Comparative analysis of targets of the *Hox* gene *Ultrabithorax* in *Drosophila melanogaster* and *Apis mellifera*

Using tools and techniques of data analytics, statistics, computational & information sciences and bioinformatics to analyze high-throughput data

By:

Abhijit Awadhiya
IISER Pune

Under the aegis of:

L S Shashidhara
Professor and HoD
Biological Sciences
IISER Pune

Table of Contents

Certificate 5

Declaration 7

Abstract 9

List of figures 11

List of tables 12

Acknowledgements 13

Introduction 15

Comparative analysis of targets of the Hox gene Ultrabithorax in Drosophila melanogaster and Apis mellifera 15

Specific objectives of the project: 17

Materials and methods 19

Preparation of the updated database and determination of the functionality of the given targets: 22

Preparation and need of “background” distribution: 27

Documentation of pairwise BLAST outputs for the pairs of polypeptides: 30

Statistical Analysis 34

Results: 35

Frequency distribution plots for “All Genes” dataset: Parameter: % Bit/AL 36

Frequency distribution plots for “All Genes” dataset: Parameter: % Bit/Min_len 37

Wing development-related groups separated using flybase summary information: Frequency distribution plots for “Wing Only” dataset: Parameter: % Bit/AL 38

Frequency distribution plots for wing development-related (“Wing Only”) dataset separated using Flybase gene summary information: Parameter: % Bit/Min_len 39

Comparison of wing development-related groups separated using “GO Biological Process” information: 40

Choo: Bit score / Aligned length 40

Pavan: Bit score / Aligned length 41

Choo: Bit score / Minimal length 42

Pavan: Bit score / Minimal length 43

Choo: %age Identities 44

Pavan: %age Identities 45

Choo: %age Positives 46

Pavan: %age Positives 47

Choo: %age Gaps 48

Pavan: %age Gaps 49

Mann-Whitney Rank Sum Test on data points of various data sets – with wing development-related genes separated using Flybase “GO Biological Process” information 50

1. Bit/Aligned length 50
2. Bit score / Minimum length of the two polypeptides 52
3. %Identities in between the two polypeptides 54
4. %Positives in between the two polypeptides 56
5. %Gaps in between the two polypeptides 58

Discussion 61

Conclusions 63

Future directions 63

References 63

Certificate

This is to certify that this dissertation entitled "Comparative analysis of targets of the *Hox* gene *Ultrabithorax* in *Drosophila melanogaster* and *Apis mellifera*: Using tools and techniques of data analytics, statistics, computational & information sciences and bioinformatics to analyze high-throughput data" towards the partial fulfillment of the BS-MS dual degree program at the Indian Institute of Science Education and Research, Pune represents original research carried out by Abhijit Awadhiya at IISER Pune under my supervision during the academic year 2013-2014.

Date: May 5th, 2014

(L S Shashidhara)
L S Shashidhara
Professor and HoD
Department of Biological Sciences
Indian Institute of Science Education
and Research Pune (IISER Pune)

Declaration

I hereby declare that the matter embodied in the report entitled “Comparative analysis of targets of the *Hox* gene *Ultrabithorax* in *Drosophila melanogaster* and *Apis mellifera*: Using tools and techniques of data analytics, statistics, computational & information sciences and bioinformatics to analyze high-throughput data” are the results of the investigations carried out by me at the Department of Biological Sciences, Indian Institute of Science Education and Research Pune (IISER Pune), under the supervision of L S Shashidhara, Professor and HoD, Biological Sciences and the same has not been submitted elsewhere for any other degree.

Date: May 5th, 2014 (Abhijit Awadhiya)

Abstract

It is now widely accepted that evolution at the level of a family of highly conserved (from insects to human) genes popularly known as *Hox* genes has led to the diversity in animal body plan that we see now. *Hox* genes are master control genes, which function by regulating the expression of downstream target genes. *Hox* gene *Ultrabithorax* (*Ubx*), first discovered in *Drosophila*, is supposed to differentiate the development of flight appendages in second and third thoracic segments in all insect species. However, the function of Ubx protein itself does not appear to have evolved amongst the diverse insect groups, although there are significant differences in Ubx sequences amongst various insect groups. This suggests that in the dipteran lineage, certain wing patterning genes have come under the regulation of Ubx.

The current work aims at identifying how the protein sequences of targets of Ubx have evolved between *Apis* and *Drosophila*. It involves comparing the protein sequences of targets of Ubx in *Drosophila* and *Apis*. Some proteins are targets of Ubx in *Drosophila* or *Apis* alone and some are common to both. We also analyzed a subgroup of these proteins, which are previously shown to be required for wing development in *Drosophila*. Bioinformatics work was designed to determine the extent to which protein-coding regions have diverged over 250 million years between the two species, enabling us to make testable predictions on the relative contribution of changes in protein-coding sequences of targets of Ubx in specifying haltere in *Drosophila* as against hind wing in *Apis*.

We observe that, against the intuitive expectations, targets of Ubx that are common to both *Apis* and *Drosophila* are relatively less conserved compared to the average homology across all proteins between the two species. Interestingly, degree of conservation at protein level of *Apis*-specific and the *Drosophila*-specific targets of Ubx were not any different from the average homology across all proteins between the two species. These common targets appear to have fast evolved and thereby may have

acquired new functions in *Drosophila*. Our work suggests this as an additional factor causing the evolution of haltere in *Drosophila*.

List of figures

1. Phylogenetic tree displaying the divergence of the various insect orders
2. Ubx is sufficient to induce a wing-to-haltere transformation in Drosophila
3. Schematic view of the various targets of Ubx in *Apis* and *Drosophila*
4. A sample BLAST output
5. Frequency distribution plots for “All Genes” dataset:
    a. Parameter: % Bit/AL
    b. Parameter: % Bit/Min_len
6. Frequency distribution plots for “Wing Only” dataset, separated using Flybase summary information:
    a. Parameter: % Bit/AL
    b. Parameter: % Bit/Min_len
7. Comparison of wing development-related parameters separated using “GO Biological Process” terms
    a. Parameter: % Bit/AL
    b. Parameter: % Bit/ML
    c. Parameter: % Identities
    d. Parameter: % Positives
    e. Parameter: % Gaps

List of tables

1. Tabulation of BLAST outputs
2. Tabulation of BLAST output generated parameters for comparison of various groups
3. Mann-Whitney Rank Sum Test on data points of various data sets – with wing development-related genes separated using Flybase “GO Biological Process” information
    a. Bit/Aligned length
    b. Bit score / Minimum length of the two polypeptides
    c. %Identities in between the two polypeptides
    d. %Positives in between the two polypeptides
    e. %Gaps in between the two polypeptides

Acknowledgements

I wish to acknowledge my mentor, Dr. L. S. Shashidhara, for his unparalleled patronage, steering, tutelage, sagacity, assistance, and consistent encouragement. He has indeed been the rationale for the work performed. His meticulous insights and consummate wisdom have been definitive in the design and accomplishment of the task.

I wish to acknowledge Dr. M. S. Madhusudhan for his valuable inputs towards many critical aspects of analysis.

I wish to acknowledge Dr. Girish Ratnaparkhi, member TAC for exceptional moral and spiritual support.

I wish to acknowledge Dr. K. N. Ganesh, director of IISER Pune, who, as my faculty advisor has always inundated me with affection and immaculate guardianship.

I wish to acknowledge my fellow lab members for their cooperation and help, in particular Sh. Naveen Prasad for providing me with the initial data sets.

I wish to acknowledge the faculty and staff members at the Indian Institute of Science Education and Research who have made my association with the institute outstanding.

Abhijit

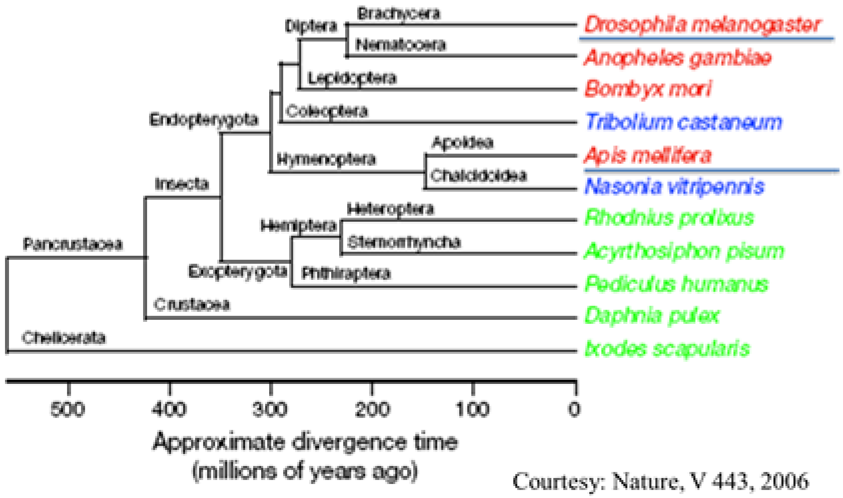
Introduction

Comparative analysis of targets of the Hox gene Ultrabithorax in Drosophila melanogaster and Apis mellifera

The anterior-posterior body plan of the developing embryo is controlled by a conserved set of genes called as the Hox genes. The Hox proteins encoded by these genes determine the segmental identity and specify the various segmental parts such as the legs or the wing. Such genes characteristically contain a conserved genetic domain called as the homeobox.

The body-plan of the dipteran *Drosophila melanogaster*, commonly called as the fruit fly consists of two wings in the second thoracic segment and two halteres in the third thoracic segment. Whereas the body-plan of the hymenopteran *Apis mellifera*, commonly called as the honeybee consists of four wings – two in the second thoracic segment and two smaller wings in the third thoracic segment. Other peculiar characteristics of most of the hymenopterans include a thin waist and their habitation in socially complex colonies. As opposed to the fly, the bee, a hymenopteran, has no halteres, which are the balancing organs functionally similar to the gyroscope, aiding in flight and navigation. Surprisingly, this contrast is made possible by the same master control gene – the *Hox* gene *Ultrabithorax* (*Ubx*). The Ubx protein controls the expression of its downstream target genes, by itself functioning as a transcription factor.

In the *Drosophila*, the Ubx is expressed in the third thoracic segment (T3) and the first abdominal segment (A1), where it suppresses the wing fate. The Ubx is known to determine its body plan by specifying the number of wing and legs.



Courtesy: Nature, V 443, 2006

Figure 1: Phylogenetic tree displaying the divergence of the various orders of insects. The *Drosophila* belongs to the order diptera, whereas the honeybee belongs to the order hymenoptera. Both these orders have diverged ~300 million years ago.

The *Hox* genes are responsible for the body-plan across diverse species from insects to mammals. In *Apis* and *Drosophila*, they differentiate the development of flight appendages in second and third thoracic segments. The function of the *Hox* protein Ubx seems conserved (unpublished data in our lab has demonstrated functional conservation of Ubx derived from diverse organisms such as *Apis* in *Drosophila* is sufficient to induce a wing-to-haltere transformation*).* This is in spite of the fact that there are significant differences in Ubx sequences amongst the various insect groups. The Ubx has been reported to operate in tandem with a plethora of signaling pathways (e.g. Wg, Dpp etc.) [7-12]. It provides it the capability to control a number of developmental processes, especially related to cellular differentiation, patterning, cellular growth and proliferation. Thus, it behaves as a “master regulator”, specifying the body plan of the insects, thereby has the capacity to influence the evolutionary processes leading to speciation. The sequencing of the *Apis* genome, having been completed recently [1], provided us with an opportunity to commence with this task.

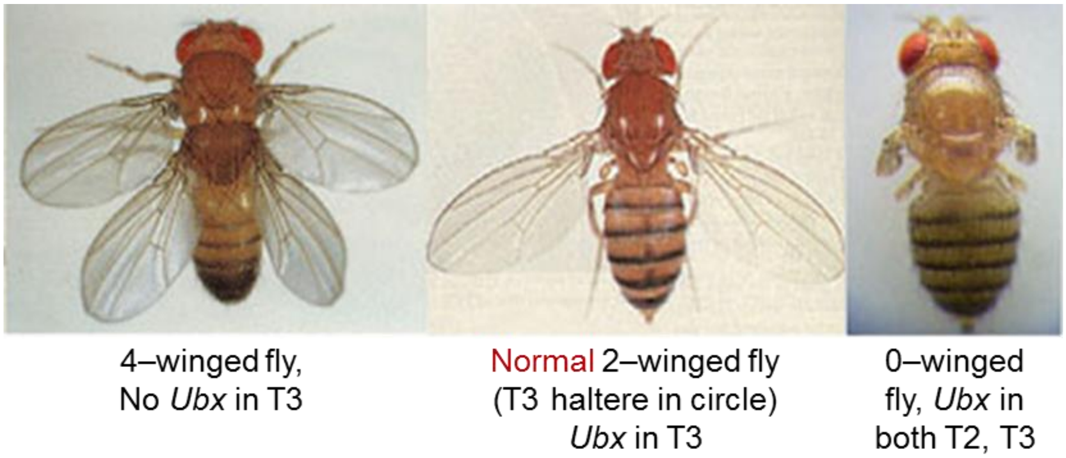


Figure 2: Lewis (1978) demonstrated that Ubx is sufficient to induce a wing-to-haltere transformation in *Drosophila*, irrespective of the thoracic segment involved (T2 or T3). Picture courtesy: Dr. Shashidhara's fruit fly portal.

The work in our laboratory in this direction involves identifying those genes that have come under the influence of Ubx specifically during dipteran evolution. It involves identifying direct targets of Ubx from different insect groups such as *Drosophila* (Agrawal et al., 2011), *Apis* and *Bombyx* (silkworm). The latter two are four-winged insects.

The current project involves comparing the genes that are targets of Ubx in *Drosophila* or *Apis* alone and that are common to both. In this project, we intend to estimate the levels of conservation between the protein-coding genes by comparing the polypeptides encoded by those genes. This, in turn, would help in appreciating the significance of the same in the evolution of the two-winged fly.

Specific objectives of the project:

1. Preparation of detailed data base outlining the identity and functional details of the genes, which are targets of Ubx in *Drosophila* and in *Apis.*

2. Documenting the respective polypeptides encoded by those genes.

3. Estimating the homology index and various other parameters indicating conservation/divergence for the polypeptides encoded by such genes regulated by Ubx in *Drosophila* or *Apis* alone, or which are common to both.

4. Estimating the average homology index of the polypeptides encoded by the genes in the three groups mentioned above.

5. Repetition of the exercise for all genes in each group as well as for different ontology groups, specifically for those genes that are directly involved in wing development.

Using suitable statistical tools and techniques to compare the polypeptides encoded by such sets of genes viz. the three groups as mentioned above, and repetition of the same for specific ontology group viz. related to wing development.

Thus, this exercise provides an estimate of relative contribution of changes in the protein sequences of wing patterning genes as also the acquisition of new target genes by Ubx in the evolution of haltere in *Drosophila melanogaster*.

Materials and methods

Experimental data for the targets of Ubx in the hind–wing of *Apis mellifera* as ascertained from a ChIP–Seq assay was available (unpublished data with Mr. Naveen Prasad who worked in our laboratory under Dr. Shashidhara), along with the ChIP–chip and ChIP–array data containing the targets of Ubx in *Drosophila melanogaster* (Pavan et. al., 2011, Choo et. al. 2011, respectively).

The target sites of Ubx binding in the genome of these insects were identified thought suitable technologies. The genes lying within a certain range (2000 base pairs upstream and downstream) of those target-binding sites were supposed to be the target genes of Ubx. Such genes were also identified via suitable bioinformatics packages.

BioMart facility of the EnsEMBL (http://metazoa.ensembl.org/biomart/martview) was utilized to provide the corresponding *Drosophila* ortholog for each *Apis* target and vice versa.

The PubMed (www.ncbi.nlm.nih.gov/pubmed/) was used to document the details of the *Apis* genes, their functions, and the details of the polypeptides encoded by them. For the fly genes, the Flybase portal (www.flybase.org) was used.

Overall, three groups of Ubx targets were distinctively discernible:

1. The targets that were specific to *Apis* hind–wing: These targets were those that were targeted by the Ubx only in *Apis* and not in *Drosophila.* However, owing to extraordinary conservation of the genetic information across species, most of these *Apis*–specific targets had a corresponding *Drosophila* ortholog: which just was not located within an influential range of the binding site of the Ubx in flies.

2. The targets that were specific to *Drosophila* and were not targeted in *Apis.*

3. Those targets those were common to both – *Apis* as well as *Drosophila.* Such targets were under the influence of the Ubx in both the insects.

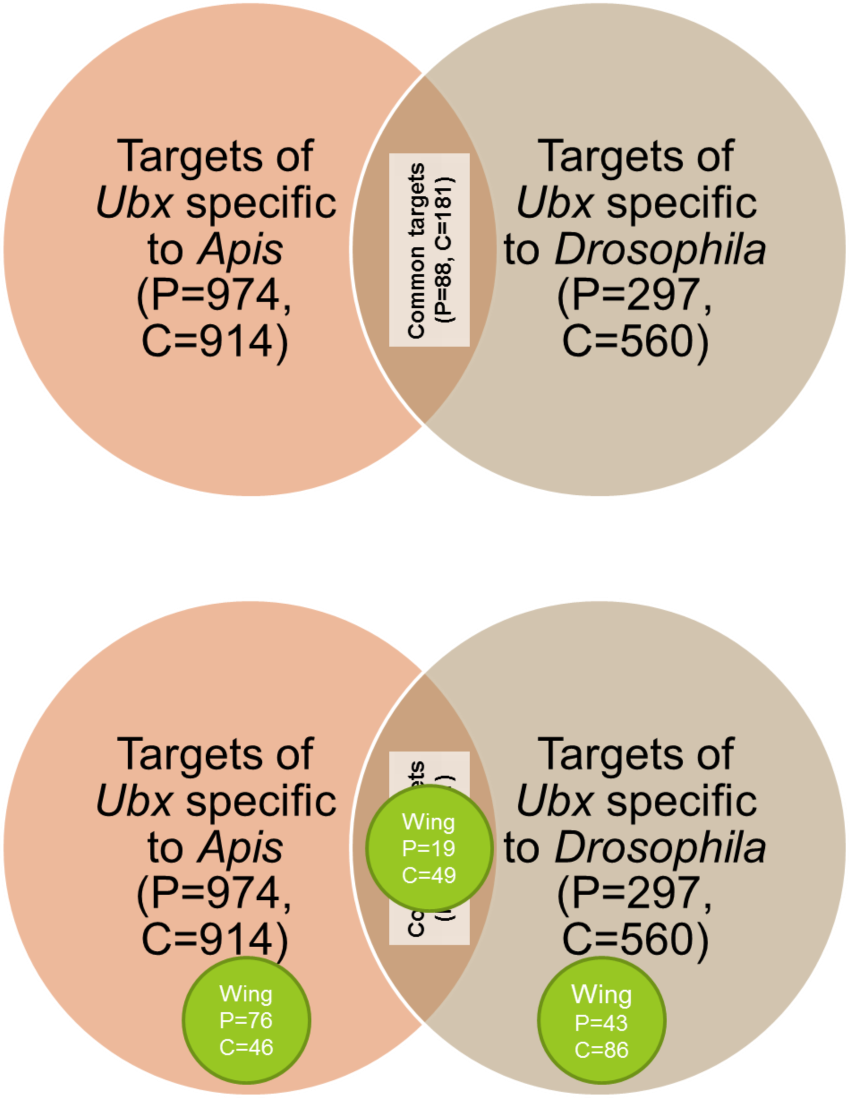

Targets of *Ubx* specific to *Apis* (P=974, C=914)
Common targets (P=88, C=181)
Targets of *Ubx* specific to *Drosophila* (P=297, C=560)
Targets of *Ubx* specific to *Apis* (P=974, C=914)
Targets of *Ubx* specific to *Drosophila* (P=297, C=560)
Wing P=19 C=49
Wing P=76 C=46
Wing P=43 C=86

Figure 3: Venn diagram depicting the various targets of Ubx in *Apis* and *Drosophila*. The subsets of genes known to be involved in wing development have been depicted in green circles. The numbers in brackets denote the sample size. P denotes the comparison between the *Drosophila* targets identified by Pavan et al. (2011) and the *Apis* targets identified by Naveen (unpublished data in our lab). C denotes the comparison between the *Drosophila* targets identified by Choo et al. (2011) and the *Apis* targets identified by Naveen (unpublished data in our lab).

The information pertaining to the targets of Ubx in *Drosophila* comes from two sources: Pavan et al. (2011) and Choo et al. (2011), and similar information pertaining to the targets of Ubx in *Apis* comes from an unpublished data in our lab. All of these were obtained through high-throughput experiments:

1. Chip-seq for *Apis* targets (by Sh. Naveen Prasad, who worked in our lab at IISER Pune),

2. Chip-chip for *Drosophila* targets (by Pavan in our lab at IISER Pune), and,

3. Chip-array for *Drosophila* targets (by Choo in the lab of Dr. White in the University of Cambridge).

Thus, **six data sets were provided,** which were in turn created from comparison of the data from three data sets as outlined above:

A. Pavan (*Drosophila*) vs. Naveen (*Apis*):

   1. *Drosophila*–specific targets of Ubx (from Pavan et al.-reported data) and their corresponding *Apis* orthologs (which are not targeted by Ubx in bee, as ascertained from Naveen-reported targets of Ubx in bee)

   2. *Apis*–specific targets of Ubx (from Naveen-reported targets of Ubx in bee) and their corresponding *Drosophila* orthologs (which were not targeted by Ubx as per Pavan et al.-reported fly data)

   3. Targets of Ubx common to both – *Drosophila* (listed in Pavan et al.-reported data) as well as their corresponding *Apis* ortholog (listed in Naveen-reported targets of Ubx in bee)

B. Choo (*Drosophila*) vs. Naveen (*Apis*): In the similar fashion as above, following data sets were provided using comparison of Choo et al.-reported fly data and Naveen-reported targets of Ubx in bee:

   1. *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs

2. *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs
3. Targets of Ubx common to both – *Drosophila* as well as *Apis*

The data sets mentioned above contained the Beebase IDs and the Flybase IDs of the genes.

Preparation of the updated database and determination of the functionality of the given targets:

The functional description of the genes belonging to *Drosophila* is vast and nearly exhaustive, it being the choice of the geneticists for decades. The *Apis* genome sequence having been completed only in 2006 (*Nature, 2006*), and it being a relatively less popular model, the information pertaining to the functionality of genes thereof is limited. Thus, the functional description of the corresponding ortholog of *Drosophila* is taken as the standard for determining the functionality of the *Apis* targets.

The progression of the work towards creation of the database was done in various stages, owing to the complexity and manual updating of the honey bee data available at hand, most of which had gone obsolete due to progression in the honeybee assembly by the Hymenoptera Sequencing Consortium. The culmination of each stage was symbolized the by creation of a new version of the data sets.

Stage one (Version 1):

The name, symbol, synonyms, and functional description of the *Drosophila*-specific IDs were penned–down using the Flybase portal (www.flybase.org), along–with the proteins encoded by them. Updates for the information too were performed in case-by-case basis.

Challenges:

The prime challenge was posed by the fact that a multitude of the entries mentioned in the original data sets had become obsolete. Moreover, when a single obsolete ID is replaced by several new IDs each with a distinctive function, then the functionality of the obsolete ID needs to be manually matched with the currently available options most suitable and fitting to the case.

The functionality information was gathered initially in the form of “GO Biological Process” but was subsequently replaced by the Flybase summary information – the latter being more exhaustive and detailed.

The gene name, gene symbol, and the gene synonyms too were codified from the Flybase.

Genes regulating the wing development were identified using the functionality information as revealed by the Flybase summary description and these genes along with their *Apis* counterparts were listed in separate data sets, suffixed as “Wing_Only”.

Thus, total of 24 (6X2 + 6X2) data sets were produced in this stage.

Stage two (Version 2):

Identification and documentation of appropriate polypeptide sequences encoded by the *Drosophila* genes:

In this stage, the polypeptides encoded by the corresponding genes were identified and stored in the form of individual text files.

This information towards the fly genes was collected from the Flybase.

Challenges – The following challenges were encountered:

1. A multitude of fly genes encoded for more than one polypeptide: In such cases, the longest polypeptide sequence was taken to ensure maximum “coverage” in BLAST.

2. In many cases, no single (“unique”) longest sequence was available: In such cases, the most recent amongst all sequences of longest length was taken.

3. In several cases, no polypeptide information was available, mostly because the genes happened to be non–protein coding genes. For the non–protein coding genes, their corresponding transcriptions were separately documented.

Stage three (Version 3):

In this stage, the name, symbol, and functional description of the *Apis* genes were documented from the PubMed, along with their polypeptide sequences.

Challenges – The foremost challenge was that about a third of the *Apis* genes were inconspicuous in the PubMed, having become obsolete, And unlike the "automatic ID updater" feature available in the Flybase, no similar feature is available for the bee, which could update the bee IDs in one go.

Moreover, inconsistencies in the data originally provided too were noticed, most conspicuous being the various "gaps". The gene functionality descriptions for *Apis* and *Drosophila* too did not match in a multitude of cases.

While documenting the polypeptide sequences of the proteins encoded by the bee genes available at PubMed, the issue similar to the one encountered for *Drosophila* posed a challenge: that a multitude of bee genes encoded for more than one polypeptide. In such cases too, the longest polypeptides were documented.

Stage four (Version 4):

Inconsistencies were found between the original data sets and the then current data from the BioMart due to updates in the records and coming-up of newer versions of bee genome assemblies.

Challenges – Bioinformatics, as we know, happens to be a continually evolving science. Upon close investigation and scrutiny, it was identified that there were certain errors and inconsistencies in the original data set supplied with respect to the then current data available with the BioMart. Moreover, there were certain gaps and some mismatches in functionality information between *Apis* and *Drosophila* genes and irreproducibility of certain information from other methods. All of these could be attributed to the consistent updates and progressions in the field of bioinformatics. The original data set supplied had been generated upon automated online conversion of Flybase IDs to their corresponding bee counterparts through the EnsEMBL / BioMart portal. Moreover, those Flybase IDs themselves were generated upon automated conversion of the original *Apis* IDs as revealed by the experimental data using the same portal. It is understood that upon regeneration of the ortholog information, a host of gene IDs are produced which may not have been a part of the original data set.

Thus, fresh data sets listing only the targets of Ubx specifically to *Apis*, *Drosophila* and common to both were provided to be built–upon for creating new data sets.

Such new data that consisted only of the Ubx targets in the original experiment was used to generate corresponding ortholog information and thus a new "master data set" was generated, containing all the relevant information pertaining to bee genes and their corresponding fly orthologs. The entries mentioned therein were searched upon the various (12) data sets in a one–by–one fashion, and the unmatched / extra data was pruned to create reproducible information.

Stage five (Version 5):

Manual search was done for the new genes for the discontinued *Apis* gene IDs documented in the original data set, along with their corresponding *Drosophila* orthologs.

Challenges – A great share of the original data that had become obsolete required manual updates. For the discontinued *Apis* gene IDs, the polypeptide sequences encoded by the discontinued IDs were documented from EnsEMBL, which was found to be hosting those discontinued polypeptides itself being running the Assembly 3.0 Amel (as opposed to PubMed which was running on an updated Assembly 4.5 Amel).

Such polypeptide sequences of discontinued bee gene IDs (extracted from EnsEMBL) were then utilized in BLAST operations as against the whole–RefSeq protein database of *Apis mellifera* to identify the then current proteins and thereby the genes from PubMed corresponding to each of the discontinued ones.

The nucleotide sequences of such “new” genes which came upon the discontinuation of the old ones too were documented, and using suitable bioinformatics tools, it was identified as to which of them actually corresponded to the Ubx binding site (as ascertained by Naveen’s experiment) on the actual bee genome.

Those genes that happened to be within ±2000bp of the actual binding site were taken to be the true representatives of the discontinued genes.

Subsequently, the *Drosophila melanogaster* orthologs of these new *Apis* gene IDs were searched for, and the exercise of finding the gene symbol, name, synonyms, functions, summary information and polypeptides encoded by them was performed. Such information was documented and appropriately inserted into the data sets.

In a multitude of instances, the BioMart portal of EnsEMBL could not be used to get the corresponding ortholog of the given *Drosophila* genes, due to non-availability of information.

Hence, for such cases, the PubMed database was used to generate the corresponding *Apis* orthologs for the given *Drosophila* genes. It was performed in two ways:

1. First, via performing protein BLAST of the polypeptide sequences encoded by the *Drosophila* genes against the RefSeq protein database of *Apis mellifera*, and,

2. Second, via performing the protein-protein BLAST of the discontinued polypeptide sequence of the corresponding discontinued ortholog of the fly (i.e. *Apis*) as against the then current RefSeq protein database of the *Apis* proteome available with the NCBI. The required data was available with EnsEMBL since it was running on a previous bee assembly.

Thus, two *Drosophila*-specific data sets were produced: one produced through the first way has been named in the usual fashion, whereas the one produced via the second way has been named by prefixing the term “New” to its name and by labeling its version as “5.2” in place of “5” in the database created for analytical purpose by us.

Preparation and need of “background” distribution:

A random-number generator was used to pick out 1000 pairs of orthologs of *Apis* and *Drosophila*, and the task of documenting their corresponding longest polypeptides was performed similar to the operation done before for the various data sets.

The objective of the same was to have a benchmark distribution against which the varied data sets could be compared. For, these randomly picked-up pairs could be compared with the Ubx targets and ontology-specific Ubx targets to calculate their relative deviation from a random distribution of pairs.

The significance of background distribution lies in it being a random distribution of polypeptides encoded by gene orthologs of *Apis* and *Drosophila*, irrespective of the fact if the same are targets of Ubx or not. Thus, it was expected to highlight the influence of Ubx-influenced genes as against randomly picked-up gene pairs.

The comparison of the pairs in this background population with the targets of Ubx as well as its subset having role in wing development was performed statistically to express numerically the similarities and differences as well as their statistical significance.

The test of significance is important in the sense that it identifies if the purported similarities or differences could occur due to accident or chance; thereby establishing if the claimed result is indeed significant or not.

Thereupon, three more background distributions each with 1000 pairs of genes and polypeptides were prepared in a similar fashion to test them further for their robustness. The same were added to the list of the following 16 data sets for further operations:

All Genes: Pavan (*Drosophila*) vs. Naveen (*Apis*):

1. *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
    a. Version 5
    b. Version 5.2 (as discussed before)
2. *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs
3. Targets of Ubx common to both – *Drosophila* as well as *Apis*

All Genes: Choo (*Drosophila*) vs. Naveen (*Apis*):

1. *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
    a. Version 5
    b. Version 5.2 (as discussed before)
2. *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs
3. Targets of Ubx common to both – *Drosophila* as well as *Apis*

Wing Only – Those pairs relevant to wing formation: Pavan (*Drosophila*) vs. Naveen (*Apis*):

1. *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
    a. Version 5
    b. Version 5.2 (as discussed before)
2. *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs
3. Targets of Ubx common to both – *Drosophila* as well as *Apis*

Wing Only – Those pairs relevant to wing formation: Choo (*Drosophila*) vs. Naveen (*Apis*):

1. *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
   a. Version 5
   b. Version 5.2 (as discussed before)
2. *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs
3. Targets of Ubx common to both – *Drosophila* as well as *Apis*

In addition to the 16 data sets mentioned above, certain additional combinations of data sets were made and plotted, but they have not been included in the statistical analysis in the current work since the information contained within them was diluted. Nevertheless, they deserve mentioning. They were the following:

Wing Only – *Apis* vs. *Drosophila*: viz. *Apis* + Common vs. *Drosophila* + Common:

Pavan (*Drosophila*) vs. Naveen (*Apis*):

1. Common + *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
2. Common + *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs

Choo (*Drosophila*) vs. Naveen (*Apis*):

1. Common + *Drosophila*–specific targets of Ubx and their corresponding *Apis* orthologs
2. Common + *Apis*–specific targets of Ubx and their corresponding *Drosophila* orthologs

Documentation of pairwise BLAST outputs for the pairs of polypeptides:

All the relevant information towards the aforementioned task had been documented, viz. recording the corresponding polypeptide sequences IDs, the complete sequences in text files, the length of the sequences, etc.

Thereupon the NCBI Blast gateway was used for documenting the scores of the BLAST outputs for each of the pairs in the various data sets and the sets of background distribution. There were four sets of backgrounds each having 1000 pairs of polypeptides.

The “default” parameters were selected within the “compare two polypeptide sequences” feature.

In BLAST, the first sequence is called as the “Query” sequence, whereas the second one is called as the “Subject” sequence. The “Subject” is aligned as against the “Query”, and relevant output parameters are provided.

Hence, if the subject and query sequences are interchanged, then the output and the final alignment too are prone to significant changes.

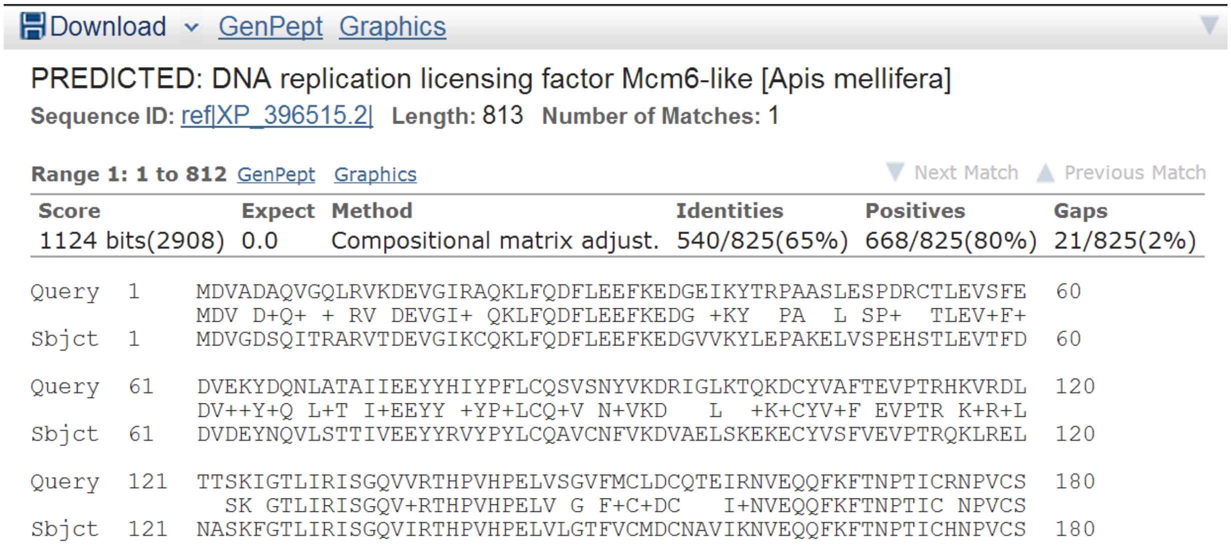
Therefore, for every polypeptide pair, two individual BLAST operations were performed by interchanging the “Query” and “Subject” sequences. The corresponding outputs from both the operations for every pairs were penned-down. The same was performed for all the 16 data sets and for the 4 background distributions.

The following nomenclature was assigned in the data sheets, which were separately prepared for documenting BLAST scores:

1. When *Apis* polypeptide is the “Query” and *Drosophila* polypeptide is the “Subject”, we called it as “forward BLAST” → Scores were tabulated.

2. When *Drosophila* polypeptide is the “Query” and *Apis* polypeptide is the “Subject”, we called it as the “reverse BLAST” → Scores were tabulated.

Thereupon, all the parameters of the output were averaged for the “forward” and “reverse” operations to come up with final averaged score for each pair of polypeptides in the data sets and the background.



PREDICTED: DNA replication licensing factor Mcm6-like [Apis mellifera]
Sequence ID: ref|XP_396515.2| **Length:** 813 **Number of Matches:** 1

```
Score            Expect  Method                        Identities      Positives       Gaps
1124 bits(2908)  0.0     Compositional matrix adjust.  540/825(65%)    668/825(80%)    21/825(2%)

Query  1    MDVADAQVGQLRVKDEVGIRAQKLFQDFLEEFKEDGEIKYTRPAASLESPDRCTLEVSFE  60
            MDV D+Q+ + RV DEVGI+ QKLFQDFLEEFKEDG +KY  PA  L SP+  TLEV+F+
Sbjct  1    MDVGDSQITRARVTDEVGIKCQKLFQDFLEEFKEDGVVKYLEPAKELVSPEHSTLEVTFD  60

Query  61   DVEKYDQNLATAIIEEYYHIYPFLCQSVSNYVKDRIGLKTQKDCYVAFTEVPTRHKVRDL  120
            DV++Y+Q L+T I+EEYY +YP+LCQ+V N+VKD   L  +K+CYV+F EVPTR K+R+L
Sbjct  61   DVDEYNQVLSTTIVEEYYRVYPYLCQAVCNFVKDVAELSKEKECYVSFVEVPTRQKLREL  120

Query  121  TTSKIGTLIRISGQVVRTHPVHPELVSGVFMCLDCQTEIRNVEQQFKFTNPTICRNPVCS  180
              SK GTLIRISGQV+RTHPVHPELV G F+C+DC   I+NVEQQFKFTNPTIC NPVCS
Sbjct  121  NASKFGTLIRISGQVIRTHPVHPELVLGTFVCMDCNAVIKNVEQQFKFTNPTICHNPVCS  180
```

Figure 4: A sample BLAST output

Thus, in the end the following entities were available for analysis for every singular pair of polypeptide for every data set and background distribution:

1. Scores:
a. “Bit” score, e.g. “1124 bits” in the figure above, and,
b. “Raw” score (the entity that follows the “Bit” score in brackets, e.g. “2908” in the figure above.
2. The e-value (expectation value, which is the most likely value of a random variable), e.g. “0.0”
3. The identities, in absolute terms e.g. “540” for 540/825
4. % Identities, in terms of %age e.g. 65%
5. The positives, in absolute terms e.g. “668” for 668/825
6. % Positives, in terms of %age e.g. 81%
7. The gaps, in absolute terms e.g. “21” for 21/825

8. % Gaps, in terms of %age e.g. 3%
9. Aligned length, e.g. “825” i.e. the total aligned positions as computed by BLAST
10. Minimal length: The sequence length of the smaller of the two polypeptides too was documented to ascertain the degrees of similarities / differences in terms of the smallest sequence.

Further, to quantify the degree of similarities / differences, the following parameters were used:

1. Bit score/Aligned length: To assess the bit score per unit length of aligned positions. Here bit score represents the average of the “forward” and the “reverse” bit scores. Similarly, the aligned length too is the average of the “forward” and “reverse” aligned lengths.
2. Bit score/Minimal length: To assess the bit score per unit length of the smallest polypeptide among the two.
3. Raw score/Aligned length: To assess the “raw” score per unit length of aligned positions.
4. Raw score/Minimal length: To assess the “raw” score in terms of the smallest polypeptide.
5. % Bit score/Aligned length for the data sets divided in suitable intervals: To assess the bit score per unit length of aligned positions, expressed in terms of relative frequency.
6. % Bit score/Minimal length for the data sets divided in suitable intervals: To assess the bit score in terms of the smallest polypeptide, expressed in terms of relative frequency.
7. % raw score/Aligned length for the data sets divided in suitable intervals: To assess the “raw” score per unit length of aligned positions, expressed in terms of relative frequency.
8. % raw score/Minimal length for the data sets divided in suitable intervals: To

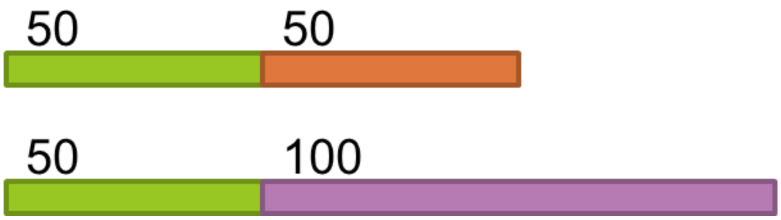
assess the “raw” score in terms of the smallest polypeptide, expressed in terms of relative frequency.
9. %P: %Positives
10. %I: %Identities
11. %G: %Gaps

The last three quantities are provided by BLAST itself, and are in terms of the percentage of absolute number of positions as against the aligned length.

The bit score and the “raw” score are the final scores that take into account the variable positive scores for identities and positives using the substitution matrix (BLOSUM62), as well as the penalty for gaps or differences.

Thus, we had 11 parameters for each of the 16 data sets and the 4 background distributions to be analyzed.

Rationale behind using Bit score/Minimal length:



As depicted in the figure above, consider a case wherein two polypeptides of varying lengths (viz. 100 and 150 amino acids respectively) are to be compared; with the first 50 amino acid positions of the two being identical and the rest having diverged. The BLAST shall align the first 50 identical positions, and hence, using the parameter bit score/aligned length, we might land up with the result stating that they are 100% identical (50/50), since the aligned length shall only be the length of first 50 identical positions. Nevertheless, using the parameter bit score/minimal length, we get a more realistic estimation of 50% identity (minimum of 100 and 150 is 100). Thus, for ensuring verity for those pairs of polypeptides wherein some protein motifs have remained identical and other positions have diverged, the latter method of comparison would prove more useful and apt.

Statistical Analysis

Purpose of statistical tests:

The statistical tests were deployed for two purposes:

To ascertain the degree of similarity / differences between the plotted charts of frequency distributions for various data sets and background distributions.

To ascertain the degree of similarity / differences between the various testable parameters of the data itself for the various data sets and background distributions.

For the first purpose, Chi-square test was found to be the most appropriate. The other tests were ANOVA, Correlation and Wilcoxon signed rank test.

For the second purpose, only the Mann-Whitney Rank Sum Test (on SigmaPlot) was found to be most appropriate. Owing to differences in sample sizes and lack of normality for most of the data sets and the background distributions, other tests were not found suitable for the purpose.

The p-value for significance for the tests was set at <0.05.

Results:

Hereby, to qualify the observed variations, we present the collated results obtained upon analyzing the various parameters and performing the statistical tests. Since the default value for significance for Mann-Whitney Rank Sum Test was taken to be **p < 0.05**, the data set comparisons yielding the p-values lower than 0.05 were taken to be statistically significantly different to be a part of the population compared against which it was compared.

Salient observations: The result for all the comparisons follows below. The salient observations could be summarized as:

- The overall rates of evolution of polypeptides encoded by the genes in the various groups were compared as against each other as well as the background.
- The background that consists of randomly picked-up pairs of orthologs was taken to be a representative of overall mean rate of evolution of polypeptides between *Apis* and *Drosophila*.
- The various groups of data sets represent the genes that have been targeted by the Ubx since last 300 million years. The rate of evolution of the polypeptides encoded by such genes in every group was compared as against the other organism-specific or ontology-specific groups or the background to determine at what **comparative rate** were these polypeptides evolving (viz. faster or slower rate) against the various groups.

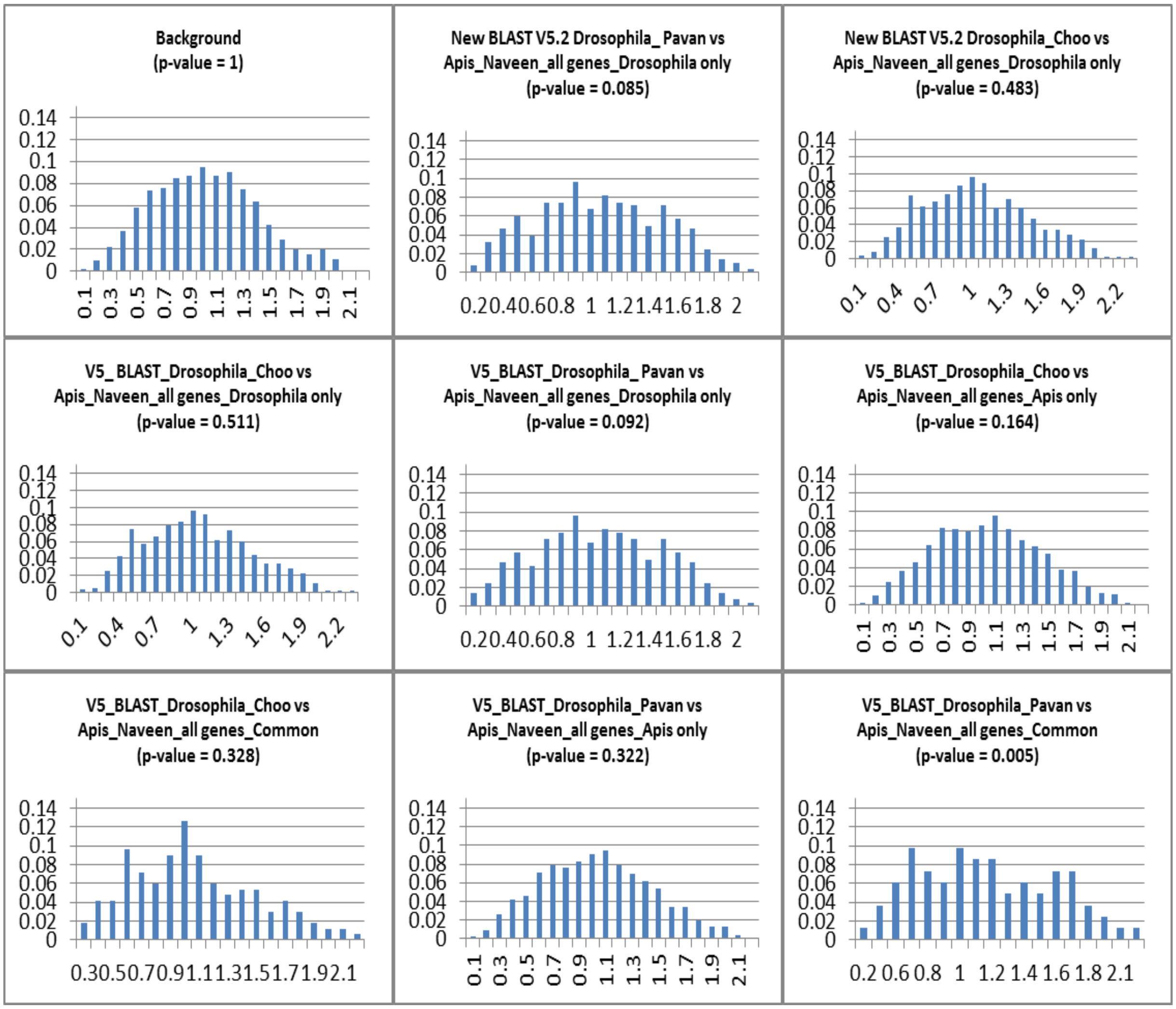
We notice that irrespective of the parameter deployed for comparison, the polypeptides encoded by the wing development-related genes that have been commonly targeted by the Ubx in *Apis* and *Drosophila* have been evolving faster, hence they must have diverged during the course of evolution, with the dipteran targets acquiring additional functions. On the contrary, the polypeptides encoded by wing development-related organism-specific genes are evolving slowly; hence, they have remained more conserved during the course of evolution.

Frequency distribution plots for “All Genes” dataset: Parameter: % Bit/AL

p-value of Mann-Whitney test against the background is in brackets.

We observe that **the distribution of “Common targets” of Ubx with differ respect to the background & other groups.**

The targets of Ubx common to fly (using Pavan et al., 2011 data) and bee show statistically significant variation in distribution using Mann-Whitney Rank Sum Test (p-value = 0.005). This observation suggests that the polypeptides encoded by the genes commonly targeted by Ubx in bee and fly are evolving faster than the genes in the other groups or the background.



Background (p-value = 1)

New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only (p-value = 0.085)

New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only (p-value = 0.483)

V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only (p-value = 0.511)

V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only (p-value = 0.092)

V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only (p-value = 0.164)

V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common (p-value = 0.328)

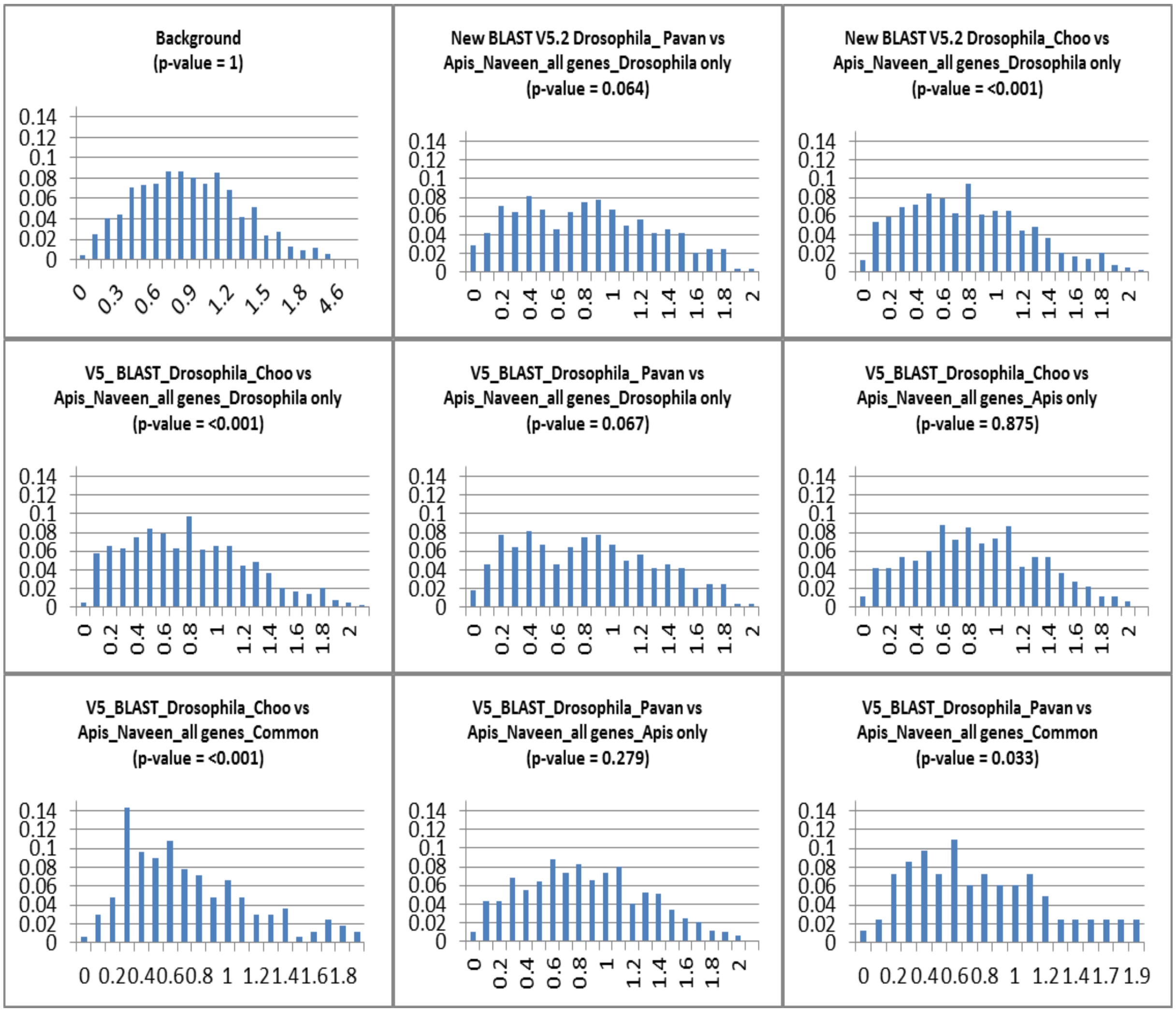
V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only (p-value = 0.322)

V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common (p-value = 0.005)

Frequency distribution plots for “All Genes” dataset: Parameter: % Bit/Min_len

p-value of Mann-Whitney test against the background is in brackets.

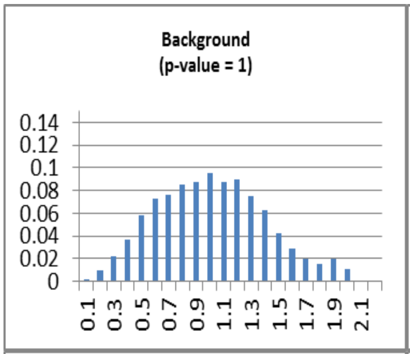
Statistically significant differences in the population distribution for the “Common targets” against background using data reported by both – Choo (2011) and Pavan (2011) are noticed along with distributional skewness towards lesser conservation indicate faster evolution of these groups against the background. Interestingly, the Choo-reported (2011) *Drosophila*-specific groups too differ from the background population (p < 0.001) using this parameter.

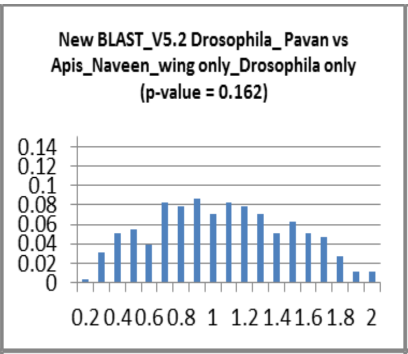


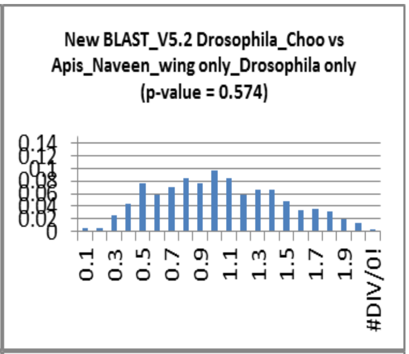
Wing development-related groups separated using flybase summary information: Frequency distribution plots for "Wing Only" dataset: Parameter: % Bit/AL

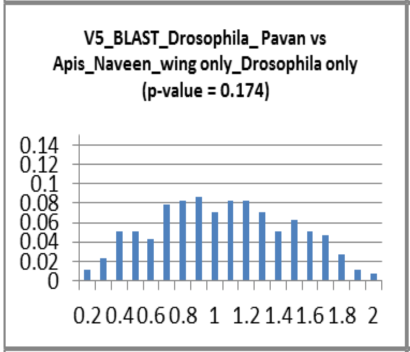
p-value of Mann-Whitney test against the background is in brackets.

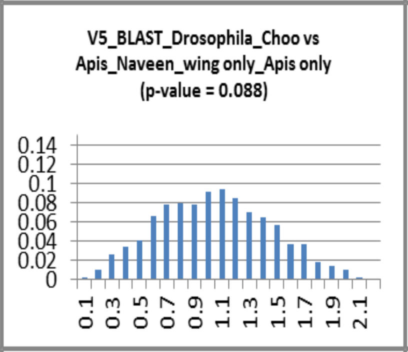
Statistically significant variation in distribution (Mann-Whitney Rank Sum Test p-value = 0.004) for the "Common targets" common to *Apis* and *Drosophila* (using Pavan et al., 2011) is noticed, suggesting that the polypeptides encoded by the genes in this group (commonly targeted by Ubx in bee and fly) are evolving faster than the genes in the background population.

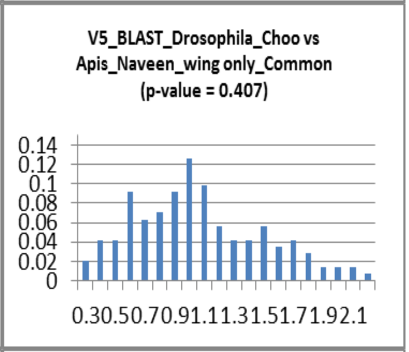

Background
(p-value = 1)

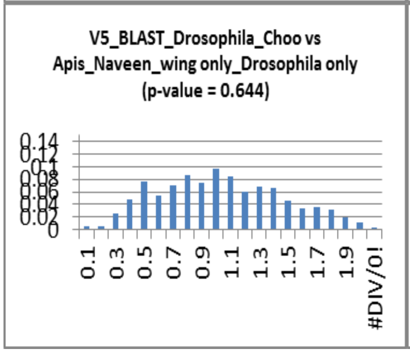

New BLAST_V5.2 Drosophila_ Pavan vs Apis_Naveen_wing only_Drosophila only
(p-value = 0.162)

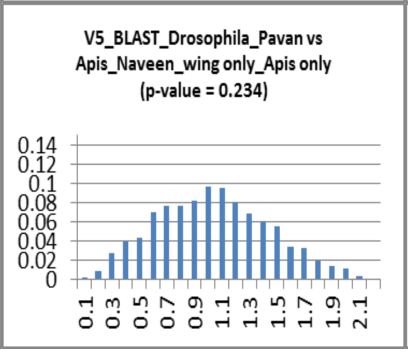

New BLAST_V5.2 Drosophila_Choo vs Apis_Naveen_wing only_Drosophila only
(p-value = 0.574)

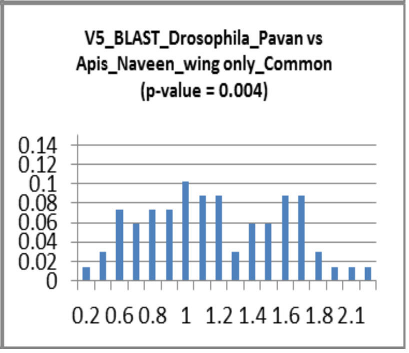

V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_wing only_Drosophila only
(p-value = 0.174)


V5_BLAST_Drosophila_Choo vs Apis_Naveen_wing only_Apis only
(p-value = 0.088)


V5_BLAST_Drosophila_Choo vs Apis_Naveen_wing only_Common
(p-value = 0.407)


V5_BLAST_Drosophila_Choo vs Apis_Naveen_wing only_Drosophila only
(p-value = 0.644)

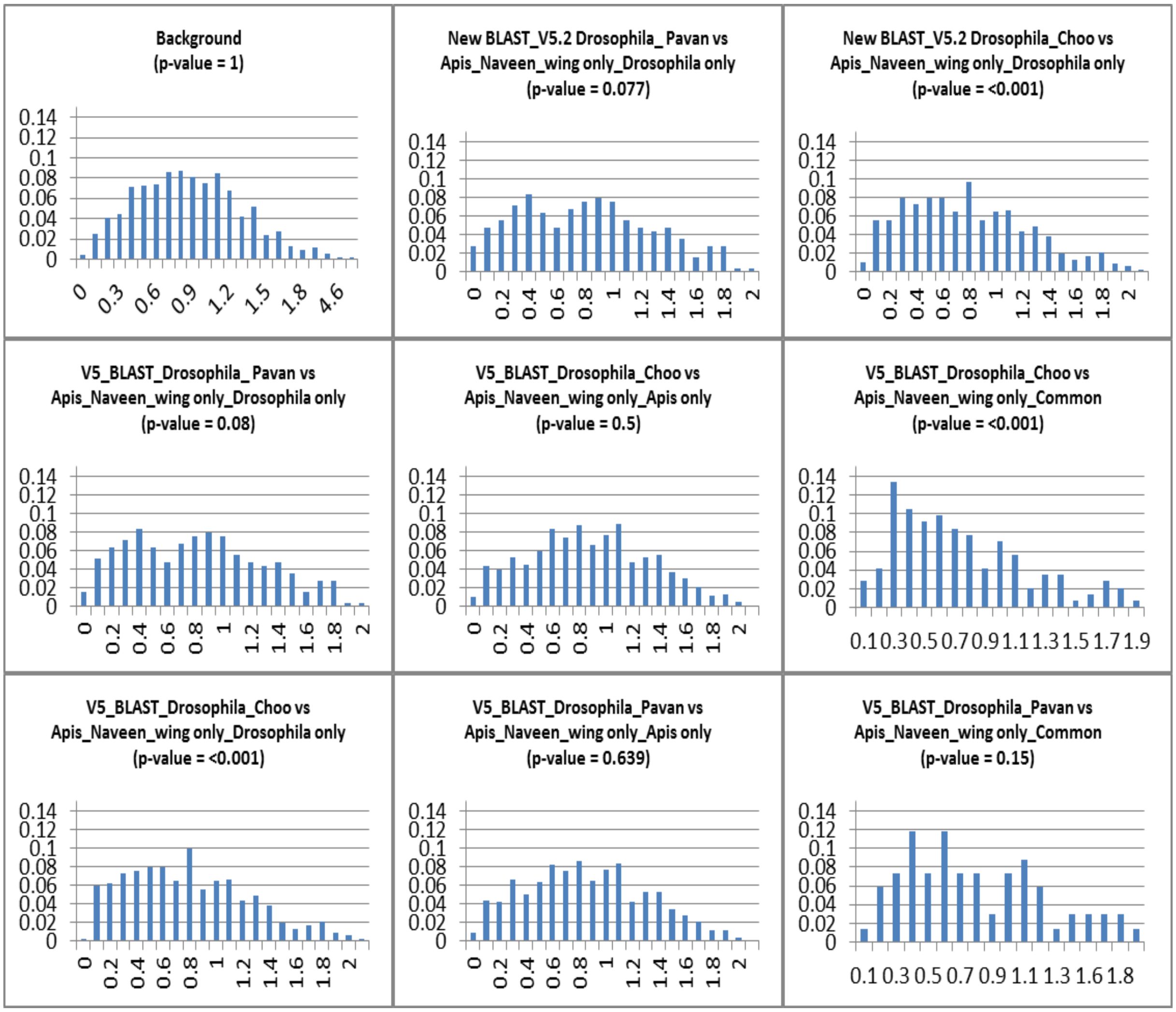

V5_BLAST_Drosophila_Pavan vs Apis_Naveen_wing only_Apis only
(p-value = 0.234)


V5_BLAST_Drosophila_Pavan vs Apis_Naveen_wing only_Common
(p-value = 0.004)

Frequency distribution plots for wing development-related (“Wing Only”) dataset separated using Flybase gene summary information: Parameter: % Bit/Min_len

p-value of Mann-Whitney test against the background is in brackets.

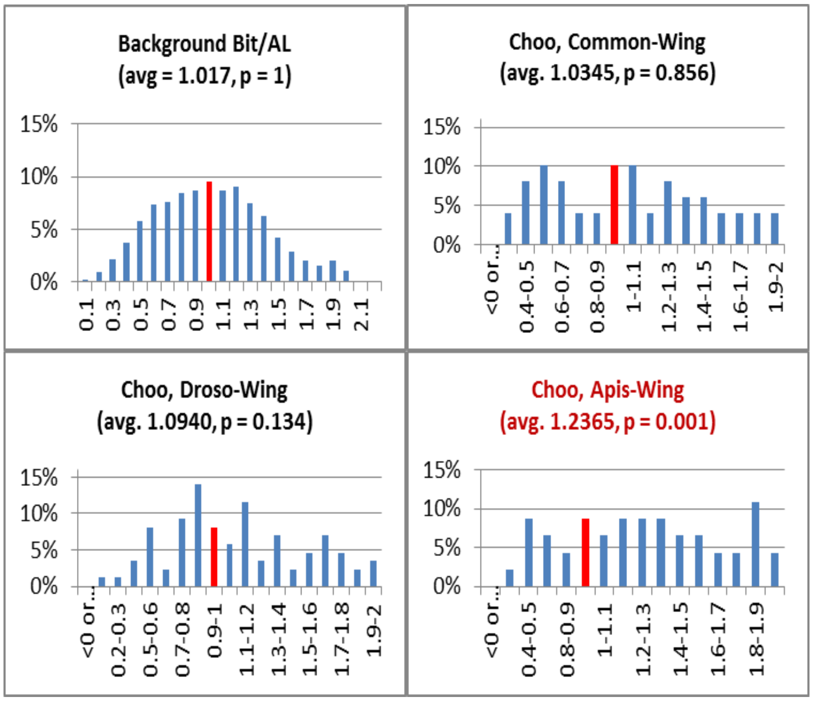
Using the parameter %Bit score/Minimal length of the polypeptides, the wing development-related *Drosophila*-specific and common targets of Ubx (using Choo et al., 2011 data) appear to be evolving faster, showing statistically significant variation (p < 0.001 in every case) in distribution against the background population.



Comparison of wing development-related groups separated using “GO Biological Process” information:

The genes targeted by Ubx in *Apis* and not in *Drosophila* (using data reported by Choo et al., 2011) that are known to be involved in wing development process appear slowly evolving (showing higher conservation) as compared to the background population distribution. Interestingly, using the parameter Bit score/Aligned length, the wing development-related common targets of Ubx are rapidly evolving as compared to its organism-specific counterparts (bee- or fly-specific groups), but appear slowly-evolving when compared with the background population distribution.

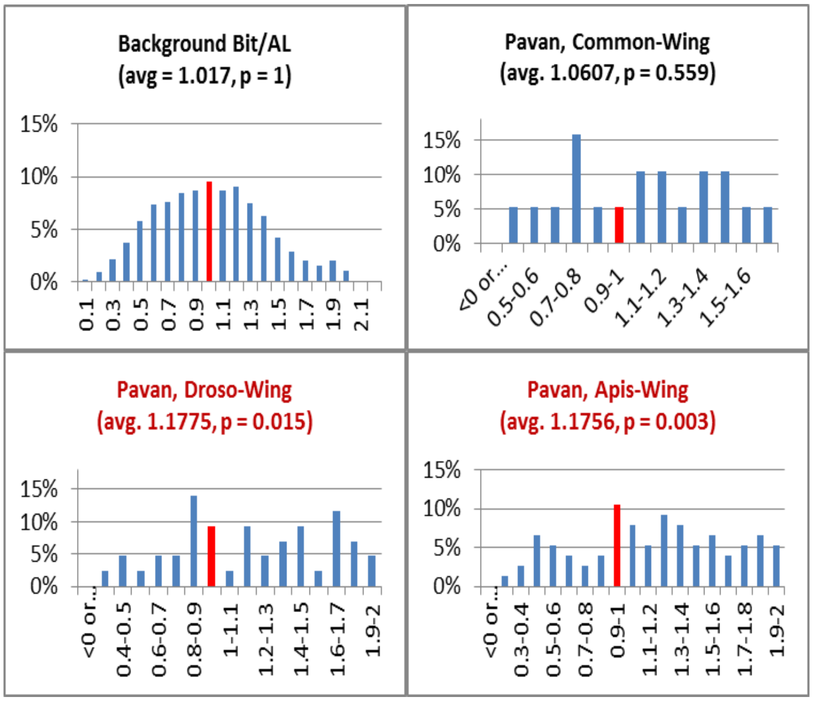
Choo: Bit score / Aligned length



Pavan: Bit score / Aligned length

Statistically significant differences in population distribution as compared with the background are observed for the *Apis*-specific and the *Drosophila*-specific wing-development related targets of Ubx using the data for fly targets from Pavan et al, 2011.

These groups have average scores higher than any other group, followed by concentration of frequency towards the higher conservation scores (right-end), indicating slower evolution of this group. Interestingly, the genes that are commonly targeted by Ubx in bee and fly scores least in this parameter (apart from the background) indicating that amongst the gene groups that are targets of Ubx, those commonly targeted in fly and bee are relatively faster evolving than the other groups.



Background Bit/AL
(avg = 1.017, p = 1)

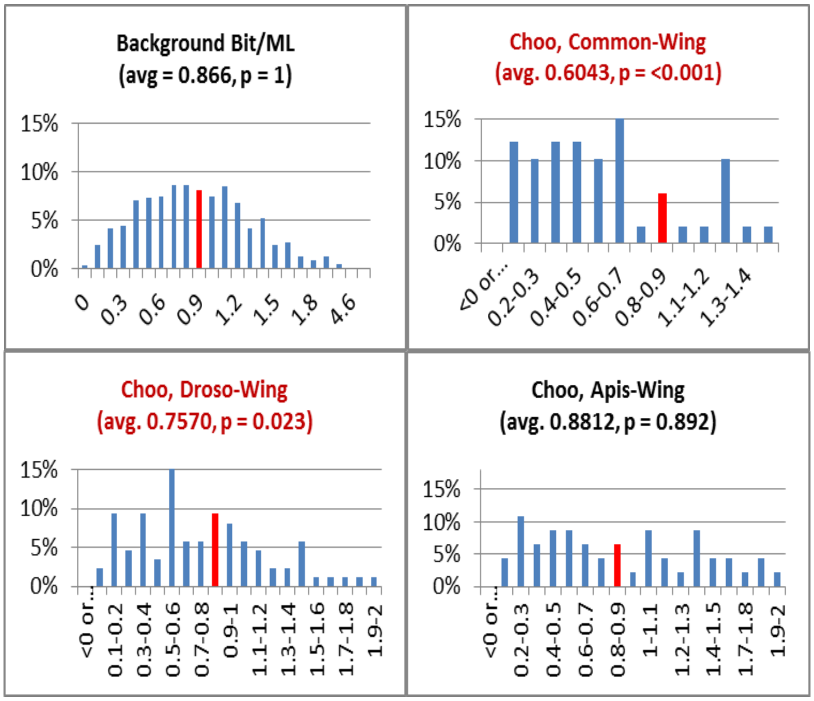
Pavan, Common-Wing
(avg. 1.0607, p = 0.559)

Pavan, Droso-Wing
(avg. 1.1775, p = 0.015)

Pavan, Apis-Wing
(avg. 1.1756, p = 0.003)

Choo: Bit score / Minimal length

Distributional skewness towards the scores of lesser conservation (hence more divergence) are seen for the “Common targets”, which lack tailing and have the least overall average scores as compared with any other group. This indicates that using this parameter, i.e. Bit score/Minimal length of the polypeptides, the wing development-related common targets are rapidly evolving when compared against any other group. Interestingly, the frequency bars are too stacked towards the lower conservation side, with the maximum score ending at 1.4 (as against 2 and 4.6 for fly/bee specific targets or background, respectively). Statistically significant differences in distribution of population for common targets and *Drosophila*-specific targets is notices (p < 0.001 and p = 0.023 respectively). However, the wing development-related *Drosophila*-specific group has overall average score are not as low as the wing development-related common targets.

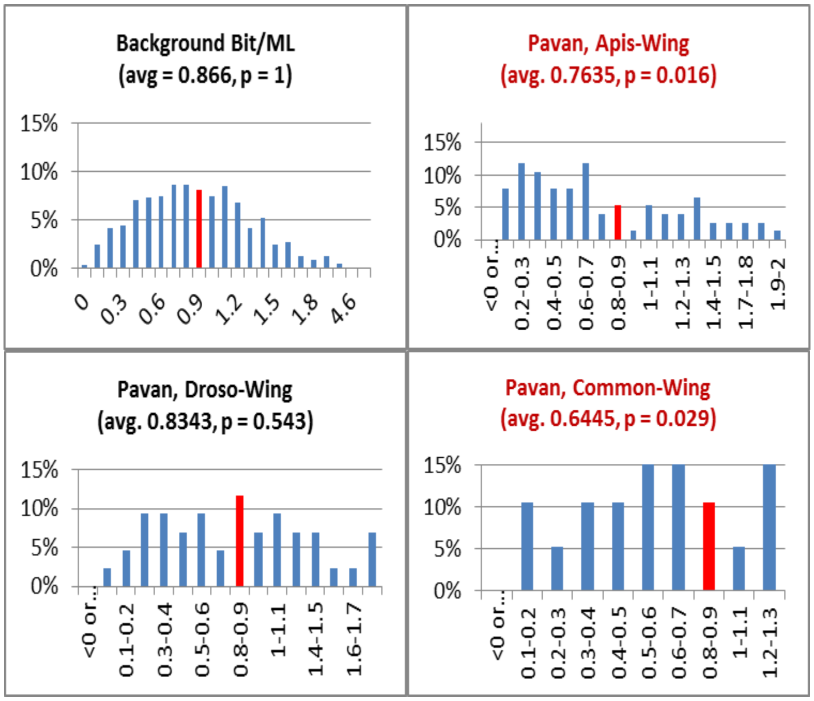


Pavan: Bit score / Minimal length

Only two groups (related to wing development) are statistically significantly different when compared with the background distribution of the population, viz. the targets common to fly and bee and the bee-specific targets with p = 0.029 and p = 0.016 respectively.

The "Common targets" score the least for this parameter compared to any other group (including the background), the highest score for this group being only 1.3.

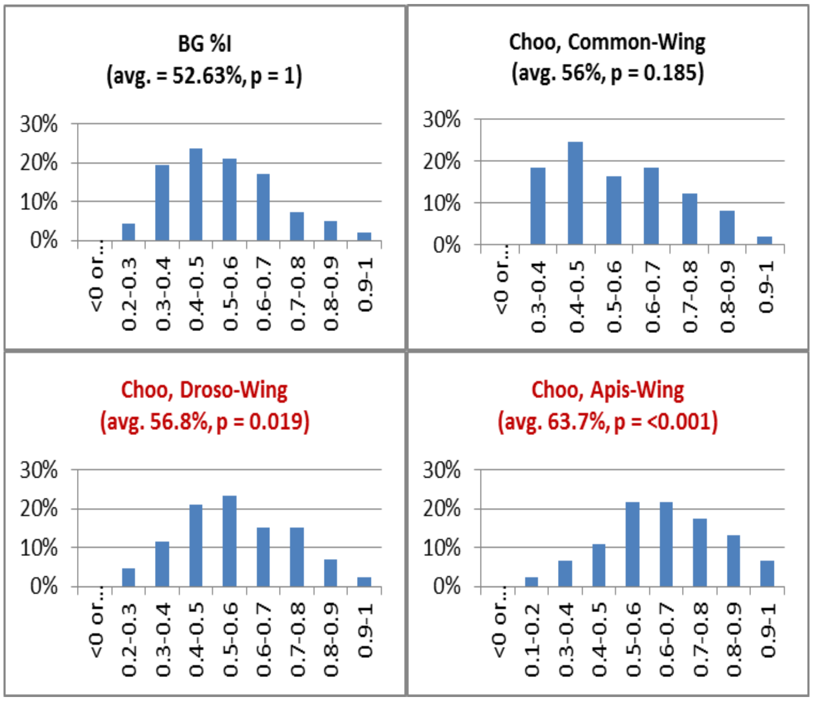
This observation suggests that among the wing development-related subgroups, the one with commonly targeted genes appear to be evolving at a faster rate. Extrapolating the analysis, the fly- or the bee-specific targets appear to be slowly evolving.



Choo: %age Identities

Distributional skewness for “Common targets” towards the leftward end of low conservation (denoting lesser conservation or faster rate of evolution) is seen. This observation is accompanied with a reverse pattern of the inclination and slope of the plot as well as its accompaniment with lesser average %age figures amongst all the groups (except the background).

In contrast, the *Apis-* or *Drosophila*-specific sets show higher overall average scores as well as a statistically significant difference (p = <0.001 and p = 0.019 respectively) in terms of population distribution from the background. This should indicate that *Apis-* or *Drosophila*-specific sets are evolving at a slower rate as compared to the common targets that are involved in wing development. Conversely, the common targets are evolving faster than the genes in the other data sets (except the background) using the parameter %Identities.

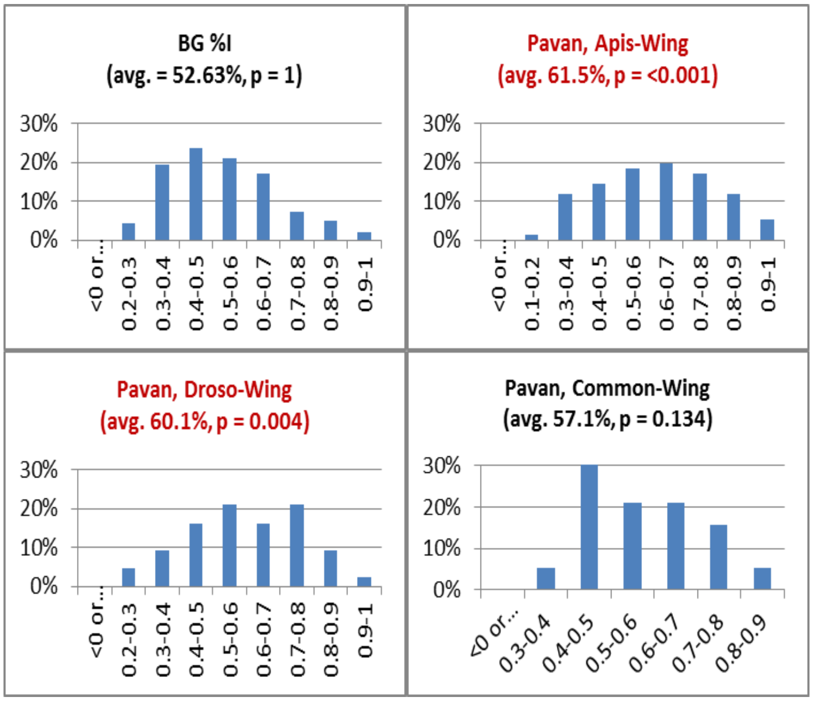


Pavan: %age Identities

This observation recapitulates the previous observation for Choo comparison.

The skewness of the distribution for the “Common targets” towards scores of lower conservation denoting a faster rate of evolution of the targets is discernible. The reverse pattern of slope so observed against the other groups, is accompanied by lesser %age figures for the common targets, which is the least amongst every other group except the background.

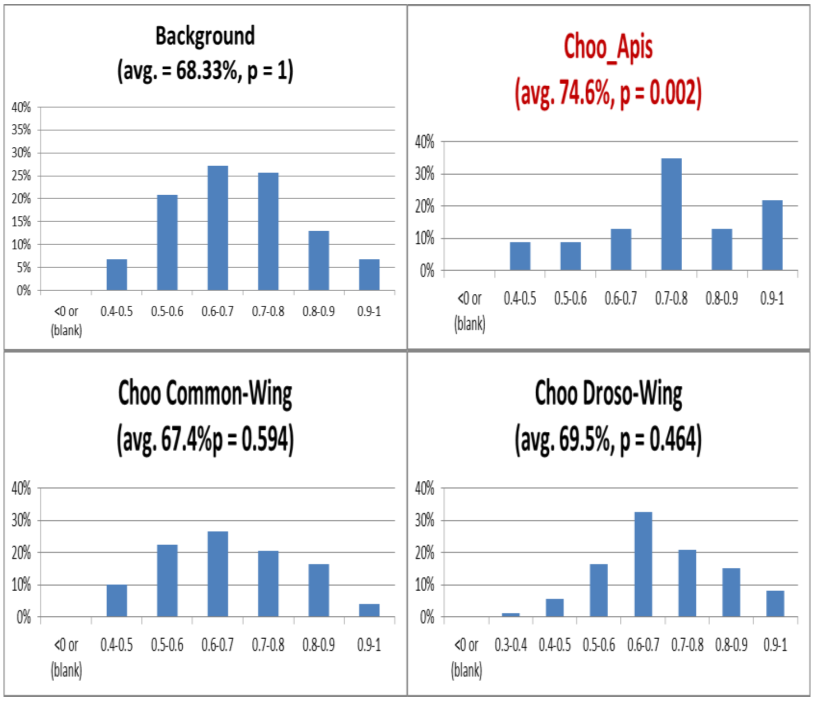
In contrast, the *Apis-* or *Drosophila*-specific groups show higher overall average scores accompanied with statistically significant differences in terms of distribution from the background (with p = <0.001 and p = 0.004 respectively). This indicates that the targets in the *Apis*- or *Drosophila*-specific groups are evolving at a slower rate (hence have higher identities) as compared to the common targets. Conversely, the common targets known to be involved in wing development are evolving faster than the genes in the *Apis*- or *Drosophila*-specific groups.



Choo: %age Positives

The %Positives denote the extent of “functional conservation” or “functional similarities” wherein the substitution of amino acids with functionally similar amino acids is counted along with identical positions. The “Common targets” show the least %age average figures, signifying functional divergences in the ortholog pairs in this group. Interestingly, this score is lesser in comparison with the background (though to a miniscule extent).

The higher overall average scores for *Apis*-specific wing development-related group accompanied by its statistically significant difference in terms of distribution from the background population signify that this group is slowly evolving in terms of “functionalities” as compared with the background distribution.



Background (avg. = 68.33%, p = 1)

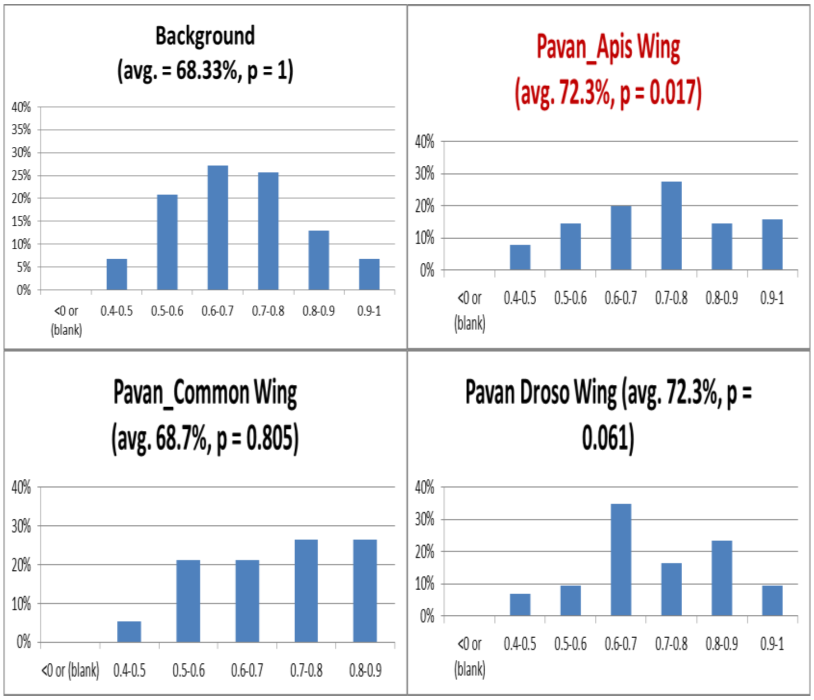
Choo_Apis (avg. 74.6%, p = 0.002)

Choo Common-Wing (avg. 67.4%p = 0.594)

Choo Droso-Wing (avg. 69.5%, p = 0.464)

Pavan: %age Positives

This comparison recapitulates the earlier comparison using the fly targets obtained from Choo et al. (2011). In terms of “functional similarities”, the *Apis*-specific group remains more conserved (or slowly evolving, accompanied by statistically significant difference from the background with p = 0.017) and average %Positives at 72.3%. They are slowly evolving in terms of functionalities when compared with the other groups (viz. common targets and the *Drosophila*-specific targets).


Background (avg. = 68.33%, p = 1)

Pavan_Apis Wing (avg. 72.3%, p = 0.017)

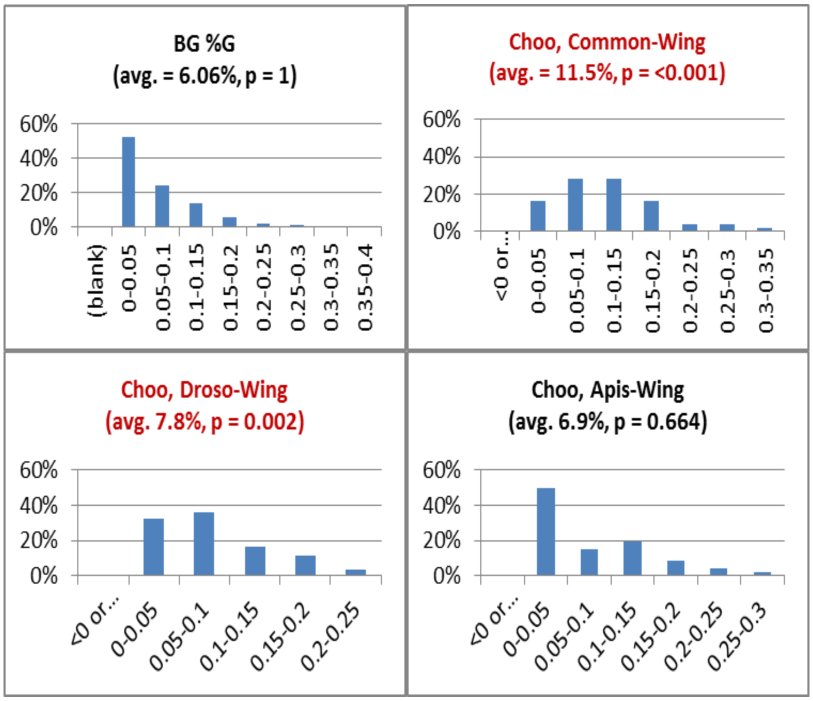
Pavan_Common Wing (avg. 68.7%, p = 0.805)

Pavan Droso Wing (avg. 72.3%, p = 0.061)

Choo: %age Gaps

The most convincing observation with a direct significance comes with the analysis of %Gaps. The gaps in the alignment are the most palpable evidences for divergence, since these entities in the alignment could not be filled-in. The gap openings as well as gap-extension both involve heavy penalties. They, thus, tend to be avoided by any heuristic algorithm such as the BLAST.

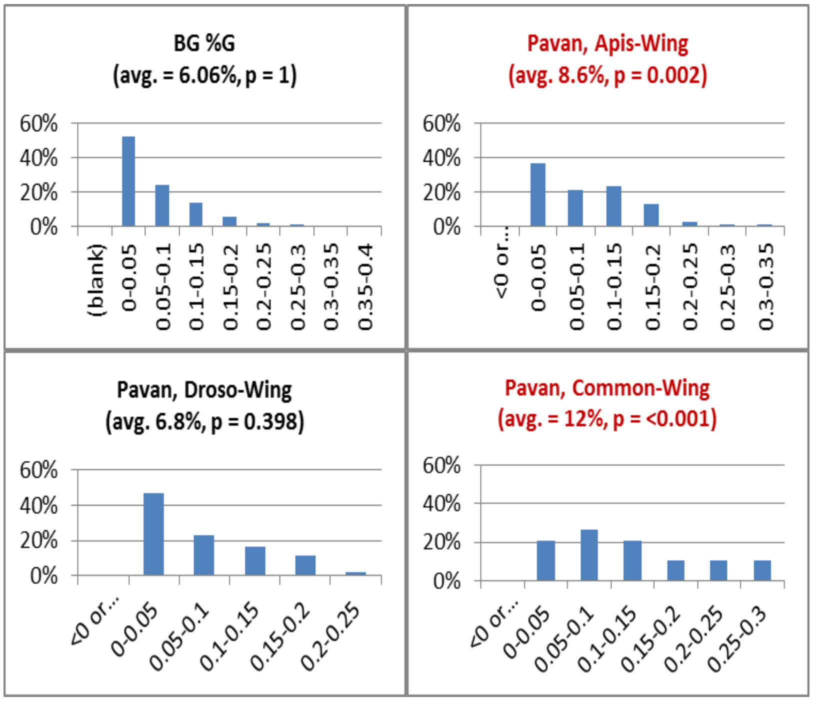
We notice that the differences in average %Gaps amongst various groups is stark: 11.5% gaps in the wing development-related common targets as compared to 6% in the background, or ~8% and ~7% for fly- and bee-specific groups respectively. Both fly-specific targets as well as common targets are statistically significantly different from the background (with p <0.001 and p = 0.002 respectively). This indicates that the common targets known to be involved in wing development are evolving at a higher rate than any other group.



Pavan: %age Gaps

The comparison using the fly targets reported by Pavan et al. (2011) recapitulates comparison using fly targets reported by Choo et al. (2011) for the common targets. The wing development-related targets of Ubx common to both bee and fly are statistically significantly different from the background, with the highest (12%) average incidences of gaps as compared to any other group.

Nevertheless, the *Apis*-specific wing-development related targets too show relatively higher incidences of gaps accompanied by a statistically significantly different distribution when compared with the background population. Overall, this indicates that the common targets are evolving at a rapid pace, followed by the *Apis*-specific targets of Ubx while using %Gaps parameter.



Mann-Whitney Rank Sum Test on data points of various data sets – with wing development-related genes separated using Flybase "GO Biological Process" information

1. *Bit/Aligned length*

Comparison between Naveen-reported (*Apis*) and Pavan et al.-reported (*Drosophila*) data:

Pavan	1	2	5	8	9	10	13	16	17
1	1	0.06	0.061	0.645	**0.015**	**0.015**	**0.007**	**0.003**	0.559
2	0.06	1	0.986	0.114	0.238	0.134	0.082	0.07	0.967
5	0.061	0.986	1	0.114	0.233	0.129	0.078	0.066	0.965
8	0.645	0.114	0.114	1	**0.025**	**0.024**	**0.011**	**0.005**	0.607
9	**0.015**	0.238	0.233	**0.025**	0.999	0.575	0.426	0.534	0.514
10	**0.015**	0.134	0.129	**0.024**	0.575	0.997	0.87	0.927	0.285
13	**0.007**	0.082	0.078	**0.011**	0.426	0.87	0.997	0.932	0.222
16	**0.003**	0.07	0.066	**0.005**	0.534	0.927	0.932	0.999	0.364
17	0.559	0.967	0.965	0.607	0.514	0.285	0.222	0.364	0.988

Comparison between Naveen-reported (*Apis*) and Choo et al.-reported (*Drosophila*) data: **Wing development-related bee-specific group (slowly evolving, higher conservation) statistically significantly differs ($p < 0.05$) from most other groups.**

Choo	1	3	4	6	7	11	12	14	15
1	1	0.313	0.33	0.583	0.471	0.134	0.092	**0.001**	0.856
3	0.313	1	0.974	0.599	0.945	0.335	0.255	**0.004**	0.859
4	0.33	0.974	1	0.616	0.933	0.329	0.25	**0.004**	0.868
6	0.583	0.599	0.616	1	0.698	0.202	0.141	**0.001**	0.995
7	0.471	0.945	0.933	0.698	1	0.413	0.32	**0.008**	0.817
11	0.134	0.335	0.329	0.202	0.413	0.999	0.89	0.072	0.411
12	0.092	0.255	0.25	0.141	0.32	0.89	0.999	0.086	0.349
14	**0.001**	**0.004**	**0.004**	**0.001**	**0.008**	0.072	0.086	0.997	**0.034**
15	0.856	0.859	0.868	0.995	0.817	0.411	0.349	**0.034**	0.997

Legends for data sets:

1: bg
2: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
3: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
4: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
5: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
6: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only
7: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common
8: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only
9: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common
10: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
11: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
12: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
13: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
14: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Apis only
15: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Common
16: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Apis only
17: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Common

2. *Bit score / Minimum length of the two polypeptides*

Comparison between Naveen-reported (*Apis*) and Pavan et al.-reported (*Drosophila*) data:

Pavan	1	2	5	8	9	10	13	16	17
1	1	0.058	0.061	0.118	**0.021**	0.543	0.544	**0.016**	**0.029**
2	0.058	1	0.96	0.38	0.436	0.805	0.803	0.332	0.178
5	0.061	0.96	1	0.403	0.423	0.815	0.814	0.32	0.171
8	0.118	0.38	0.403	1	0.11	0.863	0.863	0.072	0.066
9	**0.021**	0.436	0.423	0.11	0.999	0.542	0.538	0.646	0.366
10	0.543	0.805	0.815	0.863	0.542	0.997	1	0.405	0.234
13	0.544	0.803	0.814	0.863	0.538	1	0.997	0.405	0.234
16	**0.016**	0.332	0.32	0.072	0.646	0.405	0.405	0.999	0.536
17	**0.029**	0.178	0.171	0.066	0.366	0.234	0.234	0.536	0.988

Comparison between Naveen-reported (*Apis*) and Choo et al.-reported (*Drosophila*) data: **Wing development-related "Common targets" group (fast evolving) statistically significantly differ (p < 0.05) from most other groups.**

Choo	1	3	4	6	7	11	12	14	15
1	1	**<0.001**	**<0.001**	0.498	**<0.001**	**0.023**	**0.023**	0.892	**<0.001**
3	**<0.001**	1	0.934	**0.005**	0.096	0.572	0.576	0.319	**0.005**
4	**<0.001**	0.934	1	**0.007**	0.085	0.541	0.545	0.333	**0.004**
6	0.498	**0.005**	**0.007**	1	**<0.001**	**0.048**	**0.048**	0.942	**<0.001**
7	**<0.001**	0.096	0.085	**<0.001**	1	0.564	0.56	0.099	0.058
11	**0.023**	0.572	0.541	**0.048**	0.564	0.999	0.998	0.25	**0.043**
12	**0.023**	0.576	0.545	**0.048**	0.56	0.998	0.999	0.252	**0.043**
14	0.892	0.319	0.333	0.942	0.099	0.25	0.252	0.997	**0.01**
15	**<0.001**	**0.005**	**0.004**	**<0.001**	0.058	**0.043**	**0.043**	**0.01**	0.997

Legends for data sets:

1: bg
2: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
3: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
4: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
5: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
6: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only
7: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common
8: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only
9: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common
10: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
11: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
12: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
13: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
14: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Apis only
15: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Common
16: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Apis only
17: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Common

3. *%Identities in between the two polypeptides*

Comparison between Naveen-reported (*Apis*) and Pavan et al.-reported (*Drosophila*) data:

Pavan	**1**	**2**	**5**	**8**	**9**	**10**	**13**	**16**	**17**
1	1	**0.044**	**0.039**	0.315	**0.002**	**0.004**	**0.002**	**<0.001**	0.134
2	**0.044**	1	0.99	0.171	0.084	0.056	**0.03**	**0.003**	0.472
5	**0.039**	0.99	1	0.163	0.085	0.055	**0.029**	**0.003**	0.465
8	0.315	0.171	0.163	1	**0.006**	**0.009**	**0.004**	**<0.001**	0.208
9	**0.002**	0.084	0.085	**0.006**	0.999	0.551	0.41	0.248	0.8
10	**0.004**	0.056	0.055	**0.009**	0.551	0.997	0.873	0.652	0.487
13	**0.002**	**0.03**	**0.029**	**0.004**	0.41	0.873	0.997	0.786	0.397
16	**<0.001**	**0.003**	**0.003**	**<0.001**	0.248	0.652	0.786	0.999	0.306
17	0.134	0.472	0.465	0.208	0.8	0.487	0.397	0.306	0.988

Comparison between Naveen-reported (*Apis*) and Choo et al.-reported (*Drosophila*) data: **Wing development-related *Apis*-specific targets (slowly evolving) differ from most other groups by degrees of statistical significance (p < 0.5).**

Choo	**1**	**3**	**4**	**6**	**7**	**11**	**12**	**14**	**15**
1	1	0.055	0.055	0.423	**0.041**	**0.019**	**0.01**	**<0.001**	0.185
3	0.055	1	0.995	0.212	0.493	0.175	0.121	**<0.001**	0.555
4	0.055	0.995	1	0.213	0.491	0.175	0.121	**<0.001**	0.552
6	0.423	0.212	0.213	1	0.124	**0.043**	**0.025**	**<0.001**	0.312
7	**0.041**	0.493	0.491	0.124	1	0.406	0.31	**0.002**	0.897
11	**0.019**	0.175	0.175	**0.043**	0.406	0.999	0.89	**0.018**	0.68
12	**0.01**	0.121	0.121	**0.025**	0.31	0.89	0.999	**0.023**	0.589
14	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.018**	**0.023**	0.997	**0.019**
15	0.185	0.555	0.552	0.312	0.897	0.68	0.589	**0.019**	0.997

Legends for data sets:

1: bg
2: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
3: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
4: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
5: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
6: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only
7: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common
8: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only
9: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common
10: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
11: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
12: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
13: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
14: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Apis only
15: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Common
16: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Apis only
17: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Common

4. %Positives in between the two polypeptides

Comparison between Naveen-reported (*Apis*) and Pavan et al.-reported (*Drosophila*) data:

Pavan	**1**	**2**	**5**	**8**	**9**	**10**	**13**	**16**	**17**
1	1	0.244	0.237	0.683	**0.049**	0.061	**0.033**	**0.017**	0.805
2	0.244	1	1	0.373	0.252	0.176	0.114	0.092	0.863
5	0.237	1	1	0.366	0.252	0.174	0.112	0.09	0.862
8	0.683	0.373	0.366	1	0.072	0.079	**0.045**	**0.024**	0.871
9	**0.049**	0.252	0.252	0.072	0.999	0.638	0.516	0.593	0.47
10	0.061	0.176	0.174	0.079	0.638	0.997	0.88	0.936	0.321
13	**0.033**	0.114	0.112	**0.045**	0.516	0.88	0.997	0.943	0.262
16	**0.017**	0.092	0.09	**0.024**	0.593	0.936	0.943	0.999	0.333
17	0.805	0.863	0.862	0.871	0.47	0.321	0.262	0.333	0.988

Comparison between Naveen-reported (*Apis*) and Choo et al.-reported (*Drosophila*) data: **Wing development-related *Apis*-specific targets (slowly evolving, high conservation) outstand with statistical significance (p < 0.05).**

Choo	**1**	**3**	**4**	**6**	**7**	**11**	**12**	**14**	**15**
1	1	0.8	0.775	0.551	0.885	0.464	0.365	**0.002**	0.594
3	0.8	1	0.981	0.795	0.991	0.57	0.462	**0.004**	0.566
4	0.775	0.981	1	0.82	0.993	0.573	0.464	**0.004**	0.559
6	0.551	0.795	0.82	1	0.835	0.623	0.513	**0.003**	0.467
7	0.885	0.991	0.993	0.835	1	0.622	0.513	**0.007**	0.537
11	0.464	0.57	0.573	0.623	0.622	0.999	0.902	**0.029**	0.387
12	0.365	0.462	0.464	0.513	0.513	0.902	0.999	**0.033**	0.323
14	**0.002**	**0.004**	**0.004**	**0.003**	**0.007**	**0.029**	**0.033**	0.997	**0.015**
15	0.594	0.566	0.559	0.467	0.537	0.387	0.323	**0.015**	0.997

Legends for data sets:

1: bg
2: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
3: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
4: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
5: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
6: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only
7: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common
8: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only
9: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common
10: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
11: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
12: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
13: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
14: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Apis only
15: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Common
16: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Apis only
17: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Common

5. %Gaps in between the two polypeptides

Comparison between Naveen-reported (*Apis*) and Pavan et al.-reported (*Drosophila*) data: **The "Common targets" known to be involved in wing development (fast evolving) statistically outstand, so do wing development-related bee-specific targets (slow evolving) with statistical significance (p < 0.05).**

Pavan	1	2	5	8	9	10	13	16	17
1	1	0.73	0.559	0.161	0.342	0.398	0.275	**0.002**	**<0.001**
2	0.73	1	0.856	0.571	0.523	0.52	0.386	**0.009**	**<0.001**
5	0.559	0.856	1	0.722	0.602	0.574	0.431	**0.011**	**<0.001**
8	0.161	0.571	0.722	1	0.726	0.666	0.512	**0.011**	**0.001**
9	0.342	0.523	0.602	0.726	0.999	0.844	0.707	0.079	**0.006**
10	0.398	0.52	0.574	0.666	0.844	0.997	0.883	0.222	**0.016**
13	0.275	0.386	0.431	0.512	0.707	0.883	0.997	0.277	**0.019**
16	**0.002**	**0.009**	**0.011**	**0.011**	0.079	0.222	0.277	0.999	0.091
17	**<0.001**	**<0.001**	**<0.001**	**0.001**	**0.006**	**0.016**	**0.019**	0.091	0.988

Comparison between Naveen-reported (*Apis*) and Choo et al.-reported (*Drosophila*) data: **Wing development-related "common targets" statistically outstand from other groups.**

Choo	1	3	4	6	7	11	12	14	15
1	1	**<0.001**	**<0.001**	0.787	**<0.001**	**0.002**	**<0.001**	0.664	**<0.001**
3	**<0.001**	1	0.929	**<0.001**	0.157	0.286	0.21	0.447	**<0.001**
4	**<0.001**	0.929	1	**<0.001**	0.174	0.308	0.227	0.425	**<0.001**
6	0.787	**<0.001**	**<0.001**	1	**<0.001**	**0.005**	**0.003**	0.719	**<0.001**
7	**<0.001**	0.157	0.174	**<0.001**	1	0.97	0.909	0.175	**0.002**
11	**0.002**	0.286	0.308	**0.005**	0.97	0.999	0.887	0.193	**0.003**
12	**<0.001**	0.21	0.227	**0.003**	0.909	0.887	0.999	0.152	**0.004**
14	0.664	0.447	0.425	0.719	0.175	0.193	0.152	0.997	**0.001**
15	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.003**	**0.004**	**0.001**	0.997

Legends for data sets:

1: bg
2: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
3: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
4: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Drosophila only
5: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_all genes_Drosophila only
6: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Apis only
7: V5_BLAST_Drosophila_Choo vs Apis_Naveen_all genes_Common
8: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Apis only
9: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_all genes_Common
10: New BLAST V5.2 Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
11: New BLAST V5.2 Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
12: V5_ BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Drosophila only
13: V5_BLAST_Drosophila_ Pavan vs Apis_Naveen_WING_ONLY_Drosophila only
14: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Apis only
15: V5_BLAST_Drosophila_Choo vs Apis_Naveen_WING_ONLY_Common
16: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Apis only
17: V5_BLAST_Drosophila_Pavan vs Apis_Naveen_WING_ONLY_Common

Discussion

The *Hox* protein Ubx causes both the subtle differences between forewing and hind wing in *Apis* and specification of a new organ, haltere, in the place of hind wing in *Drosophila*. To investigate the evolutionary changes at the molecular level, our laboratory has experimentally identified targets of Ubx in *Apis* and *Drosophila*. Some of these targets are common to both the species, while majority are species-specific targets of Ubx. Here, we have compared protein-by-protein the degree of conservation of all genes for which orthologs are present in the two species. We then calculated frequency distribution of the homology scores to determine patterns, if any, amongst various subgroups of targets of Ubx. 4000 pairwise protein homology scores of proteins randomly picked up from the database was used as the background for all statistical tests.

Following were the main results:

1. Except targets that are common to both *Apis* and *Drosophila*, all other targets showed no difference in their relative conservation between the two species when compared to the background sequences.
2. Targets that are common to both *Apis* and *Drosophila*, showed lesser degree of homology compared to the background suggesting that these proteins are evolving faster. This is true whether we use bit score / Average protein length or minimum protein length or % identities or % gaps. Interestingly, even when we considered only wing development-related genes for analysis, we observed higher degree of divergence amongst the proteins that are targets of Ubx in both *Apis* and *Drosophila*. In this, we had somewhat different trends with different data sets.
    a. When the wing development-related genes were identified on the basis of summary information available in the Flybase, proteins in the “Pavan” data set (*Drosophila* targets were identified in our lab) showed significantly more divergence from the background sequences, and no such contrast is seen for the “Choo” data set wherein the (*Drosophila* targets which were identified in the laboratory of Rob White).

b. When the wing development-related genes are identified using the “GO Biological Process” information, which, too, is available in the Flybase, proteins in both Pavan data set and the Choo data showed more divergence from the background sequences. This observation was more pronounced for those wing development-related targets that are common targets of Ubx in bee and fly.

Conclusions

Proteins that are targets of Ubx in both *Apis* and *Drosophila* appear to have fast evolved and thereby may have acquired new functions in *Drosophila*. Ubx-mediated modulation of their expression patterns may lead to different consequences in *Drosophila* than in *Apis*. Our work suggests this as an additional factor causing the evolution of haltere in *Drosophila*. Other factors, which are not analyzed here, are (i) novel genes as targets of Ubx in *Drosophila* (for these no *Apis* homologue would be available) and (ii) additional regulatory features in common targets making their regulation by Ubx in *Drosophila* different from their regulation in *Apis*.

Future directions

The future work includes similar comparison of targets of Ubx in *Drosophila* and Silkworm (*Bombyx*), as well *Apis* and *Bombyx*. This helps to understand general trend in the evolution of targets of Ubx and specific features of targets of Ubx in *Drosophila* leading to the evolution of haltere.

References

1. Honey Bee Genome Sequencing Consortium (2006), *Nature* 443, 931–948
2. Agrawal, P., et al. (2011). *Scientific Reports* 1, DOI: 10.1038/srep00205
3. Lewis, E.B. 1978. *Nature* 276, 565-570
4. Choo et al. (2011). *PLoS ONE*, 6(4): e14778. doi:10.1371/journal.pone.0014778
5. Slattery et al (2011), *PLoS ONE, doi:10.1371*
6. Makhijani K et al. (2007), *Dev Biol* 302: 243–255.
7. Pallavi SK et al. (2006), *Dev Biol* 296: 340–352
8. Shashidhara, L. S. et al. (1999), *Dev. Biol.* 212, 491–502 (1999).
9. Carroll SB et al. (1995), *Nature* 375: 58–61.
10. Carroll SB (1995), *Nature* 376 (6540):479–85. doi:10.1038/376479a0
11. Weatherbee SD et al. (1998), *Genes Dev* 12: 1474–1482.