations were practiced on fifth instar larvae; however we observed that at this stage the wing venation patterns were prominent. Previously, our lab had used the late third instar larval *Drosophila* wing/haltere discs to carry out ChIP experiments, as many regulatory genes are known to contribute to pattern formation at this stage, which is critical for further differentiation of wing/haltere disc into adult wing/haltere.

In order to find a comparable stage of *Bombyx* wing buds, we dissected out wing primordia on each day from second to fifth instar to observe and determine the stage that would be appropriate to explore role of Ubx in hindwing development. The dissection was performed by making an incision in the center of the segment 2 for fore wing and segment 3 for the hind wing (Fig 2.6a) and then the wing bud which in the fourth instar looks like a tissue globule with a translucent white color was identified (Fig 2.6c). The wing bud is attached to the body by a trachea that passes from anterior to posterior through the proximal end of the bean shaped wing disc. The wing bud was released by cutting the tracheae with minimal amount of it remaining in the bud. The bud was then transferred to PBS with protease inhibitors on ice.

Expression patterns of few developmental genes that regulate wing disc development in *Drosophila* have been studied in Lepidoptera, mostly through butterfly as a model system. Expression of some of the developmental markers (Fig. 2.6 d,f; Dll and Wg) is also known through a preliminary study on *Bombyx* wing buds (Singh et al.2001). Based on these studies and the direct observations of the morphology of wing buds in *Bombyx*, we decided to use the late fourth instar of the *Bombyx* larva, as an equivalent of late third instar larval wing imaginal disc in *Drosophila* for Chromatin Immuno-precipitation (ChIP).

2.9 Generation and validation of Antibodies

In order to carry out a ChIP experiment, we had to raise antibodies specific to Ubx in *Bombyx*. The sequence of the *Drosophila* Ubx homolog in *Bombyx* was published by a Japanese group studying homeotic deletion mutants (Ueno et al,

1992). The sequence was analyzed in comparison to sequences of other known insect Ubx homologs, to clone and express specific regions of Ubx protein.

From multiple alignment (ClustalW) analysis with the known insect Ubx homologs, it was found that first seven amino acids of the protein and the C-terminal Homeodomain motif are highly conserved. In the first step, cDNA was synthesized with primers (forward Bom NdeI and reverse Bom Rev; for sequences see Table 2.2 in Appendix C2) corresponding to these regions, which would amplify the *Bombyx* N terminal region along with the initial homeodomain region including the YPWM region as well. In the second step the reverse primer (Bom Rev2) was replaced with another primer designed to exclude both the conserved YPWM and the initial homeodomain region and the insert for the expression of *Bombyx* specific N-terminal Ubx protein was obtained. This insert was cloned into a pET15b protein expression vector with a 6X Histidine tag.

Antibodies generated earlier in our laboratory against the N terminal Ubx protein of *Precis coenia* were also used in this study. The N terminal region of the *Precis* Ubx has 98% sequence identity to the N terminal region of *Bombyx* Ubx (Fig 2.8a). These antibodies were used in experiments to standardize Immunohistochemistry and Immunoblotting experiments in the initial attempts.

2.9.1 Protein Expression and Purification

The *Bombyx* N terminal Ubx expression construct was transformed into *E.coli* BL21 DE3 strain, $CaCl_2$ competent cells, by heat shock method. The transformants carrying the desired expression plasmid were used for expression. A single transformed colony was inoculated into 5ml LB Amp broth and incubated at 37°C in orbital shaker incubator at 250 rpm and grown overnight. This pre inoculum was further inoculated into 1000 ml of Terrific broth (TB, with 100µg/ml Ampicillin selection) and incubated in an orbital shaker incubator at 37°C, 250 rpm till the OD Abs600 reached 0.5. The culture was induced with Iso Propyl Thio Galactoside (IPTG) to final concentration of 1mM and incubated at 37°C, 250 rpm for 4 hours in an orbital shaker incubator. The cultures were harvested by centrifugation at 8,000rpm. Protein over expression

was checked by 12% SDS PAGE. The cell pellet was resuspended in 10ml of Lysis buffer (100mM NaH_2PO_4, 100mM Tris Cl, 8 M Urea, pH 8). This suspension was subjected to sonication (50 KHz, 2min pulse) in ice just until the suspension turned transparent. The sonicated suspension was centrifuged at 15000rpm for 15min.

As the *Bombyx* N-terminal Ubx protein was expressed with a 6x His-tag, it was purified by using a Qiagen® Ni-NTA agarose column. The Ni-NTA agarose was centrifuged to obtain a pellet, which was washed in PBS (pH8), three times. 500ul of Ni-NTA pellet was suspended in the bacterial lysate. The tubes were sealed and placed in rotor for uniform Ni-NTA -protein binding at room temperature for 1 hour. The column bound protein suspension was allowed to settle under gravity in a gravity flow column. This column was washed with 5 column volumes of wash buffer (100mM NaH_2PO_4, 100mM Tris Cl, 8M Urea, pH 6.3). Then 2 sets of elution buffers (pH 5.3 and then pH 4.3) were used twice each to collect the elutions in separate microcentrifuge tubes and stored at 4C. These elutions were then analysed by SDS-Polyacrylamide gel electrophoresis.

The protein obtained after NiNTA elution had a higher molecular weight (~ 60KDa) band, which remained after NiNTA column purification. This could not be gotten rid of even with different standardizations (like imidazole elution/wash time and pH). The protein was hence subjected to further purification by gel excision and elution (described in Kosman et al, 1998), where the protein was run in a preparative well in a large gel (22 X 17 cm area and 3cm gel thickness). The gel after run was immersed in cold 0.2M KCl solution at 4C for 2 minutes, just when the highly intense protein bands turned white due to precipitation with KCl the gel was removed from the solution and the protein band of interest was excised. This band was cut longitudinally into smaller fragments and placed along with tris glycine SDS running buffer (with 0.2% SDS, in a dialysis tubing (12 KDa cut off). This setup was placed in horizontal gel electrophoresis unit with running buffer and run at 120 V for 2 hours. The gel pieces were then removed and placed in the buffer solution. This mix was dialyzed against sterile MiliQ® water with a step-down gradient of

running buffer to replace the buffer with water completely at 4ºC overnight. This dialyzed protein solution was concentrated by centrifugation in vacuum (on a speed-vac) from an initial volume of 3 ml to final volume of 1 ml. The protein was checked on SDS-PAGE followed by Western blot hybridization (with anti-His and anti- *Precis* Ubx antibodies).

The purified protein was run on a SDS-PAGE and the expected 17 KDa band was excised and subjected to in-gel trypsin digestion and MALDI-ToF mass spectroscopy analysis to confirm its identity. The band at higher molecular weight (~60KDa) was also subjected to MALDI analysis to confirm its identity as it was also detected by Ubx-specific antibodies on Western blot hybridization.

2.9.2 Immunization and collection of antisera

The electro-eluted and purified *Bombyx* N-terminal UBX protein (about 400µg), which was dialyzed against water and concentrated by using Amicon® centrifugation, was mixed thoroughly with equal volumes of Freund's complete adjuvant and injected subcutaneously into a healthy rabbit as the first immunization booster. Subsequently after a month the next booster was given with same amount of protein, now mixed with equal volumes of Freund's incomplete adjuvant. After seven days first test bleed was taken, a serum was prepared after the overnight coagulation at 4ºC.

This serum was tested against purified recombinant protein and was found to work at 1:5000 dilutions. Two days hence (9th day post second booster) 15 ml of first bleed was taken and processed to obtain the 1st set of anti *Bombyx* N-terminal Ubx antisera. The processing was done by coagulating the blood collected, at 4 ºC overnight and centrifuging at 5000g for 10 minutes. The blood debris was pelleted down and the supernatant serum was collected, aliquot and stored at -80ºC and -30ºC freezers. One month after the bleed another booster of 400µg protein was mixed with Freund's incomplete adjuvant and administered to the animal. 10 days after booster-3, 15 ml of bleed was taken from the rabbit, processed and stored as above.

The antisera was validated by immunoblotting using post induced bacterial lysate expressing the recombinant Ubx protein, purified N terminal protein, embryonal and larval lysates. The antisera were also used to test by immunohistochemistry on fore and hind wing buds of *Bombyx*.

2.9.3 Purification of antisera

In order to use the antibody for a ChIP experiment, we needed a highly specific and pure form of the antisera with very high titer. We used a Protein-A column to purify the IgG fraction from the antisera to be employed in a ChIP experiment. Prosep-A Protein-A column from Milipore® Montage was used to purify the anti Ubx antiserum raised in rabbit. The purification was carried out according to the manufacturer's instructions. PROSEP-A media was equilibrated with 10 mL binding buffer by centrifuging the spin column at 500g for 5 minutes. 10 ml serum was then pre-cleared by filtering through a 0.22 μm filter. The filtered sample was then diluted 1:1 (v/v) in binding buffer. The diluted serum was loaded on the spin column and centrifuged at 150 g for 20 minutes at 4ºC. The spin column was then washed with 20 ml of binding buffer by centrifuging the spin column at 500 g for 5 minutes at 4ºC. The bound IgG was now eluted with 10 mL elution buffer EB2 (higher pH) into a fresh centrifuge tube containing 1.3 mL neutralization buffer to bring the sample to neutral pH by centrifuging the column for 5 minutes at 500g at 4ºC. Elution was done once more with elution buffer EB1 to see if any antibody elutes at low pH. These IgG fractions were tested on SDS-PAGE and stained with Comassie Brilliant Blue (CBB) dye.

The IgG fraction was concentrated to 1ml volume using Amicon® Ultra 15 centrifugal device with 30000 NMWL. The concentrated antibodies were stored at -80ºC for long term storage; the working stock was kept at 4ºC. The antibodies were quantified by measuring their absorbance at 280 nm on a NanoDrop® spectrophotometer.

The purified antibodies were validated by Western blot hybridization against purified N-terminal protein, embryonal, total larval and wing disc lysates.

2.9.4 Validation of the Antibodies by Western blot hybridization

Using protocol as described below, the IgG purified anti-N terminal *Bombyx* Ubx was tested by Western blot hybridization to detect Ubx before proceeding to any ChIP experiment. Lysates of *Bombyx* forewing, hindwing, larva (1st instar), embryonal lysate and purified *Bombyx* N-terminal Ubx protein was blotted on membrane to validate the titre and specificity of the antibodies.

Even before the actual *Bombyx* antibody could be raised and tested, availability of antibodies against butterfly (*Precis coenia*) Ubx allowed us to detect the Ubx protein in larval and bacterial lysates. The N terminal *Precis* Ubx region is 98% similar to the N terminal *Bombyx* protein, and all the antigenic sites are exactly the same as tested by the *in silico* antigenicity (NCBI) tool. We also used antibodies against the *Drosophila* N-terminal Ubx as negative control as this region of Ubx is not conserved between the two species.

The larval, adult, and embryonal lysates were prepared by crushing 10 larvae/embryos/adults/ in 100 µl of RIPA lysis buffer with Roche® Protease inhibitor cocktail. 40 wing buds were used to make the wing bud lysate in 20µl lysis buffer. The lysate were left on ice for 20 minutes for complete lysis and sonicated for 15 minutes at maximum wattage with a 30 sec on/off cycle on Diagenode® Bioruptor® water bath sonicator. The lysates were boiled for 10 minutes on a heating block with SDS loading buffer, and 5µl of such lysate was loaded on a 12% SDS PAGE gel. Both the blots included lanes loaded with *Bombyx* embryo and larval lysate, *Drosophila* larval lysate, purified *Bombyx* N terminal protein, and BSA (negative control). SDS-PAGE electrophoresis was carried out as per the standard procedure.

After electrophoresis was completed, the proteins were transferred onto a PVDF membrane to detect the protein by immune blotting. In order to carry out the transfer, the PVDF membrane was first equilibrated in methanol for 5 min at room temperature then both gel and membrane were rinsed well in the Western blot transfer buffer. The gel and membrane were then assembled in a sandwich with Whatman® filter paper pads and the protein was transferred in a submerged transfer apparatus at a constant current of 90mA for 12 hours at 4ºC. The

membrane was then removed and then blocked with 3% BSA in TBST for three hours at RT. Primary antibodies were added to the blot at 1:2500 dilution and kept at room temperature on shaker for 2 hours. The blot was then washed three times with TBST for 20 minutes each. Secondary antibody (HRP conjugated, anti-Rabbit) was added in TBST containing 3% BSA at 1:10,000 dilution and incubated for 2 hours at room temperature. The blot was then washed three times with TBST for 20 minutes each. The blot was layered with 500 µl of activated Milipore® Immobilon™ chemi-luminescent HRP Substrate and visualized on Fujifilm® LAS-4000 chemi-luminescent imager.

2.9.5 Validation of anti-Ubx antibodies by Immuno-histochemistry on *Bombyx* wing buds

Immuno-histochemistry with fourth instar larval buds was attempted initially with modifications on protocols used on butterfly wing buds and *Drosophila* wing discs. Experiments carried out in Japan with the protocol from Dr. Fujiwara lab worked better and is as described below.

Both fore and hind wing buds were dissected from fourth instar Daizo larvae and collected in PBS with protease inhibitors on ice. They were fixed in 4% formaldehyde (500 µl) at room temperature for 30 minutes on a rotospin rotator at RT. The buds were washed in 1ml of 0.5% PBTx for 30 minutes at RT. The buds were then washed three times in PBS with 0.01% Saponine (PBSS) for 10 minutes each at RT. The buds were blocked in 1.5% (500 µl) Roche® blocking solution for 30 minutes at RT on a rotospin rotator. The buds were left overnight in primary antibodies (anti N-terminal *Bombyx* Ubx antisera c37) at 1:500 dilution in PBSS at 4ºC. The buds were washed three times in 500 µl PBSS for 20 minutes each at RT on a rotospin rotator. The buds were then incubated in Alexa Fluor® anti rabbit 488 secondary at a dilution of 1:100 and 0.5 µl DAPI in 1ml PBSS for 1 hour, the tubes were kept stationary at RT. The buds in foil covered tubes were washed three times in 500 µl PBSS for 20 minutes each at RT on a rotospin rotator and stored at 4ºC. The buds were mounted in 50% glycerol with elevated coverslip and imaged on a confocal microscope.

Now that the antibodies were purified and validated by two methods, they were used to carry out the ChIP experiment.

2.10 Chromatin Immuno-Precipitation (ChIP)

Once the standardization at the first level of larval stage, tissue and antibody were completed, the next level pertained to the actual ChIP experiment. As this was the first ChIP experiment on Lepidopteran wing buds, Fixing, Shear size of chromatin, reduction in pulldown noise and the pulldown conditions were the components that needed to be standardized next.

2.10.1 Standardization of ChIP conditions

We encountered two problems when we ventured to carry out a ChIP experiment on the *Bombyx* wing buds, the first one being whether the DNA we are analyzing is in good quantities for a ChIP-seq and the second whether the modified protocol in use indeed does an efficient pulldown. As we did not have any known targets of Ubx in *Bombyx* and the promoter regions not very well defined in the genome, we could not ascertain the quality of ChIP before actually sequencing it. Hence, we carried out certain experiments to answer these questions and to systematically to sort the issues before actually going ahead with the ChIP and the subsequent more expensive high throughput sequencing.

1. Confirmation of the quantity of DNA for the ChIP pulldown

Though we used a good quantity of tissue ($9x10^6$ cells) in the initial experiment for a pulldown, we noticed that quantity of DNA was very low (about 40 ug is used in a ChIP experiment in general but we obtained around 5 ug which was insufficient). The initial experiments involved the usage of GE® Protein-A sepharose CL4B which was blocked in BSA. However it is known that these sepharose beads do carry some noise in the form of non-specifically bound chromatin to the beads. We used this defect to our benefit in order to detect if there is indeed enough *Bombyx* DNA in the starting input material, to amplify DNA fragments and if it is enough for a ChIP pulldown. We designed primers

against the coding regions the gene *Spalt (BmSal)* in *Bombyx*, in order to amplify material from post-ChIP processing and see if the DNA can give amplicons and if these amplicons are indeed the expected sequences.

PCR was carried out to amplify these regions in all three ChIP processed segments, the input, pulldown and negative control. The amplified DNA was then run on a 1% Agarose gel electrophoresis to excise the faint DNA amplicons. This amplified DNA was purified using Qiagen gel elution kit according to manufacturer's instructions and the DNA obtained was sequenced to confirm the identity of the amplified fragment.

2. Validation of the modified ChIP protocol

In order to validate the modified protocol for ChIP we used anti *Bombyx* GATA factor antibodies (A gift from Prof K. Iatrou, Athens) to do a ChIP pulldown and verify the known targets of BmGATA factor from wing bud lysates through PCR. The BmGATA factor is known to bind to chorion gene promoter region in the chromatin from a study on *Bombyx* ovarian follicles (Papantonis & Lecanidou, 2009). Two primer sets were designed against *Hcp13*A/B gene pair promoter region (*Hcp13*) and *Erp1*A/B promoter region (*Erp1*) to verify the ChIP pulldown. We used these primer sets in a low-cycle PCR to see if we can validate for ChIP pulldown in wing buds on BmGATA pulldown against a normal IgG negative control.

3. Other standardizations

ChIP generally requires fine tuning in standardizing the experiment and differs between different systems and based on various factors.

The first important step to keep in mind is optimization of the fixing conditions, as it determines the crosslinking strength. As we encountered inefficiency in antibody penetrance and there were doubts on the fixing of wing buds as well, we resorted to nuclear purification of the wing buds. This step ensures that fixing is done efficiently and also a cleaner ChIP where there is no protein

surplus from the cytoplasmic fraction. The nuclei were checked by DAPI just after nuclear extraction.

The second step that is of concern after fixing is the sonication. Firstly there was a choice between probe sonicator and a water bath based sonicator. As the efficiency of a water bath sonicator is better than the probe based sonicator and comparatively lesser amount of heating of sample occurs, we used Diagenode® Bioruptor ® XL (UCD500) for sonication. After many attempts on the Bioruptor®, the shear-size was brought to at a range of 100-300 bp, which is suitable for ChIP sequencing.

The third measure that we had to take was due to the noise we saw in the pilot experiments with GE® Protein-A Sepharose CL4B. In spite of blocking with BSA the matrix is known to carry a lot of background chromatin. When the final DNA extraction is done by Phenol Choloroform extraction, the organic remains may be difficult to completely get rid of, which affects the sensitive sequencing protocols. To overcome both these issues we switched to Invitrogen® Magnify™ Dynabead™ based Magnetic ChIP kit, which included Protein-A column for immuno-precipitation step and also magnetic bead based DNA extraction step.

With these measures of standardization done, we proceeded to the actual ChIP experiments on fore- and hindwing buds of *Bombyx* with the IgG purified polyclonal anti *Bombyx* N- terminal Ubx antibodies.

2.10.2 Chromatin Immuno-precipitation protocol

The protocol used for ChIP experiment was modified from the study on a modified ChIP protocol for BmGATA factor in *Bombyx* ovarian follicles (Papantonis & Lecanidou, 2009). The experiment involved extraction of nuclei from wing buds, sonication, immuno-precipitation and DNA purification. The protocol is divided into smaller sections:

1. Preparation of Chromatin

A. Nuclei preparation

80 fore- and hindwing buds each were dissected out from mid-fourth instar Daizo larvae by making incisions along the center of T2 (fore) and T3 (hind) segment. During the dissections they were stored on ice in 500µl PBS with protease inhibitors. The wing buds were then washed with HEPES low salt buffer by gently inverting 3-4 times. Then the wing buds were homogenized in 500 µl of fresh HEPES low salt buffer using a motorized homogenizer with autoclaved pestle. The lysate was then passed through a 100 micron filter to remove large debris and to isolate the nuclei. The flow through was collected and allowed to lyse on ice for 20 minutes. Centrifugation at 3000g for 5 minutes at 4°C was used to to pellet the nuclei from the lysate, which were verified with DAPI stain in 2 µl of sample. The pellet was washed gently once each with HEPES low salt buffer, centrifuged, followed by a wash with PBS with protease inhibitors (PBSpi). Freshly prepared 500µl of 1% formaldehyde (Sigma®) in PBSpi was added to the pellet and pellet resuspended. Fixing was allowed occur for 12 minutes at RT on a rotospin rotator. The fixing was stopped by adding Glycine to a final concentration of 1.25 M for 5 minutes at RT. The nuclei were pelleted and washed with PBSpi.

B. Lysis and sonication

The ChIP experiment was performed using a modified Invitrogen® Magnify™ ChIP kit protocol with the reagents provided in the kit. The washed nuclei pellet was lysed using nuclei 200µl of Invitrogen® Magnify™ lysis buffer (with Invitrogen® Magnify™ Protease inhibitors). The tubes were left on rotator for 20 minutes at RT for complete lysis to occur and a sample was tested with DAPI stain for complete lysis ie, the absence of intact nuclei. The lysate was sonicated on Diagenode® Bioruptor XL for a total time of 15 minutes with ’55 sec on /60 sec off’ cycle at high power. The sonicated lysate was centrifuged at 4°C for 10 minutes at 18000g. The supernatant contained the sonicated chromatin which was split into two tubes, one for input (40µl) and one portion (160µl) for the

ChIP experiment with negative control. Both the tubes were snap-frozen in liquid nitrogen and stored at -80°C.

2. Immunoprecipitation

A. Binding antibodies to the magnetic protein-A/G Dynabeads®

Dynabeads® were resuspended by pipetting up and down gently. As the ChIP was done with both fore- and hindwing samples, four tubes were prepared and to each tube 100 µl of cold dilution buffer and 20 µl of resuspended beads were added. The tubes were placed on a magnetic rack for 1 min to remove the buffer and fresh dilution buffer was added. Antibodies were added to each of these tubes. Two tubes each with purified anti N-terminal *Bombyx* Ubx antibodies and negative control normal Rabbit IgG. The tubes were flicked to mix and rotated on a rotospin rotator for six hours at 4°C. Then the tubes were placed on a magnetic rack to remove unbound antibodies and one more gentle wash was given with 100 µl of dilution buffer.

B. Diluting the Chromatin and processing

In two tubes, 50 µl each of hind wing chromatin was diluted with 150 µl of dilution buffer, one for anti-Ubx pulldown and one for IgG negative control. Similarly forewing chromatin tubes were also prepared. To these tubes 20 µl of antibody-bead complex was added and kept overnight on rotospin rotator at 4°C.

All the tubes were placed on the magnetic rack and the unbound supernatant was discarded. The tubes were then washed three times with IP buffer 1 (low salt wash) and twice with IP buffer 2 (high salt wash) for 5 minutes each at 4°C.

After the final IP buffer 2 wash, the beads and the inputs kept aside were resuspended in 50 µl of de-crosslinking buffer with 2.5 µg RNAse (Roche®) and incubated at 37°C for two hours. Then the tubes were heated at 65°C on a thermal block for 8 hours for de-crosslinking. The tubes were vortexed thoroughly to dislodge the complexes and then kept on magnetic rack to aspirate the supernatant that contained the antibody enriched chromatin. At this step 10 µl of the elute was retained for ChIP western. This chromatin was treated with 1

µl of Magnify™ Proteinase K at 55°C for two hours to get rid of the proteins. After this step the tubes were cooled on ice for 5 minutes and preceded for DNA purification.

C. DNA purification

DNA purification beads were resuspended by brief vortexing. A 70 µl DNA purification mixture was prepared with 20 µl of resuspended DNA purification beads and 50 µl of DNA purification buffer for each tube. To each of the six samples 70 µl of the DNA purification bead was added and pipetted gently to mix and the tubes incubated at RT for 5 minutes. The tubes were placed on magnetic rack for a minute and then the supernatant was discarded. The beads were washed twice with 150 µl DNA wash buffer by pipetting 5 times.

Post-wash, the beads were pelleted on the magnetic rack, the wash buffer was removed and 150 µl of DNA elution buffer was added, pipetted 5 times to mix. The mixture was incubated at 55°C for one hour on a thermal block. The contents were cooled placed again on magnetic rack and the eluate was stored in a fresh tube. Once again 150 µl of DNA elution buffer was added incubated for one hour at 55°C and a second elute was recovered. The two elutes were pooled and concentrated by speed vac centrifugation from 300 µl to 50 µl.

The DNA so obtained was quantified on a Nanodrop® or Qubit® spectrophotometer and analyzed with a bioanalyzer. Based on the schedule of the sequencing run it was sometimes stored at -80°C or sent to high throughput Illumina sequencing directly.

Results and Discussion

2.11 Silkworm race used for the study

Silkworms (*Bombyx mori*) for this study were maintained at Centre for Sericulture Research and Training Institute (CSR&TI) in Mysore, Karnataka. The eggs were obtained from Central Sericultureal Germplasm Research Centre (CSGRC), Hosur for Daizo and C108 races. The silkowrms were maintained on Mulberry (*Morus alba*) leaves till pupation.

When many silkworms are needed for the experiment, the eggs were sequentially released over a week to keep a multiple staged culture, which would yield the desired number of fourth instar larvae continuously.

The race Daizo was regularly obtained from the cultures maintained at Centre for Sericulture Research and Training Institute in Mysore. It was cultured four times a year and the fourth instar larvae were shipped to IISER Pune for experiments in our laboratory.

The eggs of races Daizo and C108 were obtained for extracting genomic DNA to sequence and compare both the exonic and the variable intron regions to the genome databases. Genomic DNA was extracted from the Daizo and C108 silkworm eggs. It was run on a 1% agarose gel and visualized before proceeding to PCR amplification. The Daizo extraction was a smear whereas the C108 genomic DNA was a single band. Genomic DNA from Daizo and C108 were used to amplify the region covering two intronic regions (Ci1, Ci2) of *cubitus interruptus*; and the regions covering three introns (Act1, Act2, Act3) of *Actin C4*. The amplified DNA was run on a 1% agarose gel and the amplicons were eluted using Qiagen® PCR purification kit. Each amplicon was sequenced with both forward and reverse primers. The sequences obtained were analyzed on a chromatogram file and the initial and final sequence that did not have a good base calling were clipped off before comparing to the SilkDB BGI silkworm database using the BLAST tool available in their database.

Both exon and the more variable intron regions were found to be very similar to the sequences in the genome databases, with identity of at least above 92% for most of the regions sequenced (Fig 2.5). For all the subsequent ChIP experiments, Daizo was used.

2.12 Identifying silkworm wing buds and the appropriate larval stage for ChIP

Expression patterns of few developmental genes that regulate wing disc development in *Drosophila* have been studied in Lepidoptera, mostly through butterfly as a model system. Expression of some of the developmental markers is also known through a preliminary study on *Bombyx* wing buds (Singh et al., 2001). Larvae in late second instar to early fifth instar larvae were dissected and the wing buds were identified to observe the morphological development of the organ as the larvae developed.

The wing discs/buds of *Bombyx* develop as flat bi-layered epithelial buds that resemble miniature adult wings. They are located just below the cuticle in the center of the segments T2 and T3. A trachea passes through the wing bud at the proximal end of the wing bud. The wing buds are small in the first four instars and in the fifth and last instar they grow rapidly. The wing venations are clearly visible from the late fourth instar onwards. This bud like mode of wing development is an ancestral mode which is also common to Hymenoptera (Macdonald et al., 2010).

Based on the earlier gene expression studies and direct observations of the morphology of wing buds in *Bombyx*, we decided to use the late fourth instar of the *Bombyx* larva as an equivalent of the late third instar larval wing imaginal discs in *Drosophila* (Fig 2.6).

2.13 Generation of Antibodies

2.13.1 Protein Expression and Purification

cDNA corresponding to the N-terminal region of Ubx (excluding the homeodomain) was obtained by RT-PCR and was sub-cloned into pET15b

vector. The clone was sequence-verified using T7 sequencing primer as well as using the primers that were used to amplify the insert from the total cDNA. The sequence was 100% identical to the expected sequence and was found suitable for protein expression. The protein was expressed in *E. coli* BL21DE3. The expression conditions with respect to IPTG concentration and temperature were first standardized before attempting large-scale protein expression. The expressed protein was purified with NiNTA column. The protein was re-purified by electro-elution by a method described in Kosman et al, 1998. Post electro-elution, a pure single band protein was observed of the expected size 19KDa (Fig 2.7).

As *Precis* Ubx has 98% sequence similarity (at the protein level) with *Bombyx* Ubx, we used previously raised antibodies against the *Precis* Ubx to confirm that the recombinant protein expressed is indeed *Bombyx* Ubx by Western blot hybridization (Fig 2.8B).

To further confirm the expressed and purified protein, MALDI-ToF was performed on SDS-PAGE eluted bands. The MALDI peptide peaks obtained had good intensity and matched the theoretically predicted peptide mass fingerprint (PMF) of the N-terminal *Bombyx* Ubx.

A 60 KDa protein was also detected after purification, which was also found to be Ubx protein by both MALDI-ToF analysis and by the Western blot hybridization. It was observed that on storage the 19 KDa band reduced in intensity over time, while the intensity of the 60 KDa band increased. This can be explained if the protein, on long-term storage, form covalent linkage to appear as a complex at a molecular weight of 60 KDa. Therefore for immunization, freshly eluted proteins were prepared and immediately used to raise the antibodies.

2.13.2 Immunization and collection of antisera

One rabbit was immunized and after three boosters, two sets of 25 ml antisera were collected. The serum obtained was tested against the purified protein to

test its specificity. The serum was able to detect the purified protein at a dilution of 1:5000 on Western blot hybridization.

2.13.3 Purification of antisera

10 ml of antisera was purified on a protein-A column to obtain IgG fraction for ChIP. This purified antibody was concentrated to 1ml and showed better sensitivity against the antigen (i.e. Ubx) in both purified and in larval lysate forms. The purified antibody was quantified on Nanodrop® spectrophotometer and had a concentration of 48 mg/ml. This purified form was tested before every ChIP by Western blot hybridization against larval lysate and by Immuno-histochemistry.

2.13.4 Validation of antibodies by Western blot hybridization

Western blot hybridization was carried out on *Bombyx* larval, embryo and *Drosophila* larval and adult lysates. We observed clear single band at 27 KDa in both *Bombyx* larval lysate and embryo lysates (Fig 2.8C). The antibodies did not cross react to the *Drosophila* larval Ubx or BSA pure protein, showing that the N-terminal antibodies are specific to the *Bombyx* Ubx. However, in the embryo lysate, three upper bands were consistently observed at above 60KDa. *Drosophila* N-terminal Ubx antibodies failed to detect the *Bombyx* Ubx in the same lysates, while they clearly detected *Drosophila* Ubx (Fig 2.8B). These experiments proved that the N terminal region is antigenically unique to *Bombyx* and that the antbodies raised from this region is specific to *Bombyx* Ubx.

Purified antibodies showed higher titer and more specificity than the antisera and they were able to detect Ubx in hind wing bud lysate of *Bombyx*. They were hence found to be suitable for ChIP (Fig 2.9D-E).

2.13.5 Validation of the Antibody by Immuno-histochemistry of *Bombyx* wing buds

The outer peripodial layer makes the wing bud in larval stages very difficult to stain with antibody, although DAPI can penetrate this layer and stains the nuclei

very efficiently. We used a modified protocol from Dr. Fujiwara's lab (University of Tokyo) to successfully see Ubx expression in *Bombyx* wing buds. We observed that Ubx is expressed all over the hindwing bud, whereas it is expressed only in the single outer layer, presumably the peripodial membrane, in the forewing buds (Fig 2.10 A-B). The absence of Ubx in forewings is reported for *Precis* (Fig 2.11B, Warren et al.1994). Their immunostaining protocol involved removal of the peripodial membrane, and may be the reason for complete absence of Ubx in the forewing buds.

The distribution of Ubx in fore- and hindwing buds of *Bombyx* is very similar to the Dipteran Ubx expression (Fig 2.11A). In Diptera, Ubx is expressed throughout the haltere disc, while the wing disc has Ubx only in the peripodial membrane (Fig 2.11C), which does not contribute to wing development. *Tribolium* also has Ubx only in the T3 appendage and not in T2 (Whitney et al, 2005). While in *Apis*, which is an ancestral form to Lepidoptera, Coleoptera and Diptera, the forewing buds also express Ubx (Prasad N, 2013). This suggests that Ubx expression from the forewing buds may have been lost since the divergence of insect species from Hymenoptera.

2.14 Chromatin Immuno-Precipitation (ChIP)

No ChIP had been reported previously on any lepidopteran wing buds; hence the protocol was modified from other studies and further standardized.

2.14.1 Standardization of ChIP conditions

The first standardization step was to assure that good quantity of *Bombyx* DNA was present in the lysate prepared for ChIP. The DNA isolated from ChIP samples were first used to amplify *Bm spalt* gene region spanning 472bp (Fig 2.12C-D). The amplified DNA was sequenced and verified by BLAST against the SilkDB database and was found to be in the *Spalt* region which is present in the nscaf2589 scaffold in the position 6,654,523 to 6,680,893 of the genome (Fig 2.12E). This helped us to confirm that the DNA present in the starting lysate is indeed enough in quantity and can be amplified.

As the fixation was found to be insufficient in the immune-histochemistry experiments on wing buds, nuclear extraction was followed in the ChIP protocol to primarily ensure that the fixation occurs efficiently and also there would be less noise as no excess protein from the cytoplasmic fraction would contaminate the lysate (Fig 2.12B).

Sonication conditions were standardized on the Bioruptor XL waterbath sonicator by trying sonication at different conditions and then purifying the DNA from chromatin to check the shear-size on gel. Parameters like time, wattage and pulse on/off time were modified in every trial to achieve the right sized smear of 100-300bp. The parameters were adjusted to reduce unsheared genomic DNA to get a better ChIP pull-down and sequencing. As shown in Fig 2.13 A-B, A chromatin shear with a molecular weight range between 100-350 bp, which was obtained after treating the fixed nuclear lysates to a total time of 15 minutes with 55 sec on-60 sec off cycle at high power, was found to be ideal for a ChIP-seq protocol.

Initial experiments were carried out with Protein-A sepharose, which was found to be inefficient. In a ChIP-on-chip experiment these beads are blocked with BSA and Salmon sperm DNA. Salmon sperm DNA could not be used in our ChIP-seq protocol, as it would be a contaminant in sequencing. After many modifications, finally we decided to use the Invitogen® Magnify™ Protein-A/G magnetic Dynabeads® as they do not carry background chromatin bound to the beads.

Phenol:Chloroform purification of DNA in the final step of ChIP retains organics when purified with the traditional methods. Hence the magnetic Invitogen® Magnify™ DNA purification beads were used as an alternative to get efficient DNA extraction without any contaminants. This DNA was found to comply with the quality norms needed in a ChIP sequencing reaction.

The major concern for the ChIP experiment was the consistent availability of fourth instar larvae. The silkworms are generally reared four-five times a year and a window of only three days in fourth instar larva is suitable for ChIP experiment. Hence whenever the larvae were available, Chromatin was prepared

by the methods described above and stored at -80ºC as separate input and experiment chromatin. The nuclei were checked after purification by DAPI stain and imaged under a Zeiss® Axiovision™ fluorescent microscope as a quality measure to confirm efficient nuclear prep (Fig 2.12B). The shear-size was verified only for the first time when the standardization of sonication was successful, from then on identical conditions were used to make the chromatin.

ChIP experiments with antibodies against *Bombyx* GATA-binding protein were carried out to test the above-described protocol as this protein has at least two known targets. DNA subjected to our ChIP protocol showed enrichment for amplicons of the right size for Erp1 (134bp) and Hcp13 (200bp) (Fig 2.12A).

ChIP-Western was attempted on two occasions and the appropriate band was detected (data not shown). However the quality of blot is not of the standards as in the regular Western blot hybridization carried on larval/wing bud lysates.

DNA enriched by ChIP was quantified mostly using a Nanodrop® spectrophotometer and were always found to be in quantities appropriate for high throughput Illumina® sequencing.

Illumina® GAII genome analyzer was used for the high throughput sequencing. Single end 36 bp long sequencing was done on pulled down DNA, the negative control (no primary antibody, pull down with normal Rabbit IgG) and the input DNA used for ChIP. On an average above 20 million reads were obtained with acceptable quality for each run. These reads obtained were used to align to the genome and locate the genes that may be controlled by Ubx by binding to these regions.

Summary

To summarize, this chapter describes the following

1. Ubx is expressed throughout the hindwing of *Bombyx mori*, while in the forewing it is expressed only in the outer peripodial membrane.
2. Mid fourth instar *Bombyx* wing bud was found to be the appropriate stage to carry out ChIP experiments to understand the role of Ubx in wing development. This stage is equivalent to the late third instar in *Drosophila*.
3. Polyclonal antibodies specific to *Bombyx* Ubx was raised, purified and validated for carrying out ChIP experiments.
4. A protocol was developed for Chromatin Immuno-precipitation and the same was standardized for wing buds and the DNA purified after enrichment was sent for Illumina® sequencing to identify the targets of Ubx in *Bombyx*.

In the next chapter, analyses of the ChIP sequencing reads have been described. It involves a series of quality control measures, alignment of the reads to the genome, identifying the bound regions of chromatin (peaks) and the identification of genes that may be associated with those regions.

Plates
Chapter 2

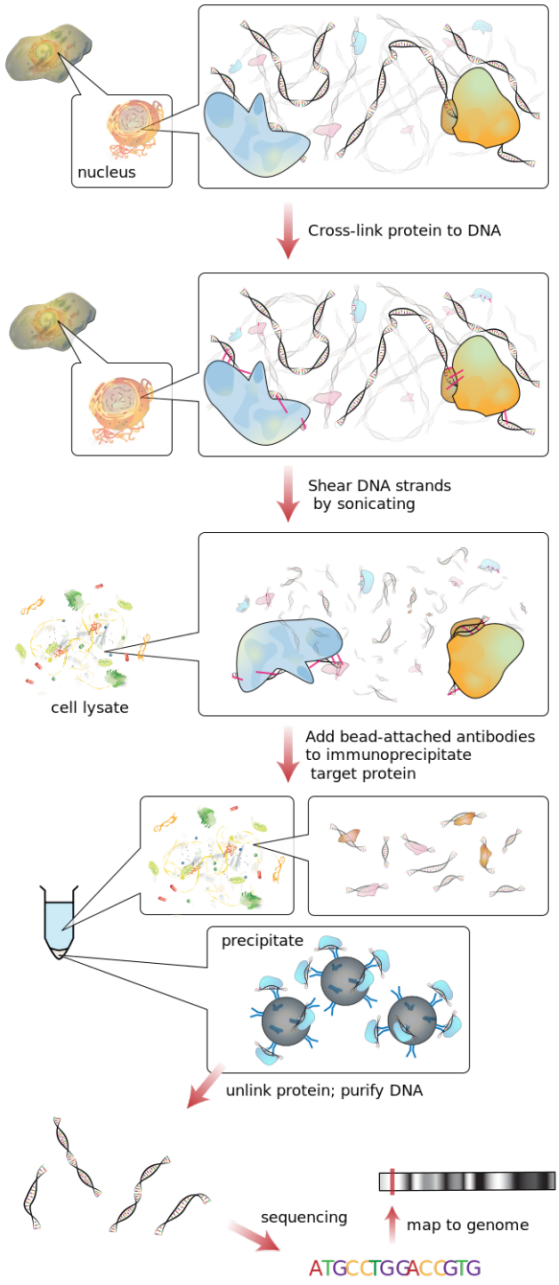


Figure 2.1. An overview of the Chromatin Immunoprecipitation (ChIP) method.

ChIP is done to study DNA-protein interactions by precipitating protein crosslinked to DNA in its chromatin state. It involves (i) crosslinking of the proteins to DNA in the chromatin state using chemical cross linking agents. (ii) shearing of the crosslinked chromatin to fragments of appropriate size. (iii) immuno precipitation of the Protein-DNA complex by using protein specific antibodies. (iv) DNA extraction and sequencing or PCR/microarray based quantitation. (Image: Wikimedia commons).

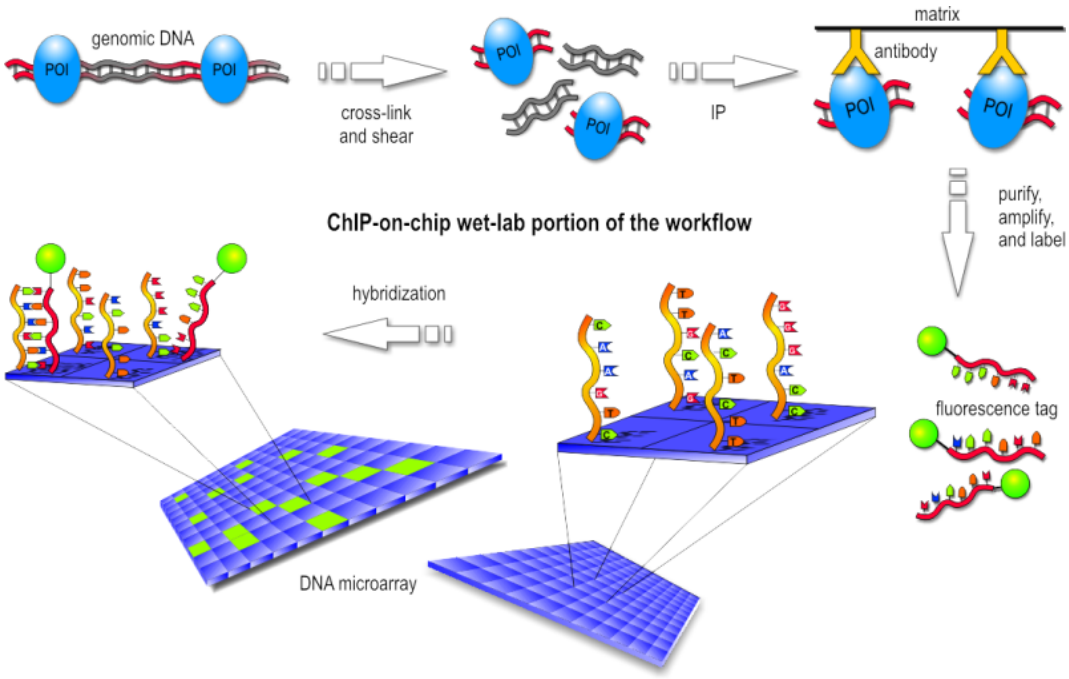
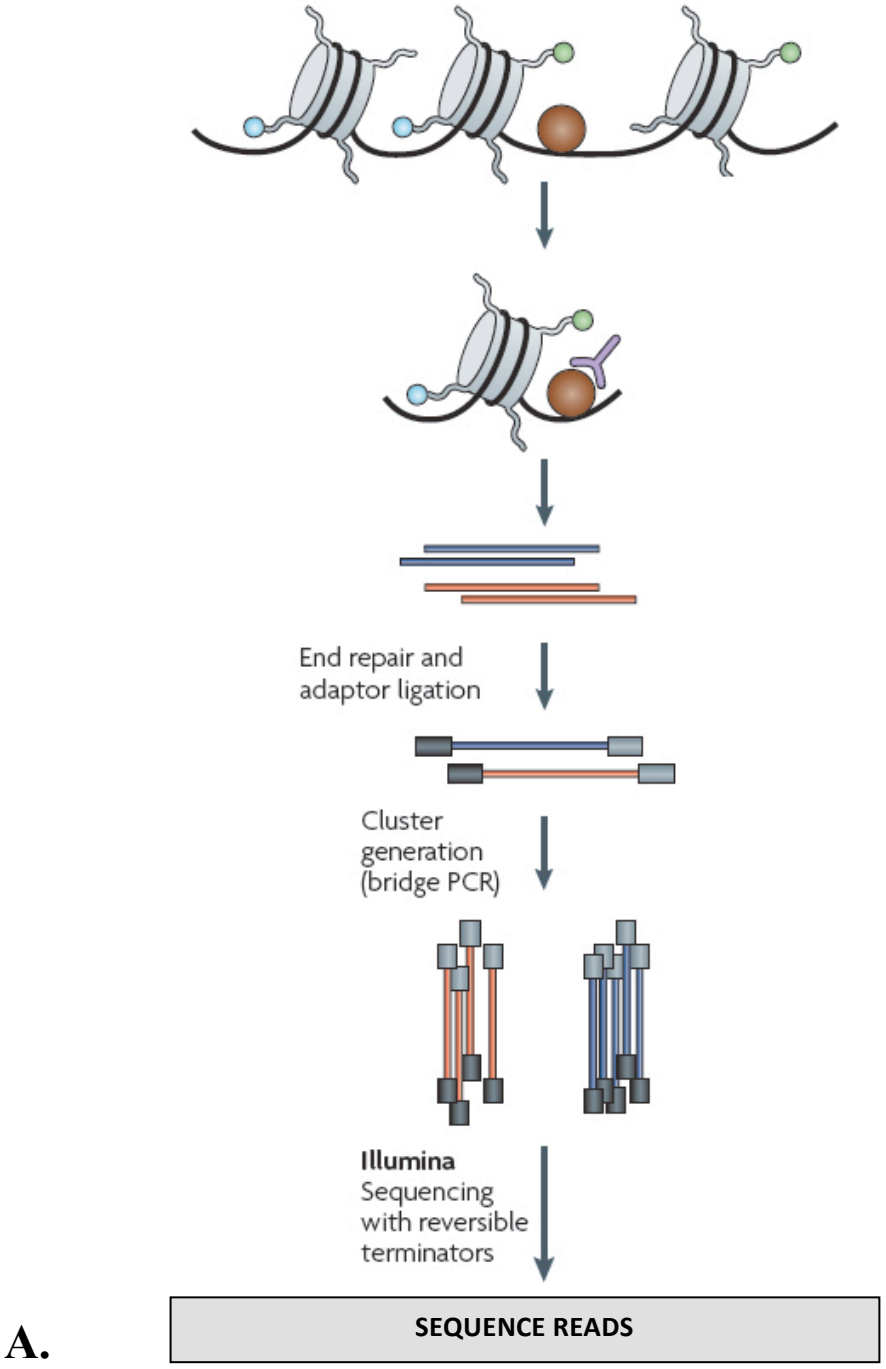


Figure 2.2. An overview of ChIP-chip methodology

Before the advent of next generation sequencing, ChIP samples were hybridized to high resolution microarrays to identify the up/down-regulated genes. The top row shows the ChIP pull down experiment. The purified DNA obtained from a ChIP experiment is labeled with separate fluorescent dyes for control and experiment. These samples are hybridized to a microarray chip containing genome wide regulatory sequences. The excess probes are washed off and the chip is scanned in a genome analyzer. The ratio of fluorescent intensity between control and experiment probes is calculated to obtain relative binding of protein to a regulatory region.



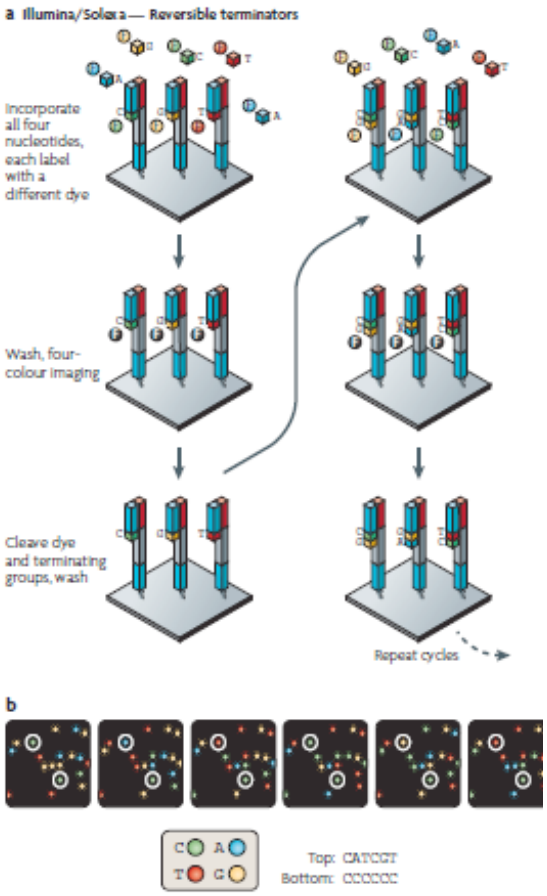
A.

Figure 2.3 A. Illumina® high throughput sequencing chemistry

Illumina® genome analyzer is based upon the principle of sequencing by synthesis chemistry. In this method the ChIP enriched DNA is ligated to adapters and loaded onto a flow cell containing complementary adapters. (Image: Park, 2009, Metzer 2010)

A. Basic flow chart of Illumina sequencing

B



C

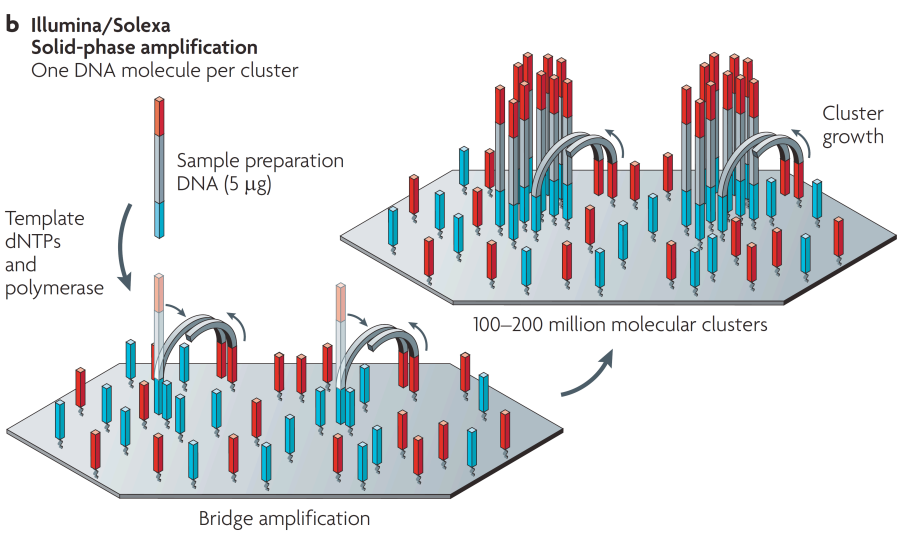


Figure 2.3 B/C. Illumina® high throughput sequencing chemistry

Adapters bridges are formed from the DNA fragments onto the flow cell. The bridges are amplified to form clusters that are subjected to sequencing. The detection of the nucleotide is done through four color cyclic reversible terminator chemistry. (Image: Park, 2009, Metzer 2010)

B. Immobilization of adapter ligated DNA fragments onto to flow cell and bridge formation.

C. Nucleotide detection by reversible terminator chemistry.

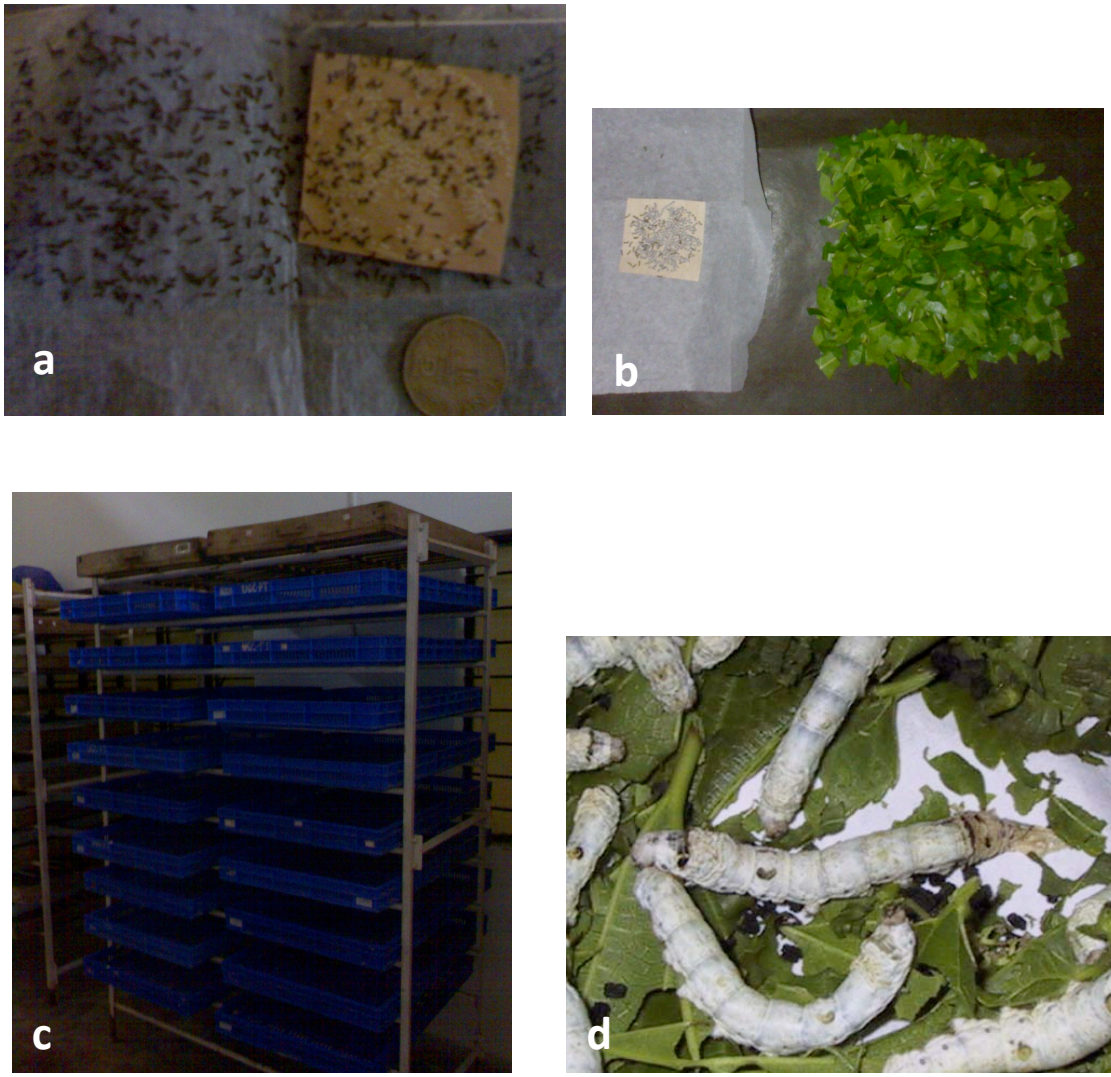


Figure 2.4 Maintenance of silkworm:

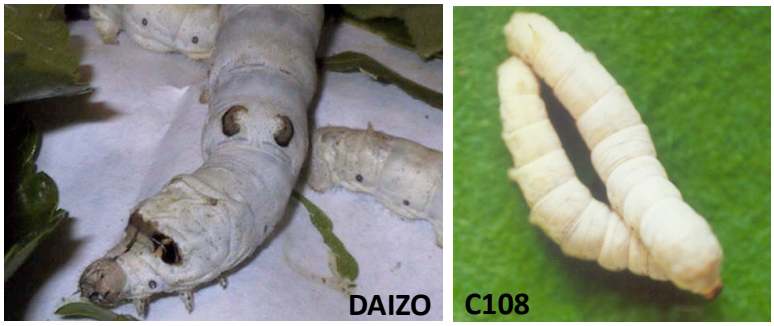
Silkworms are cultured on mulberry leaves, which are initially finely cut and fed to emerging larvae and later fed as whole leaves. They are grown and maintained in trays in temperature controlled rooms. The fourth instar larvae were cultured in large quantities for ChIP experiments.

a- emergent first instar larvae from the egg-laying

b- finely cut mulberry leaf bed for the first instar larvae.

c- temperature and humidity controlled rooms to grow the larvae.

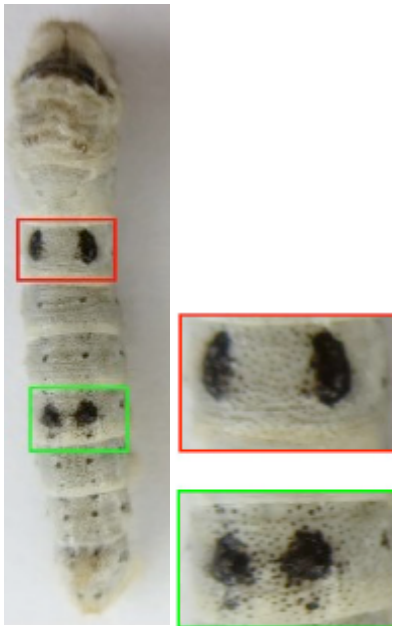
d- Fourth instar Daizo larvae feeding on mulberry.



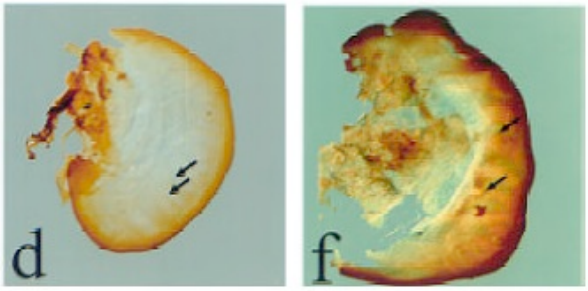

```
>nscaf2829 /length=4415533 /lengthwogaps=3966860
          Length = 4415533

 Score =  539 bits (272), Expect = e-152
 Identities = 296/304 (97%)
 Strand = Plus / Minus


Query: 88      ccccgaggaacaccccgtcctgctcactgaggctcccctcaaccccaaggccaacaggtg 147
                |||||||||||| |||||||| ||||||||||||||||||||||||||||||||||||
Sbjct: 1498172 ccccgaggaacatcccgtccttctcactgaggctcccctcaaccccaaggccaacaggtg 1498113


Query: 148     agttcatcgatgccggactatgcactttgcctctcggccggtgggccgttatcgaccgtt 207
                || |||||||| | |||||||||||||| ||||||||||||||||||||||||||||||
Sbjct: 1498112 agctcatcgattctggactatgcacttcgcctctcggccggtgggccgttatcgaccgtt 1498053


Query: 208     atctaacgagtgactttgttctgtttcagagagaagatgacccagatcatgttcgaaaca 267
                |||| |||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1498052 atctgacgaatgactttgttctgtttcagagagaagatgacccagatcatgttcgaaaca 1497993


Query: 268     ttcaacacgcccgccatgtacgtcgccatccaagccgtgctctcgctgtacgcgtccggt 327
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1497992 ttcaacacgcccgccatgtacgtcgccatccaagccgtgctctcgctgtacgcgtccggt 1497933


Query: 328     cgtaccaccggtatcgtgctggactccggcgacggtgtctcccacaccgtacccatctac 387
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1497932 cgtaccaccggtatcgtgctggactccggcgacggtgtctcccacaccgtacccatctac 1497873
```

Figure 2.5 Comparison of sequences from locally available races against genome database.

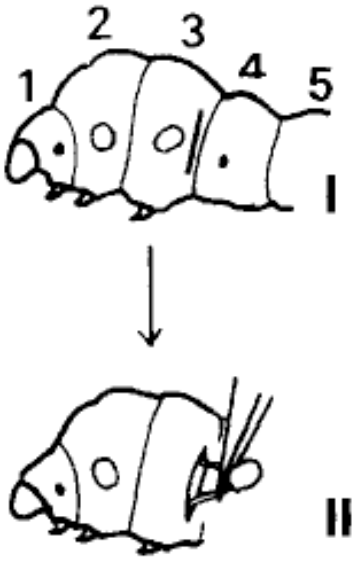
As the ChIP-seq method relies heavily on genome database for identification of genes, we compared the intron and exon sequences of two selected genes from the Daizo and C108 races to that of the genome databases (from Daizo p50T Japan and Dazao China). It was found that even the variable inron regions of locally sequenced races match closely to the genome database. Shown above is a fragment of *Actin C4* of Daizo (local race) gene as the query sequence showing almost identical sequence to the genome database subject.



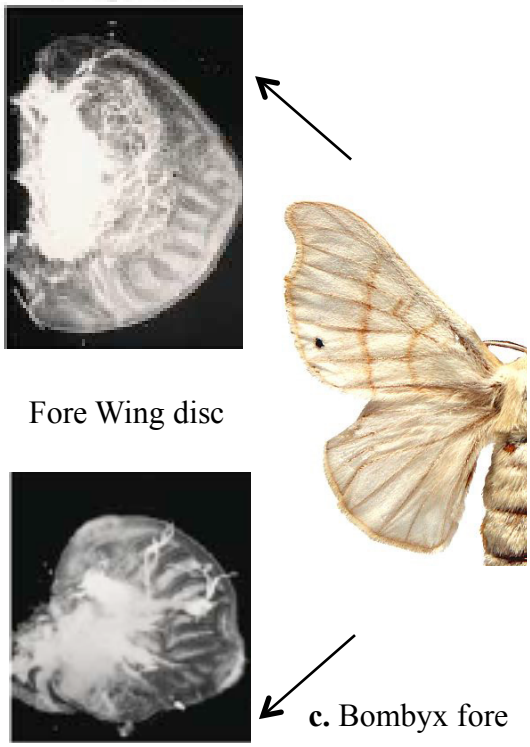
a. Daizo larvae with segmental markings



Expression in Bombyx V instar larvae
(d) Dll
(f) Wg



b-Dissection of wing buds



Fore Wing disc

Hind Wing disc

c. Bombyx fore and hind wings

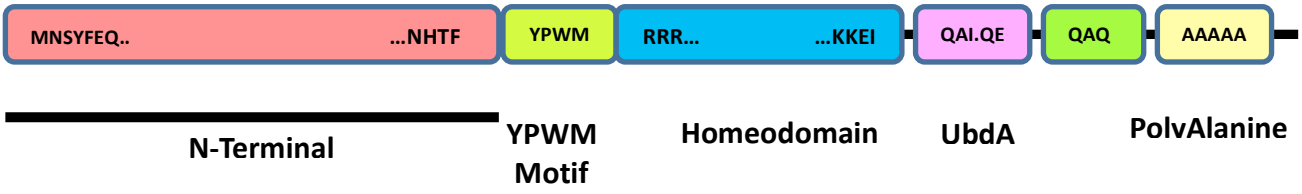
Figure 2.6 Identification of wing bud and larval stages in *Bombyx*:

a- Characteristic features of Daizo larvae are the Eyespot, Crescent markings on T3, and twin spots on segement 8. (From Nie et al, 2014)

b- Schematic showing method to dissect wing buds from fifth instar larvae. (From Hojyo and.Fujiwara, 1997)

c- Panel showing *Bombyx* fore and hind wing buds imaged at 10X resolution in brightfeld

d/f- Expression of developmental marker genes in wing discs of *Bombyx* (From Singh et al., 2001)



a- *Bombyx* Ubx protein: schematic showing motifs

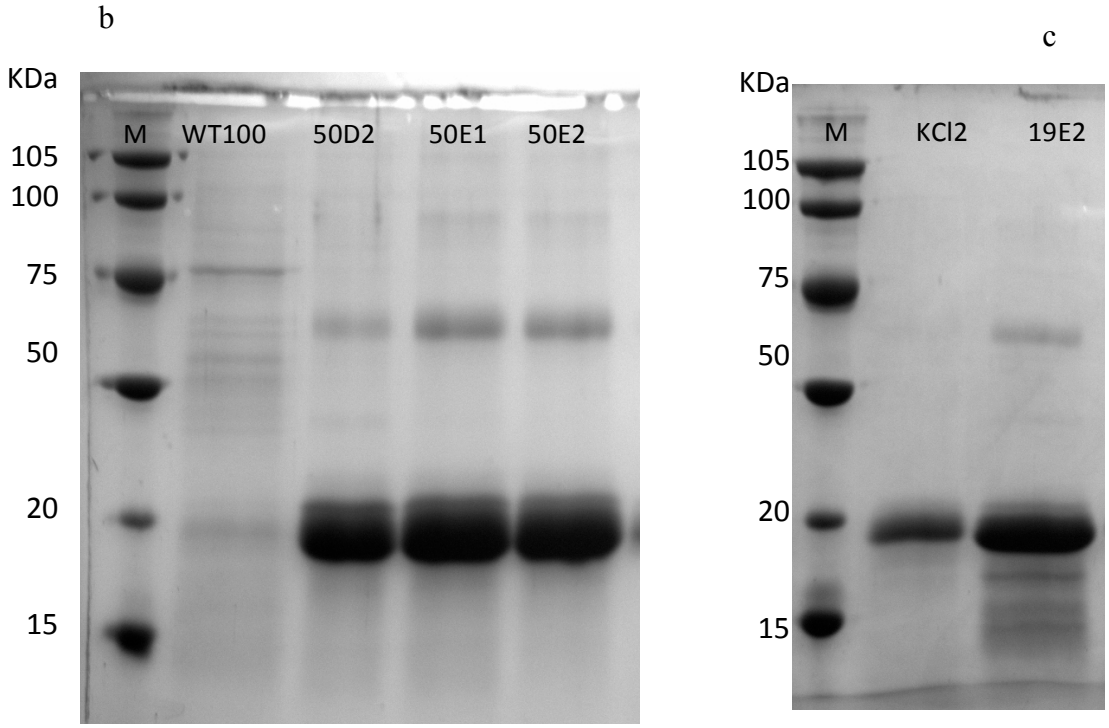


Figure 2.7 Expression and purification of *Bombyx* N terminal Ubx Protein

a- The domains in *Bombyx* Ubx full length protein: The region marked N terminal was used to raise the antibody. The N terminal protein used did not have the conserved homeodomain and YPWM motifs.

b-NiNTA purified protein eluates run on SDS-PAGE shows the expected *Bombyx* N terminal Ubx protein of size 19KDa. An additional 60KDa band was observed, which was lost after protein purification by electroelution.

c- Protein eluates run on a SDA-PAGE gel after further purification by electro elution to eliminate the 60KDa fragment. M is the marker. KCl2 is protein which was electro-eluted and purified by the KCl method. 19E2 is the pre elctroelution protein (NiNTA) elute.

A

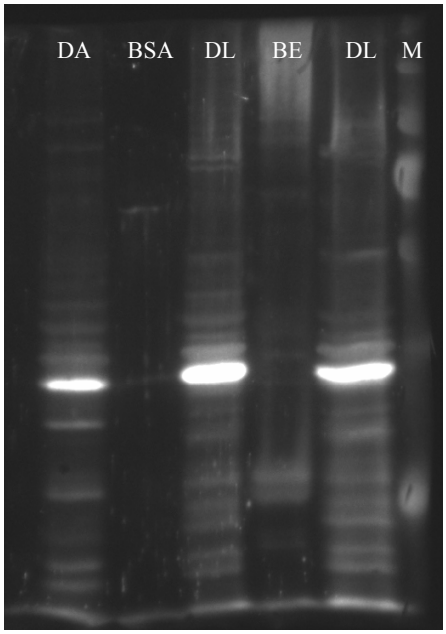
>*Bombyx* Ubx protein translated

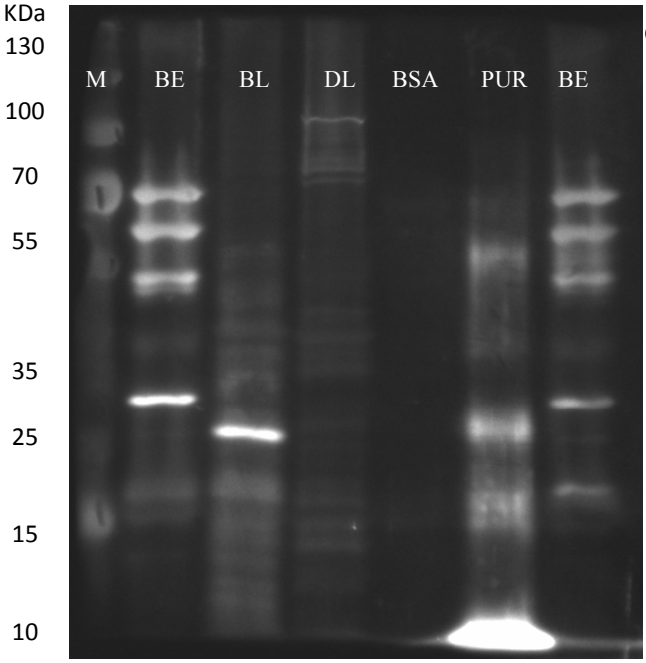
MNSYFEQGGFYGAHGVHQGGGGGDQYRGFPLGLTYAQPHALHQPRPQDSPYDASVAAA
CKLYAGEQQYPKADCSKPGGEQQNGYGGKEAWGSGLGALVRPAACTPEARYSESSSPGR
ALPWGNQCALPGSAASAAQPVHQQPTNHTFYPWMAIAGANGLRRRGRQTYTRYQTLELE
KEFHTNHYLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEIQAIKELNEQEKQAQAQK
AAAAAAAAAAAAQGHPEH

>*Precis (Junonia)* Ubx

MNSYFEQGGFYGAHGVHQGGGGGDQYRGFPLGLTYAQPHALHQPRPQDSPYDASVAAA
CKLYAGEQQYAKADCSKAGGEQQNGYGGKEAWGSGLGALVRPAACTPEARYSESSSPGR
ALPWGNQCALPGAAASAQPVQHQPTNHTFYPWMAIAGANGLRRRGRQTYTRYQTLELE
KEFHTNHYLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEIQAIKELNEQEKQAQAQK
AAAAAAAAAAAAQGHPEH

B





C

Figure 2.8 Western blot hybridization to show specificity of antibodies to *Bombyx* N-terminal Ubx

A- Comparison of Ubx proteins of *Bombyx mori* and *Precis (Junonia) coenia* using the NCBI tool 'Antigenic'. The blue boxes represent the predicted antigenic epitopes with the red lettered amino acid being the key antigenic amino acid. The proteins are 98% identical in sequence, also reflected in the antigenicity profile and antibody reactivity. Box in green shows the homeodomain.

B- Western blot hybridization with anti-*Drosophila* N terminal Ubx antibody (1:2500). Shows that it detects Ubx in *Drosophila* Adult (DA) and larval (DL), lysates while not reacting to Bombyx embryo lysate (BE) or BSA.

C- Western blot hybridization with antibodies against N terminal *Precis* Ubx (1:2500). Shows that the antibodies detect Ubx in *Bombyx* larval (BL) and embryo (BE) lysates, but not in *Drosophila* larval lysate (DL) or BSA. The purified N terminal *Bombyx* Ubx protein (PUR) however seems to be degraded from 19KDa to less than 10KDa fragments.

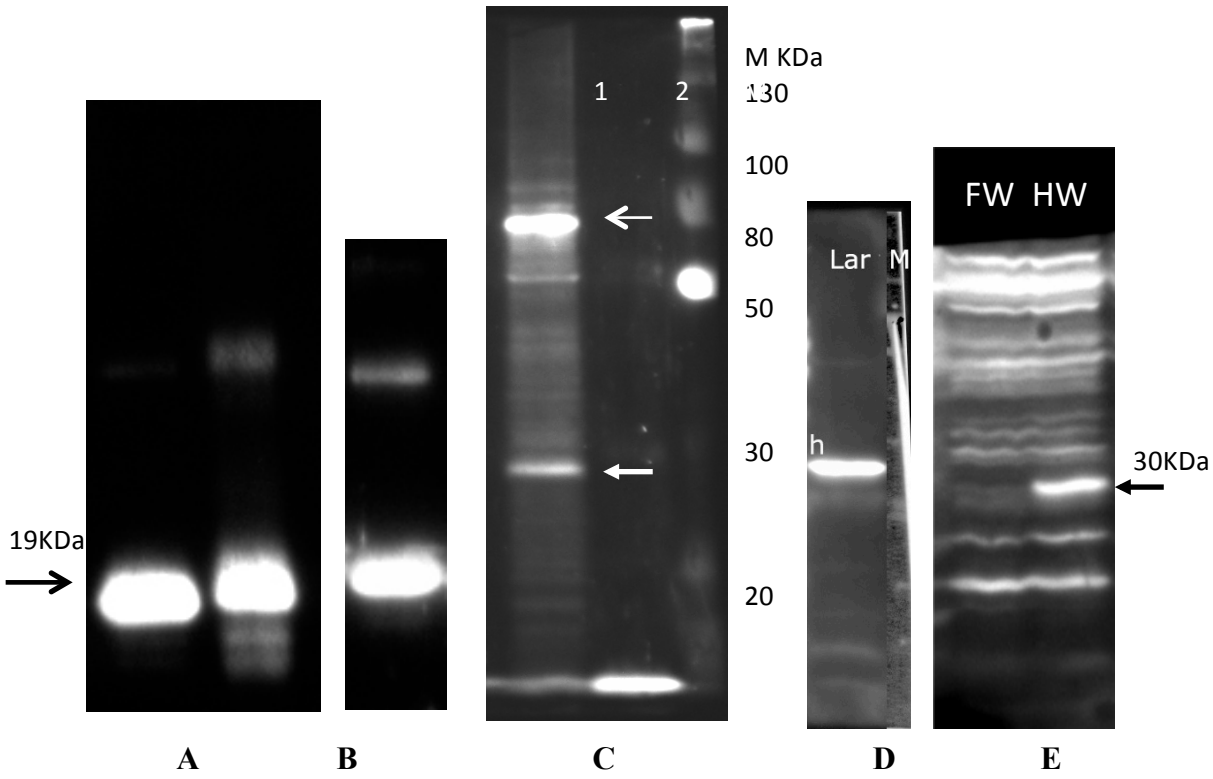


Figure 2.9 Detection of purified and endogenous *Bombyx* Ubx protein by Western blot hybridization

A- N-terminal *Bombyx* Ubx protein (19KDa) detected with anti-N-terminal *Bombyx* Ubx antisera. At 1:10000 dilution.

B- N-terminal *Bombyx* Ubx protein (19KDa) detected with anti-N-terminal *Precis* Ubx antisera. At 1:5000 dilution.

C- *Bombyx* Ubx from larval lysate (1) and purified, but degraded (2) protein detected by anti-N-terminal *Bombyx* Ubx antisera. 1:2500

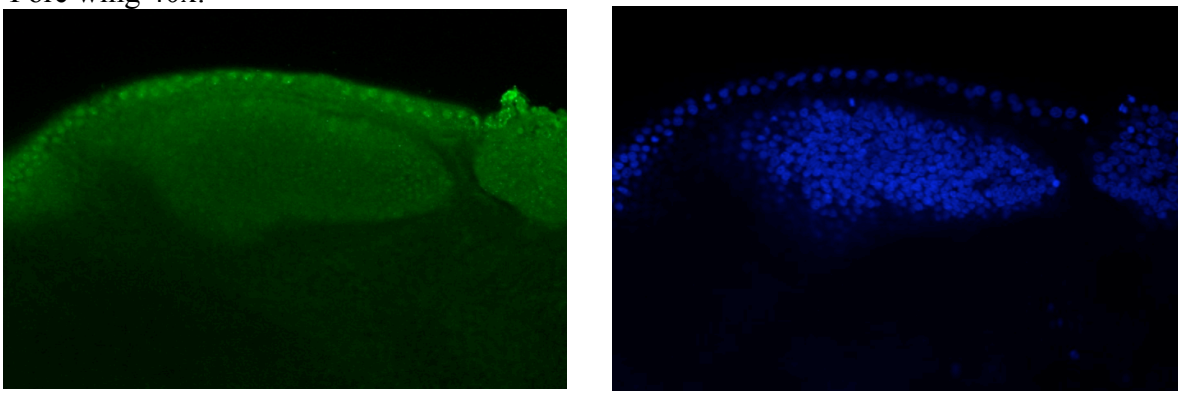
D-Bombyx Ubx from larval lysate (27KDa) detected by purified anti-N-terminal *Bombyx* Ubx antibody 1:2500

E- *Bombyx* Ubx from fore- and hindwing bud lysates detected by purified anti-N-terminal *Bombyx* Ubx antibody 1:2500. The fore wing does not show detectable amounts of Ubx expression.

A

Fore wing 40x:

Ubx DAPI



Hind Wing 40x

Ubx DAPI

B

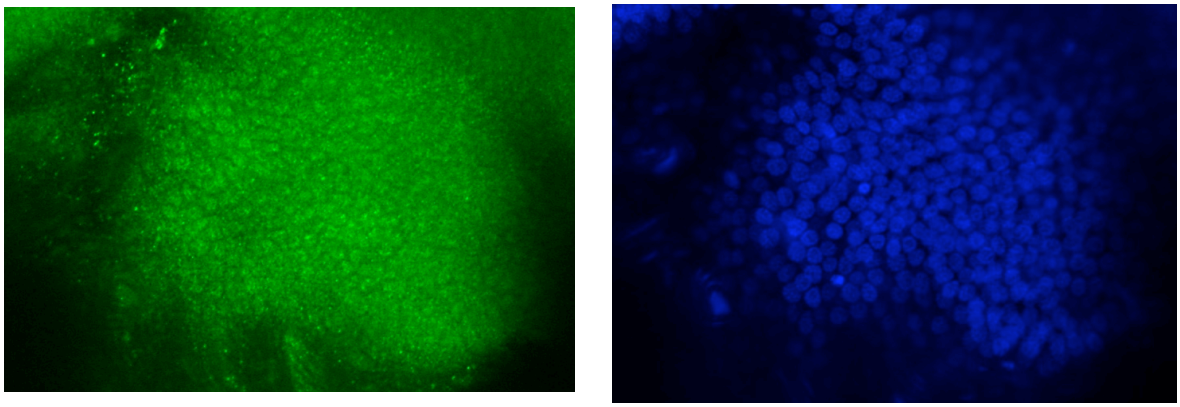


Figure 2.10 Immuno Histo-chemistry (IHC) with *Bombyx* wing buds

IHC with Bombyx wing buds show that the hind wing has good amount of UBx expression while in the forewing buds, only the peripodial membrane expresses Ubx.

A- *Bombyx* forewing bud stained with anti-*Bombyx* N-terminal Ubx antisera and DAPI. Only the peripodial membrane is stained. Dilution 1:100

B- *Bombyx* hindwing bud stained with anti-*Bombyx* N-terminal Ubx antisera and DAPI. Dilution1:100

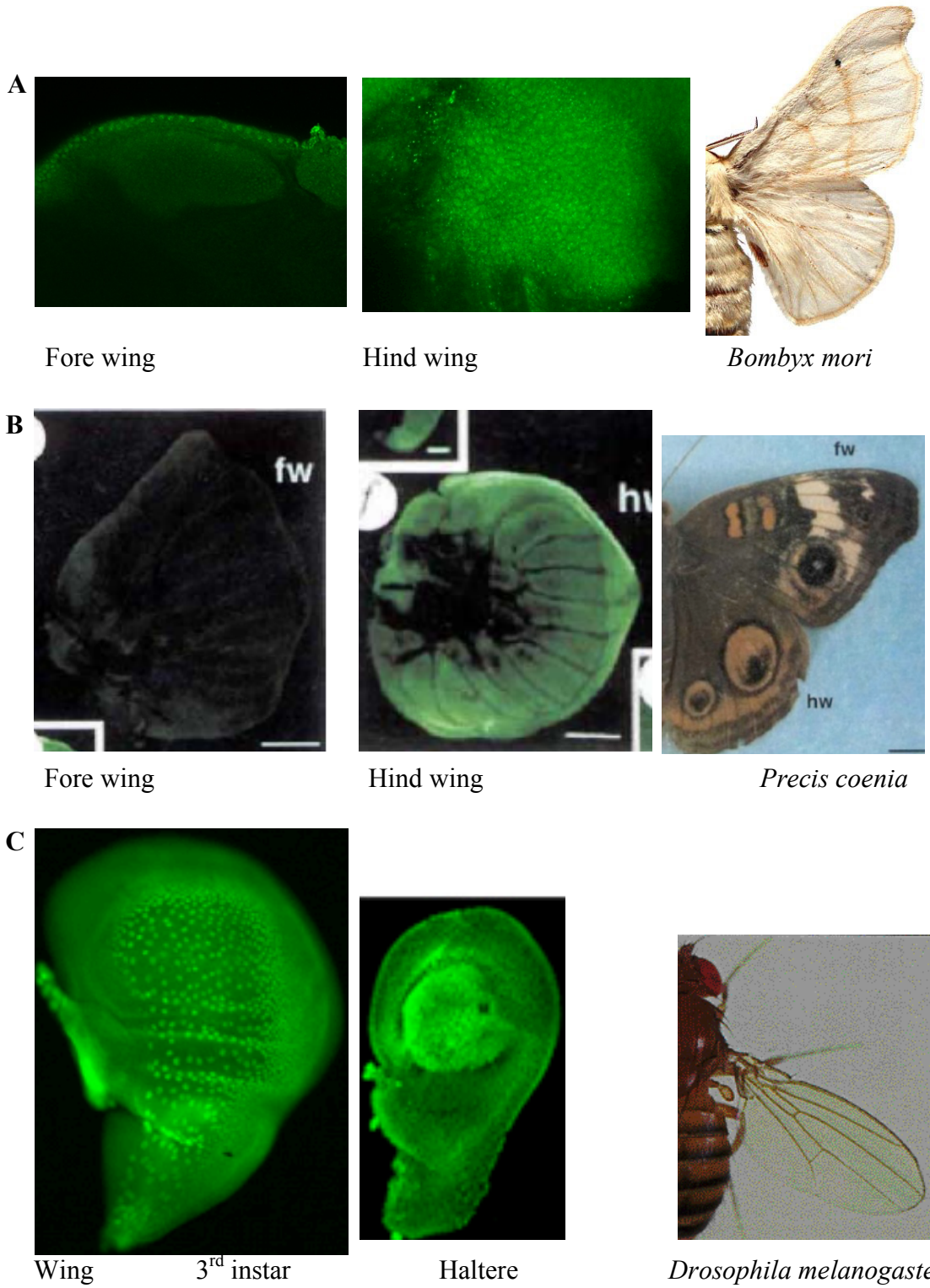


Figure 2.11 Comparison of Ubx expression in the hind and fore wing appendages of insects

A- Fore- and hind wings of *Bombyx* stained with anti-*Bombyx* N-terminal Ubx antisera

B- Fore- and hind wings of *Precis coenia* stained with anti-*Precis* N-terminal Ubx antisera.

C- Wing and haltere of *Drosophila* stained with anti-*Drosophila* N-terminal Ubx antisera

D- *Bombyx*, in spite of having morphologically similar fore and wing buds shows Ubx expression pattern similar to *Drosophila*.

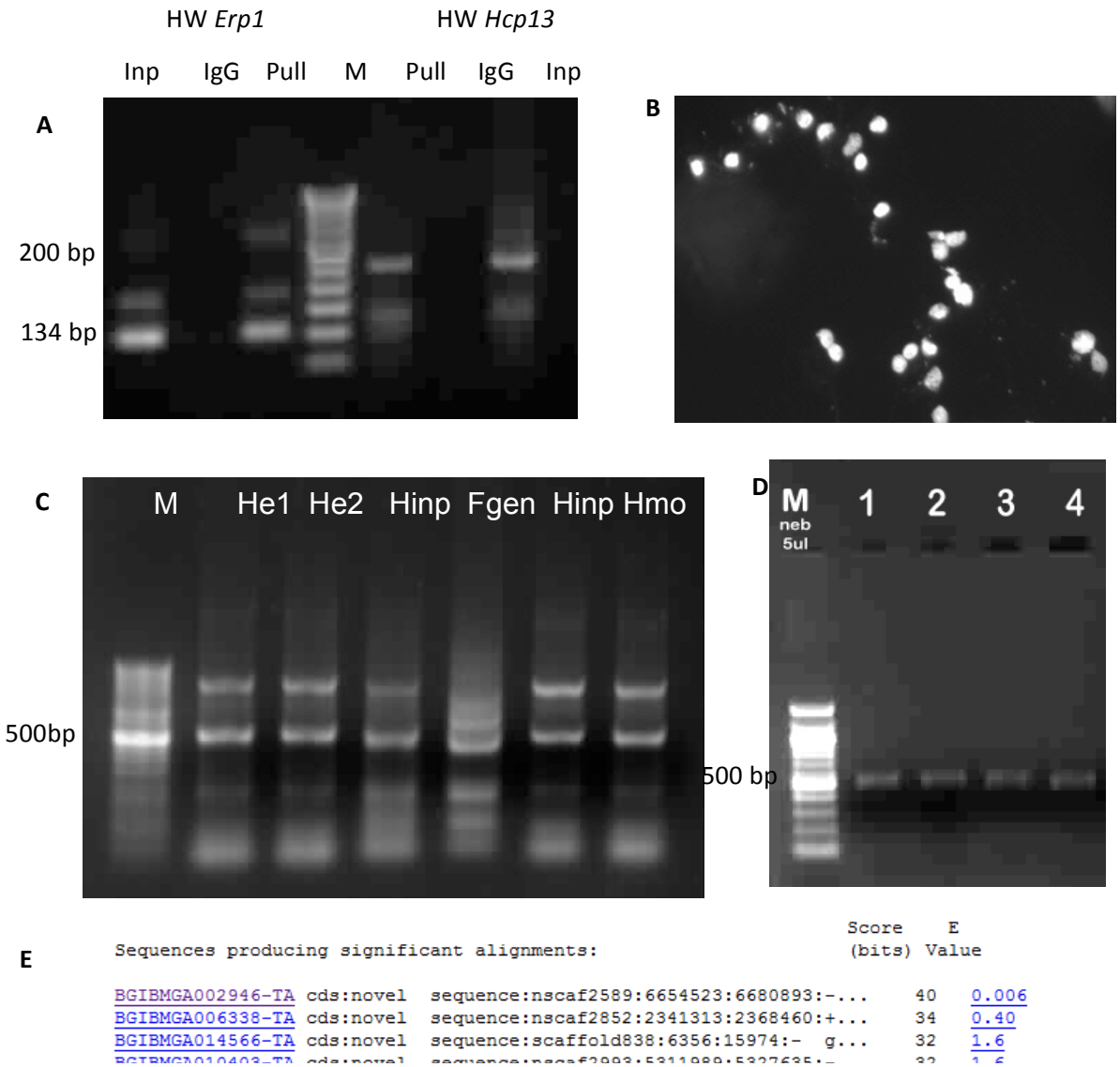


Figure 2.12 Standardization of Chromatin immuno-precipitation from *Bombyx* wing buds

A – Two known targets of *Bombyx* GATA were used for post ChIP detection of ChIPped fragments. Erp1 and Hcp13 primers were used to amplify the regions from experiment and controls to validate the ChIP protocol. The expected 134bp (*Erp1*) and 200 bp(*Hcp13*) were obtained in the input and the experiment lanes but not in the IgG lanes.

B – DAPI stained nuclei after HEPES low salt nuclei isolation, which were imaged as a control for ChIP before proceeding chromatin preparation.

C – Amplification of *Bombyx* Spalt from ChIP samples used to detect the presence of *Bombyx* DNA in ChIP samples.

D- Spalt amplicons excised from gel purified and run before sequencing.

E – Amplifid Spalt fragments were sequenced and BLASTed against the SilkDB databse to verify the ChIP DNA. The results show the right scaffold and region to which the gene Spalt belongs.

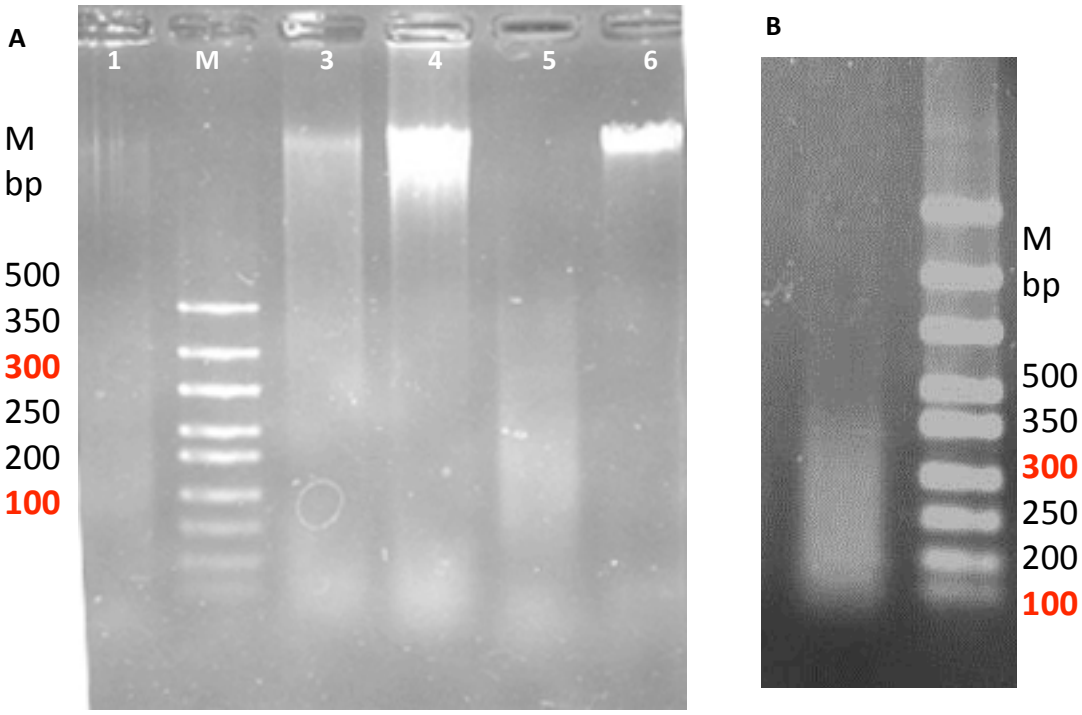


Figure 2.13 Standardization of Chromatin immuno-precipitation: Sonication/ChIP-Western

A – Chromatin shearing was standardized by varying wattage and pulse cycles on a sonicator to obtain the appropriately sized shear size. Initial attempts did not yield good quantities of DNA and had remains of high molecular weight genomic DNA which was gotten rid by longer sonication runs.

B – Final condition (total time of 15 minutes with 55 sec on-60 sec off cycle at high power) with appropriately sheared chromatin with size range of 100-300bp.

Chapter 3

Analysis of DNA sequences enriched in ChIP

Introduction

The modern day high throughput sequencing methods generate unprecedented amounts of data and they go hand in hand with the rapid strides that computational field makes every day. The raw data generated by all the genome analyzers is in the order of gigabytes per machine per run, and this trend seems to be on a upwardly path. The storage and processing of genomic data has become a major challenge in data management. The huge amount of data generated by the genome analyzers poses a huge challenge to the developers of software and more efficient algorithms.

The raw data obtained after high throughput sequencing run for a ChIP-seq analysis come as images from the genome analyzer. A base caller converts the data in the form of images to sequence tags, which can be aligned to the genome (Park, 2009). Base calling is accompanied by reliability statistics, which determine the quality of sequencing read. ChIP-seq generally generates gigabytes of short read data with sequence tags of 35-100 bases in length.

3.1 Alignment to the genome

Genome aligners are used to map the sequence tags generated, onto the genome, which is a computationally intensive process that could take hours to days depending on the computational abilities, length and the complexity of the genome. As the demand with such alignments are high and time consuming, algorithms are developed, which are a balance between accuracy, speed, memory and flexibility and which allow a certain amount of mismatch due to sequencing errors or due to natural difference with the reference genome. Hence no one aligner is suitable for all needs and they have to be chosen according to the kind of application and the data to be analyzed. Most of these algorithms are accompanied by software, which is in the form of plug in for statistical language R or standard compiled languages like C or python. These programs are not very user friendly, often times relying on command line based approach to run and analyze using the program.

The two most commonly used aligners are Burrows-Wheeler Aligner (BWA) and Bowtie. BWA is an efficient aligner to use in cases of reads which are

greater than 200 bp like those from a PacBio® genome analyzer (Li and Durbin, 2010). Bowtie is an extremely fast mapper that is based on an algorithm originally developed for fast file compression. Bowtie is a memory-efficient alignment program, particularly useful to map short sequence reads, while it does not perform very well with long reads. Bowtie uses the Burrows Wheeler (BW) index for aligning sequences to the genome with a novel quality backtracking algorithm that permits mismatches (Langmead et al, 2009). Bowtie index allows large texts to be searched efficiently in a small memory footprint. Bowtie allows usage of many processor cores of a computer processor simultaneously to improve the speed of the alignment. As this kind of alignments need less of accuracy and more speed to scan through the entire genome while aligning, Bowtie aligns and reports at least one definitive alignment for each read, but if the best match is not an exact match the quality of the alignment is reduced. As we used ChIP data with single end 36 bp reads from an Illumina® genome analyzer, which were short reads and large in number, Bowtie was the application of choice for genome alignment.

3.2 Identification of peaks

One of the popular applications of ChIP-seq is to identify the DNA binding regions of the proteins. However ChIP-seq tags represent only the ends of the ChIP fragments, instead of precise DNA binding sites (peaks). Secondly, ChIP-seq data exhibits biases based on chromosome region and copy numbers. These issues can be overcome only if the genome of the organism in question is sequenced to a great depth and also if a control ChIP-seq sample is available for test comparison, which enable better mapping of the experiment ChIP-seq tags.

Peak calling uses data from the ChIP-seq tags and the control tags (which are from input DNA), and generates a list of enriched regions that are ordered by false discovery rate (FDR) as a statistical measure. Peak calling algorithms are of different types, which use a variety of approaches such as a window based approach, overlap based approach or hidden markov model based approach. There are numerous peak-calling programs available on public portals, some of the popular ones being MACS, Homer-FindPeaks, QuEST, PeakSeq, CisGenome and SISSRs. Each of them varies in the kind and number of peaks

they can identify. Higher number of peaks gives the experimenter scope to narrow down further by refining on other parameters, if only a few peaks are identified one is left with very few options.

Model-Based Analysis of ChIP-Seq data (MACS) addresses the issue of biases and gives robust and high-resolution ChIP-seq peak predictions (Zhang et al, 2008). ChIP-seq tags represent the ends of the fragment in a ChIP-DNA library and are often shifted towards the 3' direction to better represent the binding site, but this 'shift size' is unknown to the experimenter. As ChIP-DNA fragments are likely to be sequenced from both ends, the tag density around a binding site should show a bimodal enrichment pattern. MACS takes the advantage of this bimodal distribution of tags to empirically model the shifting size to determine the binding sites.

In an experiment with controls, MACS empirically estimates the false discovery rate (FDR) for each detected peak. At each p-value, MACS uses the same parameters to find ChIP peaks over control and by swapping control peaks over ChIP (Fig 3.1). The empirical FDR is defined as Number of control peaks/Number of ChIP peaks. This FDR estimate is more robust than calculating FDR by randomizing tags along the genome. However, when tag counts from ChIP and controls are not balanced, the sample with more tags often gives more peaks even though MACS normalizes the total tag counts between the two samples and FDR here may become unreliable (Zhang et al, 2008).

3.3 Tools used in ChIP-seq analysis workflow

The enormous amount of data generated in a ChIP-seq is handled by different programs till it can be used to analyze as interpretable data. At every step, there are assortments of tools that help convert formats, index large amounts of data and do quality control analysis.

3.3.1 Quality control: FastQC

Before analyzing sequence to draw biological conclusions, some quality control check is essential, ensuring there are no problems or biases in the data. The first program that is used even before the alignment is the quality control analysis,

which assesses the quality of the reads obtained from a sequencer, based on various parameters. A quality check on the nature of the data generated by a genome analyzer conveys a meaningful startup estimate on the data and its reliability. Measures like GC content, repeats and base call quality can tell us how well the experiment has worked, mostly with respect to the success of sequencing. Most sequencers generate a quality control report but they are focused on identifying sequencer related problems, whereas FastQC aims to spot problems, which may originate either from sequencer or in the starting library material.

FastQC is a user friendly publically available high throughput sequencing quality assessment tool developed by Babraham institute. It analyzes ChIP-seq reads in FastQ, BAM and SAM formats and reports many parameters, which help assess the quality of sequencing and an overview of the reliability of the experiment. It can be run either in the terminal in a command line mode or on a java based user friendly interface.

3.3.2 Data management: SAM-Tools

Other tools that assist the ChIP-seq data analysis workflow are the Sequence Alignment Map (SAM) tools and the Browser Extensible Data (BED) tools. SAM format is the default generic format in which large nucleotide sequence alignments are stored, as in case of the ChIP-seq read-genome alignment. As there are different kinds of data generated in different genome analyzers, these storage formats bring in the unity to process data in a uniform way to analyze them using commonly available public tools. The SAM format helps in realizing efficient mapping as it allows working on large nucleotide data without loading the whole alignment into the computer memory and it also indexes by genomic position to retrieve the locations while running a process. Binary Alignment Map (BAM) format is the compressed binary version of the SAM format; it is a relatively smaller file. SAM-tools provide various utilities to manipulate alignments including sorting, merging, indexing, converting and generating alignments in required formats (Li et al, 2009).

3.3.3 Feature comparison: BED-Tools

BED-tools comprise of a software suite for comparison, manipulation and annotation of genomic features in a data format called the BED format and General Feature Format (GEF). BED tools enable genome arithmetic: that is, set theory on the genome, like allowing one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED and GFF/GTF. The BED file format provides a flexible way to define the data lines that are displayed in an annotation track. Basic BED file has three essential fields (chromosome, start and end) and nine optional ones. The number of fields per line must be consistent throughout any single set of data. The order of the optional fields is binding; lower-numbered fields must always be populated if higher-numbered fields are used. A GFF or a GTF file format is used to manipulate and analyze genome annotation features and is an extension of a basic (name, start, end) tuple that can be used to identify a substring of a biological sequence. It supports comparison of alignments in BAM, BED and GFF formats. These efficient tools are useful to compare and manipulate large genome-wide sequencing datasets and these datasets to genome databases (Quinlan & Hall, 2010). The program suite BED-tools is written in the programming language C++ and it is open source software developed on a UNIX platform. BED tools facilitate routine genomics and pipelines that can quickly answer intricate questions of large genomic datasets.

3.3.4 Combined array of tools: Galaxy

Galaxy is an open web-based platform for genomic research; it provides a plethora of tools in a user friendly open sharable platform for performing accessible and reproducible genomic science (Goecks et al, 2010). These tools are an assembly of many programs from various open source command line based resources like BED-tools and SAM-tools, efficiently combining multiple tools in an analysis workflow. Galaxy provides a web based platform that needs no computational expertise, it can be used without being impeded by problems ranging from tool installation and necessity of command line knowledge.

3.3.5 Visualization of Data: IGV viewer

Analysis of ChIP-seq data which are very large and computing intensive has become a rate limiting step in many genome wide studies. Although most analysis is automated, human interpretation and judgment of the biological meaning of such data is essential for gaining insight and elucidating complex biological phenomena. Integrative Graphics Viewer (IGV) is a freely available high performance viewer that can efficiently handle large genome wide data sets, while providing a smooth and intuitive user-friendly interface at all levels of genome resolution. It supports both array and next generation sequencing data, allowing researchers to visualize and explore their own data with custom genome datasets with genomic annotations. It endows the user an ability to view data in many genomic regions simultaneously in adjacent panels to compare and correlate the binding events across the genome and between different experiments. It is written in Java programming and runs on all popular operating systems (James & Jill 2012).

3.3.6 Identification of Homologs: Ensemble Metazoa- Biomart

'Ensembl genomes' is an integrative resource for genome scale data sets from various organisms. It is divided into protists, bacteria, fungi, plants and the invertebrate metazoa. Biomart grew out as an extension of the Ensembl genomic data-mining tool. It now links to more than 40 databases, which enable us to understand complex biology by studying different kind of genomic data in comparison with many organisms (Baker, 2012). Biomart is an open source, free tool, which provides a single interface to access and analyze genomic data from many different organisms (Kasprzyk, 2011). It hosts an updated resource of genomes and their annotations along with other features like variations, array expression etc. It allows download of tables and homologs via FTP for each release, so one can convert different kind and organism related data easily through the Ensembl Biomart interface.

3.4 Silkworm genome databases

Silkworms have been used as the source of silk to make textile from ancient times; they have also been extensively domesticated for silk production.

Silkmoth has become a useful model to study Lepidoptera, as they can be cultured in large numbers and manipulated with many tools. It also is the first representative genome of a lepidopteran insect to sequenced, as this is important not only to identify ways to improve silk production, but also to agriculture in general, as many moths that are crop pests belong to this order. The genome of silkworm was published in 2004, independently by two groups based in China (Xia et al, 2004) and Japan (Mita et al, 2004). The Chinese group used the silkworm race Dazao to sequence the genome, while the Japanese group used a closely related race Daizo p50T for the genome sequencing. The data from these two studies are stored in two databases accessible online, namely the Chinese SilkDB and the Japanese Kaikobase. The manual annotation of the genome and the complete sequencing of the BAC clones are still ongoing at the time of writing this thesis. Both these were used in the computational downstream analysis of the ChIP-seq data.

3.4.1 SilkDB BGI silkworm genome database

The silkworm database SilkDB is a web based repository and knowledge base for the curation, integration and study of the silkworm genomic data (Wang et al, 2005). It provides an integrated representation of genome wide sequence assembly, transposable elements, clusters of ESTs and other features of the *Bombyx* genome, whose sequencing was accomplished by the groups mostly based in China. It provides a comprehensive knowledge base about the silkworm genome and related information in systematic graphical ways.

The genes in the genome were predicted based on a gene finder algorithm called Beijing Genome Institute (BGI) Gene finder (BGF), which uses GenScan and FgeneSH and each predicted gene has the ID starting with BGIBMGA followed by six digit identification code. In SilkDB, tools such as map view provides both an information source and a comparative analysis platform to work on silkworm and other insects from a genomic perspective. It has local tools to explore the genome graphically at the chromosome and scaffold levels; it also had local analysis tools like BLAST in the database. Silk DB also hosts a range of compiled genomic datasets that can be downloaded and used to analyze silkworm data.

3.4.2 Kaikobase silkworm database

Kaikobase is an integrated genome database with map viewers and tools to display results and data at the level of nucleotide sequence, gene, scaffold and chromosome (Shimomura et al, 2009). It was built as the genome data sequenced by the two projects in China and Japan were insufficient to build long genomic scaffolds and unambiguous annotation of the silkworm genome. Kaikobase is a joint collaborative effort from both genome groups to merge and assemble both the genome datasets and it is hosted at the National Institute of Agro-biological Sciences, Tsukuba, Japan. The kaikobase uses an open source GBrowse genome browser to host the graphical representation of the silkworm data for scaffold and chromosome maps with genes and annotation detail. The gene regions found by kaikobase gene prediction program are supplemented with different supporting data for validation of the gene region like mRNA, EST and full length cDNA. It also is cross referenced to the SilkDB BGI gene ID along with EST and protein information of the gene. It uses NCBI-BLAST software for the sequence search function. Thus, it provides comprehensive detail on the gene region and is the best tool available for efficient utilization of the silkworm genome information for functional and applied genomics.

Materials and Methods

All the analyses were done using freely available open source programs on a computer with Linux (Ubuntu 12.04 LTS) as the operating system.

3.5 Quality control analysis of ChIP-seq reads

FastQC was used to ascertain the quality of the reads obtained from single end 36 bp sequencing from an Illumina® genome analyzer. It was run on a java based graphic interface and the results were obtained in an html file, which can be opened in any web browser. The read files obtained were in the "fastq" format and were in compressed "tar" format. They were directly loaded onto the FastQC program to obtain the detailed analysis of the reads and the corresponding graphs. The program shows three indicators for each parameter tested: acceptable (green), warning (orange) and failed (red). FastQC analyzes the following parameters of a sequence read file;

1. Per-base sequence quality: It shows a box whisker plot for each position of a nucleotide in a read. The central red line is median value, yellow box is inter quartile range, upper and lower whiskers represent 10% and 90% points, respectively and the blue line represents the mean quality.

The Y axis shows the quality scores, the higher the score better is the base call. The quality of calls degrades as the run progresses, so base calls invariably fall at the end of a read. These can be trimmed if they fall below the phred score threshold set for the experiment.

2. Per-base sequence quality scores: It allows us to see if a subset of sequences has universally low quality values. These should represent only a small percentage for the QC to be fine.

3. Per-base sequence content: It plots the proportion of each base position in a read for which each of the four bases has been called. In a random library there would be no difference between the base distributions at each nucleotide in the read, so the plot should run parallel with each other without much variation. A

strong bias indicates contamination in the library or a systematic sequencing problem.

4. Per-base GC content: It shows GC content at each base position in a read. In a random library a horizontal line is expected as there would be no differences of GC at different bases. This reflects the overall GC content of the underlying genome.

5. Per-sequence GC content: It plots the GC content across the length of each sequence in a read and compares to modeled normal distribution of GC content. In a random library a normal distribution is expected with central peak corresponding to overall GC content of the genome. Unusual shape may indicate contamination.

6. Per-base N content: When a sequencer is unable to make a base call with sufficient confidence it substitutes an N at that base. The percentage of base calls at each position at which N was called is plotted. It is seen at the end of the sequence and not a very common error.

7. Sequence length distribution: If the sequence fragments generated are of different lengths it will be reflected in this plot which shows the distribution of fragment sizes in the read. Generally simple graph with a peak at the single size (like 36 bp) is plotted.

8. Sequence duplication levels: In a diverse library most sequences will occur only once in the final set. A low level of duplication means very high level of coverage of the target sequence, but high level is more like an indication of some enrichment bias like that of PCR amplification bias. It only considers first 200000 sequences for the analysis, which is sufficient to get an impression of the duplication levels. Sequences more than 10 are also placed in the 10 category, so it's not unusual to see a small rise in this category. A big rise means high levels of duplication, which could indicate the insufficient quantity of starting DNA for the sequencing.

9. Overrepresented sequences: In a normal sequencing run, no individual sequence is overrepresented in the set. It either means, it is biologically significant or indicates contamination in the library. This module lists all the

sequence that make up more than 0.1% of the total by analyzing the first 200000 sequences. It also looks for matches to a database of common contaminants including adapters used to make the library clusters. Before interpreting the overrepresentation of a particular region in genome, a thorough analysis to ensure that they are not originating from a contaminant is very important. Adapter sequences can be trimmed off before the downstream analysis if found to be overrepresented.

10. Overrepresented K-mers: This module spots an increase in any exact duplication; however it does not work if the reads are long with poor sequence quality or partial sequence appearing at different places within the sequence. It counts enrichment of every 5-mer within the sequence of the library. It calculates an expected level (from 20% reads) at which a k-mer should have been seen based on base content of the library as a whole and uses actual count to calculate an observed/expected ratio for that k-mer. It shows any general enrichment or if a pattern of bias is seen at different points over your read length.

3.6 Creating index of the genome with Bowtie

A fasta file (466Mb) of the *Bombyx mori* genome was downloaded from the SilkDB database. It was used for ChIP read alignment to the genome using the program Bowtie (version 0.12.7 valentine). The first step to aligning the reads is to create the index of the genome sequences. The tool bowtie-build was used to create an index of the genome.

```
./bowtie-build silkgenome.fa silk_index
```

Where 'bowtie-build' is the command used, "silkgenome.fa" is the fasta file of the *Bombyx* genome and silk_index is the prefix name given to the index.

3.7 Alignment of the ChIP-seq reads to the genome

The ChIP reads in the compressed archive (tar) were decompressed to obtain the fastq or txt file which generally in the order of 1-10 Gigabytes in size. This file is then aligned to the indexed genome using bowtie (version 0.12.7 valentine).

```
./bowtie silk_index -q ChIP_reads_file.fastq -M 1 -S
-5 3 -v 3 aligned_output_file.sam -p3
```

'bowtie' is the command used. "silk_index" is the genome index prefix to be used for the alignment. -q indicates the query input file in fastq format. "ChIP_reads_file.fastq" is the file which contains the sequencing reads in fastq format. The customization -M 1 suppresses all alignments for a particular read if more than 1 alignment exists for it and reports one random alignment. -S trims initial bases from high quality (left) end of each read before alignment. -5 3 trims three bases from high quality left end of each read before alignment. -v 3 reports the alignments with most 3-mismatches. The output "aligned_output_file.sam" is obtained in the sam format. -p3 allows usage of 3 CPU core processors for carrying out this alignment which runs for 3-4 hours in this conditions or overnight when the default single core is used.

The samples untreated input, ChIP experiment with anti-N terminal *Bombyx* Ubx antibody and the Rabbit normal IgG negative controls were all aligned to the genome separately and treated exactly through the same analysis. The output of the bowtie alignment is the sam file and the statistics, which show the alignment percentages (Table 3.1).

3.8 Conversion of file format and indexing

SAM-tools was used to convert the file to a sorted BAM file and indexed for visualizing the alignments in viewer (IGV). BAM file is smaller than the SAM format and is useful for the downstream analysis.

```
samtools view -bS aligned_output_file.sam | samtools
sort - aligned_output_file
```

samtools view extracts/prints all alignments in SAM to BAM format. -b indicates that the output is required in BAM format and S indicates input is in SAM format. The output of the bowtie alignment "aligned_output_file.sam" is used as input here to convert to BAM file. The unix command line function pipe '|' allows running of the intermediate file as the input for the next program, here the output file of 'samtools view' to be run simultaneously as input of 'samtools sort' to reduce time, memory and storage space. 'samtools sort' sorts alignments by leftmost coordinates and outputs a final sorted BAM file with the name as

given like here '-aligned_output_file', it adds the '.bam' suffix to the file automatically.

```
samtools index aligned_output_file.bam
```

'samtools index' indexes sorted BAM alignment for fast random access. An index file "aligned_output_file.bam.bai" is created. The index file is necessary to visualize the alignments in BAM format in IGV viewer to compare between input, experiment and negative controls and to see the quality of the reads when aligned.

3.9 Visualization with IGV viewer

IGV allows visualization of the alignments in indexed BAM files, where different such alignments can be compared in a single window. Input, experiment and negative control alignments were compared to get a visualization of the distribution of the reads and the location of peaks. The genome annotation gff file can be added as a track apart from the three tracks as above to visualize the known genes and the location of peaks from it. IGV is an open source java based program which has a user friendly interface to upload the genome, annotation and alignment files. It is invoked from the command line by using either of the following commands while within the IGV folder:

```
java -Xmx750m -jar igv.jar
sh ./igv.sh
```

Once the graphical user interface is started an indexed genome file (not the ebwt, but a BLAST '.fai' index file) is uploaded along with the gff annotation file to IGV. The alignment files are then uploaded to compare and analyze the reads, peaks (binding regions) and genes.

3.10 Peak calling using MACS

The program MACS (version 1.4.2) was used to find the peaks (binding regions of Ubx) from the ChIP enriched DNA sequences.

```
macs14 -t aligned_output_experiment/negative_control-
file.bam -c aligned_input_control-file.bam -g
3.11e+08 --keep-dup=2    -
nNameofpeaksfile_exp_vs_input
```

'macs14' is the command. '-t' is the treatment file or the file from which the enriched regions are to be identified. This can either be the experiment bam file or the negative control bam file. The normalization of the above files are done with '-c' the input bam file, which acts as a control to normalize the pulldown files. '-g' indicates the effective genome size of the organism in question. It is defined as the genome size that can be sequenced. It is less than the actual genome size. Due to the repeats in the genome, mapable regions are reduced. Generally an effective size 75% of the genome is considered. The silkworm genome is approximately 5.14e+08 bp in size, a reduction of around 75% yields a value of 3.11e+08 which was used. '--keep-dup=2' It controls the behavior of the program to duplicate sequence tags at the same location, coordinate and strand. The default option makes MACS calculate the maximum tags at exact same location based on binomial distribution using 1e-5 as p-value cutoff. Here when '2' is specified, at most 2 tags were reported for the same location. '-n' names the string with a prefix to all files created.

Autocorrelation analysis was done using HOMER to cross check the ChIP data.

3.11 Identification of genes associated with the peaks

Once the peaks were identified, an FDR cutoff was applied and the peaks that overlapped with the negative control (normal Rabbit IgG) were deleted by using the galaxy tool "operate on genomic intervals-subtract". Two replicates of ChIP experiment on fore and hind wing with input and negative controls were carried out. The peaks found from the two replicates by normalizing against input were intersected with each other to find the peaks that are common between the two replicates. The resultant peak file, which is IgG filtered and common to both replicates in BED format was used to find the nearest genes that might be regulated by Ubx were identified by using two methods (intersect and fetch). Both SilkDB and Kaikobase genome assemblies were used for the identification of the nearest genes.

3.11.1 Galaxy Fetch- closest non-overlapping feature for every interval (peak)

The web based tool Galaxy was used to identify the genes around the peaks. 'Fetch closest non-overlapping feature' for every interval, which belongs to 'Operate on Genomic Intervals' tools was used to identify the genes from a *Bombyx* gtf annotation file (BGI SilkDB annotation acquired from ensemble database). The BED file containing the peaks after applying cutoffs and IgG filtering was uploaded onto the galaxy database along with the annotation file. Then the program compares both the files and for every interval in the interval dataset (peak dataset BED file from ChIP experiment), this tool fetches the closest non-overlapping upstream and downstream (one nearest each) features from the features dataset (*Bombyx* annotation in gtf format).

The shortest distance between the ChIP pulldown coordinates and the gene feature coordinate was measured and reported as the distance between the binding site (peak) and gene by using a simple python code (Appendix 3). The list was sorted into genes that lie within 2000 bp, 5000 bp, 10000 bp and beyond-10000 bp from the peak.

3.11.2 BED-tools slop-intersect method to find the genes near peaks

The exact location of a gene feature is not very accurate with the silkworm genomes as found in some examples (Fig 3.2); hence we also used an alternate approach to identify the genes. This method involves extending the peak region identified both sides by a certain number of nucleotides and then intersecting this extended fragment with the annotation feature gtf file. This will ensure that no genes are missed due to inaccuracy of the identification of features like the gene-intron regions (Fig 3.2).

The BED-tool slop was used to extend the peak region by certain number of nucleotides considering the limiting boundaries of the genome scaffold lengths.

```
bedtools slop -i Peak_file.bed -g silkgenome.genome -
b 2000 > slop_peakfile_2k.bed
```

'bedtools slop' is the command. '-i' the input BED file from the MACS output with scaffold regions and start-end coordinates. '-g' is the genome length limits

file that enlists the start to end coordinates of each genome scaffold. '-b' is the length of nucleotides to be extended. > 'slop_peakfile_2k' is the output file name.

The peaks were extended by 2000bp on both sides and these extended genome regions were intersected with the gene annotation feature gff file to locate the nearest gene spanning the extended peak region.

```
bedtools intersect -a slop_hw1200_2k.bed -b
Bombyx_mori.gtf   -wao >2k_anno.bed
```

'bedtools intersect is the command. '-a' is the input file, extended peak region (slop) BED file. '-b' is the genome annotation file in gtf format. '-wao' writes the original A and B entries plus the number of base pairs of overlap between the two features. > name is the output file name.

The intersection shows one peak intersecting many gene features, as a single gene is described by all the coding and non-coding features. The intersected files were then sorted to retain only one matching feature using the Microsoft excel function "=IF(COUNTIF)", which retains only one match pair per peak. Later a python code was written to solve this issue (Appendix 3).

The regions that did not have any gene upto 2000 bp were filtered and extended to 5000 bp to find the genes and similarly to 10000bp for the ones that did not have genes within 5000bp. The regions that did not have genes up to 10000 bp were filtered out and subjected to fetch non-overlapping intervals as described in 3.10.1.

The genes from these two methods were analyzed and were found to have almost identical gene sets with intersect genes set almost being a subset of the fetch. The few exceptions that were found were added to the fetch set and this pooled gene set was used to proceed further to identify the genes and homologs. After identifying all the peaks and their corresponding gene regions, there were still some peaks for which no gene could be assigned.

3.12 Annotation of genes associated with the peaks

The list of genes associated with the peaks was generated by using the BED files of ChIP enriched fragments (pulldown with antibodies against Ubx) and the *Bombyx* gtf annotation file from SilkDB. This list of genes was based on the annotation from SilkDB, which displays the BGI IDs and the detail based on the SilkDB genome. However, the data that reinforce the identity of the gene by using EST library, full-length cDNA and mRNA is accessible from the kaikobase silkworm genome database. BGI IDs are cross-referenced in the kaikobase database; hence we could use unique BGI ID of the identified gene near a ChIP peak to locate its corresponding ID in kaikobase data. This linking was done by a python program to compare and merge excel files based on unique IDs (Appendix C3). This allows us to get comprehensive detail on a particular gene with many features from two independent databases that will enable us to narrow down the reliable gene identity and function.

Apis and *Drosophila* homologs for these genes were obtained by using the Ensembl metazoan biomart download tool. The BGI IDs are the *Bombyx* identifiers in Biomart, hence these IDs were used to mine out the corresponding known homologs. As the butterfly genomes of *Danus plexipus* (Monarch butterfly) and *Heliconius melpomeme* (Postman butterfly) were sequenced by the time we did later part of this work, we also included the homologs in butterfly in addition to that of *Drosophila* and *Apis*. Many times one silkworm gene ID corresponds to more than one homologue in these insect orders. The homologs in *Drosophila* were used as the common medium to compare across the insect orders as it is the most studied insect genome with reliable and detailed annotation.

Results and Discussion

Sequence data files generated after deep sequencing of DNA pulled down by using anti-Ubx antibodies from fore- and hindwing buds of *Bombyx* were subjected to further analysis as described below.

3.13 FastQC quality control analysis of reads

First, the FastQC program was used to assess the quality of all sequence files of both the replicates. The results are as follows.

1. Per-base sequence quality: The per base sequence quality was found to be very good in most of the reads, except in one or two where the last base dipped in quality, but still within the acceptable phred scores (>30). Sometimes the sequence run was longer as per the availability of the flow cells. In such cases, trimmed data showed even better sequencing quality. Otherwise, no trimming was necessary for most of the datasets as all were of good sequencing quality.

2. Per-base sequence quality scores: It was always found to be a single peak at the expected 36 bp length for all the samples.

3. Per-base sequence content: It was observed that the GC content was uniform throughout the read in all datasets. Those datasets, in which GC content was found to be distorted and biased, were discarded. Fresh sequencing was done for those samples. Final data files used were in the allowed category of this parameter.

4. Per-base GC content: The GC content of silkworm genome is around 32% as per the genome sequencing reports (Mita et al, 2004). We found in most samples the GC content was seen to be around 40 %. There was no bias in any of the samples processed post sequencing and this was within acceptable limits.

5. Per-sequence GC content: was found to be a good normal distribution curve in all cases.

6. Per-base N content: The sequencing quality was found to be very good without any N content as the quality and base calling was of high quality.

7. Sequence length distribution: there was no distortion in the sequence length and it always showed a peak at the expected 36 bp region.

8. Sequence duplication levels: Some variation was observed in the sequence duplication levels even in the input, indicating the presence of repeats throughout the silkworm genome, which is known to have large number of repeat regions as compared to other insects. In the ChIP pulldown samples, greater duplication levels were observed probably due to smaller amount of starting DNA material for library preparation. However, the duplicates were taken care of in the downstream bioinformatics processing programs.

9. Overrepresented sequences: None of the reads that were used for the final analysis had any overrepresentation.

10. Overrepresented K-mers: In all the reads that were processed as final replicates, no overrepresented K-mers were observed.

The FastQC reports of the hind wing datasets are appended in figures 3.6 to 3.11.

3.14 Creating index of the genome with Bowtie

Bowtie 0.12.7 was used to create the index for genome for the alignment and the program bowtie-build created these files when run.

silk_index1.ebwt, silk_index2.ebwt, silk_index3.ebwt, silk_index4.ebwt, silk_index1.rev.ebwt, silk_index2.rev.ebwt

These files were placed in the folder with the bowtie executable files. These files can be transferred to different systems just by copying and pasting at the desired destination.

3.15 Alignment of the ChIP-seq reads to the genome

Bowtie 0.12.7 was used to align the ChIP experiment, negative control and input control reads for both fore and hind wings. The output was in aligned files in SAM format. These files were converted to BAM, formatted, sorted and indexed for downstream bioinformatics processing. To improvise the alignment, SAM-tools rmdup was used, which removes all the duplicate reads. However this idea was shelved as the Bombyx genome is known to have a large number of repeats and there is no consensus on if duplicates are due to their natural occurrence in the genome or due to amplification artifact. Two duplicates were retained for the analysis, while the default is 3 allowing some stringency for the alignment while taking care of the possibility of PCR duplicates, standardized after many runs with different parameters. The alignment results for both the replicates and the six files from fore and hind wings are summarized in table 3.1.

3.16 Visualization with IGV viewer

The sorted and indexed BAM files were visualized as a comparative visualization between input, pull-down using antibodies and negative controls (blank and IgG) along with genome and annotation files (Fig 3.3). A comparative visualization between these three sets allowed us to visualize the highly enriched and true peaks in comparison with the IgG and input controls. It also helped to locate the nearest genes from annotation for some genes. After the identification of peak, the bed file was used to identify the peak in the viewer. This visualization helped us identify the regions that were inconsistent between genome databases, genes that not yet validated but are known targets of Ubx in *Drosophila*, and the discontinuity of genome scaffolds.

3.17 Peak calling using MACS

The aligned and sorted BAM files of ChIP experiment and negative control were normalized to aligned and sorted Input file to identify the peaks. This was done for both hind and fore wing datasets by using MACS 1.4.2. Different parameters were altered and tested and the two replicates were treated to same parameters to identify the common peaks.

Initially, MACS generated a warning on not being able to build a model with the peaks. However, the FDR and later the annotation helped ascertain that the ChIP-seq has worked and quality of the data is satisfactory. A autocorrelation analysis of the datasets was done on by running them on HOMER. This exercise suggested that the ChIP data had good autocorrelation. As our data is from a new genome with many gaps and large number of repeats, the predictability based on the quality parameters of ChIP-seq data available so far from literature for other organisms may not be directly applicable.

MACS generated four tsv files as output. First a peak file with the scaffold coordinates of the peak, a unique MACS peak ID, length of peak, number of tags in the peak, %FDR, fold enrichment and 10*log10pvalue. A similar negative peak file was generated, which identifies the peaks, when input and experiment files are swapped to find peaks. A BED file was generated, which has the peak locations in terms of scaffold/ chromosome, start and end of peak and the -10*log10pvalue. A 'summits.bed' file was created which contains the peak summit locations of every peak with the height of fragment pileup. This file was useful in motif finding.

The negative control was chromatin pull-down with a non-specific anti-rabbit normal IgG and the peaks were obtained for this sample too by normalizing it with input control. The peaks obtained by the IgG negative control were considered nonspecific and if any of these peaks were found in the experimental datasets, they were deleted in order to only retain experiment-specific peaks. The post IgG filtered peaks were written to a new extended BED format file, with the peak location coordinates, FDR and fold enrichment values. Some peaks were found to have negative start coordinates. They were first converted to zero as this interferes with the functioning of downstream annotation programs.

Different FDR cutoffs (5, 10, 15 and 20) were applied to the hind wing data set and the number of genes in each set was compared (Fig 3.4). The difference in the number of peaks between FDR 15 and 20 was found to be the least, while the difference was larger between other consecutive sets. Therefore maintaining a good balance between the number of peaks and stringency, 15 % FDR was

used as the cutoff. Within this cutoff, 1128 peaks in hind wing and 340 peaks in fore wing were identified.

The number of peaks identified in each dataset is tabulated in Table 3.1.

3.18 Identification of genes associated with the peaks

As the annotation and coverage of the silkworm genome is not complete and some well-known genes are yet to be mapped, it was kept in mind that more genes would allow exploration and narrowing down of the relevant and confident genes rather than high stringency and lesser genes to work with in the first place. Therefore, two approaches (fetch and slop-intersect) to identify putative target/s of Ubx around a peak were used.

The genes from the fetch and slop-intersect methods were analyzed and were found to have almost identical gene sets, but the fetch set had more number of genes. As the genome coordinates of genes are not very reliable (Fig 3.2), we wanted to retain all genes possible in the given distance set and therefore used the list provided by fetch method. A small number of genes were identified only by slop-intersect method. These genes were added to the fetch set and this pooled gene set was used to for further analysis. After this exercise, only 28 peaks were left with no genes assigned and they were ignored.

870 genes for the hind wing and 245 genes for the fore wing dataset were identified after pooling the identification by both the methods (Fig 3.5).

3.19 Annotation of genes and identification of homologs

The gene identification was done primarily by mapping the reads to SilkDB genome and using SilkDB annotation. As the kaikobase database had more data to validate the identity of the gene like mRNA, full-length cDNA library, and EST database, we fetched kaikobase information for every BGI gene ID. This information was useful later to determine the identity of the gene better and also to study the biological relevance of a given gene being target if Ubx

The genes identified were used as query on Ensembl metazoan biomart to obtain the fly base IDs to determine the gene function as the fly genes are very well studied. Ensembl was also used to identify homologs of other insect such

as *Apis* (Honey bee), *Danus plexipus* (Monarch butterfly), and *Heliconius melpomeme* (Postman butterfly). Often, genes predicted in silkworm did not have confirmatory evidence (such as EST data), in such situations existence of a homolog from a different insect helped to ascertain the gene identity.

The features after linking the kaikobase data for every gene ID were as follows; location of the peak on the SilkDB scaffold, peak start and end coordinate, MACS peak ID, fold enrichment, %FDR, -10*log10pvalue, nearest gene scaffold, start and end coordinate of gene, distance of gene from the peak, BGI gene ID, exons, overlap of the gene to the extended peak (only in intersect data), kaikobase BMgn gene ID, kaikobase 'gene' ID, gene location in chromosome, start, end, kaikobase scaffold of gene, gene start-end, gene length, evidence for the identity of the gene from full length cDNA and mRNA, gene function based on homology, fly base protein ID, fly base CG number, protein family conserved domains (Pfam and Hmmer3), InterPro ID, GO (Gene ontology) biological process, GO cellular component, GO molecular component, tissue library from which the EST related to this gene was obtained, tissue in which the gene was expressed, number of ESTs in Wing, Wing disc, Embryo and cell, homologs from biomart databases of *Apis mellifera*, *Danus plexipus*, *Heliconius melpomeme* and *Drosophila melanogaster,* Fly base gene ID, Fly base CG number, Fly base gene symbol and Fly base gene name. Wherever, a BGI gene ID corresponded to more fly or other homologs and they were written into separate lines in an excel sheet.

81 fly homologues for putative targets of Ubx in forewing and 548 fly homologs for the putative targets of Ubx in hindwing were identified. These homologs were used in the comparative analysis described in the next chapter (Fig 3.5). Analysis included identification of kinds of developmental processes targeted by Ubx in *Bombyx* as against in *Drosophila* and possible evolutionary trend in *Bombyx* lineage as against *Drosophila* lineage from the ancestral *Apis* lineage.

Summary

This chapter described the analysis of sequences (reads) of the ChIP-enriched DNA (pulled down using anti-Ubx antibodies) from the *Bombyx* fore- and hindwings. Different quality control measures and the alignment of these reads to the silkworm genome were described. The aligned files were then used to find out the binding regions (peaks). The gene associated with the peaks were mined and annotated with homology-based information from various databases.

In the forewing dataset, 340 peaks were identified, with 245 genes associated to the peaks of which 81 have the homologs in *Drosophila*. In the hindwing dataset, 1128 peaks were identified of which 870 were associated with genes, of which 548 have homologs in *Drosophila*.

The next chapter (Chapter 4) describes the detailed comparative analysis of the targets of Ubx in *Bombyx, Apis and Drosophila*. It also describes a Gene Ontology (GO) analysis of the targets of Ubx in three insects and a comparison amongst them.

Plates
Chapter 3

Table 3.1 Alignment statistics of the ChIP-seq reads to the *Bm* genome

	number	%	number	%	number	%
	Hind wing set 1					
Reads	**Input**		**Bm Ubx ChIP**		**IgG control**	
Total	40546523		25046981		34538471	
uniquely aligned	20997202	51.79	10014365	39.98	3220871	9.33
failed	9297654	22.93	9491311	37.89	29286972	84.80
aligned to repeats	10251667	25.28	5541305	22.12	2030628	5.88
	Hind wing set 2					
Reads	**Input**		**Bm Ubx ChIP**		**IgG control**	
Total	27450288		24760478		34373212	
uniquely aligned	10044114	36.59	4277866	17.28	2336155	6.80
failed	5115628	18.64	16120520	65.11	29585342	86.07
aligned to repeats	12290546	44.77	4362092	17.62	2451715	7.13

	Fore wing set 1					
Reads	**Input**		**Bm Ubx ChIP**		**IgG control**	
Total	39479183		32814413		12771532	
uniquely aligned	24447194	61.92	15400384	46.93	4483048	35.10
failed	2477537	6.28	6379401	19.44	4331327	33.91
aligned to repeats	12554452	31.80	11034628	33.63	3957157	30.98
	Fore wing set 2					
Reads	**Input**		**Bm Ubx ChIP**		**IgG control**	
Total	24100558		29280813		25416392	
uniquely aligned	11301510	46.89	3077446	10.51	2263430	8.91
failed	3760440	15.60	24269918	82.89	22137585	87.10
aligned to repeats	9038608	37.50	1933449	6.60	1015377	3.99

The above tables are the statistics obtained for the ChIP-seq reads after aligning with Bowtie for two replicates (set1 and 2) each of fore- and hindwing data. The table depicts, for each replicate read, the total number and percentage of reads obtained after the sequencing run (Total), the reads that aligned uniquely to a genome region (Uniquely aligned), which are considered as the useful ChIP reads, the failed reads and the reads that aligned to repeat regions in the genome (aligned to repeats).

Input- input control sequenced and used as a normalization control, ideally should cover maximum regions of genome without any bias for any regions.

Bm Ubx ChIP- sequences obtained from ChIP experiment carried out with N terminal *Bombyx* Ubx against wing bud lysates. Should contain peaks.

IgG control- sequences obtained from control experiment with normal IgG sera to perform the ChIP pulldown.

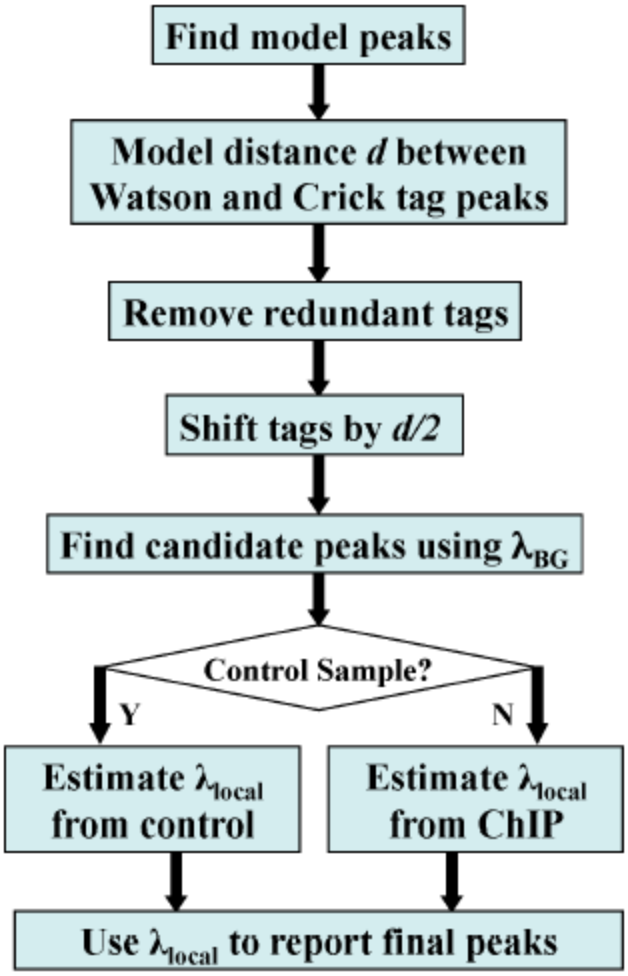


Figure 3.1 MACS flow chart explaining the method it uses to identify peaks

The flowchart depicts the algorithm that MACS uses to identify peaks and assign the enrichment values over control. (Source: Angelini C, EMBO practical course)

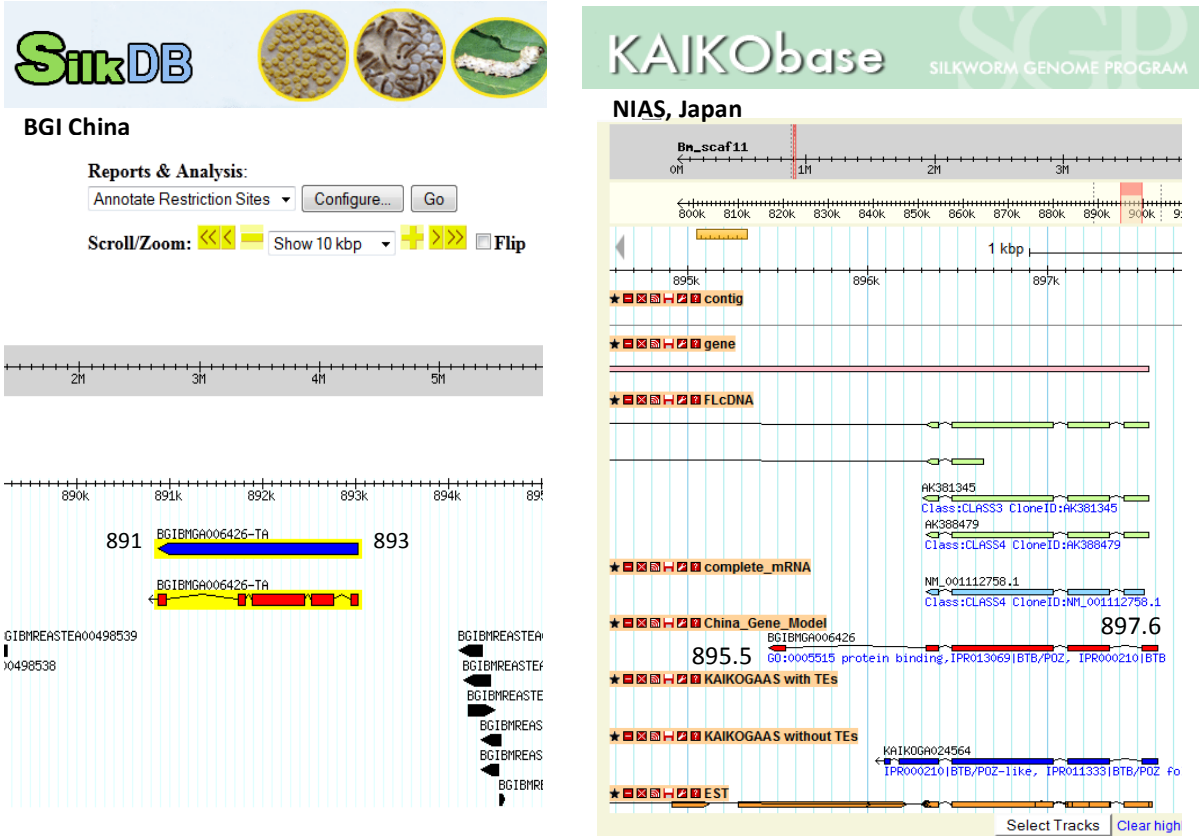
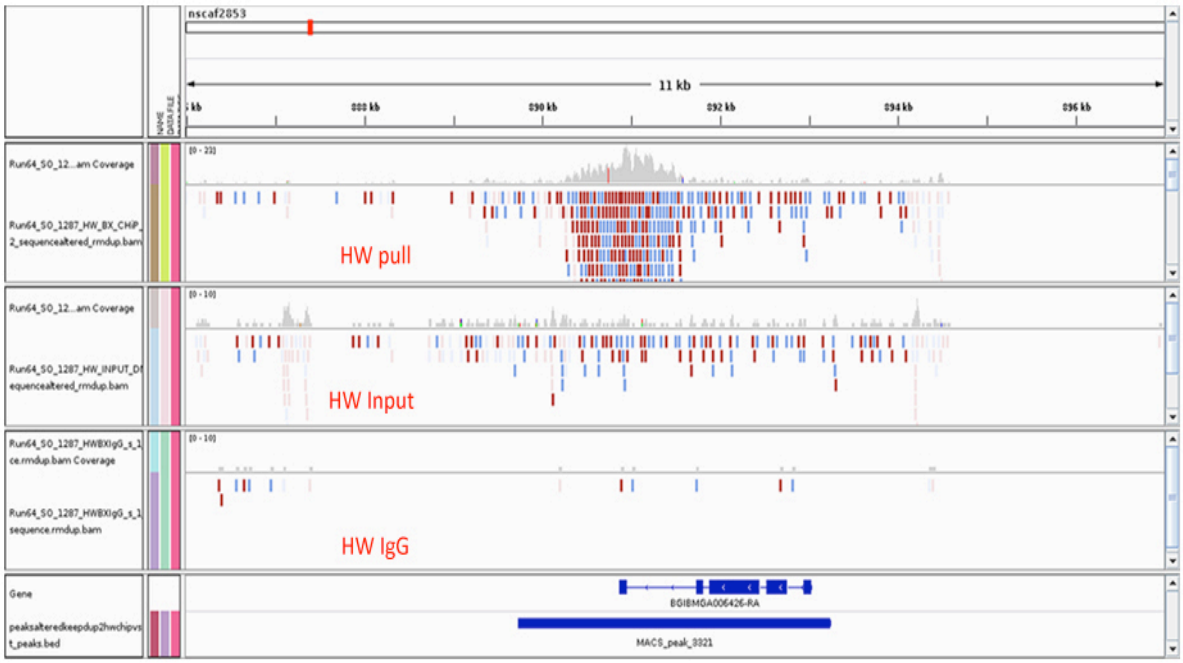


Figure 3.2 Silkworm genome databases.

The two panels above depict the genome browsing visualization of the two silkworm genome databases namely, the SilkDB and Kaikobase. Both the databases were used to identify and annotate the genes. As an example these panels show the location of the gene "*modifier of Mdgn4*". What can be observed here is that the coordinates in both the genomes are different and hence the correlation of this gene to a peak is difficult. The same (putative) gene in SilkDB shows a longer gene between 891-893Kbp while Kaikobase shows a complete gene with longer gene region with intron and exons.

The fetch method for finding the gene nearest to a peak does not allow overlap with the gene region. Hence in the case of silkdb, as the peak we identified overlapped a little region of the gene, this gene was not listed as a target. But the intersect method allows a gene overlap. Hence the intersect method was used as a complimentary method to the fetch-method to get a complete coverage of the targets. Kaikobase is reliable for gene identification with a variety of evidence for identifying a gene region, like Full length cDNA (starting with the name AK) and mRNA (starting with NM). The BGI ID is the common link to both and also displayed in the kaikobase in red.



A

B

10,000bp

PEAK/
Ubx binding region

10,000bp

Target gene

Target gene

Figure 3.3 Visualization with IGV viewer and assignment of genes to peaks.

A- IGV visualization of the peak corresponding to the gene *Modifier of Mdgn4* in a comparative panel with pull-down, input and negative (IgG) control. The input is flat (no specific enrichments) throughout and the negative controls had very few tags. The last two panels show the gene from annotation gtf file and the peak from a BED file from MACS.

B- All genes found within 10Kb of the peak identified were assigned to the peak and used for further analysis.

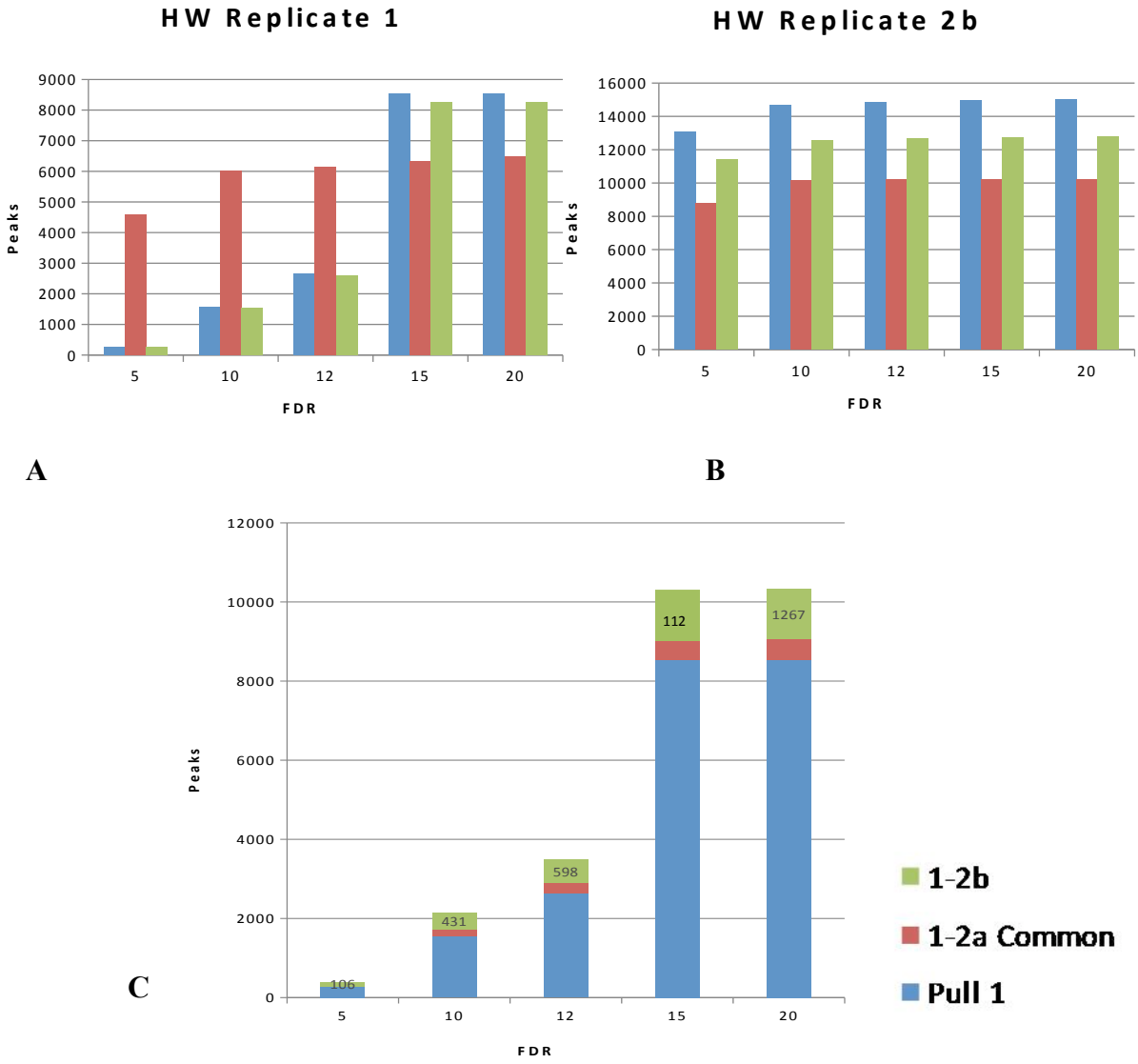


Figure 3.4 Number of Peaks in replicate 1 and 2 of hind wing

A- Peaks in replicate 1 of ChIP seq with *Bombyx* hindwing. The increase in the number of pulldown peaks from 10 to 15% FDR is drastic hence 15% FDR was chosen as the cut off.

B- Peaks in replicate 2 of ChIP-seq with *Bombyx* hindwing

C- The common peaks between replicate 1 and 2 are shown in green. The entire bar represents all the peaks. The brown bars are from the first sequencing run of the second replicate, which was not of acceptable quality the same sample was re-sequenced and constitutes the second replicate.

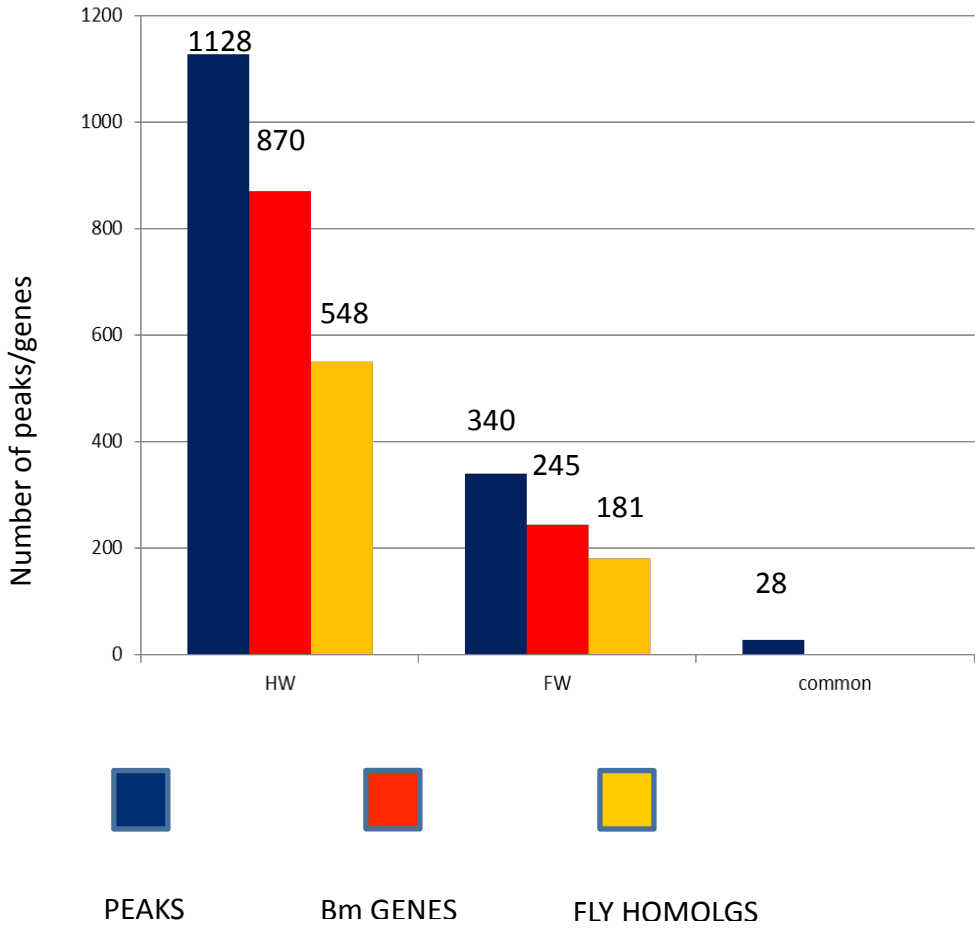
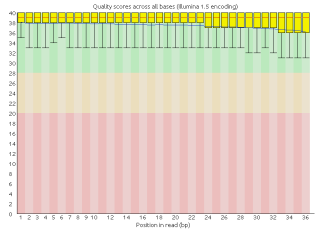
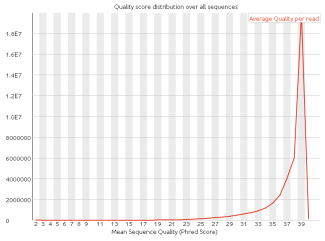


Figure 3.5 Peaks common between two IgG filtered replicates

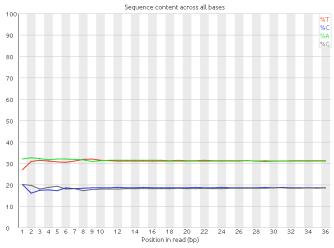
The distribution of peaks, assigned genes and corresponding fly homologs for the forewing (FW), hindwing (HW) and common (to both FW and HW) datasets at 15% FDR. The genes in hind wing are higher in number as Ubx is present only in the peripodial membrane of the forewing while it is expressed throughout the hind wing.



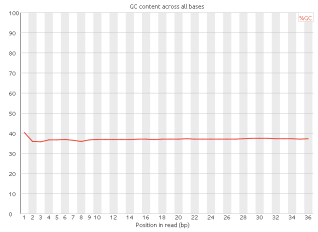
A. Per base sequence quality



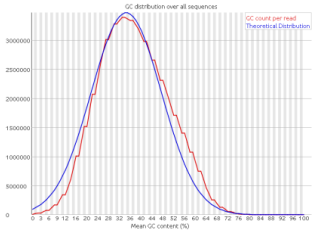
B. Per sequence quality score



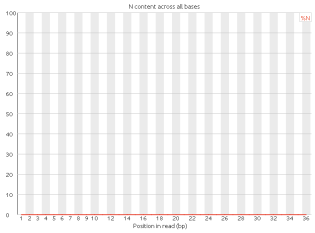
C. Per base sequence content



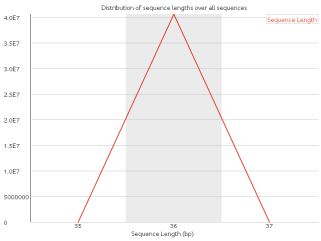
D. Per base GC content



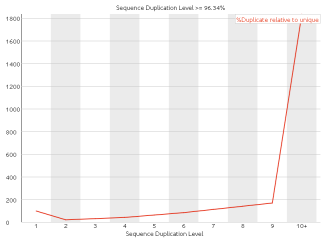
E. Per sequence GC content



F. Per base N content

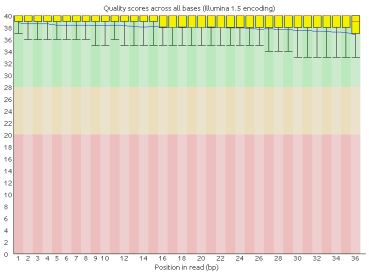


G. Sequence length distribution

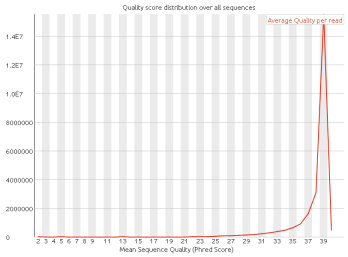


H. Sequence duplication levels

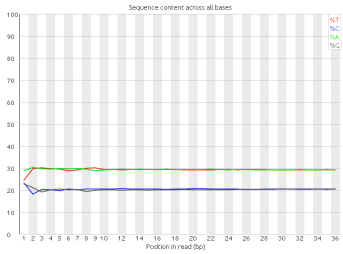
3.6 FastQC quality report of hindwing Input dataset 1



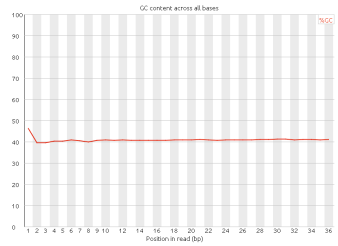
A. Per base sequence quality



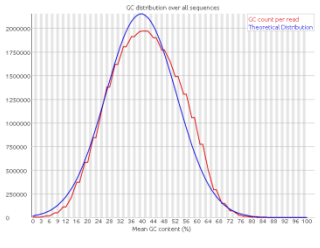
B. Per sequence quality score



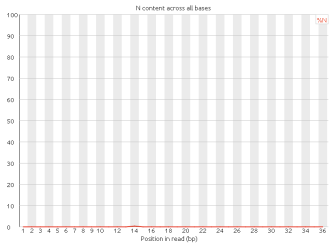
C. Per base sequence content



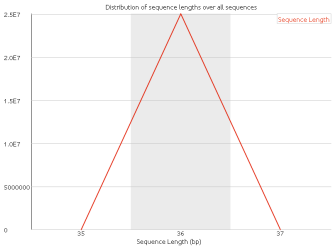
D. Per base GC content



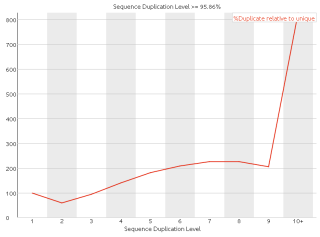
E. Per sequence GC content



F. Per base N content

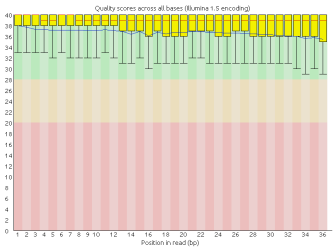


G. Sequence length distribution

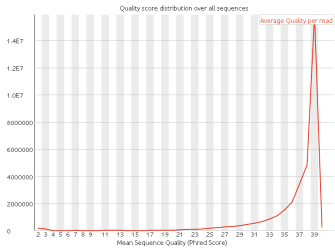


H. Sequence duplication levels

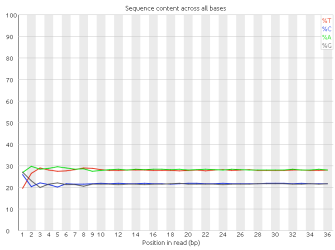
3.7 FastQC quality report of Ubx ChIP dataset 1 for hindwing



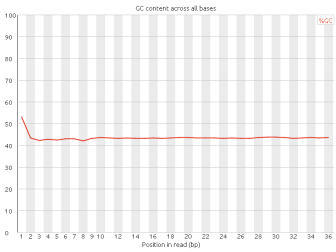
A. Per base sequence quality



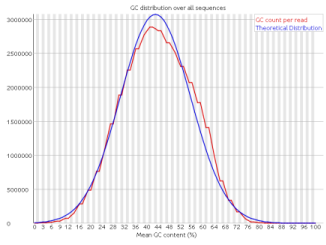
B. Per sequence quality score



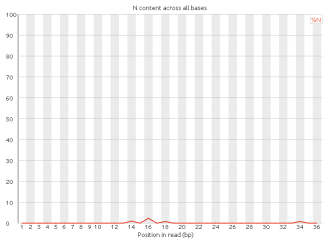
C. Per base sequence content



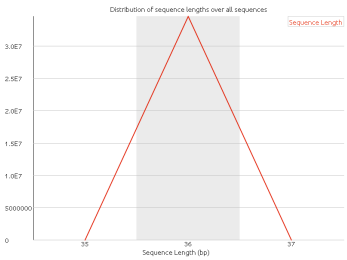
D. Per base GC content



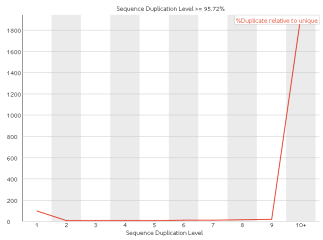
E. Per sequence GC content



F. Per base N content

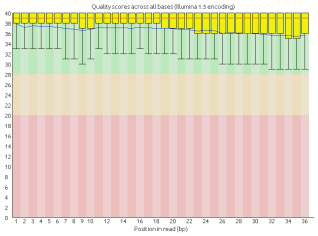


G. Sequence length distribution

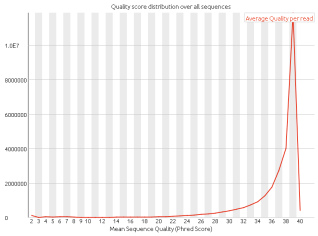


H. Sequence duplication levels

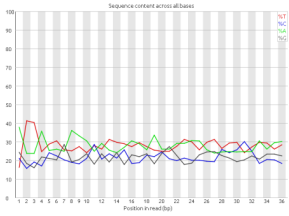
3.8 FastQC quality report of IgG negative control dataset 1 for hindwing



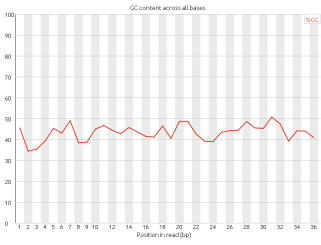
A. Per base sequence quality



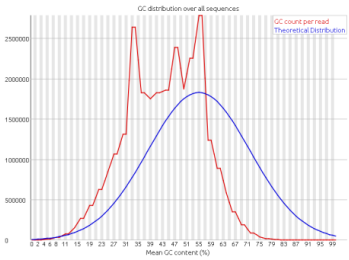
B. Per sequence quality score



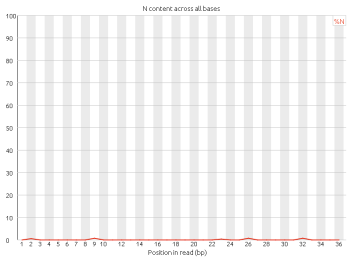
C. Per base sequence content



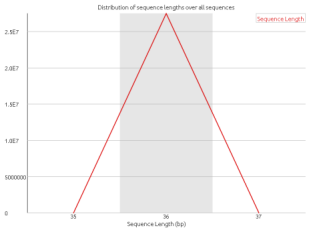
D. Per base GC content



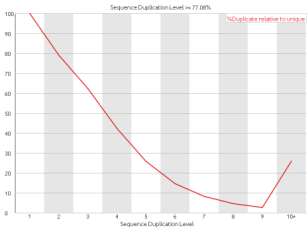
E. Per sequence GC content



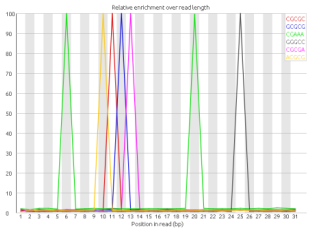
F. Per base N content



G. Sequence length distribution

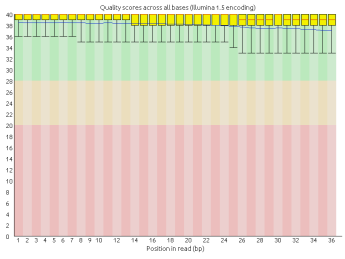


H. Sequence duplication levels

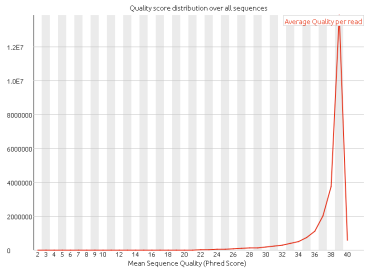


I. Kmer content

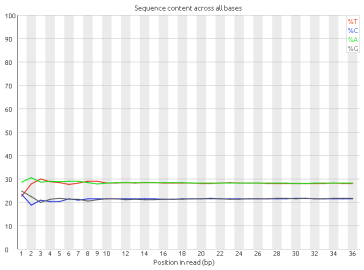
3.9 FastQC quality report of hindwing Input dataset 2



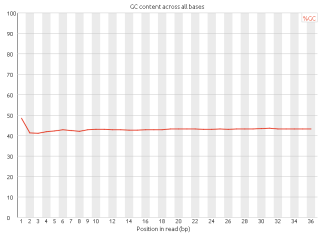
A. Per base sequence quality



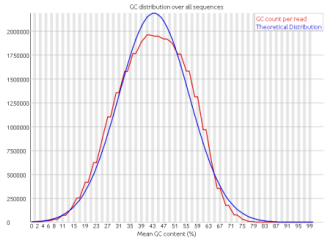
B. Per sequence quality score



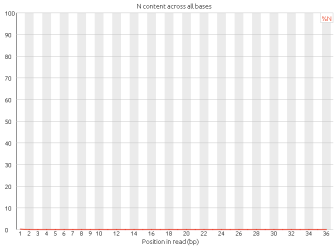
C. Per base sequence content



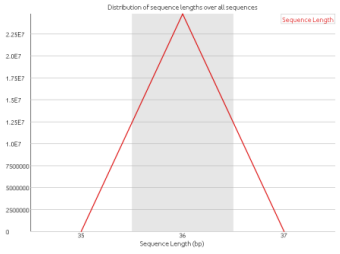
D. Per base GC content



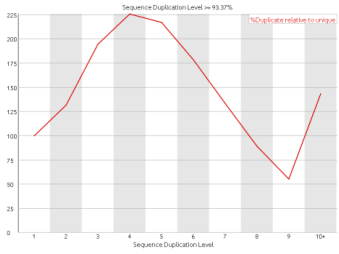
E. Per sequence GC content



F. Per base N content

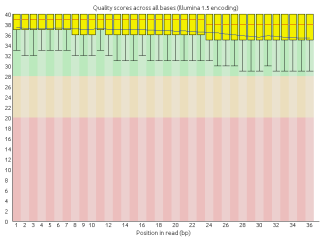


G. Sequence length distribution

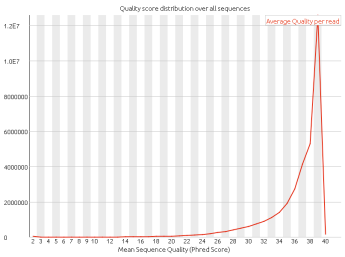


H. Sequence duplication levels

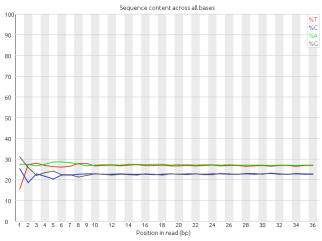
3.10 FastQC quality report of Bm Ubx ChIP dataset 2 for hindwing



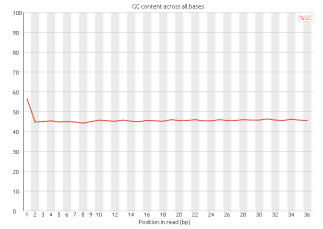
A. Per base sequence quality



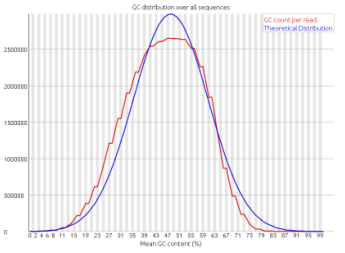
B. Per sequence quality score



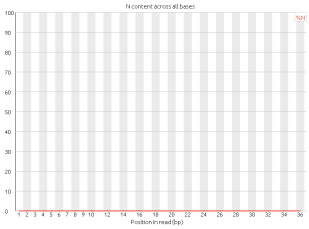
C. Per base sequence content



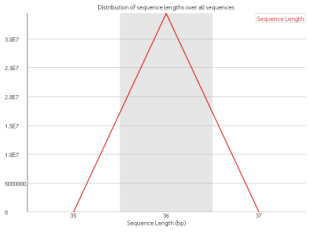
D. Per base GC content



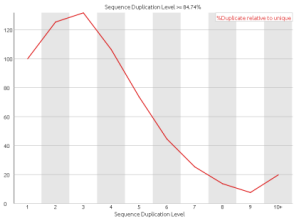
E. Per sequence GC content



F. Per base N content



G. Sequence length distribution



H. Sequence duplication levels

3.11 FastQC quality report of IgG negative control dataset 2 for hindwing

Chapter 4

Comparison of targets of Ubx from *Bombyx*, *Drosophila* and *Apis*

Introduction

4.1 Comparative genomics

Since the advent of sequencing of complete genomes of organisms, comparative genomics has emerged as the favorite tool to understand the biological meaning of the excessively large body of the sequence data available. It involves the comparison of various genomic features between two or more organisms. Comparative genomics allows us to apply the modern evolutionary theory, which assumes that two genomes under comparison had a common ancestor and, therefore, origin and biological information of almost every base in the organism may be explained.

Evolution is a combination of two natural processes, the random mutations and the forces of natural selection that shape the organism by selecting traits and genes responsible for the same. The selection forces eliminate deleterious mutations (negative selection) or increase the frequency of the mutant alleles that allow a gain in fitness (positive selection) or they may not exert any effect, if the mutations are neutral (neutral selection). This process of mutation and selection is reflected in the base pair differences in the genomes of the organisms. Although the evolution of an organism may be theoretically deduced by comparative genome analyses; our understanding of the evolutionary processes is inadequate to explain all real-life observations. This is the reason why most work has focused on the analysis to gain insights into function of conserved sequences, which, in general, means understanding the negative selection (Ureta-Vidal et al, 2003).

In order to obtain an insight into the evolution of Dipteran haltere, we compared the targets of Ubx in the hindwing appendages of the three insects studied in our lab (*Drosophila*, *Apis* and *Bombyx*).

4.2 Direct targets of Ubx from *Drosophila* and *Apis*

In the current study, we identified targets of Ubx in *Bombyx* hindwing, as part of a larger effort to understand the role of Ubx in shaping the hindwing appendage by comparison with the targets of Ubx with data available from our

lab and other labs on other insects. Amongst the relatively large number of studies on the development of haltere in *Drosophila* are the more recent ChIP-on-chip experiments to identify direct targets of Ubx (Choo et al, 2011, Slattery et al, 2011 and Agrawal et al, 2011). Slattery et al. (2011) used antibodies against the full length Ubx, containing the homeodomain, which amounts to some non-specificity, and hence their data was not used for comparison in this study. Choo et al. (2011) used a Ubx::YFP protein trap line and antibodies against YFP for the ChIP, while previous studies in our lab (Agrawal et al., 2011) used polyclonal antibodies against a N-terminus fragment of Ubx to identify Ubx-specific targets. Data from these two studies were used in comparative analyses of direct targets of Ubx in this study.

Our laboratory has recently identified direct targets of Ubx in the hindwing of *Apis* using polyclonal antibodies against a N-terminus fragment of Ubx (Naveen, 2013) and this data is also used for the comparative analyses in this study. These gene lists allowed us to compare and analyze direct targets of Ubx from three different insect groups (*Apis*, *Bombyx* and *Drosophila*) with diverse hindwing morphology. As wing and haltere development is well studied in *Drosophila* and large number of targets of Ubx are functionally characterized in the context of wing development in *Drosophila*, the comparative analyses were focused on identification of evolutional mechanism leading to haltere development. Therefore, only those targets of Ubx for which fly homologues exist were considered for analyses.

4.3 Gene Ontology databases

The next generation sequencing era has generated enormous amount of genome-wide data at rates that exceed the knowledge of functions of genes and regulatory elements. To convert the genome-wide data to meaningful biological information, they need to be analyzed and annotated with the help of known genes in the closely related organisms. Many databases, which are publically available today, such as KEGG, Panther, Ensembl, Swiss-prot and DAVID, focus on the annotation and curation of functional data on each gene by such comparisons.

Gene ontology (GO) is a bioinformatics initiative to consolidate all gene and protein representations from various species to understand the biological function of the genes and their products. A GO database generally maintains and develops gene and gene product attributes in a species independent manner. The ontology covers three aspects of biology: (i) biological process: operations or sets of molecular events pertaining to functioning of cells, tissues, organs and organisms, (ii) cellular component: parts of the cell or its environment, and (iii) molecular function: the activities of a gene product at the molecular level, such as binding or catalysis. A typical database also puts efforts to annotate genes and gene products and disseminate this data to the researchers to use them in finding new genes and functions in many species. It also maintains tools to analyze the data and visualize it in an enrichment analysis or building a gene network. In summary, a GO analysis describes how a gene product behaves in a cellular context.

Considering the amount of data present in a number of databases, one would need a user friendly interface that facilitates transition from large scale genomic data to meaningful biological interpretation. Database for Annotation, Visualization, and Integrated Discovery (DAVID), is one such bioinformatics database with tools to mine through the biological data associated with genes and gene products (Jr et al, 2003). It gives descriptive data with intuitive graphic displays and visualization tools to analyze and visualize data. DAVID rapidly annotates and summarizes gene and protein lists according to shared categorical data for Gene Ontology, protein domain, and biochemical pathway membership. It provides functional classification into biological processes, pathways, cellular location and molecular function of gene products. It also describes biochemical pathway maps and conserved protein domain architectures, while being linked to a rich source of biological annotation. DAVID is suitable for functional annotation and analysis of human, mouse, rat or fly genomes. The GOTERM_BP_FAT was used for biological process (BP) while the Kyoto Encyclopedia for Genes and Genomes (KEGG) term was used for pathway analysis.

4.4 Visualization through BioVenn and Circos

One of the most popular methods to visualize data relationships like overlap and exclusion between data sets is the Venn diagram. Biovenn is a user-friendly online tool to generate area-proportional Venn diagrams from lists of biological identifiers (Hulsen et al, 2008). It supports a wide range of identifiers from biological databases.

Circos is software, which allows effective visualization of complex data and information such as genomic data through a circular layout (Krzywinski, 2009). The visualization is ideal to explore relationships between multiple objects or positions. It allows, for example, creation of Circos plots to describe relationships between data groups or multi layered annotations of genes. Circos was primarily designed to visualize and present genomic data. However, now it finds many uses to present statistical representations. Circos was used in this work to compare all the available databases on targets of Ubx to get a holistic picture of the relationships between these data and the proportion of genes shared between them.

Materials and Methods

4.5 Comparative analysis of targets of Ubx in insects

As this study intends to understand the targets that have come under the regulation of Ubx during evolution, we only considered the genes that have homologs of each other in all the three insects for the comparative and GO analyses. As wing development in *Drosophila* is one of the most studied systems at cell and molecular levels, we used those targets of Ubx in *Bombyx* and *Apis*, for which valid fly homologues exist. Targets of Ubx identified by Choo et al. (2011; referred to as ***Drosophila* (R)**) and Agrawal et al. (2011; referred to as ***Drosophila* (P)**) using ChIP-on-chip methods were used for *Drosophila*, while Fly homologs of targets of Ubx identified by Prasad (2013) using ChIP-seq method were used as targets of Ubx in *Apis*. Unless otherwise specified, all the comparisons were for targets of Ubx in the hindwing in *Bombyx* and *Apis* and haltere in *Drosophila*. Three lists of Ubx targets (in the form fly base homologs/IDs) from *Bombyx*, *Apis* and *Drosophila* were made in text format. The tool BioVenn was used to compare the fly base IDs and to plot the Venn diagrams.

Targets of Ubx were directly compared with each other to identify targets that are common and specie-specific between the three organisms. Comparisons were made between targets of Ubx from *Bombyx* fore- and hindwing, *Bombyx* hindwing and *Apis* hindwing, *Bombyx* hindwing and *Drosophila* haltere. Comparison of targets of Ubx in *Apis* hindwing and *Drosophila* haltere has already been done in our laboratory in an earlier study (Naveen, 2013). Finally a three-way comparison was done between the insects between hind wing/haltere targets of *Bombyx*, *Apis* and *Drosophila*.

To compare all the datasets and to visualize the extent of the overlap between them, a Circos plot (Krzywinski et al. 2009) was employed. The datasets compared were targets of Ubx from *Bombyx* and *Apis* (both fore- and hindwing), *Drosophila* (*Drosophila* (R) and *Drosophila* (P)). The Circos plot was generated from a tabular representation of the overlapping number of genes between the datasets. Only one direction of overlap between any two data sets was

considered while tabulating the overlapping gene numbers, for the sake of non-redundancy and clarity of the plot. A tabular format was created as a text document and uploaded to the online table visualizer to obtain the Circos layout of the connections between the data sets. In a typical Circos plot, the rows and columns are represented by circularly arranged segments on the inner circle. The angular size of the segment is proportional to the total value (number of genes in dataset) of cells for the row or column. The cell values (number of overlapping genes) are represented by uniquely colored ribbons proportional to the value between a row and a column. Relative contribution of the individual cell values of a given row or column is encoded by circularly arranged stacked bars in the outer circle.

4.6 Gene Ontology analysis of target sets

The gene lists were uploaded to the DAVID Bioinformatics resources version 6.7 online database to obtain a Gene Ontology (GO) classification. First the fly homologs of targets of Ubx in *Bombyx* were uploaded to obtain the biological process (GO TERM_ BP_FAT) and the pathways (KEGG) grouping of the genes. A default *Drosophila* gene background was selected in the DAVID database to calculate fold enrichment and percent genes associated with a process or pathway. The detection thresholds were reduced to the minimum (gene) count of 1 from a default 2 and the ease score (a modified Fisher Exact P-Value) of 1 from a default 0.1 to include maximum number of genes and categories. The percentages of genes associated with a process were plotted in graphs for the GO terms relevant to wing growth and development. Targets of Ubx from *Apis* and *Drosophila* (*Drosophila* (R) and *Drosophila* (P)) lists were processed in the same way.

4.7 Comparative Gene Ontology analysis

To compare the percentage of genes representing each of the GO terms, a comparative analysis was carried out by plotting the percentages for all the processes and pathways between the three insects.

Extending the pairwise comparative analysis between insects (4.2), the common and species-specific gene lists were subjected to GO analysis to obtain process

and pathway gene associations. The percentages obtained in common and species-specific target lists were plotted and subjected to a pairwise comparative analysis. GO terms relevant to wing development in general, which were highly represented and also the terms specific to wing disc development were plotted and compared.

Results and Discussion

Here we report the observations of comparing targets of Ubx across three insect species. Targets of Ubx in *Bombyx* are from this study. Targets of Ubx in *Drosophila* are for from two different studies. Choo et al. (2011; referred to as *Drosophila* (R)) and Agrawal et al. (2011; referred to as *Drosophila* (P). Targets of Ubx in *Apis* are from Prasad (2013). Unless otherwise specified, all the comparisons were for targets of Ubx in the hindwing in *Bombyx* and *Apis* and haltere in *Drosophila*. Comparisons were made only for a subset of targets of *Bombyx* and *Apis*, for which fly homologues are listed in the databases.

The ChIP-chip experiments carried out earlier in our lab on *Drosophila* discs were performed using the late third instar larval discs (Agrawal et al, 2011). The development patterns of wing buds/discs in the insects are different with different development time scales. To compare the Ubx binding across insect orders we estimated the closest stage of the larval instar to that of *Drosophila* late third larval instar for each of the insects used. We also had to consider the feasibility aspect where the wing bud dissection was possible in good numbers for the chromatin preparation.

The fifth instar wing discs of the honey bee *Apis mellifera* was used for the ChIP experiments. The developmental marker cut was used to identify the stage which is equivalent to the late third larval instar of *Drosophila*. Many other developmental markers were used to analyze the expression patterns in *Apis* wing discs (Prasad N, 2013).

4.8 Comparative analysis of targets of Ubx in different insects

As described in Chapter 3, 245 targets of Ubx were identified for the forewing of Bombyx and 801 targets for the hindwing, only 43 targets were common to both. Amongst these, 181 and 548 had fly homologues, respectively for fore- and hind-wings. Significantly lesser number of targets of Ubx in the fore wing compared to the hind wing was not surprising as the forewing expresses Ubx

only in the peripodial layer. When the fly homologs of the targets in fore- and hindwing were compared, it was observed that 36 targets are common to both. *vestigial*, a pro-wing gene in *Drosophila* was a notable target found to be shared between the two wing discs.

When targets of Ubx in *Bombyx* were compared to those of *Drosophila*, we noticed that many genes essential for *Drosophila* wing development, such as *brinker*, *engrailed*, *hedgehog*, *vestigial etc*, are common to the two species. Amongst targets of Ubx only in *Drosophila* too are few genes known to have important function during wing development, such as *ten-m, vein, wingless, dpp, homothorax, Notch*. These targets may have come under the regulation of Ubx in the course of evolution and may play an important role in haltere specification.

It was observed that 19.8 % targets of Ubx in *Bombyx* were common to *Drosophila* (R), while 9.48 % of genes were common between *Bombyx* and *Drosophila* (P) (Fig 4.4 and Fig 4.7). When targets of Ubx in *Apis* were analyzed similarly, it was observed that 16.58 % were common to *Drosophila* (R) and 7.45 % were common to *Drosophila* (P) (Fig 4.14) (Prasad, 2013). We observed that 16.6 % of targets of Ubx in *Bombyx* were common to those in *Apis* (Fig 4.10). In summary, somewhat higher percentage of targets was shared between *Drosophila* and *Bombyx* as compared to *Drosophila* and *Apis*. This is reflective of the fact that Lepidopteran and Dipteran lineages diverged much later compared to Hymenopteran lineage. When all the three sets namely targets of Ubx in *Bombyx*, *Apis* and *Drosophila* (R) were compared together, the genes common to all the three data sets were very few (33), but most of these genes are well known in the context of *Drosophila* wing development (Fig 4.15), suggesting an essential role for Ubx in the hindwing development/modification in all insect groups, even when the diversification of forewing-hindwing mornphology is minimal.

To visualize the comparison between targets of Ubx across all the three species, a Circos plot was drawn (Fig. 4.16). In a circular plot showing the percentage of genes shared between each of the dataset, the Circos plot allowed us to visualize a holistic view of the comparison.

A key feature that was visually evident from the Circos plot was that, amongst the three insects only *Apis* shared a larger number of targets of Ubx between the fore- and hindwings (Fig 4.16). The forewing of *Apis,* unlike in *Bombyx* and *Drosophila,* expresses Ubx during development to the same extent seen in developing hindwing,. Therefore, the targets of Ubx are expected to be common between the two developing wings.

4.9 Gene Ontology analysis of target sets

Genes were assigned to various biological processes and pathways by using DAVID. The *Drosophila* homologs of targets of *Bombyx* were subjected to such an analysis to understand the kind of biological processes that are targeted by Ubx during hindwing development. The percentage or proportion of the genes represented in the biological process or pathway is reported by the program DAVID, this percentage is used to plot the graph (Fig 4.2). All the molecular and cellular processes that are essential in shaping the wing in *Drosophila* were represented in greater proportions.

The biological process "regulation of transcription" was found to have the maximum number of genes amongst the targets of Ubx in *Bombyx* with 13.22 % of genes of the dataset representing this GO category.

4.10 Comparative Gene Ontology analysis

The GO classification was also done for targets of Ubx in *Drosophila* (both *Drosophila* (R) and *Drosophila* (P)) and *Apis*.

Amongst all the GO categories, we observed that genes related to processes such as regulation of transcription and wing development are overrepresented in all the three insect orders studied here. However, as a general trend, a given GO category is represented in similar proportions across the three insect orders (Fig 4.3 and Fig 4.17). *Apis* lineage (Hymenoptera) branched some 350 million years ago to give rise to the branch that led to *Bombyx* and *Drosophila*. This suggests that Ubx has been targeting similar biological processes across these three very diverse insect groups for the past 350 million years. Interestingly, a trend was observed wherein the more ancestral *Apis* has the least percentage of genes represented for each of GO category. It is followed by *Bombyx*, while

Drosophila has the highest representation. The novel targets that are specific to *Drosophila* may have played a part in quantitatively increasing the differences between wing and haltere. This increase in the percentage of genes in *Drosophila* is more prominent in case of processes such as cell adhesion and regulation of growth (Fig 4.3). *Drosophila* (Dipteran) lineage further diverged from *Bombyx* (Lepidopteran) nearly 250 million years ago. This suggests that evolution of dipteran is correlated (and perhaps a main driving force) with increased number of wing development genes coming under the regulation of Ubx (Fig 4.20). Cell adhesion, proliferation and growth control are some of the developmental tools that may be regulated by Ubx to shape a globular haltere from a default flat-shaped wing state in the T3 segment.

As this study intends to understand the evolution of the hindwing appendage, targets of Ubx specific to a given insect and common between two insects were segregated and GO analysis of these genes was performed. When the percentages of these genes were plotted, it was observed that each of the GO categories was over-represented amongst the common targets as against species-specific targets (Fig 4.5-4.7). As only those genes for which fly homologues are known were considered for all these comparisons, it is likely that while different insect groups have different number of targets with large number being specific to that particular insect group, there appears to be a strong selection pressure for similar functional categories of genes to be targeted by Ubx in all lineages.

Interestingly, proportional representation of wing development-related genes amongst the targets of Ubx that are common to *Bombyx* and *Apis* was comparatively lower to the enrichment amongst the targets common to *Bombyx* and *Drosophila* (Fig 4.13). This is in contrast to the fact that hindwing morphology has not diverged much from that of forewing in *Apis* and *Bombyx* compared to morphological differences between wing and haltere in *Drosophila*. Furthermore, targets of Ubx that are common to *Drosophila* and *Apis* had higher proportional representation of genes related to wing development compared to the targets of Ubx in *Bombyx* and *Apis* (Fig 4.13). This suggests that Ubx in *Drosophila* appear to have retained as well as acquired more wing-development genes as its targets.

Summary

To summarize, this chapter described the comparative analyses of direct targets of Ubx in the hindwing of *Bombyx* and *Apis* and haltere from two ChIP-chip studies in *Drosophila* by direct comparison of the fly homologs between these insects. It was observed that *Bombyx* and *Drosophila* shared a higher percent of common targets between them as compared to *Bombyx* and *Apis*, although the wing morphologies are similar between fore- and hind-wings in *Bombyx* and *Apis*.

When targets were compared as subgroups belonging to different GO categories, it was observed that similar biological processes are regulated by Ubx in similar proportions, without much difference in the kind of processes and pathways actually regulated by Ubx.

A pairwise comparison was done between the GO-classified common and species-specific targets of Ubx. It was observed that genes belonging to each of the GO categories were represented disproportionately in higher number amongst the targets that are common to two insect species as against species-specific targets. Finally, it was observed that Ubx in *Drosophila* not only has retained wing development-related genes as its targets from lineages ancestral to the divergence of *Apis* lineage, has acquired more such genes as its targets compared to *Bombyx* lineage.

As Ubx does not appear to target any specific development pathway or biological process only in *Drosophila*, evolution of haltere could be driven by increase in the number of wing development-related genes coming under the regulation of Ubx. Although expression pattern of very few genes are studied during the wing development in Lepidopteran insects (such as *Bombyx* and *Precis*), all those (nubbin, Wg, Dll) genes have shown identical expression pattern between forewing and hindwing. As these genes are differentially expressed between wing and haltere during *Drosophila* development, diversity in the morphology in insect wings could also be attributed to the level of

evolutionary changes in the regulatory sequences. In this context, Chapter 5 describes a whole-transcriptome study on the *Bombyx* fore and hindwings and its comparison to the microarray data available in *Drosophila* wing and haltere.

Plates
Chapter 4

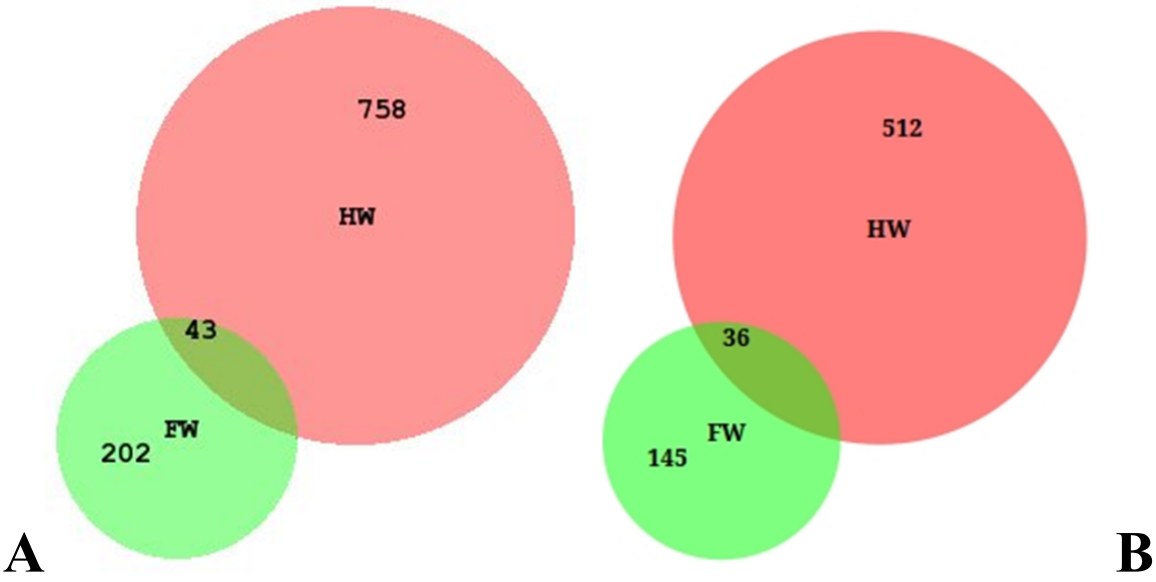


Figure 4.1 A. Venn diagram showing the extent of overlap between the Ubx targets (**BGI IDs**) in fore- and the hindwing of *Bombyx*. **B. V**enn diagram showing overlap of Ubx targets, when **only fly homologs** of the targets in fore- and hindwing were considered. Very few targets are shared between fore- and hindwings of *Bombyx* and this is expected, as the forewing does not have prominent Ubx expression.

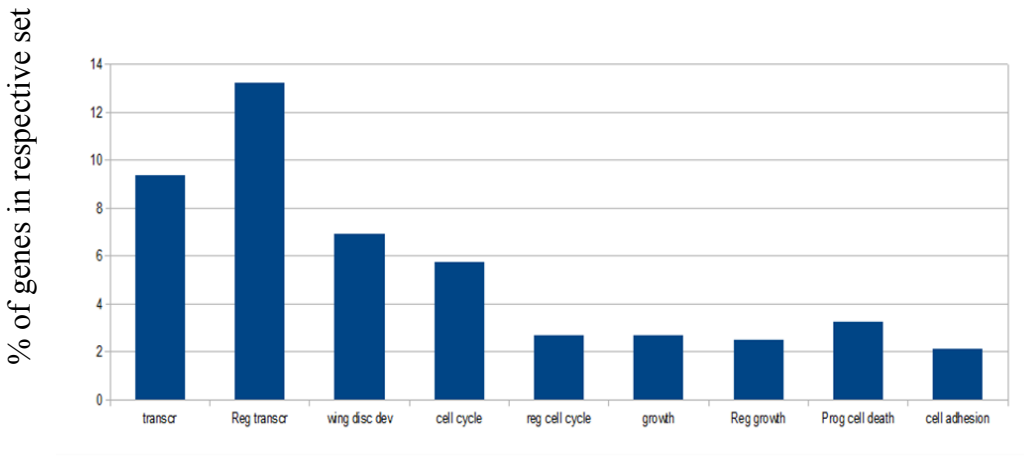


Figure 4.2 Gene Ontology analyses of targets of Ubx targets in the hindwing in *Bombyx*. Selected GO biological process (on X-axis) categories relevant to wing development are plotted in this graph. In general all the molecular and cellular processes that are essential in shaping the wing in *Drosophila* are represented in considerable proportions.

transcr = Transcription, reg = regulation of, Prog = Programmed , dev = Development

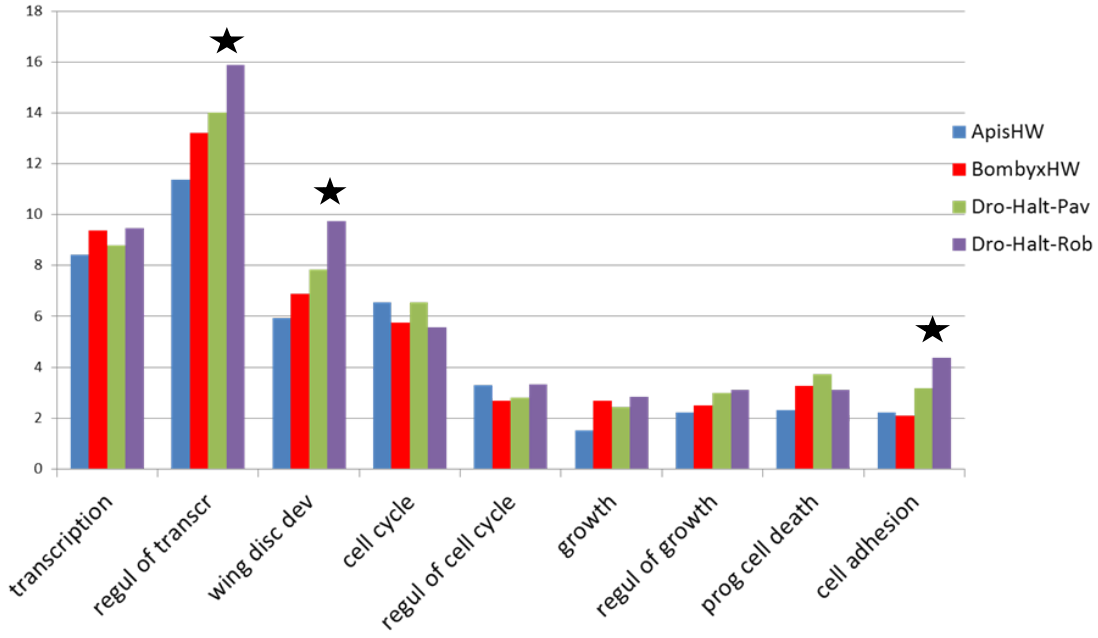


Figure 4.3 A comparative graph of the GO analysis (biological process: X-axis) of targets of Ubx in the hindwing of *Apis*, *Bombyx* and haltere in *Drosophila*. As a general trend, a given GO category is represented in similar proportions across the three insect orders; however certain processes (starred) show an increasing trend in *Drosophila.* These genes may have key roles in the specification of haltere.

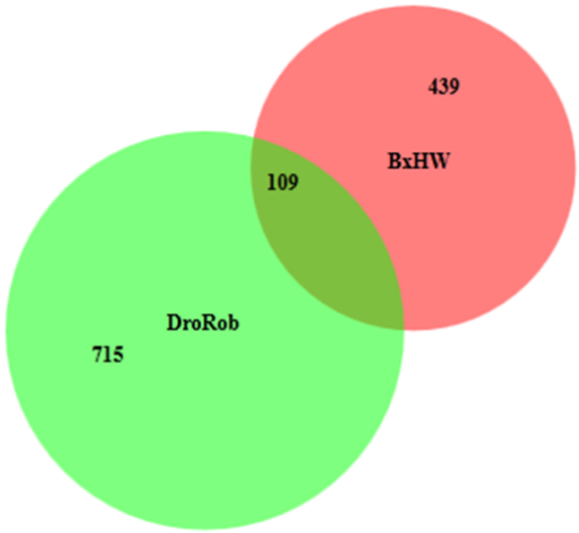


Figure 4.4 Comparison of targets of Ubx (only for which fly homologs exist are considered here) between *Bombyx* hindwing and *Drosophila* (R). 19.8 % of targets of Ubx in *Bombyx* hindwing are common.

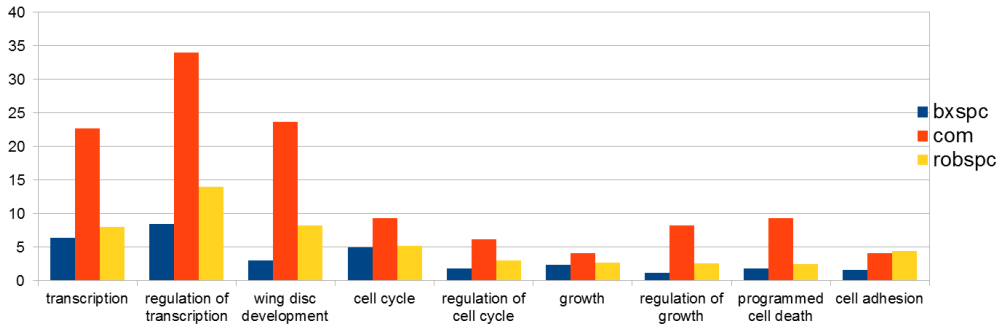


Figure 4.5 A comparative GO analysis of the targets of Ubx common to *Bombyx* and *Drosophila* R (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Drosophila*). GO categories relevant to wing development in general are plotted. Each GO category is enriched at higher proportion in common targets compared to species-specific ones. The GO terms for the categories in the graph are given below.

GO term and Biological process
GO:0006350~transcription
GO:0045449~regulation of transcription
GO:0035220~wing disc development
GO:0007049~cell cycle
GO:0051726~regulation of cell cycle
GO:0040007~growth
GO:0040008~regulation of growth
GO:0012501~programmed cell death
GO:0007155~cell adhesion

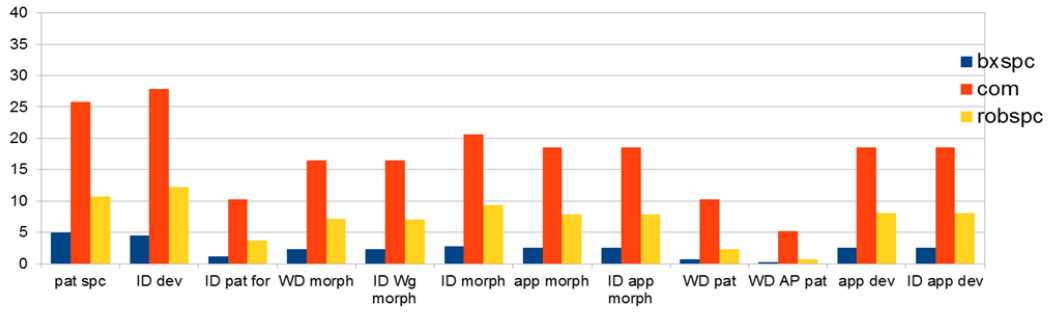


Figure 4.6 A comparative GO analysis of the targets of Ubx common to *Bombyx* and *Drosophila* (R) (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Drosophila*). GO categories relevant to wing and imaginal disc patterning and development are plotted. Each GO category is enriched in common targets compared to species-specific ones. The key for the abbreviations on the X axis is in the table below.

Abbrv.	**GO term and Biological process**
pat spc	GO:0007389~pattern specification process
ID dev	GO:0007444~imaginal disc development
ID pat for	GO:0007447~imaginal disc pattern formation
WD morph	GO:0007472~wing disc morphogenesis
ID Wg morph	GO:0007476~imaginal disc-derived wing morphogenesis
ID morph	GO:0007560~imaginal disc morphogenesis
app morph	GO:0035107~appendage morphogenesis
ID app morph	GO:0035114~imaginal disc-derived appendage morphogenesis
WD pat	GO:0035222~wing disc pattern formation
WD AP pat	GO:0048100~wing disc anterior/posterior pattern formation
app dev	GO:0048736~appendage development
ID app dev	GO:0048737~imaginal disc-derived appendage development

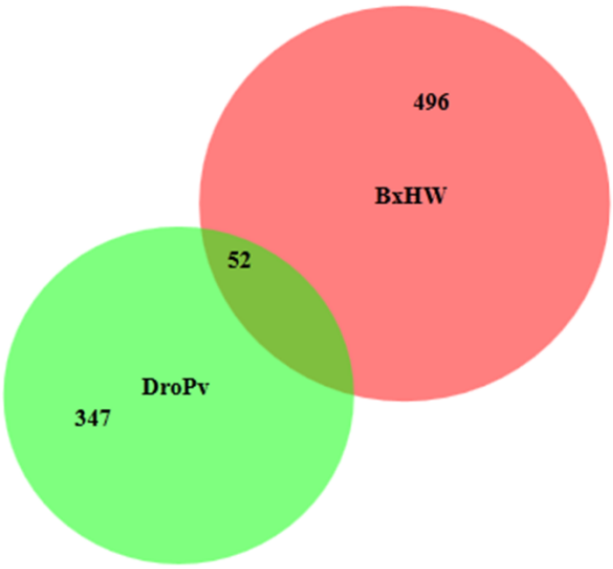


Figure 4.7 Comparison of targets of Ubx (only for which fly homologs exist are considered here) between *Bombyx* hindwing and *Drosophila* haltere *Drosophila* (P) (Agrawal et al, 2011). 9.48 % of the targets of *Bombyx* Ubx are common to those in *Drosophila*.

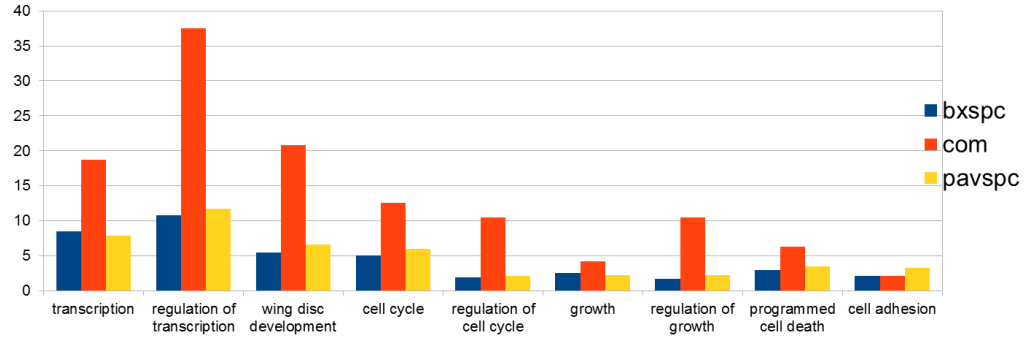


Figure 4.8 A comparative GO analysis of targets of Ubx common to *Bombyx* and *Drosophila* (P) (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Drosophila*). GO categories relevant to wing development in general are plotted. Each category is enriched in common targets compared to species-specific ones.

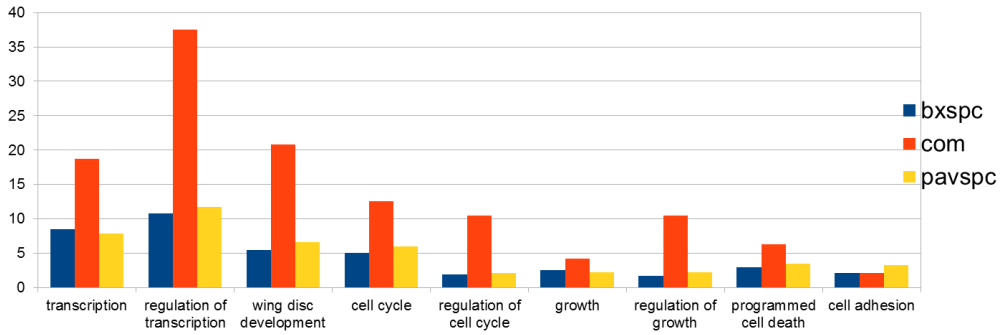


Figure 4.9 A comparative GO analysis of the targets of Ubx common to *Bombyx* and *Drosophila* (R) (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Drosophila*). GO categories relevant to wing and imaginal disc patterning and development are plotted. Each GO category is enriched in common targets compared to species-specific ones.

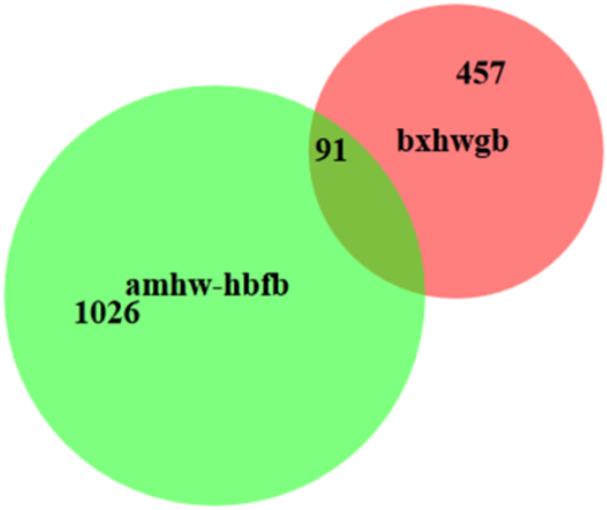


Figure 4.10 Comparison of targets of Ubx (only for those fly homologs exist are considered here) between *Bombyx* hindwing (bxhwgb) and *Apis* hindwing (amhw-hbfb) (Prasad N, 2013). 16.6 % of targets in *Bombyx* hindwing are common between the two studies.

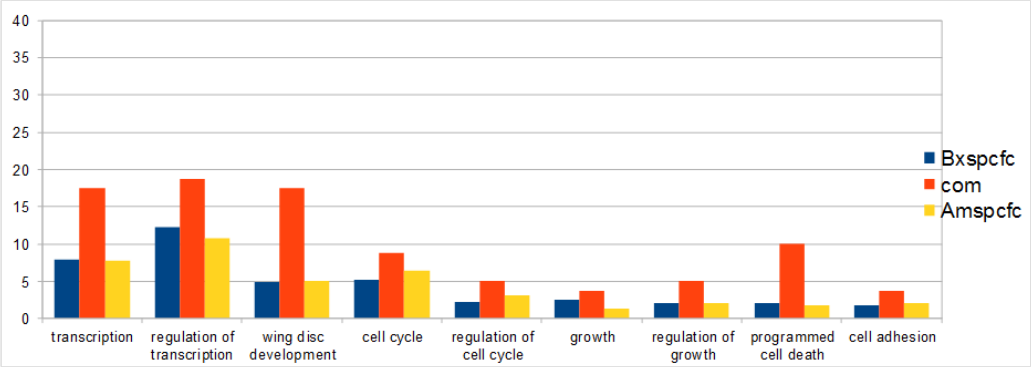


Figure 4.11 A comparative GO analysis of the targets of Ubx common to *Bombyx* and *Apis* (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Apis*). GO categories relevant to wing development in general are plotted. Each category is enriched in common targets compared to the species-specific ones. However, the enrichment is not as much as seen in targets that are common to *Bombyx* and *Drosophila*.

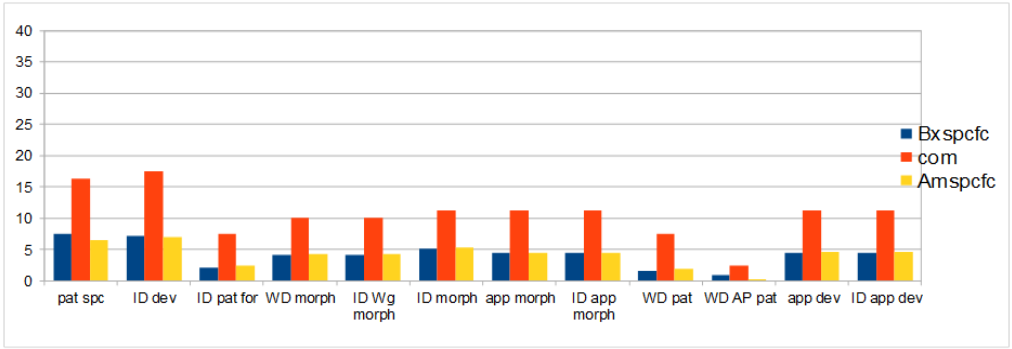


Figure 4.12 A comparative GO analysis of the targets of Ubx common to *Bombyx* and *Apis* (red) and the species-specific targets (blue: *Bombyx* and Yellow: *Apis*). GO categories relevant to wing and imaginal disc patterning and development are plotted. Each category is enriched in common targets compared to the species-specific ones.

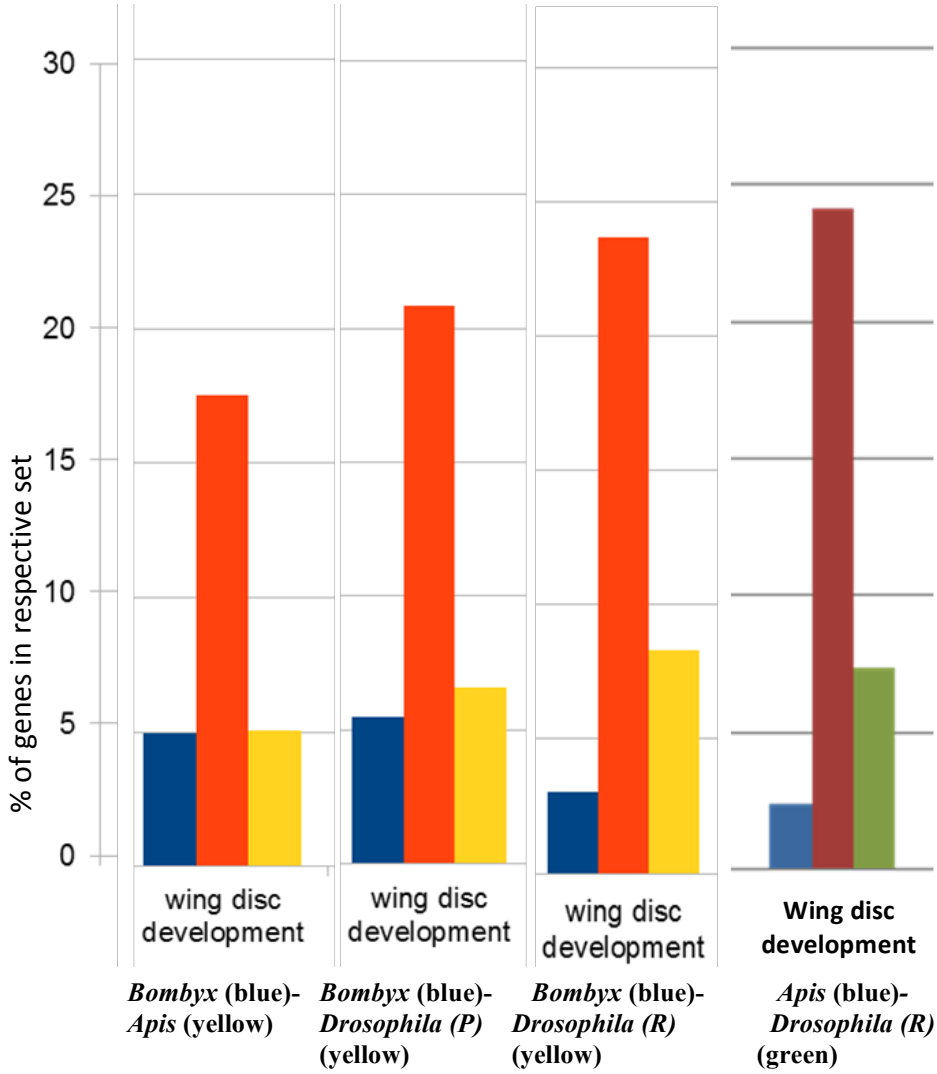


Figure 4.13 Pair-wise comparison of targets of Ubx. Only GO category related to wing development is considered here. In all pair-wise comparisons, targets that are common to two species show enrichment for wing related genes compared to species-specific targets. However, degree of enrichment is higher amongst the targets common to *Bombyx* and *Drosophila* than targets that are common to *Apis* and *Bombyx*. Interestingly, targets common to Apis and Drosophila (Prasad N, 2013) show similar levels of enrichment for wing-related genes.

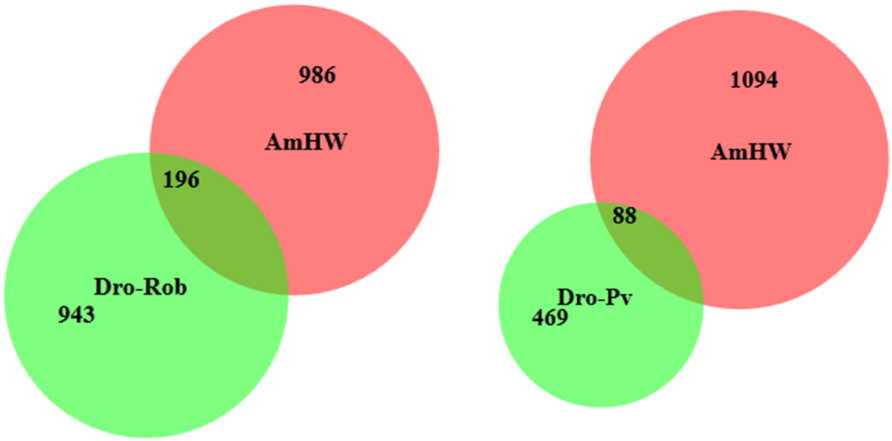


Figure 4.14 Comparison of targets of Ubx (only for which fly homologs exist are considered here) between *Apis* hindwing and *Drosophila* haltere (*Drosophila* (R) and *Drosophila* (P)) (Prasad N, 2013). 16.58 % of targets of Ubx in *Apis* are common to *Drosophila* (R) and 7.45 % to *Drosophila* (P).

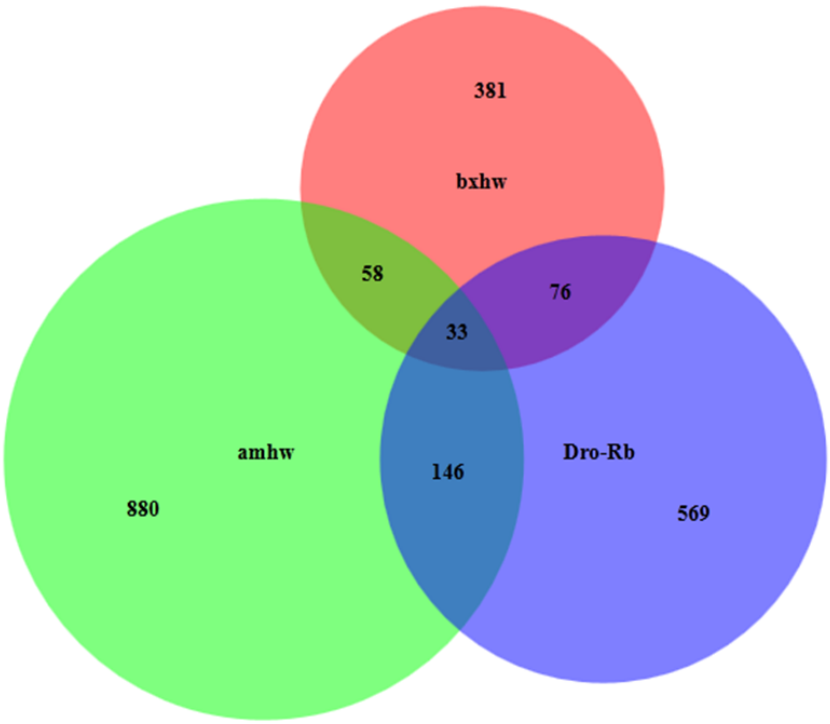


Figure 4.15 A three-way comparison of targets of Ubx (only for which fly homologs exist are considered here) in *Apis* and *Bombyx* hindwing and *Drosophila* haltere (*Drosophila* (R)). 33 genes are common to all the three insects and they are the genes known to be relevant to *Drosophila* wing development.

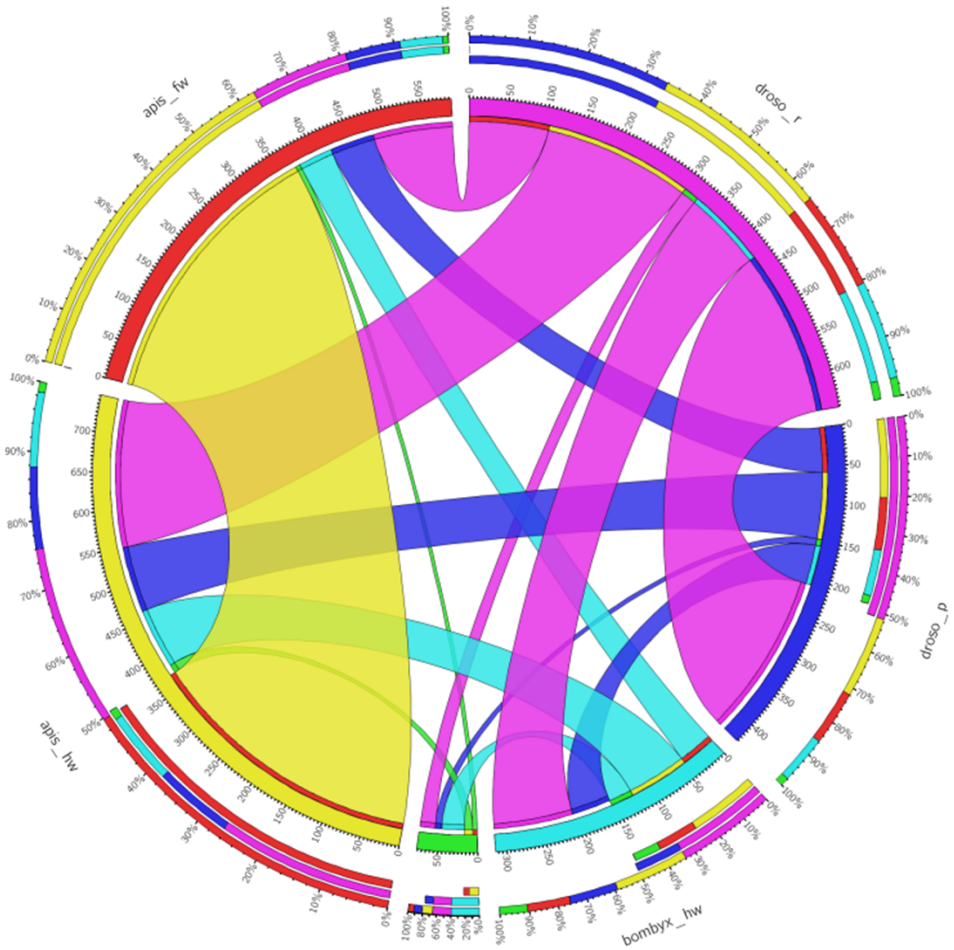


Figure 4.16 A Circos plot displaying the common targets of Ubx (only for which fly homologs exist are considered here) in forewing and hindwing of *Apis* and *Bombyx* and haltere in *Drosophila* (*Drosophila* (R) and *Drosophila* (P)). The large yellow ribbon common to forewing and hindwing in *Apis* is noteworthy. Hymenoptera, an ancestral form shows Ubx expression in both fore- and hindwings and naturally has many targets common to the two wings. In *Bombyx*, wherein there is no morphological difference between the fore- and the hindwing (as in *Apis*), Ubx is not expressed in the developing forewing (except in peripodial membrane), a situation very similar to *Drosophila*.

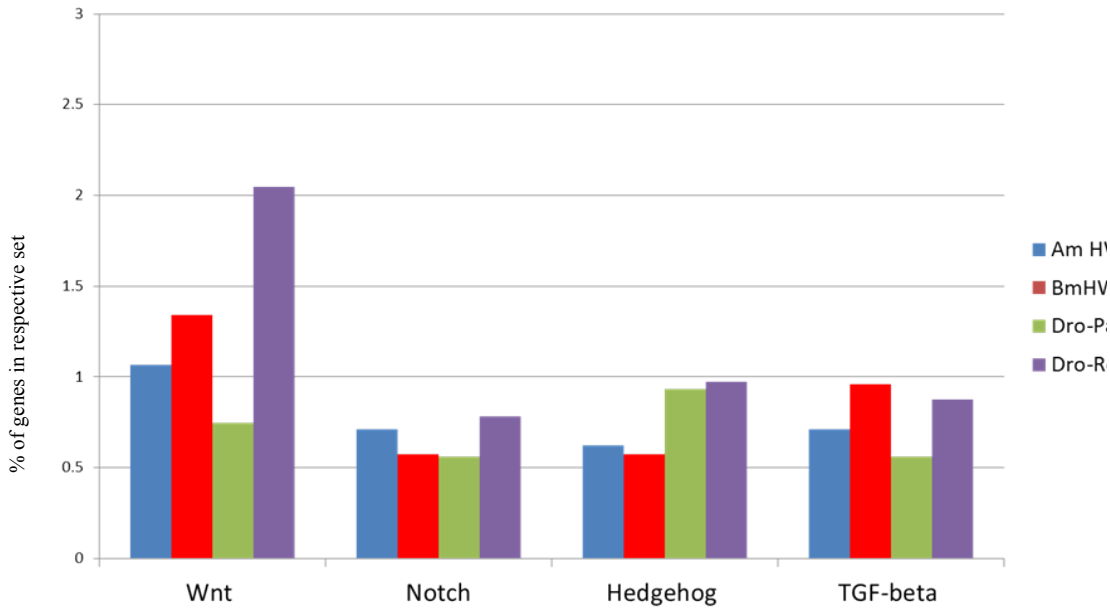


Figure 4.17 A comparative graph of GO analysis of targets of Ubx tin the four data sets. Only GO cateogies related to signaling pathways involved in wing development are considered here. The proportions of the signaling pathways (X axis) remain similar between insects. Dro-Pav = data set *Drosophila* (P). Dro-Rob = data set *Drosophila* (R).

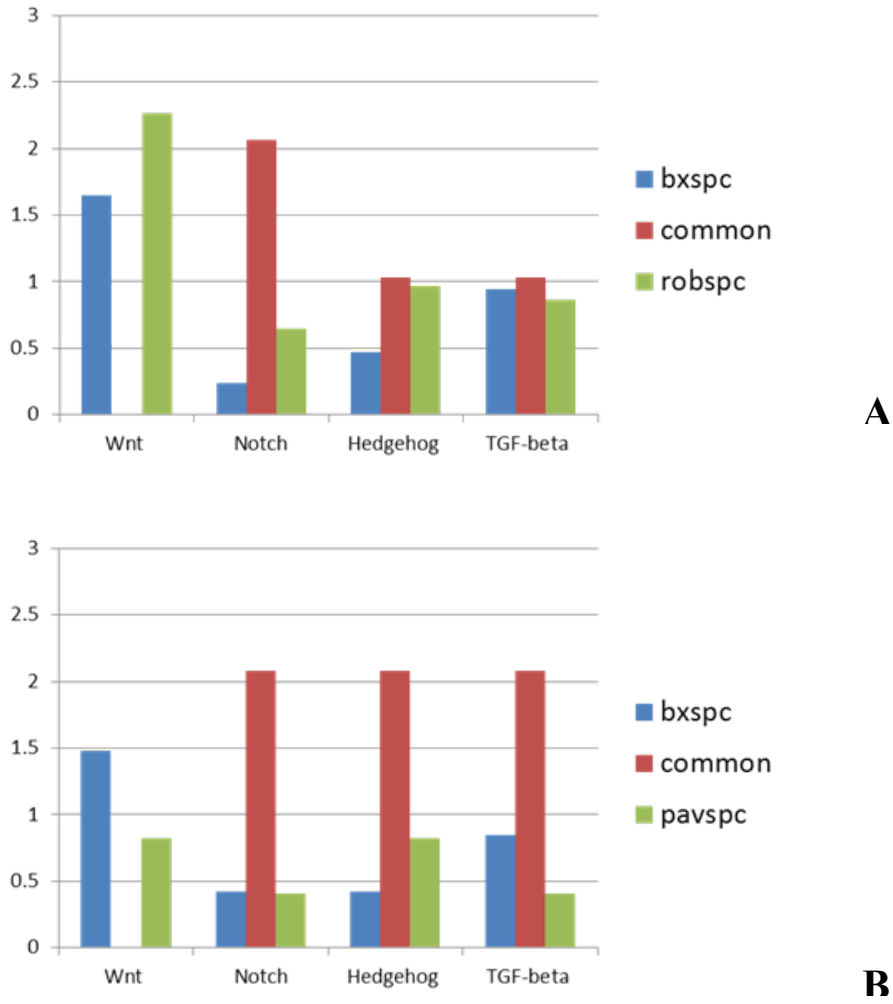


Figure 4.18 A comparative graph of GO analysis of the targets of Ubx common to *Bombyx* and *Drosophila* (red) and the species-specific targets (blue: *Bombyx* and Green: *Drosophila*). Only GO categories related to signaling pathways involved in wing development are considered here. Each category is enriched in common targets compared to the species-specific ones, except the Wnt pathway components. No representative of this pathway was seen amongst the targets of Ubx common to the two species. A: targets of Ubx in *Bombyx* vs Drosophila R. B: targets of Ubx in *Bombyx* vs Drosophila P.

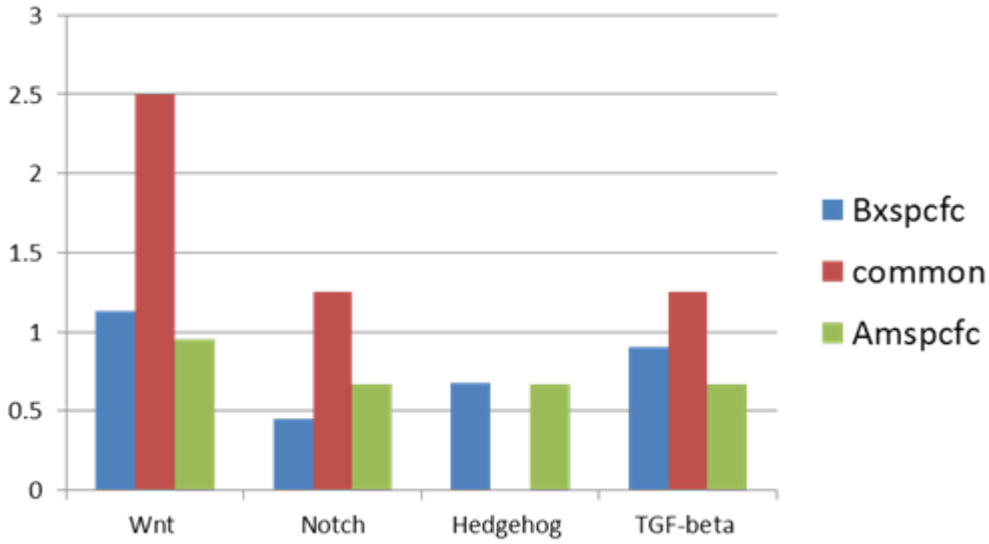


Figure 4.19 A comparative graph of GO analysis of the targets of Ubx common to *Bombyx* and *Apis* (red) and the species-specific targets (blue: *Bombyx* and Green: *Apis*). Only GO categories related to signaling pathways involved in wing development are considered here. Each category is enriched in common targets compared to the species-specific ones, except the Hedgehog pathway components. No representative of this pathway was seen amongst the targets of Ubx common to the two species.

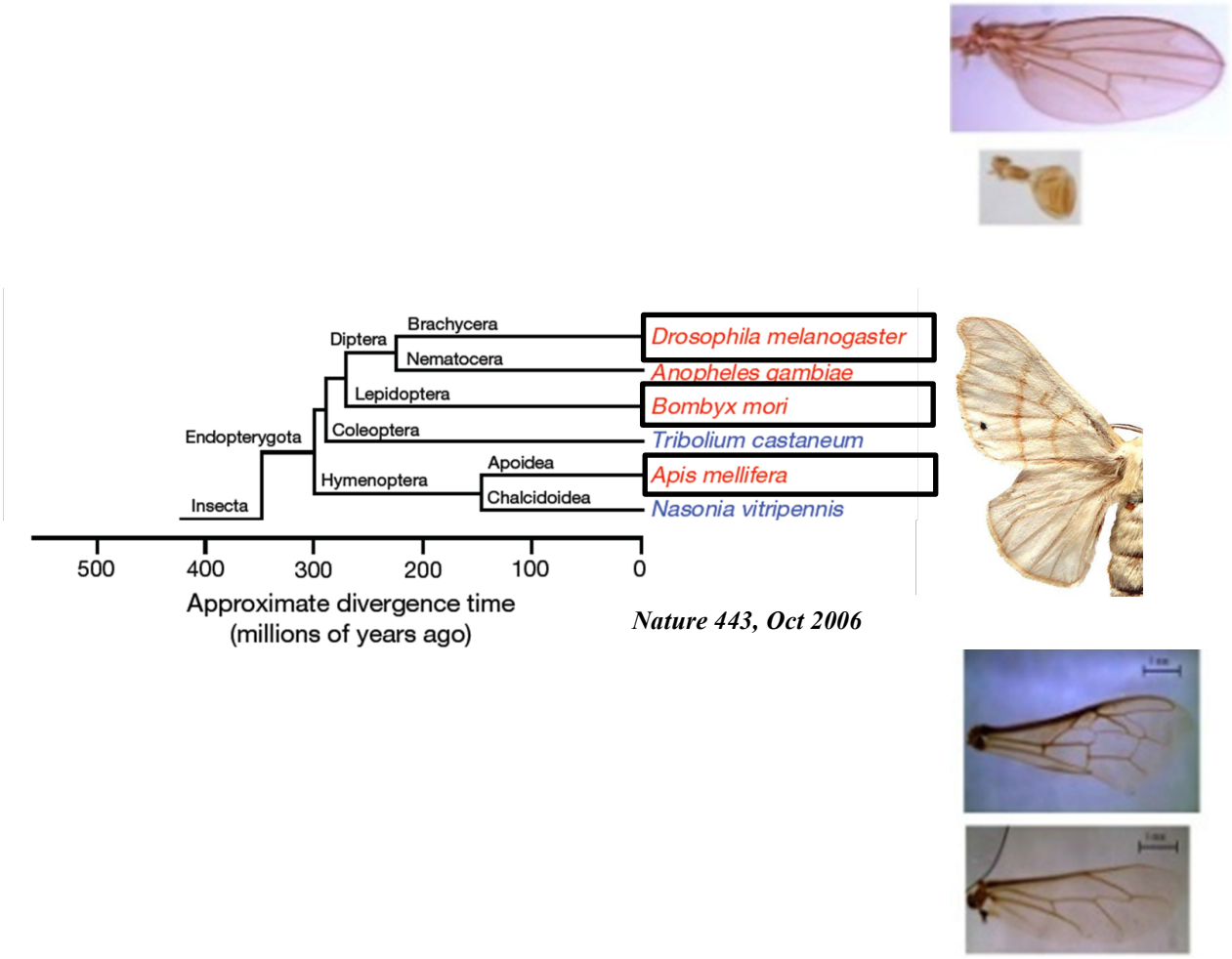


Figure 4.20 The phylogenetic tree representing the evolutionary relationships of insects and related arthropods for which either the whole genome sequences are available (red) or draft assembly of the genome sequences are available (blue) with approximate divergence times. Hymenoptera is basal to Lepidoptera, Coleoptera and Diptera.

Right-side panel: Wing and haltere are morphologically distinct in *Drosophila* (top panel), while the fore- and hindwings do not show major morphological differences in *Bombyx* (middle panel) and *Apis* (bottom panel). Images: Prasad N, 2013

Table 4.1 List of the targets of Ubx (only those which have known fly homologues) common between *Bombyx* hindwing set and *Drosophila* (R) set.

Table titles: FBID- Fly Base Identification number CGnum- Celera Genomics ID number from Flybase. SYMBOL and NAME from Flybase gene ID.

SL	FBID KEY	CG num	SYMBOL	NAME
1	FBgn0030520	CG10990	Pdcd4	Programmed cell death 4 ortholog
2	FBgn0053196	CG33196	dp	dumpy
3	FBgn0032120	CG33298	CG33298	0
4	FBgn0264442	CG43860	ab	abrupt
5	FBgn0033913	CG8468	CG8468	0
6	FBgn0028622	CG13432	qsm	quasimodo
7	FBgn0000395	CG15671	cv-2	crossveinless 2
8	FBgn0026160	CG7958	tna	tonalli
9	FBgn0041094	CG7590	scyl	scylla
10	FBgn0002733	CG14548	E(spl)mbeta-HLH	Enhancer of split mbeta, HLH
11	FBgn0029881	CG3973	pigs	pickled eggs
12	FBgn0011837	CG4070	Tis11	Tis11 homolog
13	FBgn0000097	CG3166	aop	anterior open
14	FBgn0004893	CG10021	bowl	brother of odd with entrails limited
15	FBgn0020443	CG6382	Elf	Ef1alpha-like factor
16	FBgn0010548	CG11140	Aldh-III	Aldehyde dehydrogenase type III
17	FBgn0004907	CG17870	14-3-3zeta	14-3-3zeta
18	FBgn0027499	CG12340	wde	windei
19	FBgn0000575	CG1007	emc	extra macrochaetae
20	FBgn0036030	CG6767	CG6767	0
21	FBgn0010113	CG15532	hdc	headcase
22	FBgn0039907	CG2041	lgs	legless
23	FBgn0262735	CG1691	Imp	IGF-II mRNA-binding protein
24	FBgn0014133	CG1822	bif	bifocal
25	FBgn0003975	CG3830	vg	vestigial
26	FBgn0002735	CG8333	E(spl)mgamma-HLH	Enhancer of split mgamma, HLH
27	FBgn0002631	CG6096	E(spl)m5-HLH	Enhancer of split m5, HLH
28	FBgn0262656	CG10798	dm	diminutive
29	FBgn0002973	CG3779	numb	numb
30	FBgn0000546	CG1765	EcR	Ecdysone receptor
31	FBgn0010313	CG2530	corto	corto

32	FBgn0004644	CG4637	hh	hedgehog
33	FBgn0004646	CG3039	ogre	optic ganglion reduced
34	FBgn0024250	CG9653	brk	brinker
35	FBgn0010575	CG5580	sbb	scribbler
36	FBgn0000179	CG3578	bi	bifid
37	FBgn0261618	CG42551	larp	La related protein
38	FBgn0016977	CG18497	spen	split ends
39	FBgn0031474	CG2991	CG2991	0
40	FBgn0025681	CG3558	CG3558	0
41	FBgn0016076	CG14029	vri	vrille
42	FBgn0000308	CG9553	chic	chickadee
43	FBgn0000320	CG9554	eya	eyes absent
44	FBgn0005771	CG4491	noc	no ocelli
45	FBgn0001983	CG4158	wor	worniu
46	FBgn0010300	CG10719	brat	brain tumor
47	FBgn0029092	CG11804	ced-6	ced-6
48	FBgn0001291	CG2275	Jra	Jun-related antigen
49	FBgn0262114	CG42236	RanBPM	Ran-binding protein M
50	FBgn0003396	CG7734	shn	schnurri
51	FBgn0000577	CG9015	en	engrailed
52	FBgn0033636	CG10897	tou	toutatis
53	FBgn0022764	CG8815	Sin3A	Sin3A
54	FBgn0002643	CG8118	mam	mastermind
55	FBgn0041585	CG11430	olf186-F	olf186-F
56	FBgn0034500	CG11200	CG11200	Carbonyl reductase
57	FBgn0027529	CG8920	CG8920	0
58	FBgn0020257	CG9952	ppa	partner of paired
59	FBgn0034797	CG12781	nahoda	nahoda
60	FBgn0003977	CG3496	vir	virilizer
61	FBgn0021895	CG18426	ytr	yantar
62	FBgn0010435	CG2727	emp	epithelial membrane protein
63	FBgn0020386	CG1210	Pdk1	Phosphoinositide-dependent kinase 1
64	FBgn0035101	CG1212	p130CAS	p130CAS
65	FBgn0262624	CG12026	Tmhs	Tetraspan membrane protein
66	FBgn0035445	CG12014	CG12014	0
67	FBgn0262719	CG43163	CG43163	0
68	FBgn0035953	CG5087	CG5087	0
69	FBgn0036154	CG6168	CG6168	0
70	FBgn0261381	CG42631	mtTFB1	Mitochondrial Transcription Factor B1
71	FBgn0036279	CG4357	Ncc69	sodium chloride cotransporter 69

72	FBgn0029114	CG6890	Tollo	Tollo
73	FBgn0260635	CG12284	th	thread
74	FBgn0261547	CG42665	Exn	Ephexin
75	FBgn0036732	CG7571	Oatp74D	Organic anion transporting polypeptide 74D
76	FBgn0000568	CG8127	Eip75B	Ecdysone-induced protein 75B
77	FBgn0036886	CG9300	CG9300	0
78	FBgn0014037	CG32217	Su(Tpl)	Su(Tpl)
79	FBgn0003415	CG9936	skd	skuld
80	FBgn0037120	CG11247	CG11247	0
81	FBgn0259212	CG42312	cno	canoe
82	FBgn0037305	CG12173	CG12173	0
83	FBgn0005585	CG9429	Crc	Calreticulin
84	FBgn0004595	CG17228	pros	prospero
85	FBgn0038129	CG8449	CG8449	0
86	FBgn0263396	CG16901	sqd	squid
87	FBgn0262127	CG33967	kibra	kibra ortholog
88	FBgn0002781	CG32491	mod(mdg4)	modifier of mdg4
89	FBgn0003867	CG6705	tsl	torso-like
90	FBgn0039213	CG6668	atl	atlastin
91	FBgn0039286	CG11849	dan	distal antenna
92	FBgn0002609	CG8346	E(spl)m3-HLH	Enhancer of split m3, HLH
93	FBgn0039709	CG31009	Cad99C	Cadherin 99C
94	FBgn0015221	CG1469	Fer2LCH	Ferritin 2 light chain homologue
95	FBgn0261444	CG3638	CG3638	0
96	FBgn0003079	CG2845	phl	pole hole
97	FBgn0040066	CG17437	wds	will die slowly
98	FBgn0023215	CG13316	Mnt	Mnt
99	FBgn0086899	CG34412	tlk	Tousled-like kinase
100	FBgn0000542	CG2904	ec	echinus
101	FBgn0046687	CG3171	Tre1	Trapped in endoderm 1
102	FBgn0261383	CG3125	IntS6	Integrator 6
103	FBgn0000042	CG4027	Act5C	Actin 5C
104	FBgn0029897	CG3203	RpL17	Ribosomal protein L17
105	FBgn0003447	CG32858	sn	singed
106	FBgn0030065	CG12075	CG12075	0
107	FBgn0031950	CG14536	Herp	Homocysteine-induced ER protein
108	FBgn0041210	CG1770	HDAC4	HDAC4
109	FBgn0030884	CG6847	CG6847	0

Table 4.2 List of the targets of Ubx (only those which have known fly homologues) common between *Bombyx* hindwing set and *Drosophila* (P) set.

Table titles: FBID- Fly Base Identification number CGnum- Celera Genomics ID number from Flybase. SYMBOL and NAME from Flybase gene ID.

SL	FBID KEY	CGnum	SYMBOL	NAME
1	FBgn0005771	CG4491	noc	no ocelli
2	FBgn0035445	CG12014	CG12014	0
3	FBgn0000308	CG9553	chic	chickadee
4	FBgn0000448	CG33183	Hr46	Hormone receptor-like in 46
5	FBgn0000568	CG8127	Eip75B	Ecdysone-induced protein 75B
6	FBgn0000575	CG1007	emc	extra macrochaetae
7	FBgn0001230	CG5436	Hsp68	Heat shock protein 68
8	FBgn0001942	CG9075	eIF-4a	Eukaryotic initiation factor 4a
9	FBgn0002643	CG8118	mam	mastermind
10	FBgn0002733	CG14548	E(spl)mbeta-HLH	Enhancer of split mbeta, HLH
11	FBgn0002735	CG8333	E(spl)mgamma-HLH	Enhancer of split mgamma, HLH
12	FBgn0003396	CG7734	shn	schnurri
13	FBgn0003415	CG9936	skd	skuld
14	FBgn0004644	CG4637	hh	hedgehog
15	FBgn0005612	CG3090	Sox14	Sox box protein 14
16	FBgn0010113	CG15532	hdc	headcase
17	FBgn0010300	CG10719	brat	brain tumor
18	FBgn0010313	CG2530	corto	corto
19	FBgn0010548	CG11140	Aldh-III	Aldehyde dehydrogenase type III
20	FBgn0010575	CG5580	sbb	scribbler
21	FBgn0010774	CG1101	Ref1	RNA and export factor binding protein 1
22	FBgn0011837	CG4070	Tis11	Tis11 homolog
23	FBgn0014184	CG16747	Oda	Ornithine decarboxylase antizyme
24	FBgn0020278	CG5248	loco	locomotion defects
25	FBgn0022764	CG8815	Sin3A	Sin3A
26	FBgn0023215	CG13316	Mnt	Mnt
27	FBgn0024250	CG9653	brk	brinker
28	FBgn0026160	CG7958	tna	tonalli
29	FBgn0026533	CG5935	Dek	Dek
30	FBgn0027529	CG8920	CG8920	0

31	FBgn0030719	CG9177	eIF5	eIF5
32	FBgn0033244	CG8726	CG8726	0
33	FBgn0034261	CG4966	HPS4	Hermansky-Pudlak Syndrome 4
34	FBgn0034500	CG11200	CG11200	Carbonyl reductase
35	FBgn0034743	CG4046	RpS16	Ribosomal protein S16
36	FBgn0034878	CG3941	pita	pita
37	FBgn0036154	CG6168	CG6168	0
38	FBgn0036663	CG9674	CG9674	0
39	FBgn0038872	CG5874	Nelf-A	Negative elongation factor A
40	FBgn0040066	CG17437	wds	will die slowly
41	FBgn0041094	CG7590	scyl	scylla
42	FBgn0053196	CG33196	dp	dumpy
43	FBgn0265991	CG30084	Zasp52	Z band alt spliced PDZ-motif protein 52
44	FBgn0083951	CG34115	CG34115	0
45	FBgn0086687	CG5887	desat1	desat1
46	FBgn0086899	CG34412	tlk	Tousled-like kinase
47	FBgn0260634	CG10192	eIF4G2	eukaryotic Transl ini fac 4G2
48	FBgn0260635	CG12284	th	thread
49	FBgn0261618	CG42551	larp	La related protein
50	FBgn0262114	CG42236	RanBPM	Ran-binding protein M
51	FBgn0262127	CG33967	kibra	kibra ortholog
52	FBgn0262656	CG10798	dm	diminutive

Table 4.3 List of the targets of Ubx (only those which have known fly homologues) common between *Bombyx* hindwing set and *Apis* hindwing set.

Table titles: FBID- Fly Base Identification number CGnum- Celera Genomics ID number from Flybase. SYMBOL and NAME from Flybase gene ID.

SL	FBID KEY	CGnum	SYMBOL	NAME
1	FBgn0000179	CG3578	bi	bifid
2	FBgn0000307	CG5813	chif	chiffon
3	FBgn0000319	CG9012	Chc	Clathrin heavy chain
4	FBgn0000448	CG33183	Hr46	Hormone receptor-like in 46
5	FBgn0000542	CG2904	ec	echinus
6	FBgn0000546	CG1765	EcR	Ecdysone receptor
7	FBgn0000565	CG7266	Eip71CD	Ecdysone-induced protein 28/29kD
8	FBgn0000568	CG8127	Eip75B	Ecdysone-induced protein 75B
9	FBgn0000577	CG9015	en	engrailed
10	FBgn0001197	CG5499	His2Av	Histone H2A variant
11	FBgn0002638	CG10480	Rcc1	Regulator of chromosome condensation1
12	FBgn0002643	CG8118	mam	mastermind
13	FBgn0003079	CG2845	phl	pole hole
14	FBgn0003231	CG10360	ref(2)P	refractory to sigma P
15	FBgn0003415	CG9936	skd	skuld
16	FBgn0003975	CG3830	vg	vestigial
17	FBgn0004855	CG3284	RpII15	RNA polymerase II 15kD subunit
18	FBgn0004893	CG10021	bowl	brother of odd with entrails limited
19	FBgn0004907	CG17870	14-3-3zeta	14-3-3zeta
20	FBgn0005612	CG3090	Sox14	Sox box protein 14
21	FBgn0005696	CG5923	DNApol-alpha73	DNA polymerase alpha 73kD
22	FBgn0005771	CG4491	noc	no ocelli
23	FBgn0010315	CG9096	CycD	Cyclin D
24	FBgn0010391	CG2522	Gtp-bp	GTP-binding protein
25	FBgn0011211	CG3612	blw	bellwether
26	FBgn0013531	CG18780	MED20	Mediator complex subunit 20
27	FBgn0013764	CG3924	Chi	Chip
28	FBgn0014037	CG32217	Su(Tpl)	Su(Tpl)
29	FBgn0016977	CG18497	spen	split ends
30	FBgn0020238	CG31196	14-3-3epsilon	14-3-3epsilon
31	FBgn0020257	CG9952	ppa	partner of paired

32	FBgn0023213	CG10811	eIF4G	eukaryotic translation initiation factor 4G
33	FBgn0025582	CG9677	Int6	Int6 homologue
34	FBgn0025634	CG13367	CG13367	0
35	FBgn0026079	CG6133	Nsun2	NOP2-Sun domain fam, member 2 orth
36	FBgn0026189	CG7740	prominin-like	prominin-like
37	FBgn0026418	CG6603	Hsc70Cb	Hsc70Cb
38	FBgn0027291	CG12233	l(1)G0156	lethal (1) G0156
39	FBgn0027609	CG15437	morgue	modifier of rpr and grim,
40	FBgn0027616	CG12076	YT521-B	YT521-B
41	FBgn0027654	CG2239	jdp	jdp
42	FBgn0028467	CG11070	CG11070	0
43	FBgn0028648	CG8612	mRpL50	mitochondrial ribosomal protein L50
44	FBgn0029114	CG6890	Tollo	Tollo
45	FBgn0029736	CG4041	CG4041	0
46	FBgn0030065	CG12075	CG12075	0
47	FBgn0030608	CG9057	Lsd-2	Lipid storage droplet-2
48	FBgn0030719	CG9177	eIF5	eIF5
49	FBgn0031256	CG4164	CG4164	0
50	FBgn0031985	CG8683	mon2	0
51	FBgn0032120	CG33298	CG33298	0
52	FBgn0033482	CG1371	CG1371	0
53	FBgn0034084	CG8435	CG8435	0
54	FBgn0034583	CG10527	CG10527	0
55	FBgn0035101	CG1212	p130CAS	p130CAS
56	FBgn0036165	CG7533	chrb	charybde
57	FBgn0036913	CG8334	CG8334	0
58	FBgn0037120	CG11247	CG11247	0
59	FBgn0037138	CG7145	P5CDh1	delta-1-P 5 C dehydrogenase
60	FBgn0037255	CG1078	Fip1	0
61	FBgn0037305	CG12173	CG12173	0
62	FBgn0037703	CG8165	JHDM2	JmjC domain- histoned demethylase
63	FBgn0038163	CG10841	CG10841	0
64	FBgn0038834	CG15697	RpS30	Ribosomal protein S30
65	FBgn0039205	CG13623	CG13623	0
66	FBgn0039329	CG10669	CG10669	0
67	FBgn0039830	CG1746	CG1746	0
68	FBgn0039907	CG2041	lgs	legless
69	FBgn0040068	CG7893	Vav	Vav ortholog (H. sapiens)
70	FBgn0041094	CG7590	scyl	scylla
71	FBgn0050372	CG30372	Asap1	ArfGAP with SH3,ankyrin r and PH

				dom
72	FBgn0052672	CG32672	Atg8a	Autophagy-specific gene 8a
73	FBgn0086899	CG34412	tlk	Tousled-like kinase
74	FBgn0260439	CG17291	Pp2A-29B	Protein phosphatase 2A at 29B
75	FBgn0260634	CG10192	eIF4G2	eukaryotic trans inn fac 4G2
76	FBgn0260635	CG12284	th	thread
77	FBgn0261383	CG3125	IntS6	Integrator 6
78	FBgn0262127	CG33967	kibra	kibra ortholog
79	FBgn0262656	CG10798	dm	diminutive
80	FBgn0262719	CG43163	CG43163	0
81	FBgn0262735	CG1691	Imp	IGF-II mRNA-binding protein
82	FBgn0266720	CG9474	Snap24	Synaptosomal-associated protein 24kDa
83	FBgn0266186	CG1599	Vamp7	Vesicle-associated membrane protein 7
84	FBgn0266436	CG45066	CG45066	

Table 4.4 List of the targets of Ubx (only those which have known fly homologues) common between *Bombyx* hindwing set, Apis hindwing and *Drosophila* (R) set.

Table titles: FBID- Fly Base Identification number CGnum- Celera Genomics ID number from Flybase. SYMBOL and NAME from Flybase gene ID.

SL	FBID KEY	CGnum	SYMBOL	NAME
1	FBgn0000179	CG3578	bi	bifid
2	FBgn0000542	CG2904	ec	echinus
3	FBgn0000546	CG1765	EcR	Ecdysone receptor
4	FBgn0000568	CG8127	Eip75B	Ecdysone-induced protein 75B
5	FBgn0000577	CG9015	en	engrailed
6	FBgn0002643	CG8118	mam	mastermind
7	FBgn0003079	CG2845	phl	pole hole
8	FBgn0003415	CG9936	skd	skuld
9	FBgn0003975	CG3830	vg	vestigial
10	FBgn0004644	CG4637	hh	hedgehog
11	FBgn0004893	CG10021	bowl	brother of odd with entrails limited
12	FBgn0004907	CG17870	14-3-3zeta	14-3-3zeta
13	FBgn0005771	CG4491	noc	no ocelli
14	FBgn0010575	CG5580	sbb	scribbler
15	FBgn0014037	CG32217	Su(Tpl)	Su(Tpl)
16	FBgn0016977	CG18497	spen	split ends
17	FBgn0020257	CG9952	ppa	partner of paired
18	FBgn0029114	CG6890	Tollo	Tollo
19	FBgn0030065	CG12075	CG12075	-
20	FBgn0032120	CG33298	CG33298	-
21	FBgn0035101	CG1212	p130CAS	p130CAS
22	FBgn0037120	CG11247	CG11247	-
23	FBgn0037305	CG12173	CG12173	-
24	FBgn0039907	CG2041	lgs	legless
25	FBgn0041094	CG7590	scyl	scylla
26	FBgn0086899	CG34412	tlk	Tousled-like kinase
27	FBgn0260635	CG12284	th	thread
28	FBgn0261383	CG3125	IntS6	Integrator 6
29	FBgn0262127	CG33967	kibra	kibra ortholog
30	FBgn0262656	CG10798	dm	diminutive

31	FBgn0262719	CG43163	CG43163	-
32	FBgn0262735	CG1691	Imp	IGF-II mRNA-binding protein
33	FBgn0263396	CG16901	sqd	squid

Chapter 5

Trascriptome analysis of *Bombyx* wing buds

Introduction

Transcriptome is the complete set of transcripts present in a cell; it also includes the knowledge of their quantitative levels at a specific developmental stage or for a given condition (Wang et al, 2009). The trasncriptome analysis is instrumental in understanding how genomes encode the diverse patterns of gene expression to define cell proliferation and differentiation during development (Pepke et al, 2009). Before the advent of high throughput sequencing technologies, large-scale gene expression studies were performed using hybridization arrays. The next generation sequencing-based whole genome transcriptomics as compared to hybridization methods has the advantages of high throughput, greater coverage, high resolution and low background noise (Li et al 2012). High throughput sequencing methods like RNA-Sequencing (RNA-Seq) allows mapping, annotation and quantification of the total RNA population.

5.1 RNA-Sequencing

In RNA-Seq, a population of RNA (total or fractionated, such as poly (A) + for mRNA) is converted into a library of cDNA fragments with adaptors attached to one (single end) or both ends (paired end). Each molecule is then sequenced in a high throughput genome analyzer like Illumina® Genome analyzer to obtain short sequences from the ends (Fig 5.1). The reads are typically 36-400 bp in length depending on the technology used (Wang et al, 2010). The high throughput sequencing based on Illumina/Solexa sequencing is already described in Chapter 2 (section 2.3) and it is principally the same for RNA-Sequencing.

RNA-Seq can reveal the precise location of transcription boundaries, give information on intron-exon boundaries, alternative splice forms, isoforms and also abundance of the mRNA in a given tissue and state with very little noise or background. Thus, RNA-Seq allows sequencing of the entire transcriptome in a high throughput and quantitative manner, at a single base resolution, with gene expression estimate at the genome wide scale at relatively low cost than arrays.

5.2 RNA-Seq Analysis

Sequence reads obtained from a high throughput genome analyzer are processed through a series of programs to analyze the data and extract meaningful biological interpretation from it. Like any other high throughput sequencing method, RNA-Seq also involves handling large-scale data and requires programs to analyze and mine meaningful biological information. The Transcriptome analysis of RNA-Seq data can be divided into three categories, (i) mapping and read alignment, (ii) transcript assembly and genome annotation and (iii) RNA quantification for expression level estimation.

Many programs are employed for RNA-Seq analysis; one of the most popular program suites for reference genome assisted transcriptome analysis is Tuxedo suite (includes the Bowtie, TopHat and Cufflinks programs). It is an open source program tool suite for gene discovery and comprehensive analysis of RNA-Seq data. Please refer to Fig 5.2A for a schematic overview of the programs and their functions for RNA-Seq analysis using Tuxedo suite. A flow chart describing the protocol workflow for the analysis of RNA-Seq data using these tools is shown in Fig. 5.2B. The first step in an RNA-Seq analysis is to map the sequence reads to a reference genome.

5.2.1 Mapping sequence reads using TopHat

TopHat is a software package that identifies transcription splice sites *ab initio* by large scale mapping of RNA-Seq reads. TopHat first maps non-junction reads (contained within exons) using Bowtie (described in Chapter3, section 3.1) as an alignment engine. TopHat primarily aligns using Bowtie and then breaks up reads that Bowtie cannot align on its own into smaller pieces called segments. Often these segments when processed independently will align to the genome. When several of a read's segments align to a genome far apart from one another, TopHat infers that the read spans a splice junction and estimates where the junction's splice sites are. TopHat maps the reads of each sample to the reference genome and attaches metadata to each alignment so that

downstream programs like Cufflinks and Cuffdiff can be more accurate in assembly and quantification (Trapnell et al, 2012).

5.2.2 Transcript assembly and discovery by Cufflinks

Accurate quantification of the expression level of a gene from reads requires accurate identification of which isoform of a given gene produced each read. Cufflinks assembles individual transcripts from RNA-Seq reads that are aligned. As many splice variants may be present for a certain gene in the data, Cufflinks reports a parsimonious transcriptome assembly of data. The algorithm reports it as few full-length transcript fragments (transfrags), that are needed to justify all the splicing outcomes in the data. This constitutes the assembly phase. After assembly, Cufflinks quantifies the expression level of each transfrag in the sample by using a rigorous statistical model filtering artifacts.

Merging assemblies: When working with different samples in a transcriptome analysis, it is necessary to pool the data and assemble it into a comprehensive set along with the reference genome as a uniform basis for downstream calculations in differential analysis. A program called Cuffmerge performs a reference annotation-based transcript (RABT) assembly to merge reference genome and individual sample transfrags to produce a single annotation for downstream differential analysis. The final assembly can be screened for genes and transcripts that are differentially expressed between samples using the program Cuffdiff.

Discovering new genes and transcripts is another interest in the transcriptome analysis. Cuffcompare is a program that can compare Cufflinks assemblies to reference annotation and sort out new genes from known annotated ones (Trapnell et al, 2012).

5.2.3 Differential analysis with Cuffdiff

Cuffdiff is a program that calculates the expression of two or more samples and tests the statistical significance of each observed change in expression. The

model assumes that the number of reads produce by each transcript is proportional to its abundance while accounting for the biological variability.

Cuffdiff takes an assembled (GTF) file of transcripts as input, along with the two alignment (BAM) files containing the fragment alignments for the two samples/conditions to be compared. It produces a number of output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene. Tracking the former allows one to see changes in splicing, and the latter lets one allows to find outchanges in relative promoter use within a gene.

Cuffdiff reports additional differential analysis with statistics such as fold change and p-values, beyond simple change in gene expression. It uses q-value, a corrected p-value to account for multiple testing (i.e. you are testing thousands of genes) and determine the significance of the gene expression difference. Those with q-value <0.05 are considered "significant" when biological replicates exist. It can identify genes that are differentially spliced or regulated via promoter switching. Cuffdiff also calculates the total expression level of a transcriptional start site (TSS) group by adding up the expression levels of isoforms within the group (Trapnell et al, 2012).

5.2.4 Visualization with CummeRbund

Cuffdiff generates data in the form of files that are tabular in nature with names and values of the expression levels. This kind of data is better visualized by plotting graphs. CummeRbund is an R based program for plotting graphs with various visualization tools to present the Cuffdiff data. It drastically simplifies data exploration tasks, such as plotting and cluster analysis of expression data (Trapnell et al, 2012).

5.3 Gene expression studies in haltere of *Drosophila*

This experiment to sequence the transcriptome of *Bombyx* fore- and hindwing buds was with the intention of understanding the expression levels of various genes in the two in conjunction with the ChIP data identifying taregts of Ubx in

the hindwing. As we strive to understand the evolutionary divergence of insect wing appendages, we also wanted to get an idea of the differentially expressed targets in *Bombyx* wing buds in comparison with the same in wing and haltere of *Drosophila*. Published work is available for the differentially expressed targets of Ubx in wing and haltere in *Drosophila* from two earlier works. The first one from our lab by Mohit Prasad et al. in 2006 using microarray followed by other validation experiments to identify potential targets of Ubx during haltere specification. The second source of targets of Ubx in haltere comes from the work of Pavlopoulos and Akam, 2011, where they used extensive microarray (Affymetrix®) profiling and quantitative RT-PCR to identify primary transcriptional responses to Ubx at different stages of *Drosophila* development. Results of these two studies were used to do a comparative analysis with the RNA-Seq data of *Bombyx* wing buds.

Materials and Methods

5.4 Isolation of wing buds for RNA preparation

Wing buds were isolated from fourth instar *Bombyx* larvae (Daizo race), the same stage that was used for ChIP experiment by a method as described in section 2.8. Eighty each of fore- and hindwing buds were isolated and the freshly isolated buds were collected into a microcentrifuge tube kept in liquid nitrogen all the time during isolation of wing buds. The tubes were then immediately transferred to -80ºC before isolation of RNA. The total RNA was isolated by Trizol extraction and checked for quality on a Bio-analyzer®. The libraries were prepared and paired end sequencing was done on an Illumina® platform. Both hindwing and forewing buds were sequenced from both ends with sequence read lengths of 100 bp.

5.5 Analysis of the RNA-Seq reads

5.5.1 FastQC quality control

The program described in chapter 3 (section 3.5) was used to assess the quality of the paired end 100 bp sequences obtained after the RNA-Seq on a Illumina® Genome analyzer. The read files were analyzed using FastQC. The primary focus was to see if the general quality profile of reads is up to the mark and if any trimming is required for improving the alignment to genome.

5.5.2 Trimming of read-ends

Reads in all the four files were trimmed by 10 bases each at start and end of the reads to improve the alignment to genome. The program Seqtk, which is a fast and lightweight tool for processing sequences in the FASTA or FASTQ format, was used for the trimming.

```
seqtk trimfq -b 10 -e 10 input.fastq > output.fastq
```

Where ‘seqtk’ is the program, command ‘trimfq’ trims fastq files, ‘-b 10’ trims 10 bases at the beginning, ‘-e 10’ trims 10 bases at the end, ‘input.fastq’ is the input read file and ‘output.fastq’ is the output fastq file generated.

5.5.3 Mapping reads to the genome using TopHat

The aligner from the Cufflinks Tuxedo tools suite, TopHat 2.0.8b, was used as the principle mapper to map RNA-Seq reads after quality control and trimming. This version of TopHat aligned RNA-Seq reads to the silkDB silkworm genome using the ultra-high throughput short read aligner Bowtie 2.0. The first step was to index the genome with this version of Bowtie.

```
./bowtie2-build -f silkgenome.fa silk_index2
```

Where 'bowtie2-build' is the command used, "-f silkgenome.fa" specifies the fasta file of the *Bombyx* genome and silk_index2 is the prefix name given to the index.

```
TopHat /location of bowtie2/silk2_index /location of
output/ TopHat_output/ LEFT_READ_FILE_trim.fq
RIGHT_READ_FILE_trim.fq --num-threads 3
```

Where 'TopHat' is the mapping program. 'silk2_index' is the prefix of the bowtie 2.0 index files. 'TopHat_output' is the output folder where the analysis data is to be written. '`LEFT_READ_FILE_trim.fq`' is the left end read fastq file of the RNA-Seq run and '`RIGHT_READ_FILE_trim.fq`' is the right end read fastq file of the RNA-Seq run. '--num-threads 3' allows usage of 3 threads of CPU simultaneously.

To check the alignment statistics, SAMtools was used on the file called accepted_hits.bam, present in the 'TopHat_output' folder.

```
samtools flagstat accepted_hits_hw.bam
```

Where, 'samtools' is the program. 'flagstat' is the tool that calculates the statistics and 'accepted_hits_hw.bam' is the bam file of the reads that aligned to the genome.

5.5.4 Assembling aligned reads with Cufflinks

The aligned reads for each sample (fore and hind wing aligned files) were assembled into a GTF format using Cufflinks. It also estimates the abundance of the transcripts.

```
cufflinks -o cufflinks_out_folder /location of TopHat
output/accepted_hits.bam
```

Where, 'cufflinks' is the program, '-o `cufflinks_out_folder`' is the output indicating the cufflinks_out folder. 'accepted_hits.bam is the TopHat output aligned file.

5.5.5 Calculating differential expression using Cuffdiff

The genome annotation files created after individual cufflink steps were merged into a single gft using cuffmerge program. This merged assembly provides a uniform basis for calculating gene and transcript expression in each condition. The merged assembly and reads are fed to cuffdiff, which calculates expression levels and tests the statistical significance of the observed changes.

The cuffmerge program was run to merge the fw and hw assemblies.

```
cuffmerge -o hw_fw_merge -g Bombyx.gtf -p 3 -s
/location_of_bowtie/silk2_index.fa
fw_transcripts.gtf hw_transcripts.gtf
```

Where, 'cuffmerge' is the program. '-o hw_fw_merge' is the ouput directory. '-g *Bombyx*.gtf' is the known annotation file. '-p 3' allows running program on three CPU threads. '-s silk2_index.fa' is the indexed genome fasta file. Fw and hw_transcripts.gtf are the gtf files to be merged.
This creates an hw_fw_merge directory with merged.gtf file.

Cuffdiff was used to find the differential expression between the fore- and hindwing samples.

```
cuffdiff -o cuffdiff_out merged.gtf
accepted_hits_fw.bam accepted_hits_hw.bam
```

Where, 'cuffdiff' is the program, '-o cuffdiff_out' is the output folder. 'merged.gtf' is the merged gtf file. 'accepted_hits_fw.bam' is the forewing assembled bam file and 'accepted_hits_hw.bam' is the hindwing assembled bam file.

The output of cuffdiff was used in CummeRbund program (5.4.5) to visualize the differential expression through different plots.

5.5.6 Visualizing differential expression through CummeRbund

CummeRbund an R package was used to visualize and integrate all the data produced using cuffdiff. It is useful to plot graphs to visualize the cuffdiff data generated.

On a terminal console, 'R' was entered to load the R-shell and to run R based programs.

Load the CummeRbund package into the R environment:

```
> library(cummeRbund)
```

Create a CummeRbund database from the Cuffdiff output:

```
> cuff_data < - readCufflinks('diff_out')
```

Plot the distribution of expression levels for each sample

```
> csDensity(genes(cuff_data))
```

Compare the expression of each gene in two conditions with a scatter plot

```
> csScatter(genes(cuff_data), 'C1', 'C2')
```

Create a volcano plot to inspect differentially expressed genes

```
> csVolcano(genes(cuff_data), 'C1', 'C2')
```

Create a heatmap

```
> csHeatmap(genes(cuff_data), cluster="both")
```

5.5.7 Comparing assembly to known gene annotation using Cuffcompare

The fore- and the hindwing assemblies were separately run on the Cuffcompare program which helps analyze the assembled transcripts. This program compares

the assembled fore- and hindwing files, individually to the reference annotation (*Bombyx*.gtf), separating new genes from known ones.

A file called gtf_out_list.txt that lists all of the GTF files in the working directory is created. Cuffcompare compares each assembly GTF in the list to the reference annotation file *Bombyx*.gtf.

```
cuffcompare -i gtf_out_list.txt -r Bombyx.gtf
```

Where, 'cuffcompare' is the program, '-i `gtf_out_list.txt`' is the input file containing gtf files, obtained after cufflink run on that sample, '-r *Bombyx*.gtf' is the reference annotation file.

The gene lists generated through cuffcompare were used in comparisons between fore- and hindwing datasets by using Biovenn program to plot the Venn diagrams.

5.6 Correlating ChIP-enriched targets of Ubx and gene expression in *Bombyx* hindwing buds

To understand whether targets of Ubx in *Bombyx* identified through ChIP are differentially expressed between fore- and hindwing buds, a comparison was made between the ChIP targets and the differentially expressed genes found from the RNA-Seq experiment by using the Venn diagram generator, BioVenn.

Results and Discussion

5.8 Isolation of wing buds for RNA preparation

Fourth instar *Bombyx* larvae were dissected to obtain around 80 wing buds each of fore- and hindwings. These buds were isolated into tubes maintained in liquid nitrogen and after the isolations were complete, they were stored at -80ºC. RNA was isolated by Trizol extraction and was run on a Bioanalyzer® to check for quality. A good yield of RNA was obtained for both the samples and was suitable for library preparation and sequencing (Fig 5.3). The RNA obtained from both fore- and hindwing bud tissues was used to make a library and paired end sequencing was done on an Illumina GA II with a read length of 100 bp.

5.9 Analysis of the RNA-Seq reads

The paired end 100 bp reads obtained for fore- and hindwing bud samples from the Illumina® genome analyzer were first assessed for quality measures and then processed for further analysis.

5.9.1 FastQC quality control

The program FastQC (refer sections 3.3.1 and 3.5) was used to assess the quality of the RNA-Seq reads generated from the Illumina® paired end sequencing runs for fore- and hindwing samples of *Bombyx*. Most of the parameters were found to be satisfactory for a RNA-Seq run (Fig 5.4 and 5.5). A total sequence count of 39 million reads for the forewing and 46 million reads for the hindwing sequencing runs were obtained with a GC content of 40 %.

The per base sequence quality (Phred) score of all the four runs dropped to values between 35-30 in the last 10 bases at both the ends of the reads. Per base and per sequence GC content of the reads indicated error when analyzed prior to trimming, showing fluctuations in bases 1-10, which could be attributed to lack of sequencing quality in the start and end regions of the read. Therefore, these reads were trimmed off using seqtk master program to improve the alignments.

In RNA-Seq runs, sequence duplications are a common place and hence the error reported in this parameter was not taken as an alarm and the trimmed runs were accepted for further processing to analyze the transcriptome.

5.9.2 Trimming of read-ends

The reads were trimmed off using seqtk master program, removing 10 bases from start and end of the reads to improve the alignments. As we had sequenced the transcriptome with a longer read length of 100 bp, the trimming process did not alter the sequence length significantly at the same time improved the alignment statistics. Only the trimmed reads were used for all downstream processing and transcriptome analysis.

5.9.3 Mapping reads to the genome using TopHat

The silkworm genome obtained from silkDB was first run on bowtie 2.0 build program to index the genome. Six index files were created, four with a suffix .bt2 and two with rev#.bt2. The reads were aligned as paired end sequences onto to the silkworm genome using TopHat 2.0.8b and Bowtie 2.0. TopHat aligns the reads to the genome and outputs the alignment as aaccepted_hits.bam file, apart from bed files describing splice junctions, insertions and deletions from UCSC BED track reported. The alignment statistics were obtained by running the program SAMtools flagstat on the main alignment output file, accepted hits.bam, from TopHat for fore- and hindwing reads. The alignments for fore- and hindwing reads are as shown in table 5.1.

5.9.4 Assembling aligned reads with Cufflinks

The sequence files aligned to the genome were stored as ‘accepted_hits.bam’ file in the TopHat output folder one each for fore- and hindwing reads. The ‘accepted_hits.bam’ file was used for the cufflink assembly. Cufflinks treats each paired fragment reads as a single alignment and it assembles overlapping ‘bundles’ of fragment alignments to give rise to an assembly file, called ‘transcripts.gtf’. It also estimates the abundances of the assembled transcripts.

Cufflinks estimates new sequence transcript features, estimated isoform and gene-level expression values in the generic FPKM tracking format. The 'transcripts.gtf' file can be used in programs like cuffcompare and cuffidiff to estimate expression levels or to identify new transcripts.

5.9.5 Calculating differential expression using Cuffdiff

To estimate differences in the levels of gene expression between fore- and hindwing, assemblies were first merged into a single gtf file by using the program cuffmerge. An output file with merged assembles of fore- and hindwing assemblies, called merged.gtf was obtained. The merged.gtf file was used on Cuffdiff to estimate the differences in expression levels between fore- and hindwing samples. Very few genes (59) showed differential expression between fore- and hindwing buds (the statistical significance was based on the q-value (similar to p-value) calculated by cutdiff). The list of differentially expressed genes is tabulated in Table 5.2. The output generated by cutdiff was fed to an R-based (a statistical package) program called CummeRbund to generate various plots to visualize the differential expression data.

5.9.6 Comparing assembly to known gene annotation using Cuffcompare

Cufflinks utility suite called Cuffcompare was employed to compare fore and hindwing assemblies to the known annotation gene set in *Bombyx*. Cuffcompare separates new genes from known ones and new isoforms of known genes from known splice variants. Cuffcompare, when run on each of fore- and hindwing assemblies provided us with a list of known genes and new genes. These lists were annotated using SilkDB and Biomart (as described in section 3.11) and were compared to each other using the Venn diagram generator BioVenn.

5.9.7 Visualizing differential expression through CummeRbund

The package CummeRbund was used to generate various graphical representations of the differential expression data like Density plot, scatter plot, volcano plot and differential heat maps were generated (Fig 5.6). All these

displayed the absence of any major gene expression differences between the two wing buds in *Bombyx*. While 13,485 annotated transcripts were common to both the buds, only 391 and 225 genes were unique to the fore- and hindwing buds, respectively (Fig 5.7). The absence of any gene expression differences is reflective the morphological similarities between the fore- and hindwings.

The significance cut off in cufflinks is meaningful in cases when biological replicates exist. In the current analysis, we did not have any replicates as the data was generated primarily to improve the quality of the identification of targets of Ubx. Nevertheless, when fold enrichment greater or equal to two was considered, 241 genes were found differentially expressed between the fore- and hindwing buds as opposed to 59 genes in the default cufflinks significance cut-off. Amongst the genes differentially expressed between fore- and hindwing buds are *Ubx*, *engrailed* (*en*), *cheerio* and *bent*. In *Drosophila* too Ubx is expressed only in hlatere, while *en* is not differentially expressed between wing and haltere. Interestingly, we observed increased levels of *cheerio* and *bent* in the hindwing bud, while in *Drosophila*, haltere express much lower levels of these genes compared to the wing.

5.10 Comparison of ChIP-Seq data and transcriptome data

We next asked the question, how many of the genes that are differentially expressed between fore- and hindwing buds (with at least a two-fold enrichment) are direct targets of Ubx. We observed that only 10 genes are present in both data sets (Fig 5.8; Table 5.3). This suggests a minimal role for Ubx during hindwing development, which is morphologically similar to the forewing.

5.11 Differential expression in *Drosophila* of homologues of targets of Ubx in *Bombyx*

The fly homologs of the direct targets of Ubx in *Bombyx* hindwing identified by ChIP-seq were compared with the genes that are differentially expressed between wing and haltere as described in previously published microarray

studies (Mohit Prasad et al. 2006 and Pavlopoulos and Akam, 2011) using BioVenn. We observed that homologues of many targets of Ubx in *Bombyx* (37 when compared to Mohit Prasad et al. 2006 and 65 when compared to Pavlopolous and Akam, 2011) are differentially expressed in wing and haltere (Fig 5.9, Tables 5.4 and 5.5), while those genes show no differential expression between fore- and hindwing buds in *Bombyx*. Regulatory regions of these genes may have evolved in *Drosophila* to respond more strongly to the presence of Ubx.

5.12 Future direction: Analysis of Ubx-binding regions

We have observed that very few genes are differentially regulated between fore- and hindiwngs and amongst them even fewer are directly regulated by Ubx. Interestingly, we have observed that fly homologues of many targets of Ubx in *Bombyx* are differentially expressed between wing and haltere. This suggests that these targets may have acquired additional features in their enhancer regions so that they respond differently in the fly lineage. It has been shown earlier that that Ubx binds to a core TAAT motif in *Drosophila* (Ekker et al, 1994). However, in both *Drosophila* and *Apis*, these motifs do not appear to be recognition sequences for Ubx to identify its targets in the chromatin (Agrawal et al., 2011; Prasad, 2013). Preliminary analysis of the enhancer regions bound by Ubx suggests that in *Bombyx* too, there is no consensus motif that could be referred to as recognition site for Ubx.

The future work in this direction would involve detailed comparative analysis of enhancer regions of few targets (in both *Drosophila* and *Bombyx*) around the regions where Ubx binds and identify regulatory sequences that are different between the two species. This follows functional validation in transgenic *Drosophila* of those regulatory sequences that causes a given target to be differentially expressed between wing and haltere in *Drosophila*.

5.13 Comparative analysis of targets of Ubx and differentially expressed genes between fore and hind wing appendages.

Specification of haltere fate as opposed to a default wing state by Ubx in two winged insects is a paradigm for regulation of organ specification and modification. In order to understand the specification of haltere as opposed to the default wing state in T3 segment of *Drosophila* by Ubx microarray and ChIP experiments have been done. To understand the evolution of Ubx as a Hox factor that alters fate to modify an organ, we need to understand the role of Ubx in other insect hind wing appendages as well. The exploration of the role of Ubx in regulating the downstream targets in hind wing appendage across insect orders will also elucidate those molecular players that are crucial for the development of a haltere as opposed to a wing. Comparative studies between Ubx targets identified by ChIP seq (in case of *Bombyx* and *Apis*) and ChIP-chip (in case of *Drosophila*) along with the expression patterns of fore wing/wing and hind wing/haltere allows us to understand the targets that may be relevant in haltere development.

In the chapter 4 (section 4.8) we had already established through comparative analysis that the many targets of Ubx common to all the three orders are known to be relevant in wing development and also the targets specific to the *Drosophila* set consist of many genes experimentally shown to be crucial in haltere development (Hersh and Carroll, 2005, Mohit Prasad et al. 2006, Pavlopoulos and Akam, 2011). Ubx binding data suggests that it regulates some crucial wing development genes across the three insect orders (like *en, vg*). Ubx was also found to be show significant expression in the fore wing discs of *Apis* (Prasad N, 2013) and the peripodial membrane in *Bombyx* and *Drosophila*. Thus binding of Ubx to the enhancer of a gene alone does not seem to be imparting the functionality that is necessary to modify the wing fate to that of haltere. The role of regulation around the Ubx binding seems to be key in imparting function to this factor.

The targets or the biological processes Ubx regulates in the three insect orders were not so distinct. We then verified if there is any differential expression

between the morphologically similar fore and hind wing buds of *Bombyx* through RNA seq analysis. When we carried out transcriptome studies we found that very few genes/transcripts are differentially expressed between the fore and the hind wing buds of *Bombyx*. We found 241 genes that were differentially regulated between the fore and the hind wing discs when a 2 fold change or higher cut off was applied. If the default cufflinks significance criterion was used only 59 genes were shown to be differential between the discs.
There is an evident differential expression of Ubx itself, which is ubiquitous in hind wing disc and limited to peripodial membrane of the fore wing disc in *Bombyx*, this sort of served as positive control for the transcriptome studies. As there is no obvious modification in the hind wing, the lack of differential expression between fore and hind wings in *Bombyx* is expected. Similarly in the case of *Apis*, where the expression of Ubx in the fore wing is also comparable to that of hind wing, some genes (identified by qPCR) were found show minimal or no differential expression. The fore and hind wing buds in *Apis* have Ubx expression and share a significant number of ChIP targets between them (Prasad N, 2013).

The next question we needed to address was if any of the targets were under the regulation of Ubx to show differential expression between fore and hind wings in *Bombyx*. We compared the transcriptome data with the ChIP seq data in *Bombyx* in order to find out the genes that are up or down regulated in the hind wing and also are a direct target of Ubx. These genes could be the ones that are regulated by Ubx and in response show differential expression. However of the 241 genes that are differentially regulated between fore and hind wing in *Bombyx* only 10 are actually the targets of Ubx (correlate to hind wing ChIP seq data). Interestingly *engrailed* a gene which is not differentially regulated between wing and haltere was found to be regulated by Ubx in *Bombyx* hind wings. Two other targets *bent* and *cheerio* were also regulated by Ubx but were down regulated in *Bombyx* hind wings as opposed the up regulated state in haltere.

Very few targets which are differentially expressed between hind and fore wings and also targeted by Ubx indicate a minimal role for Ubx during *Bombyx*

hindwing development, which is morphologically similar to the forewing. As we did not see that the targets of Ubx in hind wing of *Bombyx* were differentially regulated, the next question we asked was if the same genes are targeted by Ubx across insect orders but if some are differentially regulated only in Dipterans as opposed to the other orders. If they are, then these differentially expressed genes are likely to be relevant in the development of a haltere. So we did a comparative analysis between the differentially expressed genes in *Drosophila* (from studies described below) and the Ubx bound targets in *Bombyx* hind wings.

Identification of differentially expressing genes between wing and haltere discs of *Drosophila* was done in our lab (Mohit Prasad et al. 2006) and in Dr.Akam's lab (Pavlopoulos and Akam, 2011). The work in our lab was carried out using the DGRC cDNA library probes using a traditional microarray using the wild type wing, haltere and the Cbx^{hm} discs to identify the potential candidates that are relevant to Ubx function in specifying the haltere. This work identified the genes that had differential expression between wing and haltere discs and also followed up with validation of some of the resulting candidate genes that are crucial for haltere specification.

To identify the targets regulated by Ubx specify haltere, Dr.Akam's lab used the TARGET version of the GAL4/UAS system coupled with Affymetrix® *Drosophila* Genome 2.0 microarrays to profile transcriptional changes in wings after Ubx misexpression. Using temperature switch for expression of ectopic Ubx in wings they measured the transcriptional responses to Ubx during larval, pre-pupal, and pupal development. The earlier notion that Ubx only represses wing specific targets in haltere does not hold true as shown by these papers, Ubx can up or down regulate a variety of targets between haltere and wing. A list that was a combination of the targets identified in these two studies was used for a comparative analysis with the Ubx targets in hind wings of *Bombyx*.

The comparative analysis between the fore-hind wing using RNAseq suggested that a majority of the genes were not differentially regulated between fore and hind wing buds in *Bombyx*. Another comparative analysis was done to compare

the differentially expressed genes in haltere and wing to the set of genes that were targets of Ubx in hind wings of *Bombyx*. This intersection will list out the genes that are bound by Ubx in both the organisms but are only differentially regulated in *Drosophila*. We identified 37 (Mohit Prasad et al. 2006) and 65 (Pavlopoulos and Akam, 2011) genes that are targets in both the insects but are only differentially regulated in case of *Drosophila*. These genes could be the crucial ones for the development of haltere and its specification from the default wing fate. Many of these genes identified have already been validated for their role in haltere development.

To understand the relevance of binding of Ubx in the appendages of *Bombyx* and *Apis,* in which cases there is no differential regulation between the appendages we analyzed the binding motifs. The hypothesis was that variable factors that bring about the different wing appendage modification in hind wings of different insects may be at the level of regulation at the binding site. We went ahead and dissected out the enhancer elements identified in these studies particularly those targets of Ubx that were only differentially expressed between wing and haltere whilst the same was not differential between hind and fore wings.

The pro wing gene *vestigial* is differentially expressed between wing and haltere but not in fore and hind wing either in *Apis* or *Bombyx*. When the enhancer regions between these insects were compared we found the presence of ADF and MAD motifs in *Drosophila* which were not present in *Apis* (N Prasad, 2013). When we created a fly transgenic carrying the mutated version of *Drosophila* we were able to abolish the differential expression of Vg between wing and haltere. This experiment indicates that the combined effect of the binding proteins with Ubx on the chromatin may be crucial to being about the specification and modification resulting from the action of Ubx. Similar studies of looking closely at the binding regions and the motifs occurring have been done for *Bombyx*.

Apart from the comparative analysis the transcriptome for *Bombyx* was not completely explored for features like splice variants and novel genes found to

be expressed in wing discs. Genome annotation of *Bombyx* indicates that a set of genes (around 6000 in 30000 genes) identified through full-length cDNA does not correspond to the genomic data. These genes could be identified through correlation with the transcriptome data and will serve to identify new targets in the wing disc which may have been left out due to non-availability of annotation or sequence in the genome data.

Hence through our studies on all the three insects we now have a direction which looks like the regulation by Ubx as a Hox factor is dependent on the milieu of factors binding along with Ubx on chromatin. This concurs with the fact that there is no consensus binding site for Ubx across insects and that Ubx expressed in fore wings does not amount to any modification. This mechanism where Ubx regulates along with co-occurring factors could be the paradigm that may govern many other Hox factors which may regulate the downstream effectors based on the complexes they bind with in different insects and situations. Evolutionarily it is likely that the enhancer regions have evolved to host many factors that can help regulate the binding and control of targets in a lineage specific manner. However this hypothesis has to be tested and our lab is conducting experiments using *Drosophila* transgenics to explore the regulatory mechanisms across insect orders and evolutionary time.

5.14 Mechanims of Hox regulation: A discussion an evo-devo perspective of Ubx regulation in Insects.

Homeotic genes are known to control the body axis formation and the segment specific development. They are expressed in a collinear fashion in the developing embryo and the interaction between them regulates the growth and development of tissues and organs in spatial and temporal precision. The context specific expression and spatio-temporal control of the interaction is the key process that regulates the segment-wise development of an animal.

Hox genes which are the ‘selector genes’ express transcription factors that work by regulating a set of downstream targets known as ‘realizator genes’ and this is

achieved by different modes. Firstly Hox control is achieved by synergistic inputs from signaling cascades and other transcription factors acting in tandem in a spatio-temporal fashion. The next mode of regulation is at the actual binding site itself, where the ability of a Hox factor may depend on other interacting protein co-factors. Enhancer sequences to which Hox proteins bind may sport multiple binding sites to achieve regulation levels based on the monomers that can bind to it (Galant et al, 2002). The protein factors may be modified to have different motifs which alter binding to the enhancer site. These concepts are known from the studies done on individual selector genes and their mode of action but genome wide studies will add more insights into the evolution and mode of action of these genes.

Hox genes may regulate a process or an organ specification by a master control mode where they control the primary regulator genes to initiate development of a new organ. These few regulators then control realizator genes to effect in organogenesis. Ubx however seems to control the specification by modulating the wing transcriptional network at multiple levels controlling many genes which is reflected by the direct binding data we have for the three insects.

The difference in the expression alone is not enough to explain the regulation leading to the development controlled by a Hox factor, for example in the case of hind wing specification. The specification of haltere in the T3 segment in *Drosophila* is controlled by Ubx. The default state of the thoracic segments is the wing state and in T3 Ubx modifies the fate of wing to specify a haltere. But a haltere is not just the result of Ubx expression in the T3 segment as when we see in other insects the presence of Ubx in wings does not always lead to modifications. The hind wings of Lepidoptera express Ubx but still maintain a wing without much modification from the fore wing. Ubx can down regulate and up regulate downstream targets to achieve the haltere specification fate from a default wing state. Presenting a very different scenario is the case of *Tribolium* (beetle) hind wings where the default state is the elytron state and Ubx regulation helps maintain the hind wing state instead of a elytra. Hence just the expression alone does not result in the appendage modification but the interaction of the factors and downstream regulation brings in the change required.

The Hox genes have a very well conserved helix loop helix motif called the homeodomain through which they bind the DNA, however how the specificity is achieved when all the Hox factors bind to similar core motif is not clearly understood and remains a question to this day. Hox genes can achieve specificity in regulation while binding to similar regions in DNA, how this specificity is achieved in spite of binding with a well conserved motif is also known as Hox paradox.

Ubx like other factors has a well conserved homeodomain and was shown to bind to heptamers around the core TAAT motif by in vitro binding experiments (Ekker et al, 1994). However when we studied Ubx binding in *Drosophila* discs we found that the TAAT motif was widespread throughout the genome and it was not overrepresented in the ChIP enriched sequences (in other insects too). On further analysis we found that other motifs are found with the Ubx enriched sequences that may be binding along with Ubx and regulating the downstream targets. In other words Ubx depends on a milieu of factors for tissue specific regulation. Studies have also shown that Ubx interacts with certain protein co factors like exd and hth which alter binding specificities of Ubx to achieve tissue specific regulation (Galant et al, 2002).

This study is an attempt to understand the role of the Hox protein factor Ubx that controls the fate of an organ across different insect orders. The study aims to understand the regulatory modes that have evolved across different organisms. The second aspect it also focuses on is the possible mechanisms of development of the modified organ under the regulation of the master regulator in a genome-wide study. There have been studies which have explored individual hox genes and the way the protein or gene sequence has evolved across insect orders through evolution (Alonso et al, 2001). These studies have been able to identify the duplication, deletion or emergence of new genes in the hox complex to achieve the insect-order specific response to evolutionary pressure. Ours is a first of its kind of evo-devo study that involves genome wide identification of the targets of Hox factors to compare and understand the regulatory mechanisms that have evolved in organ specification and modification by comparing the genome wide binding of a Hox factor across insect orders.

Ubx as a Hox factor controlling the development of the third thoracic appendage is a classic example of organ fate determination by a single master

regulator gene. Studies in *Drosophila* had identified the differentially expressing genes between the wing and haltere. Furthermore genome wide ChIP-chip studies identified direct binding sites and the corresponding targets of Ubx in haltere.

In order to test of the hypothesis put forth from the observations in *Drosophila* we explored other insects to understand Hox regulation in general and Ubx in particular. The popular hypothesis is that Ubx may control different sets of genes in different hind wing appendages to bring modifications (Weatherbee et al, 1999). However our studies hint that Ubx could be controlling similar sets of genes but the altered regulation due to other binding factors may change the appendage development and modification.

Previous to this study no selector gene has be studied with a genome wide approach across insect groups. There was a possibility that Ubx bound different motifs and CREs to regulate a unique set of genes in different hind wing appendages. But we find that there is no consensus motif across insect orders or within the insect itself for Ubx. It is likely that Ubx is recruited with the help of a complex of proteins to its binding site. In all the insects we studied we saw that co-occurring factors on chromatin were highly enriched over background in the ChIP enriched sequences, this was true in comparison with the canonical TAAT based motifs as well. However more binding studies must be done to confirm this hypothesis.

We also looked if the set of processes or pathways Ubx regulated was different or unique to different kinds of appendages, we did not find any major deviation between the sets of Ubx targets across insect orders. The next question we posed was if Ubx is indeed bound in many cases but really not making a difference to the development what genes were up or down regulated in response to Ubx in different insects? We found that Ubx along with many other factors could bind to the enhancer elements.

We went ahead and dissected out the enhancer elements identified in these studies particularly those targets of Ubx that were only differentially expressed between wing and haltere whilst the same was was not differential between hind and fore wings. The pro wing gene *Vestigial* is differentially expressed between wing and haltere but not in fore and hind wing either in *Apis* or *Bombyx*. When the enhancer regions between these insects were compared we found the

presence of ADF and MAD motifs in *Drosophila* which were not present in *Apis* (N Prasad, 2013). When we created a fly transgenic carrying the mutated version of *Drosophila* we were able to abolish the differential expression of Vg between wing and haltere. This experiment indicates that the combined effect of the binding proteins with Ubx on the chromatin may be crucial to being about the specification and modification resulting from the action of Ubx.

Further experiments and understanding of the interactions between Ubx and these proteins in regulatory mechanisms would allow us to decipher the specificity and the complexity of Hox mediated regulation. Hence not much is known about the evolution and development of selector genes in the way they regulate downstream targets in molecular detail. The regulation by Hox genes as selectors will be best understood if we study genome wide enhancers and gene networks and the subtle interactions (Reviewed by Mann and Carroll, 2002). Our study is the first step towards understanding the Hox regulation, from an evo-devo perspective.

Summary

This chapter described the identification of the fore- and hindwing (of wing buds from fourth instar larvae) transcriptomes in *Bombyx* by using high throughput RNA-Sequencing. The paired end reads obtained from the two tissue datasets were analyzed and annotated.

Tuxedo suite was used to map the reads and estimate differences in expression levels between the transcriptomes of fore- and hindwing. It was found that very few genes (214 independent transcripts) when compared to the total genes expressed (13711 independent transcripts) were differentially expressed between the two wing buds. When compared to data from ChIP-seq (as described in previous chapters), it was observed that a very small fraction of the targets of Ubx are differentially expressed between the fore- and hindwing buds.

However, we have observed that a larger fraction of genes that are differentially expressed in *Drosophila* are direct targets of Ubx in *Bombyx*, but they are not differentially expressed in *Bombyx*. This suggests that certain genes may have evolved to respond more strongly in *Drosophila* to the presence of Ubx.

We conclude this chapter by pointing out that although large number genes are targeted by Ubx in both *Bombyx* and *Drosophila,* majority of them are differentially regulated only in *Drosophila*. It is possible that these genes may have evolved to be regulated by Ubx only in dipteran lineage. To understand the mechanism, we would compare and analyze the enhancer regions bound by Ubx in *Bombyx* and *Drosophila*.

Plates
Chapter 5

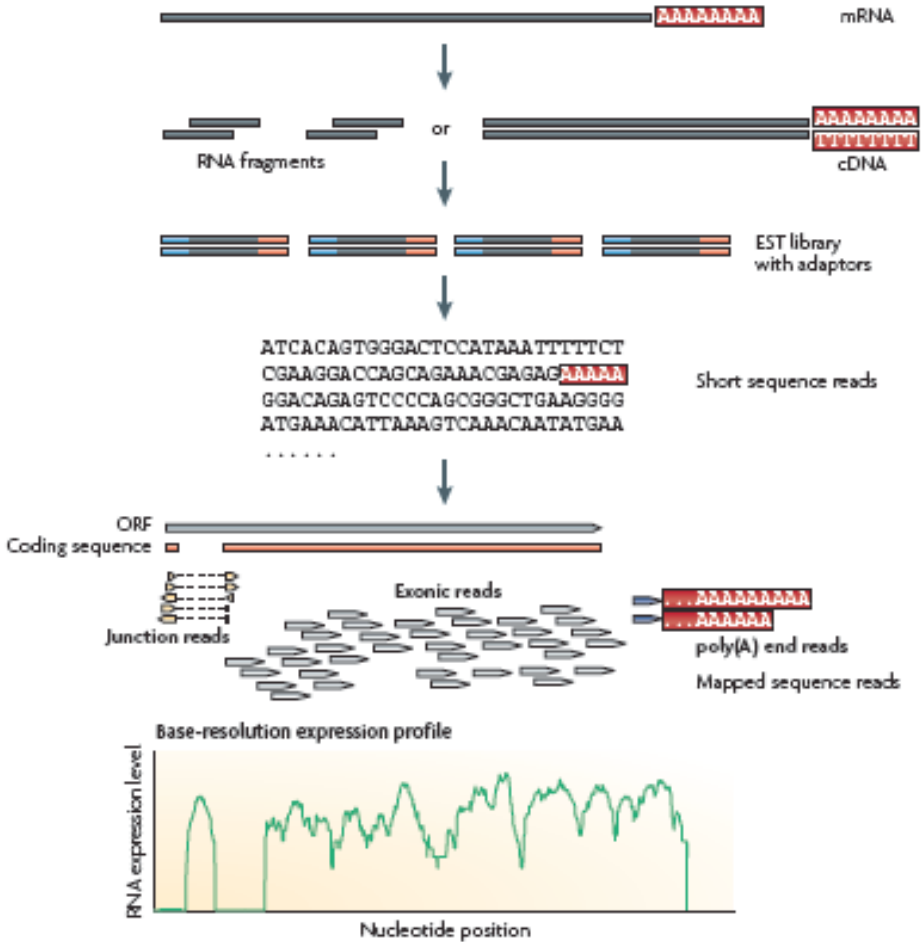


Figure 5.1 Flow-chart of RNA-Sequencing: Transcripts are first converted to a library of cDNA fragments. Sequencing adaptors are added to each cDNA fragment and short sequences from both ends are obtained from each cDNA using high throughput sequencing. The resulting sequences are aligned to a reference genome and classified as exons, junctions and end reads. These are used to generate an expression profile. (Wang et al, 2009)

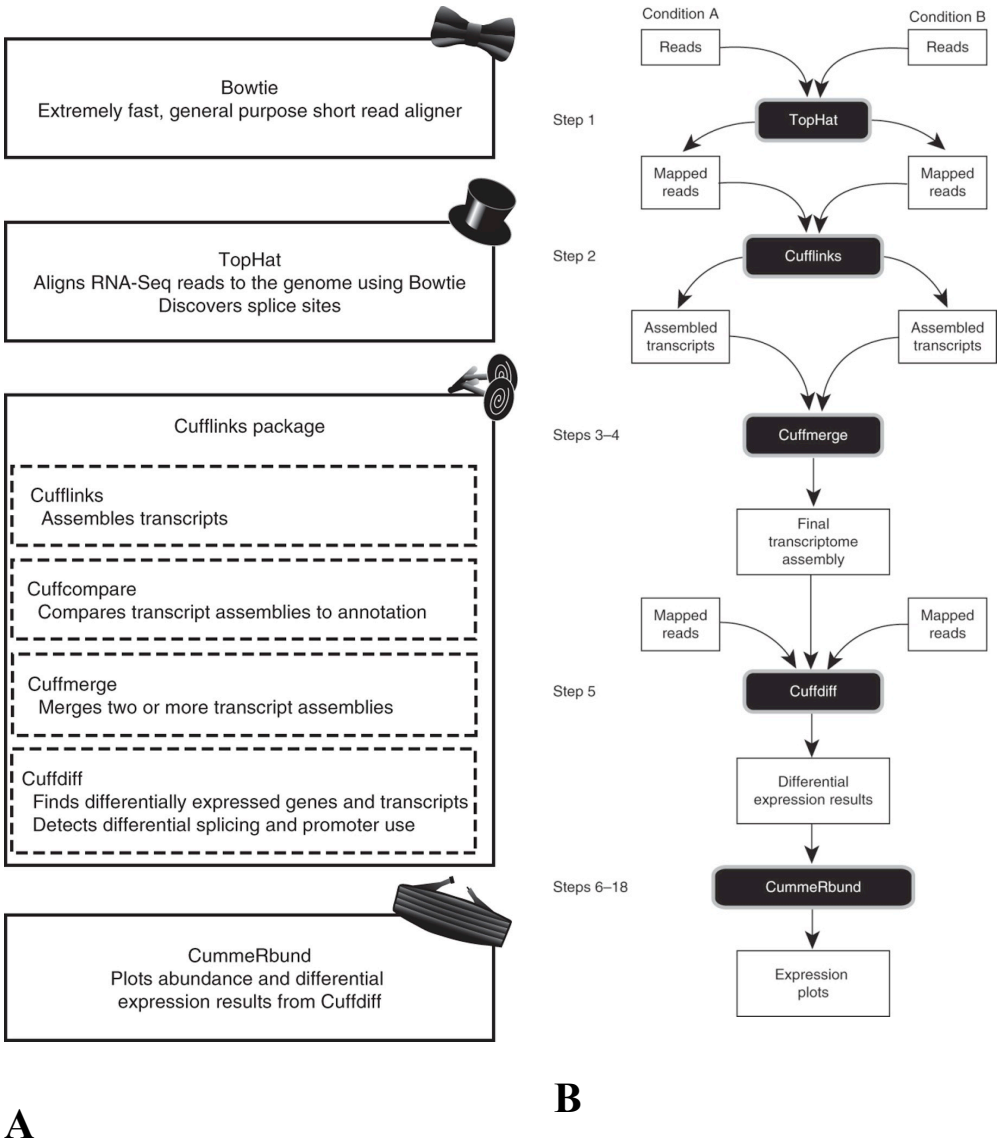


Figure 5.2 Overview of the programs (A) and protocol (B) of the Tuxedo suite used for the transcriptome analysis. (Trapnell et al, 2012)

A. The programs in Tuxedo suite allow the user to align genome using the help of Tophat/Bowtie and use an array of cufflink tools to assemble the transcripts and quantify them and even plot the data as graphs.

B. Schematic representation of the flow chart of processing RNA seq data through the Tophat- Cufflinks pipeline.

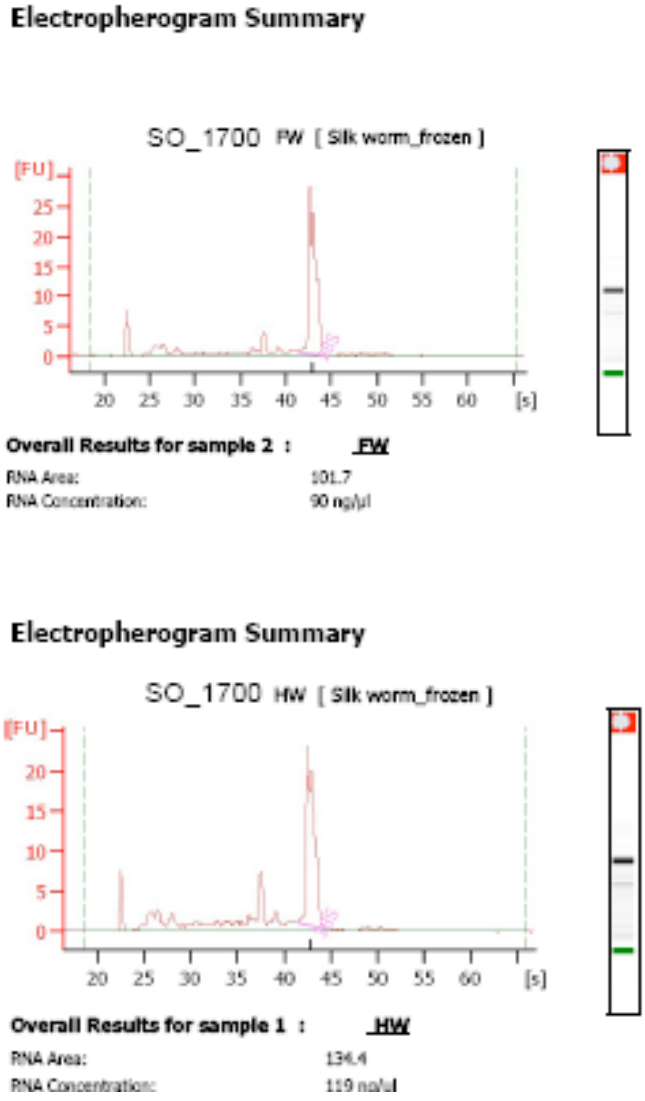
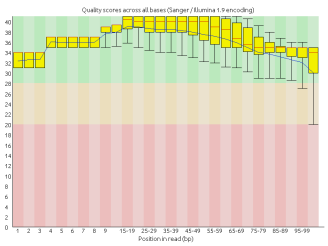
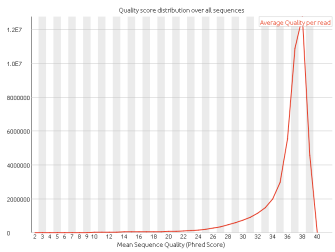


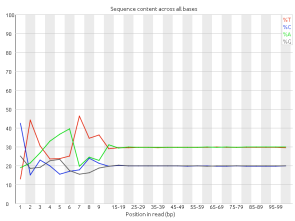
Figure 5.3 Quality tests of isolated RNA. RNA yield from *Bombyx* fore- (FW) and hindwing (HW) buds. A good yield of RNA with acceptable quality was obtained for both the samples and was suitable for RNA-Seq library preparation and sequencing. Courtesy: Genotypic ltd.



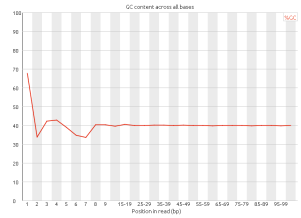
A. Per base sequence quality



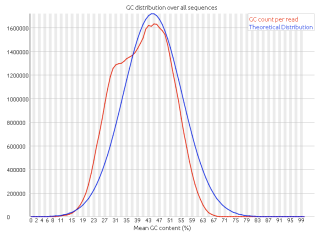
B. Per sequence quality score



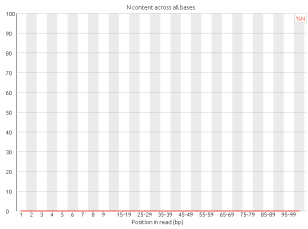
C. Per base sequence content



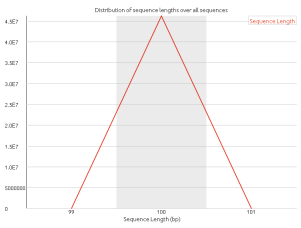
D. Per base GC content



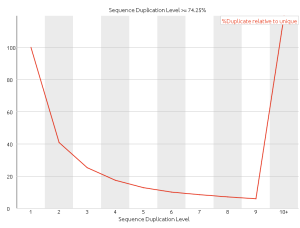
E. Per sequence GC content



F. Per base N content

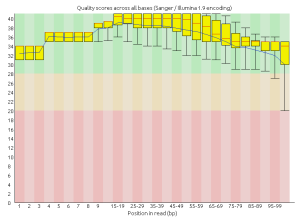


G. Sequence length distribution

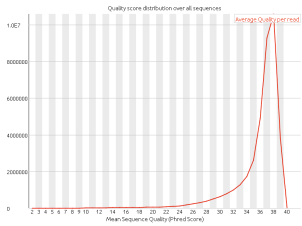


H. Sequence duplication levels

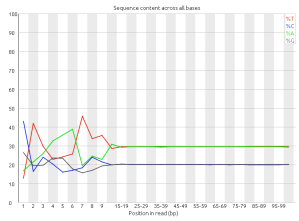
Figure 5.4 Quality control analysis of the hindwing bud RNA-Seq reads using FastQC. Both R1 and R2 reads had similar quality statistics.



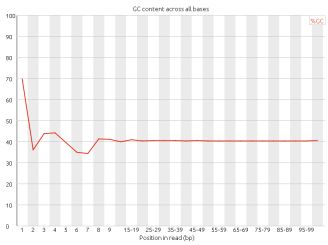
A. Per base sequence quality



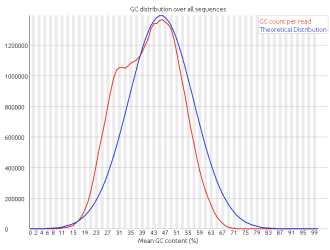
B. Per sequence quality score



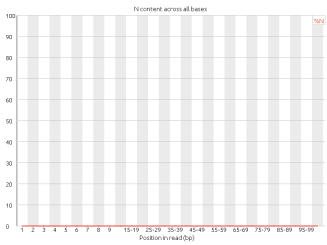
C. Per base sequence content



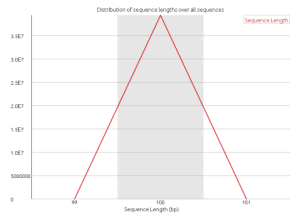
D. Per base GC content



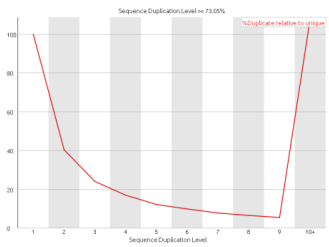
E. Per sequence GC content



F. Per base N content



G. Sequence length distribution



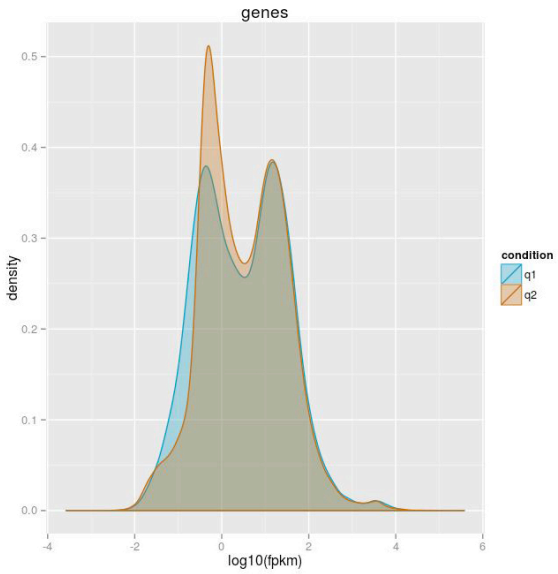
H. Sequence duplication levels

Figure 5.5 Quality control analysis of the forewing bud RNA-Seq reads using FastQC. Both R1 and R2 reads had similar quality statistics.

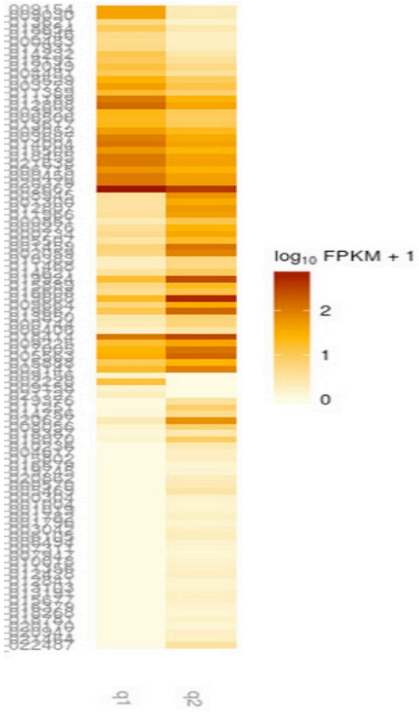
Table 5.1 Alignment statistics (using TopHat) for the sequence reads from two ends of transcripts from hindwing (A) and forewing (B). Note, near-perfect overlap of sequences from two ends suggesting good quality of the RNA-seq data.

A. HW samtools flagstat accepted_hits_hw.bam	No. of reads
in total (QC-passed reads + QC-failed reads)	112859139
duplicates	0
mapped (100.00%:-nan%)	112859139
paired in sequencing	112859139
read1	56819672
read2	56039467
properly paired (65.85%:-nan%)	74321992
with itself and mate mapped	107914516
singletons (4.38%:-nan%)	4944623
with mate mapped to a different chr	18064042
with mate mapped to a different chr (mapQ>=5)	471382

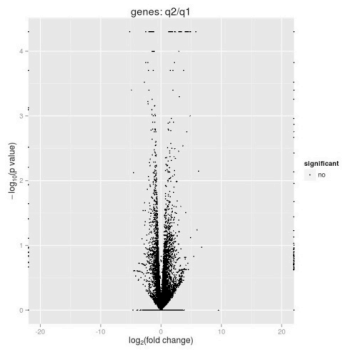
B. FW samtools flagstat accepted_hits_fw.bam	No. of reads
in total (QC-passed reads	95504600
duplicates	0
mapped (100.00%:-nan%)	95504600
paired in sequencing	95504600
read1	48135950
read2	47368650
properly paired (65.33%:-nan%)	62392010
with itself and mate mapped	91241396
singletons (4.46%:-nan%)	4263204
with mate mapped to a different chr	15100124
with mate mapped to a different chr (mapQ>=5)	303096



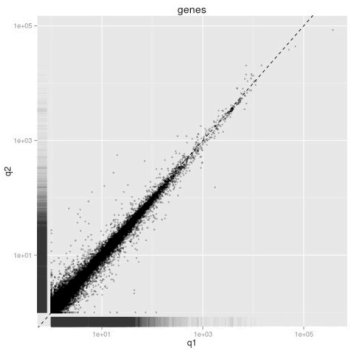
A. Density plot



B. Heat map of top 100 differential genes



C. Volcano plot



D. Scatter plot

Figure 5.6 Visualization of the transcriptome data between fore- (q1) and hindwing (q2) buds using CummeRbund. The plots show that fore- and hindwing buds express identical gene sets and to the similar quantitative levels.

Table 5.2: Differentially expressed genes between fore- and hindwings of *Bombyx* when q-value significance cut off is applied.

Bm Gene stable ID	Fly ID	Fly CG ID	FB GENE NAME
BGIBMGA000421	FBgn0040601	CG13643	-
BGIBMGA005808	FBgn0029994	CG2254	-
BGIBMGA004737	FBgn0030570	CG12540	-
BGIBMGA012968	FBgn0031089	CG9572	-
BGIBMGA009518	FBgn0032192	CG5731	-
BGIBMGA002849	FBgn0032414	CG17211	-
BGIBMGA007275	FBgn0034267	CG4984	-
BGIBMGA009872	FBgn0035620	CG5150	-
BGIBMGA010235	FBgn0038181	CG9297	-
BGIBMGA012637	FBgn0038366	CG4576	-
BGIBMGA009066	FBgn0038784	CG4362	-
BGIBMGA009872	FBgn0038845	CG10827	-
BGIBMGA013107	FBgn0040397	CG3655	-
BGIBMGA013042	FBgn0259241	CG42339	-
BGIBMGA012524	FBgn0264489	CG43897	-
BGIBMGA013945	FBgn0000044	CG10067	Actin 57B A4
BGIBMGA013945	FBgn0000045	CG7478	Actin 79B A4
BGIBMGA013945	FBgn0000046	CG18290	Actin 87E A4
BGIBMGA005812	FBgn0000116	CG32031	Arginine kinase
BGIBMGA004547	FBgn0005666	CG32019	bent
BGIBMGA013756	FBgn0034197	CG15918	Chitin deacetylase-like 9
BGIBMGA000331	FBgn0033725	CG8502	Cuticular protein 49Ac BMORCPR40
BGIBMGA004618	FBgn0031461	CG16987	dawdle
BGIBMGA012997	FBgn0030597	CG9504	Ecdysone oxidase
BGIBMGA008860	FBgn0000639	CG17285	Fat body protein 1
BGIBMGA011485	FBgn0004620	CG6992	Glutamate receptor IIA
BGIBMGA011485	FBgn0020429	CG7234	Glutamate receptor IIB
BGIBMGA011485	FBgn0046113	CG4226	Glutamate receptor IIC
BGIBMGA006693	FBgn0029167	CG7002	Hemolectin
BGIBMGA009688	FBgn0000448	CG33183	Hormone receptor-like in 46
BGIBMGA008024	FBgn0010482	CG9432	lethal (2) 01289
BGIBMGA004103	FBgn0011296	CG4533	lethal (2) essential for life
BGIBMGA000388	FBgn0260660	CG42543	Multiplexin BMORCPR146
BGIBMGA001201	FBgn0002789	CG4696	Muscle protein 20
BGIBMGA006510	FBgn0002772	CG5596	Myosin alkali light chain 1
BGIBMGA014226	FBgn0264695	CG17927	Myosin heavy chain
BGIBMGA002259	FBgn0002773	CG2184	Myosin light chain 2
BGIBMGA000613	FBgn0003149	CG5939	Paramyosin
BGIBMGA008038	FBgn0043578	CG9681	PGRP-SB1

BGIBMGA004045	FBgn0035089	CG9358	Pherokine 3 CSP1
BGIBMGA004066	FBgn0035089	CG9358	Pherokine 3 CSP5
BGIBMGA001587	FBgn0004117	CG4843	Tropomyosin 2
BGIBMGA008861	FBgn0031692	CG6514	Troponin C at 25D
BGIBMGA006937	FBgn0010423	CG9073	Troponin C at 47D
BGIBMGA006937	FBgn0010424	CG7930	Troponin C at 73F
BGIBMGA013531	FBgn0053519	CG33519	Unc-89
BGIBMGA013700	FBgn0004169	CG7107	upheld
BGIBMGA001030	FBgn0004028	CG7178	wings up A
BGIBMGA001031	FBgn0004028	CG7178	wings up A
BGIBMGA000612			
BGIBMGA000624			
BGIBMGA000736			
BGIBMGA000737			
BGIBMGA001586			
BGIBMGA002565			
BGIBMGA002747			
BGIBMGA003216			
BGIBMGA006054			
BGIBMGA008022			
BGIBMGA008023			
BGIBMGA008101			
BGIBMGA008204			
BGIBMGA008287			
BGIBMGA009573			
BGIBMGA009687			
BGIBMGA010111			
BGIBMGA010979			
BGIBMGA011077			
BGIBMGA011078			
BGIBMGA011079			
BGIBMGA012523			
BGIBMGA012796			
BGIBMGA013041			
BGIBMGA014298			
BGIBMGA014116			
BGIBMGA001862			

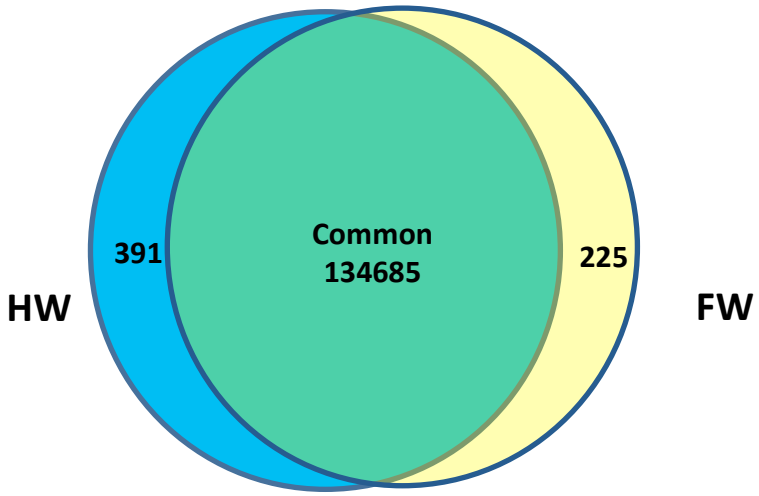


Figure 5.7 Comparison of genes expressed in forewings (FW) and hindwings (HW) in *Bombyx*.

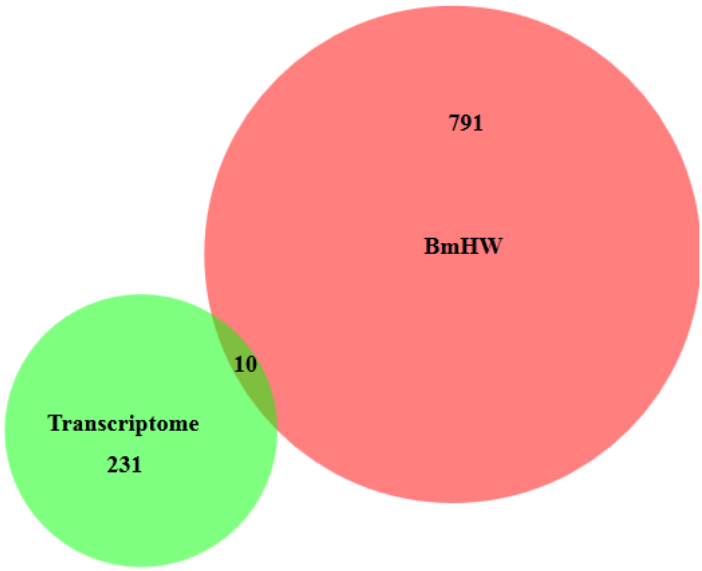


Figure 5.8 Comparison of genes that are direct targets of Ubx in *Bombyx* and are differentially expressed genes between fore- and hindwing buds.

Table 5.3 (Right panel): Genes that are direct targets of Ubx and are differential expressed between fore- and hindwing buds.

1	*Muscle protein 20*
2	*Dynein heavy chain at 89D*
3	*sidestep*
4	*bent*
5	*cheerio*
6	*engrailed*
7	*Hormone receptor-like in 46*
8	*Heat shock proteins 68-70*
9	*Heat-shock-protein-70Aa*
10	*Heat-shock-protein-70Bbb*

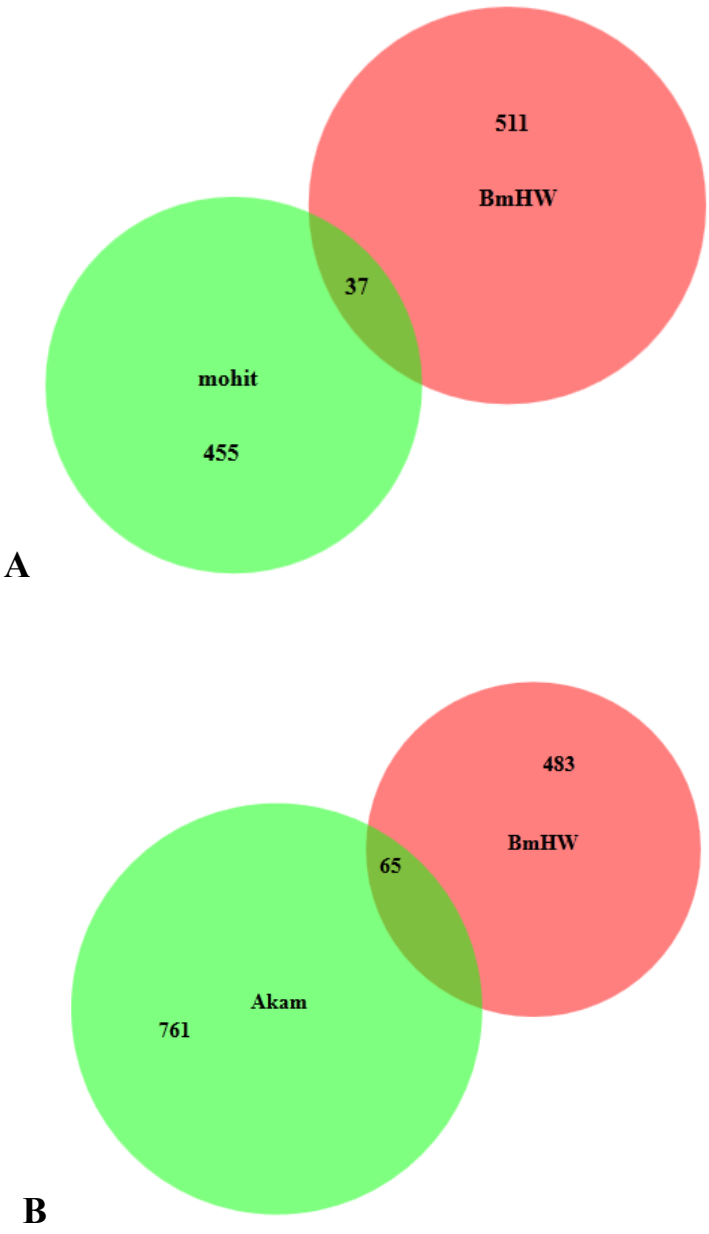


Figure 5.9 Comparison of genes that are direct targets of Ubx in *Bombyx* hindwing and genes that are differentially expressed between wing and halatere in *Drosophila* (A. Mohit Prasad et al, 2006. B. Pavlopolous and Akam, 2011).

Table 5.4: Genes that are direct targets of Ubx in *Bombyx* and their homologues are differentially expressed between wing and haltere in *Drosophila* (Mohit Prasad et al 2006).

FBID KEY	FB CG ID	SYMBOL	FB GENE NAME
FBgn0000097	CG3166	aop	anterior open
FBgn0000319	CG9012	Chc	Clathrin heavy chain
FBgn0000575	CG1007	emc	extra macrochaetae
FBgn0000591	CG8365	E(spl)m8	Enhancer of split m8
FBgn0001179	CG8019	hay	haywire
FBgn0002590	CG8922	RpS5a	Ribosomal protein S5a
FBgn0002734	CG8328	E(spl)mdelta	Enhancer of split mdelta,
FBgn0002778	CG3297	mnd	minidiscs
FBgn0004687	CG3201	Mlc-c	Myosin light chain cytoplasmic
FBgn0004893	CG10021	bowl	brother of odd with entrails limited
FBgn0004907	CG17870	14-3-3zeta	14-3-3zeta
FBgn0010113	CG15532	hdc	headcase
FBgn0010348	CG8385	Arf79F	ADP ribosylation factor at 79F
FBgn0010548	CG11140	Aldh-III	Aldehyde dehydrogenase type III
FBgn0014133	CG1822	bif	bifocal
FBgn0014184	CG16747	Oda	Ornithine decarboxylase antizyme
FBgn0015509	CG1877	Cul1	Cullin 1
FBgn0020443	CG6382	Elf	Ef1alpha-like factor
FBgn0026238	CG2944	gus	gustavus
FBgn0026761	CG3152	Trap1	Trap1
FBgn0027499	CG12340	wde	windei
FBgn0030520	CG10990	Pdcd4	Programmed cell death 4 ortholog
FBgn0031950	CG14536	Herp	Homocysteine-induced ER protein
FBgn0032476	CG5439	CG5439	-
FBgn0033095	CG3409	CG3409	-
FBgn0034644	CG10082	CG10082	-
FBgn0036030	CG6767	CG6767	-
FBgn0036337	CG11255	AdenoK	Adenosine Kinase
FBgn0037236	CG9772	Skp2	-
FBgn0038926	CG13409	CG13409	-
FBgn0039244	CG11069	CG11069	-
FBgn0039907	CG2041	lgs	legless
FBgn0040296	CG3396	Ocho	Ocho
FBgn0040340	CG11642	TRAM	TRAM
FBgn0050440	CG30440	CG30440	-
FBgn0052447	CG32447	CG32447	-
FBgn0262735	CG1691	Imp	IGF-II mRNA-binding protein

Table 5.5: Genes that are direct targets of Ubx in *Bombyx* and their homologues are differentially expressed between wing and haltere in *Drosophila* (Pavlopoulos and Akam, 2011).

FB ID	FBCG ID	SYMBOL	FB GENE NAME
FBgn0000273	CG4379	Pka-C1	cAMP-dependent protein kinase 1
FBgn0000307	CG5813	chif	chiffon
FBgn0000395	CG15671	cv-2	crossveinless 2
FBgn0000448	CG33183	Hr46	Hormone receptor-like in 46
FBgn0000591	CG8365	E(spl)m8-HLH	Enhancer of split m8, helix-loop-helix
FBgn0001981	CG3758	esg	escargot
FBgn0002631	CG6096	E(spl)m5-HLH	Enhancer of split m5, helix-loop-helix
FBgn0002633	CG8361	E(spl)m7-HLH	Enhancer of split m7, helix-loop-helix
FBgn0002733	CG14548	E(spl)mbeta-HLH	Enhancer of split mbeta, helix-loop-helix
FBgn0002735	CG8333	E(spl)mgamma-HLH	Enhancer of split mgamma, helix-loop-helix
FBgn0003257	CG3593	r-l	rudimentary-like
FBgn0003292	CG6097	rt	rotated abdomen
FBgn0003885	CG2512	alphaTub84D	alpha-Tubulin at 84D
FBgn0003888	CG3401	betaTub60D	beta-Tubulin at 60D
FBgn0003975	CG3830	vg	vestigial
FBgn0004360	CG1916	Wnt2	Wnt oncogene analog 2
FBgn0004581	CG30170	bgcn	benign gonial cell neoplasm
FBgn0004687	CG3201	Mlc-c	Myosin light chain cytoplasmic
FBgn0004779	CG1330	Ccp84Ae	Ccp84Ae
FBgn0005612	CG3090	Sox14	Sox box protein 14
FBgn0005666	CG32019	bt	bent
FBgn0010774	CG1101	Ref1	RNA and export factor binding protein 1
FBgn0011648	CG12399	Mad	Mothers against dpp
FBgn0011837	CG4070	Tis11	Tis11 homolog
FBgn0013279	CG6489	Hsp70Bc	Heat-shock-protein-70Bc
FBgn0014141	CG3937	cher	cheerio
FBgn0025885	CG11143	Inos	Inos
FBgn0026160	CG7958	tna	tonalli
FBgn0027654	CG2239	jdp	jdp
FBgn0028540	CG9008	CG9008	-
FBgn0028622	CG13432	qsm	quasimodo
FBgn0028741	CG6355	fab1	-
FBgn0029881	CG3973	pigs	pickled eggs
FBgn0030114	CG17754	CG17754	-
FBgn0030520	CG10990	Pdcd4	Programmed cell death 4 ortholog

FBgn0030608	CG9057	Lsd-2	Lipid storage droplet-2
FBgn0031256	CG4164	CG4164	-
FBgn0031310	CG4764	Vps29	Vacuolar protein sorting 29
FBgn0031322	CG5001	CG5001	-
FBgn0031816	CG16947	CG16947	-
FBgn0031976	CG7367	CG7367	-
FBgn0032120	CG33298	CG33298	-
FBgn0032132	CG4382	CG4382	-
FBgn0032782	CG9994	Rab9	Rab9
FBgn0033483	CG12919	egr	eiger
FBgn0033913	CG8468	CG8468	-
FBgn0034709	CG3074	Swim	Secreted Wg-interacting molecule
FBgn0034985	CG3328	CG3328	-
FBgn0035087	CG2765	CG2765	-
FBgn0035499	CG14996	Chd64	Chd64
FBgn0036165	CG7533	chrb	charybde
FBgn0036849	CG14079	CG14079	-
FBgn0037416	CG15592	Osi9	Osiris 9
FBgn0040251	CG6658	Ugt86Di	Ugt86Di
FBgn0041094	CG7590	scyl	scylla
FBgn0051057			
FBgn0051676	CG31676	CG31676	-
FBgn0052447	CG32447	CG32447	-
FBgn0053196	CG33196	dp	dumpy
FBgn0083919			
FBgn0086708	CG32130	stv	starvin
FBgn0261286	CG12785	Mat89Ba	Maternal transcript 89Ba
FBgn0261642			
FBgn0262127	CG33967	kibra	kibra ortholog
FBgn0262656	CG10798	dm	diminutive

Future directions

The variety and diversity in insect appendages were attributed to the Hox genes that determine segmental identity in the body plan development. The function of the master regulator Hox protein Ultrabithorax (Ubx) is a classic example of organ specification, where it alters the wing fate in third thoracic segment of *Drosophila* to specify a haltere (Lewis, 1978). The morphological differences between hindwing appendages in insects were attributed not only to the expression of the protein Ubx itself or its levels, but to the differences in the target genes it regulated (Weatherbee et al, 1998). Our studies in continuation to the efforts in last decade to understand the evolution and development of organ modifications using high throughput methods now hint that rather than the difference in the assortment of the target genes, it is the regulatory pattern of these targets by Ubx and other associated proteins that could control the organ modification.

In this study we found that many of the targets that Ubx regulates in the three insect orders (*Bombyx, Apis* and *Drosophila*) are common, which are known to play important role in *Drosophila* wing development. The differential expression of those genes between fore and hind appendages in *Bombyx* is minimal, while in *Drosophila* many of them are differentially regulated. It is likely that the binding of other transcription factors around Ubx-binding sites in the *cis*-regulatory regions may contribute to the development of haltere. The preliminary motif finding studies using MEME on ChIP data indicate that there is no clear target recognition sequence for Ubx suggesting that binding of other transcription factors in the vicinity may provide docking sites for Ubx to bind DNA. For example, GAGA factor is known to be associated in *Drosophila* with Ubx regulation in haltere (Agrawal et al, 2011). The cofactors Extradenticle (Exd) and homothorax (hth) are proteins that are known to be critical for target selectivity of various Hox proteins. The levels of regulation may be altered by the presence or absence of such cofactors. As many targets of Ubx are common amongst *Drosophila*, *Bombyx* and *Apis*, the regulation could also be in the

selection of the cofactors that either interact directly with Ubx or bind alongside Ubx during the regulation of targets.

The future work in this direction would involve detailed comparative analysis of enhancer regions of few targets (in both *Drosophila* and *Bombyx*) around the regions where Ubx binds and identify regulatory sequences that are different between the two species. This follows functional validation in transgenic *Drosophila* of those regulatory sequences that causes a given target to be differentially expressed between wing and haltere in *Drosophila*. The validation involves (i) transgenic flies expressing a reporter gene under the regulation of enhancers of targets of Ubx in *Bombyx* and enhancers of corresponding targets in *Drosophila* and (ii) mutating the enhancers of *Bombyx* such a way that it my functionally behave like those of *Drosophila* and vice-versa.

There is also a possibility of recruitment of other cofactors to regulate the expression of targets of Ubx. A proteomics approach to identify the cofactors that are bound to Ubx in the late third instar haltere disc of *Drosophila* would be a very important step in understanding the regulation of targets in *Drosophila* in coordination with cofactors. The availability of a specific antibody to *Drosophila* Ubx is an inherent advantage to such a study (Agrawal et al, 2011). However Protein Immunoprecipitation studies with systems such as haltere discs is a challenging task. Alternatively, Ubx protein trap lines (DGRC Kyoto, Choo et al, 2011) can be used to achieve effective immuno-precipitation. We have raised a UAS-Ubx-FLAG transgenic fly that can be used to overexpress Ubx for immuno-precipitation experiments. The divergence of Ubx function which may involve evolution of new cofactor partners or activity modifiers may be crucial to explain the action of organ specification by Ubx in particular and Hox proteins in general.

The comparisons done in this study between hindwing of *Bombyx* and *Apis* and haltere in *Drosophila* are informative as to what differences have evolved between the two kinds of appendages and what regulatory role Ubx is playing in this process. The wing appendages in insects are diverse and the complete understanding of this diversity through evolution and role of Ubx in its

development can be answered better if more insects are explored in the way done in this study. The role of Ubx in hindwing appendage specification is quite different in the beetles (Coleoptera), for example, Ubx represses the default elytron (a protective structure modified from wing) formation to promote the wing development in the third thoracic segment in *Tribolium* (Tomoyasu et al, 2005). The identification of targets of Ubx in hindwings of *Tribolium* may give further clues as to what regulation patterns have evolved and lead to such diversity in insect hindiwing appendages. We have raised antibodies specific to *Tribolium* Ubx, which may be used to carry out ChIP-seq to identify the targets of Ubx in the *Tribolium* wing.

All the studies done to identify the targets of Ubx in *Drosophila* haltere have been based on the ChIP-chip methodology, whereas the current studies in other insects involves the higher resolution studies based on the more advanced ChIP-sequencing. To identify the targets of Ubx in *Drosophila* haltere at a higher resolution and to make an effective comparison built on the same platform, it may be advisable to venture into a ChIP-seq study on the haltere discs to identify the targets of Ubx in *Drosophila*.

These studies should help advance the understanding of regulation of body formation and patterning by Hox genes in general and throw light on the role of Ubx and its regulation in bringing about appendage diversity in the largest animal order on earth.

Appendices

Appendix Chapter 2

2.7.2: Comparison of sequences from silkworm races to that of Genome database

Table 2.1: Primer sets used to amplify the intronic regions for the sequence comparison between sequence from locally available races and the SilkDB database.

Primer Name	Sequence (5'- 3')	Length (bp)
Bmx Ci 1F	TGCTTGTCGTTCACATGAGG	20 bp
Bmx Ci 1R	GTGGGTCTTCAGGTTTTCCA	20 bp
Bmx Ci 2F	GGTCGCACACTGGAGAGAA	19 bp
Bmx Ci 2R	GGACTGTTTTCACGTGTTTCC	21 bp
Bmx Actin4 1F	CGGCAATCGGTATCTGTTTC	20 bp
Bmx Actin4 1R	TGCTATTGCACAGCTTCGTT	20 bp
Bmx Actin4 2F	GCAATAACGAAGCTGTGCAA	20 bp
Bmx Actin4 2R	TTCTGTCCCATACCGACCAT	20 bp
Bmx Actin4 3F	GAGGCACAGAGCAAAAGAGG	20 bp
Bmx Actin4 3R	GGAGTGCGTATCCCTCGTAG	20 bp

PCR Components 50 µl reaction

10X Buffer	-5.0µl
10mM dNTP	-2.0 µl
50mM $MgCl_2$	-1.5 µl
20 µM FP	-2.5 µl
20 µM RP	-2.5 µl
Pfu Taq Polymerase	-1.0 µl
Genomic DNA template	-1.0 µl
MQ Water	-35 µl

PCR Cycle

Temp	94ºC	94ºC	56ºC	72ºC	72ºC	4ºC	Cycles
Time	2m	45s	25s	45s	5m	Hold	35

2.9 Generation and validation of Antibodies

Table 2.2: Primer sets used to amplify the N terminal region of *Bombyx* Ubx

Forward BomNdeI : 5' GGAATTCCATATGCAGGGCGGCGGT 3'

Reverse BomRev : 5' GTTGCTGTTAGCGAATGTTACAAAA 3'
(binds to initial region of Homeodomain)

Reverse BomRev2 : 5' CGGGATCCGTTCGCTCCTGCTATG 3'
(excludes YPWM and Homeodomain regions)

2.9.4 Validating the Antibody through Western blot hybridization

RIPA lysis buffer

Tris-HCl pH 8.0	50mM
NaCl	150mM
NP40 (Igepal)	1%
Sodium Deoxycholate	0.5%
Sodium Dodecyl Sulphate	0.1%
Dithiothreitol (DTT)	1mM
Protease Inhibitor cocktail	1X
PMSF	1X

Western blot transfer buffer

Tris base	25 mM
Glycine	192 mM
Methanol	10%

Western blot wash buffer TBST pH 7.6

Tris base	50 mM
NaCl	150 mM
Tween20	0.05%

2.10.1 Standardization of ChIP conditions

1. Confirmation of the quantity of DNA for the ChIP pull-down

Primer

Table 2.3: Bm Spalt 472bp

Forward IDT BmSal-js 5' GAATGCACTCCGACCCCG 3'

Reverse IDT BmSal-jas 5' GCGACGGTGATCGAGCGA 3'

PCR Components 25 µl reaction

10X Genei Taq Buffer	- 2.5µl
10mM Genei dNTP	-2.5 µl
50mM $MgCl_2$	-2.5 µl
20 µM FP	-1.0 µl
20 µM RP	-1.0 µl
Genei Taq Polymerase	-1.0 µl
ChIP pulldown/input template	-3.0 µl
MQ Water	-11.5 µl

PCR Cycle

Temp	94ºC	94ºC	55ºC	72ºC	72ºC	4ºC	Cycles
Time	2m	45s	35s	45s	5m	Hold	35

2. Validation of the modified ChIP protocol

Primers (Papantonis & Lecanidou, 2009).

1. Bm Erp 1

Forward IDT Bm Erp 1F 5' CTTAAACTTCTGAGGGC 3'

Reverse IDT Bm Erp 1R 5' CTTTGATCAATTGAGGAAC 3'

2. Bm Hcp 13

Forward IDT Bm Hcp13F 5' GTAACTAAGAATCATGTTCACCTTG 3'

Reverse IDT Bm Hcp13R 5' GAGCAGTTTCCTTGAAAATCCG 3'

PCR Components 50 µl reaction

10X Agilent® Pfu Buffer -5.0µl

10mM Genei dNTP -5.0 µl

DMSO -0.5 µl

20 µM FP -2.0 µl

20 µM RP -2.0 µl

Pfu Taq Polymerase -1.0 µl

ChIP pulldown/input template -2.0 µl

MQ Water -31.5 µl

PCR Cycle

Temp	94ºC	94ºC	58ºC	72ºC	72ºC	4ºC	Cycles
Time	3m	30s	30s	20s	5m	Hold	28

2.10.2 Chromatin Immuno-precipitation

HEPES low salt lysis buffer (made fresh, kept on ice)

10mM HEPES pH 7.8

10mM KCl

0.1M EDTA

0.5mM PMSF

1mM DTT

1x Roche® EDTA free complete-Protease inhibitor cocktail

Appendix Chapter 3

3.10.1 Galaxy Fetch closest non-overlapping feature for every interval (peak)

Python code to calculate the shortest distance between two gene region coordinate pairs.

```
infile = open('2k_intersect_distpeaks', 'rU')
lines = infile.readlines()
for line in lines:
     line = line.split()
     l_1 = [1,2]
     l_2 = [11,12]
     list_dist = []
     list_ind = []
     for i in l_1:
          for j in l_2:
               diff = int(line[i])-int(line[j])
               diff = abs(diff)
               list_dist.append(diff)
               f_ele = int(line[i])
               s_ele = int(line[j])
               list_ind.append((f_ele,s_ele))
     for i in range(len(list_dist)):
          if list_dist[i]==min(list_dist):
               print list_ind[i][0]-list_ind[i][1]
```

3.10.2 BED-tools slop-intersect method to find the genes near peaks

Python code to retain one intersection pair between peak and gene feature

```
import sys

infile=open('10k_anno_hw-15fdr.tsv', 'rU')
outfile = open('output.tsv', 'w')


gene=''
peak=''
genes=[]
count=0
total=0
for line in infile:
    total=total+1
    line=line.strip()
    linsplit=line.split('\t')
    if len(linsplit)<2: continue #skip empty lines
    if linsplit[-2]=='0': continue
    gensplit=linsplit[-2].split()
    if gensplit[2] in genes:
        continue

    else:
        count=count+1
     outfile.write(line)
     outfile.write('\n')
     genes.append(gensplit[2])
        peak=linsplit[3]
print "non_redundant_lines =" , count
print "total_lines =" , total

outfile.close()
```

3.11 Annotation of genes associated with the peaks

A. Python code to merge two excel sheets based on the common unique ID. Does not work if the database has 2 entries for a query.

```
import xlrd
import xlwt
#test file
input_file=xlrd.open_workbook("test_query.xls")
sheet = input_file.sheet_by_index(0)

#database file
compare_file=xlrd.open_workbook("database.xls")
compare_sheet = compare_file.sheet_by_index(0)


print (sheet.cell_value(1,0) ==
compare_sheet.cell_value(1,0))

#open a new XL file

wbk = xlwt.Workbook()
wt_sheet = wbk.add_sheet("harsha")
for j in range(sheet.nrows):
    for i in range(sheet.ncols):
        wt_sheet.write (j,i,sheet.cell_value(j,i))
    for k in range(compare_sheet.nrows):
        if sheet.cell_value(j,0) ==
compare_sheet.cell_value(k,0):
            for q in range(compare_sheet.ncols):
                wt_sheet.write (j,sheet.ncols+q,
compare_sheet.cell_value(k,q))
                pass
#Save the new XL file with a file name
wbk.save('test_done.xls')
```

B. Python code (improvised to add lines when more than one match is seen) to make a index library of the database tsv file and merge two tsv files based on the common unique ID.

```
def bgi_sort(col_ind):
     infile = open('geneseta_full.txt','rU')
     outfile = open('outputbgi.txt', 'w')
     list_lines = infile.readlines()
     for i in range(len(list_lines)):
          linestripped = list_lines[i].strip()
```

```
          linesplit_list   = linestripped.split('\t')
          stringofinterest=linesplit_list[col_ind]
          colstripped = stringofinterest.strip()
          colsplit_list = colstripped.split(',')
          count = 0;
          for x in colsplit_list:
               if x[:3]=='BGI':
                    count += 1
                    linesplit1_list =
linesplit_list[:col_ind]+[x]+linesplit_list[col_ind:]
     ##insert x(e.g x = 'BGI2345266')
                    ##every column of linesplit_list
written to outfile
                    ccc = 0
                    for l_column in linesplit1_list:
                         ccc += 1
                         outfile.write(str(l_column))
## l_column is a string
                         outfile.write('\t')

                    outfile.write('\n')
          if count == 0:
               linesplit_list =
linesplit_list[:col_ind]+[0]+linesplit_list[col_ind:]
               for l_column in linesplit_list:
                         outfile.write(str(l_column))
 ## l_column is a string
                         outfile.write('\t')
               outfile.write('\n')

     outfile.close()


def bgi_match(query, target, col_ind):
     infile_query = open(query, 'rU')
     infile_target = open(target, 'rU')
     outfile = open('output_matched.txt', 'w')

     ##making a query list
     ##making a target list


     dict_target = {}
     lines_target = infile_target.readlines()
     for line in lines_target:
          linestripped = line.strip()
          linesplit =linestripped.split('\t')
          keyis = linesplit[col_ind]
          if keyis in dict_target:
               dict_target[keyis].append(linesplit)
## it will be list of lists
```

```
            else:
                dict_target[keyis] = [linesplit]
        lines_query = infile_query.readlines()

        for lineq in lines_query:
            lineq = lineq.strip()
            lineq = lineq.split('\t')
            actual_query = lineq[-1]

            str_query = ''
            if actual_query in dict_target:

                for every_string in lineq:

                    str_query += '\t'+every_string

                for every_list in
dict_target[actual_query]:

                    str_target = ''

                    for every_string in every_list:

                        str_target +=
'\t'+every_string

                    complete_string =
str_query+'\t'+str_target

                    outfile.write(complete_string)
                    outfile.write('\n')
        outfile.close()



bgi_sort(9) ## 9 is 10th column(9th index) where bgi
numbers exist.

bgi_match('query.txt', 'outputbgi.txt',9)
```

BIBLIOGRAPHY

Agrawal, P., Habib, F., Yelagandula, R. & Shashidhara, L.S. (2011). Genome-level identification of targets of Hox protein Ultrabithorax in *Drosophila* , novel mechanisms for target selection. Sci. Rep. 1, 205,1-10

Akam, M. E. (1983). The location of Ultrabithorax transcripts in *Drosophila* tissue sections. The EMBO journal, 2(11), 2075.

Akam, M. E. (1985). Segments, lineage boundaries and the domains of expression of homeotic genes. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 312(1153), 179-187.

Akam, M. E., & Martinez-Arias, A. (1985). The distribution of Ultrabithorax transcripts in *Drosophila* embryos. The EMBO journal, 4(7), 1689.

Alonso C, Maxton-Kuechenmeister J, Akam M. (2001). Evolution of Ftz protein function in insects. Curr Biol, 11:1473-1478.

Angelini, D. R., & Kaufman, T. C. (2005). Comparative developmental genetics and the evolution of arthropod body plans. Annual Review of Genetics, 39, 95–119.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. New biotechnology, 25(4), 195-203.

Averof, M., & Akam, M. (1995). Hox genes and the diversification of insect and crustacean body plans. Nature, 376(6539), 420-423.

Baker, M. (2012). Quantitative data, learning to share. Nature Methods, 9(1), 39–41.

Bender, W., Spierer, P., Hogness, D. S., & Chambon, P. (1983). Chromosomal walking and jumping to isolate DNA from the Ace and rosy loci and the bithorax complex in *Drosophila* melanogaster. Journal of molecular biology, 168(1), 17-33.

Bender, W., Weiffenbach, B., Karch, F., & Peifer, M. (1985). Domains of cis-interaction in the bithorax complex. In Cold Spring Harbor symposia on quantitative biology. Cold Spring Harbor Laboratory Press. Vol. 50, 173-180

Bennett, R. L., Brown, S. J., & Denell, R. E. (1999). Molecular and genetic analysis of the *Tribolium* Ultrabithorax ortholog, Ultrathorax. Development genes and evolution, 209(10), 608-619.

Blair, S. S. (2007). Wing vein patterning in *Drosophila* and the analysis of intercellular signaling. Annu. Rev. Cell Dev. Biol., 23, 293-319.

Brakefield, P. M., Gates, J., Keys, D., Kesbeke, F., Wijngaarden, P. J., Monteiro, A., ... & Carroll, S. B. (1996). Development, plasticity and evolution of butterfly eyespot patterns. Nature, 384(6606), 236-242.

Brusca RC, Brusca GJ. (1990). The Invertebrates. Sunderland, MA, Sinauer Associates

Cabrera, C. V., Botas, J., & Garcia-Bellido, A. (1985). Distribution of Ultrabithorax proteins in mutants of *Drosophila* bithorax complex and its transregulatory genes. Nature, 318(6046), 569-571.

Capovilla, M., Brandt, M., & Botas, J. (1994). Direct regulation of decapentaplegic by Ultrabithorax and its role in *Drosophila* midgut morphogenesis. Cell, 76(3), 461-475.

Carroll SB, Laughon A, Thalley BS.(1988). Expression, function, and regulation of the hairy segmentation protein in the *Drosophila* embryo. Genes Dev. 2(7),883-90.

Carroll, S. B. (1994). Developmental regulatory mechanisms in the evolution of insect diversity, Development. 217–223.

Carroll, S. B. (1995). Homeotic genes and the evolution of arthropods and chordates. Nature 376, 479–485

Castelli-Gair, J., & Akam, M. (1995). How the Hox gene Ultrabithorax specifies two different segments, the significance of spatial and temporal regulation within metameres. Development, 121(9), 2973-2982.

Chai, C et al, (2008). A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, *Bombyx mori*. Insect Biochemistry and Molecular Biology. 38,1111-1120.

Choo SW, White R and Russell S. (2011).Genome wide analysis of the binding of the Hox protein Ultrabithorax and the Hox co factor Homothorax in *Drosophila* . Plos One 6(4), 1-14

Cohen, B., Simcox, A. A., & Cohen, S. M. (1993). Allocation of the thoracic imaginal primordia in the *Drosophila* embryo. Development, 117(2), 597-608.

Crickmore, M. A., & Mann, R. S. (2006). Hox control of organ size by regulation of morphogen production and mobility. Science, 313(5783), 63-68.

Dickinson, M. H. (1999). Haltere-mediated equilibrium reflexes of the fruit fly, *Drosophila* melanogaster. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 354(1385), 903–16.

Duan, J., Li, R., Cheng, D., Fan, W., Zha, X., Cheng, T... Xia, Q. (2009). SilkDB v2.0, a platform for silkworm (*Bombyx mori*) genome biology. Nucleic Acids Research, 1–4.

Duncan I. (1987). The Bithorax Complex. Annual Review of Genetics. Vol. 21, 285-319

Erwin, D. H., & Davidson, E. H. (2002). The last common bilaterian ancestor. Development, 129(13), 3021-3032.

Farnham, P. J. (2009).Insights from genomic profiling of transcription factors. Nature Rev. Genet. 10, 605–616.

Flower, J. W. (1964). On the origin of flight in insects. Journal of Insect Physiology, 10(1), 81-88.

Galant, R., Skeath, J. B., Paddock, S., Lewis, D. L., & Carroll, S. B. (1998). Expression pattern of a butterfly achaete-scute homolog reveals the homology of butterfly wing scales and insect sensory bristles. Current Biology, 8(14), 807-813.

Galant, R., Walsh, C. M., & Carroll, S. B. (2002). Hox repression of a target gene, extradenticle-independent, additive action through multiple monomer binding sites. Development, 129(13), 3115-3126.

Gibson, G. (2000). Evolution, Hox genes and the cellared wine principle. Current biology, 10(12), R452-R455.

Goecks, J., Nekrutenko, A., Taylor, J., & Team, T. G. (2010). Galaxy, a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11, R86

Goff, L., Trapnell, C., & Kelley, D. (2011). cummeRbund, Analysis, Exploration, Manipulation and Visualization of Cufflinks High-Throughput Sequencing Data. R package version, 1(0).

González-Reyes, A., & Morata, G. (1990). The developmental effect of overexpressing a Ubx product in *Drosophila* embryos is dependent on its interactions with other homeotic products. Cell, 61(3), 515-522.

Grenier, J. K., & Carroll, S. B. (2000). Functional evolution of the Ultrabithorax protein. Proceedings of the National Academy of Sciences, 97(2), 704-709.

Grenier, J. K., Garber, T. L., Warren, R., Whitington, P. M., & Carroll, S. (1997). Evolution of the entire arthropod Hox gene set predated the origin and radiation of the onychophoran/arthropod clade. Current Biology, 7(8), 547-553.

Hartenstein, V., & Posakony, J. W. (1989). Development of adult sensilla on the wing and notum of *Drosophila* melanogaster. Development, 107(2), 389-405.

Hayes, P. H., Sato, T., & Denell, R. E. (1984). Homoeosis in *Drosophila*, the ultrabithorax larval syndrome. Proceedings of the National Academy of Sciences, 81(2), 545-549.

Held, L.I. (2002). Imaginal discs- The genetics and cellular logic of pattern formation. Cambridge university press

Hersh, B. M., & Carroll, S. B. (2005). Direct regulation of knot gene expression by Ultrabithorax and the evolution of cis-regulatory elements in *Drosophila*. Development, 132(7), 1567-1577.

Hersh, B. M., Nelson, C. E., Stoll, S. J., Norton, J. E., Albert, T. J., & Carroll, S. B. (2007). The UBX-regulated network in the haltere imaginal disc of D. melanogaster. Developmental biology, 302(2), 717-727.

Ho, J. W., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., & Park, P. J. (2011). ChIP-chip versus ChIP-seq, lessons for experimental design and data analysis. BMC Genomics, 12(1), 134.

Hojyo, T., & Dr.Fujiwara, H. (1997). Reciprocal transplantation of wing discs between a wing deficient mutant (fl) and wild type of the silkworm, *Bombyx mori*. Development, growth & differentiation, 39(5), 599-606.

Hughes, C. L., & Kaufman, T. C. (2002). Hox genes and the evolution of the arthropod body plan. Evolution & Development, 4(6), 459–99.

Hulsen, T., de Vlieg, J., and Alkema, W. (2008) BioVenn- a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. Vol 9, 488.

Hunter, P. (2007). The nature of flight. EMBO reports, 8(9), 811-813.

James, T., & Jill, P. (2012). Integrative Genomics Viewer (IGV), high-performance genomics data visualization and exploration, Briefings in Bioinformatics. 14(2), 178–192.

Jr, G. D., Sherman, B. T., Hosack, D. A., & Yang, J. (2003). DAVID , Database for Annotation , Visualization , and Integrated, Genome Biology. Vol 4(9), R60.

Kasprzyk A, (2011).BioMart, driving a paradigm change in biological data management. doi,10.1093 bar049

Kaufman TC, Seeger MA, Olsen G. (1990) Molecular and genetic organization of the antennapedia gene complex of *Drosophila* melanogaster. Adv Genet., 27,309-62.

Kelsh, R., Weinzierl, R. O., White, R. A., & Akam, M. (1994). Homeotic gene expression in the locust Schistocerca, An antibody that detects conserved epitopes in ultrabithorax and abdominal-A proteins. Developmental genetics, 15(1), 19-31.

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Staines, D. M. (2014). Ensembl Genomes 2013, scaling up access to genome-wide data, Nucleic Acids Research, 42,546–552.

Keys, D. N., Lewis, D. L., Selegue, J. E., Pearson, B. J., Goodrich, L. V., Johnson, R. L., ... & Carroll, S. B. (1999). Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. Science, 283(5401), 532-534.

Kinsella, R. J., Ka, A., Spudich, G., Almeida-king, J., Staines, D., Derwent, P., Flicek, P. (2011). Original article Ensembl BioMarts , a hub for data retrieval across taxonomic space, Database 2011, 1–9.

Kosman, D., Small, S., & Reinitz, J. (1998). Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. Development genes and evolution, 208(5), 290-294.

Krzywinski M et al. (2009) Circos, an information aesthetic for comparative genomics. Genome Research. Vol 19,1639-1645.

Kukalová-Peck, J. (1983). Origin of the insect wing and wing articulation from the arthropodan leg. Canadian Journal of Zoology, 61(7), 1618-1669.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S, & Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research, 22(9), 1813-1831.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology.10 (3).

Lawrence, P. A., & Struhl, G. (1996). Morphogens, compartments, and pattern, lessons from *Drosophila*? Cell, 85(7), 951-961.

Lewis, E. B. (1963). Genes and developmental pathways. American Zoologist, 33-56.

Lewis, E. B. (1978). A gene complex controlling segmentation in *Drosophila*. Nature, 276(5688), 565-570.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows – Wheeler transform, Bioinformatics 26(5), 589–595.

Li, Y., Wang, G., Tian, J., Liu, H., Yang, H., Yi, Y., ... & Zhang, Z. (2012). Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA-Sequencing. PloS one, 7(8), e43713.

Macdonald, W. P., Martin, A., & Reed, R. D. (2010). Butterfly wings shaped by a molecular cookie cutter, evolutionary radiation of lepidopteran wing shapes associated with a derived Cut/wingless wing margin boundary system. Evolution & Development, 12(3), 296–304.

Makhijani, K., Kalyani, C., Srividya, T., & Shashidhara, L. S. (2007). Modulation of Decapentaplegic gradient during haltere specification in *Drosophila*. Developmental biology, 302(1), 243-255.

Martinez-Arias, A & White, R. A. (1988). Ultrabithorax and engrailed expression in *Drosophila* embryos mutant for segmentation genes of the pair-rule class. Development, 102(2), 325-338.

Masumoto, M., & Yaginuma, T. (2009). Functional analysis of Ultrabithorax in the silkworm, *Bombyx mori* , using RNAi. Development Genes and Evolution. Nov 2009

McKenna DD, Farrell BD. (2010). 9-genes reinforce the phylogeny of Holometabola and yield alternate views on the phylogenetic placement of Strepsiptera. PLoS ONE 5(7),e11887

Metzer M L. (2010). Sequencing technologies -the next generation. Nat Rev.Genet. 11, 31-46

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kana*mori*, H., … Abe, H. (2004). The Genome Sequence of Silkworm , *Bombyx mori*. Genome Biology, 35, 27–35.

Mohit, P., Makhijani, K., Madhavi, M. B., Bharathi, V., Lal, A., Sirdesai, G., ... & Shashidhara, L. S. (2006). Modulation of AP and DV signaling pathways by the homeotic gene Ultrabithorax during haltere development in *Drosophila*. Developmental biology, 291(2), 356-367.

Morata, G. (1975). Analysis of gene expression during development in the homeotic mutant Contrabithorax of *Drosophila* melanogaster. Journal of embryology and experimental morphology, 34(1), 19-31.

Nagata, T., Suzuki, Y., Ueno, K., Kokubo, H., Xu, X., Hui, C. C., & Fukuta, M. (1996). Developmental expression of the *Bombyx* Antennapedia homologue and homeotic changes in the Nc mutant. Genes to Cells, 1(6), 555-568.

Nie, H., Liu, C., Cheng, T., Li, Q., Wu, Y., Zhou, M., ... & Xia, Q. (2014). Transcriptome Analysis of Integument Differentially Expressed Genes in the Pigment Mutant (quail) during Molting of Silkworm, *Bombyx mori*. PloS one, 9(4), e94185.

Nüsslein-Volhard, C., & Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. Nature, 287(5785), 795-801.

Palopoli, M. F., & Patel, N. H. (1998). Evolution of the interaction between Hox genes and a downstream target. Current biology, 8(10), 587-590.

Papantonis, A., & Lecanidou, R. (2009). A modified chromatin-immunoprecipitation protocol for silkmoth ovarian follicular cells reveals

C/EBP and GATA binding modes on an early chorion gene promoter. Molecular biology reports, 36(4), 733-736.

Park P.J. (2009).ChIP–seq, advantages and challenges of a maturing technology. Nat Rev.Genet. 10, 669-680

Pavlopoulos, A., & Akam, M. (2011). Hox gene Ultrabithorax regulates distinct sets of target genes at successive stages of *Drosophila* haltere morphogenesis. PNAS, 108(7), 2855-2860.

Pearson, J. C., Lemons, D., & McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nature Reviews Genetics, 6(12), 893-904.

Pepke, S., Wold, B., & Mortazavi, A. (2009). Computation for ChIP-seq and RNA-Seq studies. Nature methods, 6, S22-S32

Pick, L., & Heffer, A. (2012). Hox gene evolution, multiple mechanisms contributing to evolutionary novelties. Annals of the New York Academy of Sciences, 1256, 15–32.

Prasad N. (2013). Thesis: Hox genes and evolution of arthropod body plan, A comparative analysis of targets of Ultrabithorax in *Drosophila melanogaster* and *Apis mellifera*.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools, a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–2.

Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., … Cunningham, C. W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature, 463(7284), 1079–83.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., & Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. Science, 290(5500), 2306-2309

Roch, F., & Akam, M. (2000). Ultrabithorax and the control of cell morphology in *Drosophila* halteres. Development, 127(1), 97-107.

Sanchez-Herrero, E., Casanova, J., Kerridge, S., & Morata, G. (1985). Anatomy and function of the bithorax complex of *Drosophila*. In Cold Spring Harbor symposia on quantitative biology . Cold Spring Harbor Laboratory Press. Vol. 50, 165-172

Shimomura, M., Minami, H., Suetsugu, Y., Ohyanagi, H., Satoh, C., Antonio, B., Mita, K. (2004). KAIKObase , An integrated silkworm genome database and data mining tool, BMC Genomics. 8, 1–8.

Singh MK, Singh A and Gopinathan KP. (2001). The wings of *Bombyx mori* develop from larval discs exhibiting an early differentiated state, a preliminary report. J. Biosci. 26-2,167-177

Slattery M, Ma L, Négre N, White KP, Mann RS (2011) Genome-Wide Tissue-Specific Occupancy of the Hox Protein Ultrabithorax and Hox Cofactor Homothorax in *Drosophila*. PLoS ONE 6(4), e14686

Solomon, M. J., Larsen, P. L., & Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde, Evidence that histone H4 is retained on a highly transcribed gene. Cell, 53(6), 937-947.

Tomoyasu, Y., Wheeler, S. R., & Denell, R. E. (2005). Ultrabithorax is required for membranous wing identity in the beetle *Tribolium castaneum*, 643–647.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat, discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England), 25(9), 1105–11.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562–78.

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5),

Ueno, K., Hui, C., Fukuta, M., & Suzuki, Y. (1992). Molecular analysis of the deletion mutants in the E homeotic complex of the silkworm *Bombyx mori*, Development 114, 555–563.

Ureta-vidal, A., Ettwiller, L., & Birney, E. (2003). Comparative genomics, genome-wide analysis in metazoan eukaryotes, Nat Rev Genet. Vol 4, 251-262

Vachon, G., Cohen, B., Pfeifle, C., McGuffin, M. E., Botas, J., & Cohen, S. M. (1992). Homeotic genes of the bithorax complex repress limb development in the abdomen of the *Drosophila* embryo through the target gene Distal-less. Cell, 71(3), 437-450.

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., Yu, J. (2005). SilkDB, a knowledgebase for silkworm biology and genomics. Nucleic Acids Research, 33(Database issue), D399–402.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq, a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63.

Warren, R. W., Nagy, L., Selegue, J., Gates, J., & Carroll, S. (1994). Evolution of homeotic gene regulation and function in flies and butterflies. Nature,372, 458-461.

Weatherbee, S. D., Halder, G., Kim, J., Hudson, A., & Carroll, S. (1998). Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the *Drosophila* haltere. Genes & Development. 12,1474-1482.

Weatherbee, S. D., Nijhout, H. F., Grunert, L. W., Halder, G., Galant, R., Selegue, J., & Carroll, S. (1999). Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. Current Biology, 109–115.

Weinstock, G. M., et al. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature, 443(7114), 931–949.

White, R. A. H., & Wilcox, M. (1985). Distribution of Ultrabithorax proteins in *Drosophila*. The EMBO journal, 4(8), 2035.

White, R. A., & Akam, M. E. (1985). Contrabithorax mutations cause inappropriate expression of Ultrabithorax products in *Drosophila*. Nature, 318(6046), 567-569.

White, R. A., & Wilcox, M. (1984). Protein products of the bithorax complex in *Drosophila*. Cell, 39(1), 163-171.

Wolpert L, Beddington R, Jessell T, Lawrence P, Meyerowitz E, Smith J (2002). Principles of development (3rd Ed.). Oxford university press.

Xia et al. (2004). Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx mori*). Science 306,1937-40.

Yamaguchi, J. et al. (2013). Periodic Wnt1 expression in response to ecdysteroid generates twin-spot markings on caterpillars. Nat. Commun. 4,1857

Yasukochi, Y., Ashakumary, L. A., Wu, C., Yoshido, A., Nohata, J., Mita, K., & Sahara, K. (2004). Organization of the Hox gene cluster of the silkworm, *Bombyx mori*, a split of the Hox cluster in a non-*Drosophila* insect. Development genes and evolution, 214(12), 606-614.

Zdobnov, E. M., & Bork, P. (2007). Quantification of insect genome divergence. Trends in Genetics, 23(1), 16-20.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology, 9(9), R137.

Zheng, Z., Khoo, A., Fambrough Jr, D., Garza, L., & Booker, R. (1999). Homeotic gene expression in the wild-type and a homeotic mutant of the moth *Manduca sexta.* Development genes and evolution, 209(8), 460-472.