PLOS ONE

RESEARCH ARTICLE

In-silico identification and comparison of transcription factor binding sites cluster in anterior-posterior patterning genes in *Drosophila melanogaster* and *Tribolium castaneum*

Anshika Moudgil[1], Ranbir Chander Sobti[2], Tejinder Kaur[1]*

1 Department of Zoology, DAV University, Jalandhar, Punjab, India, **2** Department of Biotechnology, Panjab University, Chandigarh, India

* tejinder10034@davuniversity.org

Abstract

The cis-regulatory data that help in transcriptional regulation is arranged into modular pieces of a few hundred base pairs called CRMs (cis-regulatory modules) and numerous binding sites for multiple transcription factors are prominent characteristics of these cis-regulatory modules. The present study was designed to localize transcription factor binding site (TFBS) clusters on twelve Anterior-posterior (A-P) genes in *Tribolium castaneum* and compare them to their orthologous gene enhancers in *Drosophila melanogaster*. Out of the twelve A-P patterning genes, six were gap genes (*Kruppel*, *Knirps*, *Tailless*, *Hunchback*, *Giant*, and *Caudal*) and six were pair rule genes (*Hairy*, *Runt*, *Even-skipped*, *Fushi-tarazu*, *Paired*, and *Odd-skipped*). The genes along with 20 kb upstream and downstream regions were scanned for TFBS clusters using the Motif Cluster Alignment Search Tool (MCAST), a bioinformatics tool that looks for set of nucleotide sequences for statistically significant clusters of non-overlapping occurrence of a given set of motifs. The motifs used in the current study were Hunchback, Caudal, Giant, Kruppel, Knirps, and Even-skipped. The results of the MCAST analysis revealed the maximum number of TFBS for Hunchback, Knirps, Caudal, and Kruppel in both *D. melanogaster* and *T. castaneum*, while Bicoid TFBS clusters were found only in *D. melanogaster*. The size of all the predicted TFBS clusters was less than 1kb in both insect species. These sequences revealed more transversional sites (Tv) than transitional sites (Ti) and the average Ti/Tv ratio was 0.75.

Introduction

To know the development processes occurring in metazoans, it is vital to comprehend the regulatory mechanics of the underlying transcriptional network. The genomic sequence of an organism contains a significant amount of information that specifies how and when genes will be expressed. Despite the availability of genome sequences for many metazoans, very little is known about how this biological data is encoded [1, 2]. Previous research on the early

development of *Drosophila melanogaster*, a model organism for more than three decades, provides an excellent context for studying the cis-regulatory modules (CRMs). CRMs are certain areas of non-protein-coding DNA, that play a significant role in controlling the expression patterns of genes to build an embryo's tissue [3]. The CRMs are composed of groups of short DNA sequences that are acknowledged and bound by certain transcription factors [4]. The enhancers, promoters, and silencers consist of cis-regulatory sequences, recognized as CRMs [5]. Enhancers and silencers are usually found upstream (5′), downstream (3′), or in the intron (or introns) of the gene they control, although they can also be found far away while promoter sequences are always present upstream to the target gene [6]. During the early stages of embryonic development, a quick cascade of gene regulation determines the segmented body pattern of *D. melanogaster* [7]. The process of Anterior-Posterior (A-P) segmentation is initiated by the maternal gene products that are present in the gradient. The bicoid protein gradient exhibiting morphogenetic features is the best-known example. A spatial-differential, concentration-dependent expression (or repression), of certain zygotic genes, is regarded to establish the bicoid protein gradient [8, 9]. The first zygotic gene to be expressed belongs to the gap genes cascade which is a well-studied system in *D. melanogaster* [10]. The gap genes control the pair-rule genes which further regulate the segment polarity genes and the Homeotic gene complex in succession [10–13]. In addition to this, the maternal genes, gap genes, pair-rule genes, and Homeotic gene complex also self-regulate themselves as shown in Fig 1. An ideal framework for researching the cis-regulatory sequences is provided by all research on the early stages of *Drosophila* development. A variety of interactions between different transcription factors (TF) and their target regulatory areas have been thoroughly defined [11, 14], and comparative investigations have demonstrated that cis-regulatory regions are often functionally conserved throughout the genus [15–17]. Previous research characterized transcription factor binding
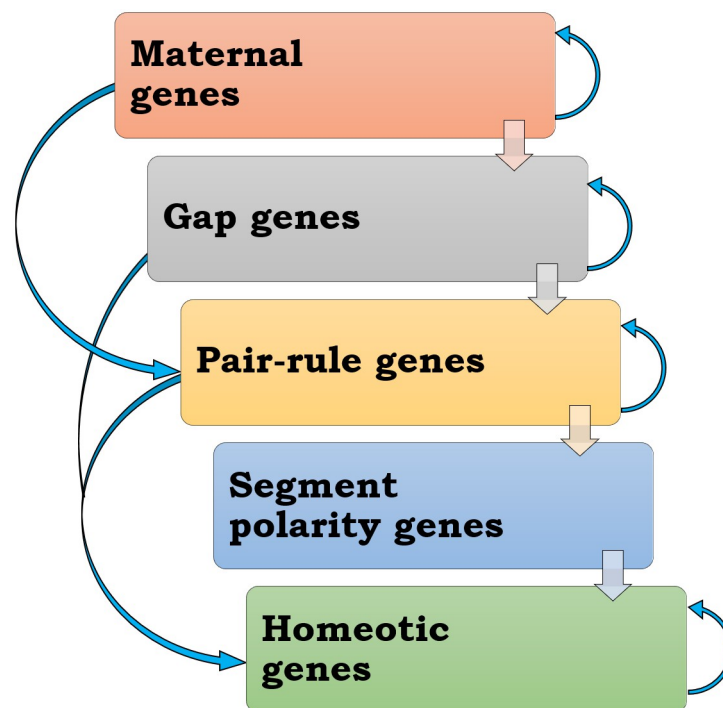

Fig 1. Schematic representation of the regulatory relationship between the anterior-posterior patterning gene cascade in *Drosophila melanogaster*.

site (TFBS) locations in the early *Drosophila* embryo and estimated the binding affinity of each factor using Position weight matrices (PWMs) [1, 18]. PWMs are a valuable method for analysing the location of potential binding sites and estimating their binding strength [19, 20]. All the research about the cis-regulatory modules is mainly focused on the *Drosophila* genus. Not much has been done on other insect species. Over the last two decades, *T. castaneum* has emerged as a potent organism to study short germ segmentation, embryonic head and leg development, metamorphosis, and in insect biology. The anterior-posterior patterning in *Tribolium* follows an ancestral route i.e. short germ embryogenesis, which is different from *D. melanogaster* which follows the latest route i.e. long germ embryogenesis [21, 22]. In this paper, we investigate the Transcription factor binding sites cluster i.e. Cis-regulatory modules in the gap genes and pair-rule genes of *D. melanogaster*, and compare them with their orthologs present in *T. castaneum*.

Material and methods

Data collection

The Translated gene sequences for the A-P patterning genes of *D. melanogaster* were downloaded from Flybase (FB2021_03) [23]. The NCBI database (pBlast) was then used to collect the A-P patterning protein sequences of *T. castaneum* orthologous to those of *D. melanogaster* [24]. The orthologous sequences with a high query-covered value and a low error value were selected. The genomic data viewer was used to access these sequences. Following that, 20 kb flanking sequences were added both upstream (-) and downstream (+) to each target gene's sequence. The gene sequences along with the additional flanking sequences were then downloaded in the FASTA format. The sequence location of genes alongwith the flanking regions are given in Tables 1 and 2 for *D. melanogaster* and *T. castaneum* respectively.

Motif collection

The JASPAR database, which is open to the public, contains position weight matrices (PWMs), of various species in six taxonomic groupings. The PWM motifs used for the present study were Bicoid, Hunchback, Caudal, Giant, Kruppel, Knirps, and Even-skipped of *D. melanogaster*. These were downloaded in meme format from JASPAR software [25].

Table 1. List of genes used for enhancer localization along with their chromosomal locations, accession numbers, gene locations, and gene sequence location with the flanking regions in *Drosophila melanogaster*.

S.No.	Gene Name	Gene chromosomal location and Accession number	Gene Location	Gene sequences along with their 20 kb flanking regions
1	Hunchback	Chr. 3R (NT_033777.3)	8691649–8694381	8671649–8714381
2	Knirps	Chr. 3L (NT_037436.4)	20692673–20694617	20672673–20714617
3	Caudal	Chr. 2L (NT_033779.5)	20769990–20783396	20749990–20803396
4	Kruppel	Chr. 2R (NT_033778.4)	25226133–25231872	25206133–25251872
5	Giant	Chr. X (NC_004354.4)	24272888–2428993	2407288–2448993
6	Tailless	Chr. 3R(NT_033777.3)	30852398–30854173	30832398–30874173
7	Even-skipped	Chr. 2R (NT_033778.4)	9979170–9980942	9959170–10000942
8	Odd-skipped	Chr. 2L (NT_033779.5)	3604856–3606494	3584856–3626494
9	Runt	Chr. X (NC_004354.4)	20690004–20697985	20670004–20717985
10	Hairy	Chr. 3L (NT_037436.4)	8676032–8678639	8656032–8698639
11	Fushi-tarazu	Chr. 3 (NT_033777.3)	6864,255–6865910	6844,255–6885,910
12	Paired	Chr. 2L (NT_033779.5)	12083091–12085750	12063091–12105750

Table 2. List of genes used for enhancer localization along with their chromosomal locations, accession numbers, gene locations, and gene sequence location with the flanking regions in *Tribolium castaneum*.

S.No.	Gene Name	Gene chromosomal location and Accession number	Gene Location	Gene sequences along with their 20 kb flanking regions
1	Hunchback	Chr. LG5 (NC_007420.3)	8101789–8103523	8081615–8123697
2	Knirps	Chr. LG3 (NC_007418.3)	4699064–4709975	4679064–4729975
3	Caudal	Chr. LG4 (NM_007419.2)	2828167–2845622	2808167–2865622
4	Kruppel	Chr. LG10 (NC_007425.3)	6495006–6502111	6475006–6522111
5	Giant	Chr. LG4 (NC_007419.2)	3765785–3766872	3745785–3786872
6	Tailless	Chr. LG2 (NC_007417.3)	11221687–11225898	11201687–11245898
7	Even-skipped	Chr. LG7 (NC_007422.5)	624611–628174	604611–648174
8	Odd-skipped	Chr. LG8 (NC_007423.3)	7672873–7679594	7652873–7699594
9	Runt	Chr. LG8 (NC_007423.3)	11253714–11257683	11233714–11277683
10	Hairy	Chr. LG9(NC_007424.3)	15627964–15629572	15607964–15649572
11	Fushi-tarazu	Chr. LG2 (NC_007417.3)	8414592–8415698	8394592–8435698
12	Paired	Chr. LG6 (NC_007421.3)	8595852–8650582	8575852–8650582

https://doi.org/10.1371/journal.pone.0290035.t002

Meme suite analysis

For the discovery of transcription factor binding site (TFBS) clusters, MCAST, an application of MEME Suite was run [26]. MCAST scans for clusters of matches to one or more nucleotide motifs in sequences [27]. The input of sequence was given in text file in FASTA format while the motifs were given in Meme format. ARR1 of *Saccharomyces cerevisiae* was used as an out-group. For the identification of TFBS, the parameters which were used in the present study were: p-value should be less than 0.005, the error value less than 5 and the gap between two TFBS should be less than 30 base pairs. The result was displayed in the HTML format and a cluster with greater motif score, and low error value was selected as pCRM.

Annotation of transcription start site (TSS)

Transcription start sites (TSS) for the A-P genes were predicted using the genome data viewer tool of NCBI in the case of *D. melanogaster* and *T. castaneum* [28].

Annotation of promotor region, exon, intron region of the gene

Putative promoter regions, exon, and intronic regions were identified using two databases Ensemble and NCBI [28, 29].

The NCBI database's genome data viewer was used to get the necessary sequence, which featured intron, exon, and promoter sequences. The intron and exon of the given sequence were represented by different colours. Exons were represented by light pink, introns by green, and the following gene by blue colour. The promotor sequence area was recovered up to 1000 bp upstream of the TSS region, including the AT-rich region and the TATA box.

The searched-for sequence, which contained the 5' flanking area, the promotor, the exon, the intron, and the 3' flanking region, was retrieved from the Ensemble website.

Sequence alignment and variation of predicted cis-regulatory modules

All the predicted pCRM were aligned using the ClustalW [30] tool in Bioedit software [31] and these aligned sequences were subjected to calculate the conserved sites, transition pairs (Ti), transversional pairs (Tv), and transition/transversion (Ti/Tv) ratio in MEGA XI software [32].

Interaction between the TFs

In the present study, the STRING database was used to predict the protein-protein interactions between the different transcription factors searched for their binding sites on the A-P patterning genes [33].

Results

TFBS clusters in gap genes of *D. melanogaster* and *T. castaneum*

The results of the MCAST analysis for identifying TFBS clusters in the gap genes (*Hunchback*, *Knirps, and Caudal)* are depicted in Fig 2. The results reveal that the location of transcription factor binding sites (TFBS) on the *Hunchback* gene is upstream to the transcription start site (TSS) in both *D. melanogaster* and *T. castaneum* (Fig 2.1). However, the results are variable in *D. melanogaster* and *T. castaneum* for another gap gene called *Knirps*, on which, the TFBS cluster is located within intron 2 of the gene in the dipteran insect and upstream in the coleopteron (Fig 2.2). As far as the *Caudal* gene is concerned, the cluster of TFBS is located downstream to the TSS in both the insect species (Fig 2.3). The cluster was found to be present within exon 1 of the gene in *D. melanogaster* and exon 3 of the gene in *T. castaneum*.

Fig 3 illustrates the results of MCAST analysis for predicting TFBS clusters on *Kruppel*, *Giant*, and *Tailless* genes. The TFBS cluster in the *Kruppel* gene is present upstream to the TSS in *D. melanogaster* and downstream to TSS in *T. castaneum* (Fig 3.1). In case of the *Giant* gene, the cluster of TFBS is present upstream to TSS in both *D. melanogaster* and *T. castaneum* as depicted in Fig 3.2. The location of the TFBS cluster in the gene *Tailless* is upstream in *D. melanogaster* while downstream to the TSS in *T. castaneum*.

TFBS clusters in Pair rule genes of *D. melanogaster* and *T. castaneum*

Fig 4 shows the results for the TFBS clusters as predicted by the MCAST software for the *Even-skipped*, *Hairy*, and *Runt* pair-rule genes in *D. melanogaster* and *T. castaneum*. The cluster of TFBS in both *Even-skipped* and *Runt* genes are present upstream of the TSS in both *D. melanogaster* and *T. castaneum* (Fig 4.1 and 4.3). Fig 4.2, shows the cluster of TFBS on the *Hairy* gene. The cluster is present upstream to the TSS in *D. melanogaster* and downstream to the TSS in *T. castaneum*.

The CRMs for *Odd-skipped*, *Fushi-tarazu*, and *Paired* genes as predicted by the MCAST software are depicted in Fig 5. Fig 5.1 exhibits the cluster of TFBS on the *Odd-skipped* gene. The cluster of is present downstream to the TSS in *D. melanogaster* and upstream in *T. castaneum*. The clusters of TFBS in both *Fushi-tarazu* and *Paired* genes are present upstream to the TSS in both the insect species (Fig 5.2 and 5.3).

The results depicted by the MCAST software, that is, location of the TSS, TFBS cluster, and size of the putative cluster in base pairs are summarized in Table 3 for *D. melanogaster* and Table 4 for *T. castaneum*. Table 5 depicts the number of transcription factor binding sites in all genes predicted by the software in both the insect species.

Multiple sequence alignment, conserved sites, transitional pairs, transversional pairs, and transition/transversion rate

All the predicted pCRMs were subjected to multiple sequence alignment using the CLUS-TALW [30] tool in Bioedit software [31]. These were further analysed for the presence of conserved sites, transition pairs, transversional pairs, and transition/transversion rate using MEGA XI software [32]. The results for the analysis are depicted in Table 6.
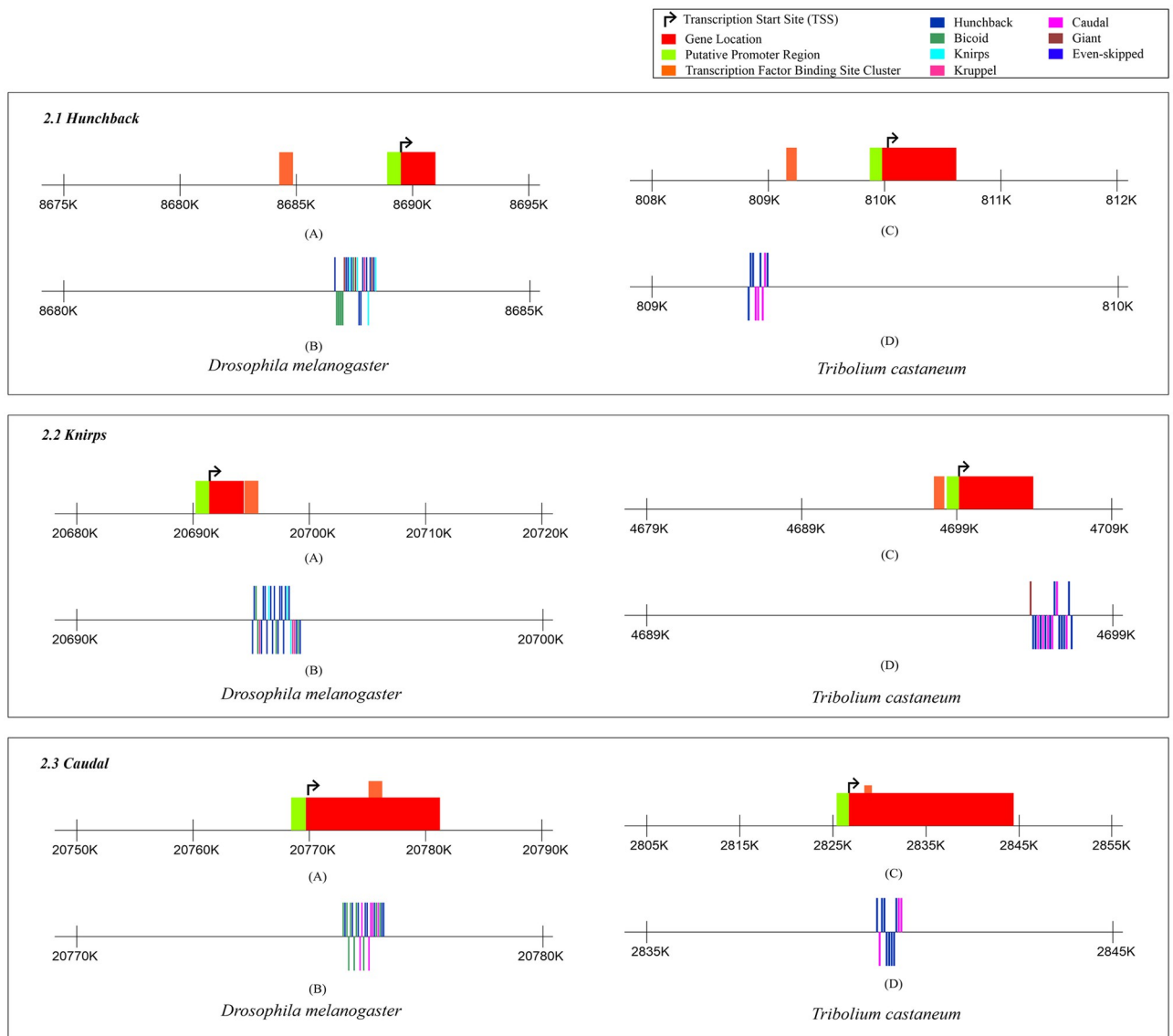
Fig 2. Figure showing the pCRM details of *Drosophila melanogaster* and *Tribolium castaneum* for different transcription factors in *Hunchback,* *Knirps,* and *Caudal*. 2.1 shows the predicted transcription factor binding site clusters in *Hunchback* for *D. melanogaster* and *T. castaneum*.A) Locations of the gene, transcription start site (TSS) and the transcription factor binding site (TFBS) cluster as predicted by the MCAST software for the *Hunchback* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Hunchback* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Hunchback* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Hunchback* for *T. castaneum*. 2.2 shows the predicted transcription factor binding site clusters in *Knirps* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Knirps* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Knirps* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Knirps* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Knirps* for *T. castaneum*, 2.3 shows the predicted transcription factor binding site clusters in *Caudal* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Caudal* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Caudal* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Caudal* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Caudal* for *T. castaneum*.

https://doi.org/10.1371/journal.pone.0290035.g002

Interaction between the TFs

Figs 6 and 7 depict the interactions between Bicoid, Hunchback, Caudal, Knirps, Kruppel, Giant, and Even-skipped transcription factors of *D. melanogaster* and *T. castaneum* respectively.
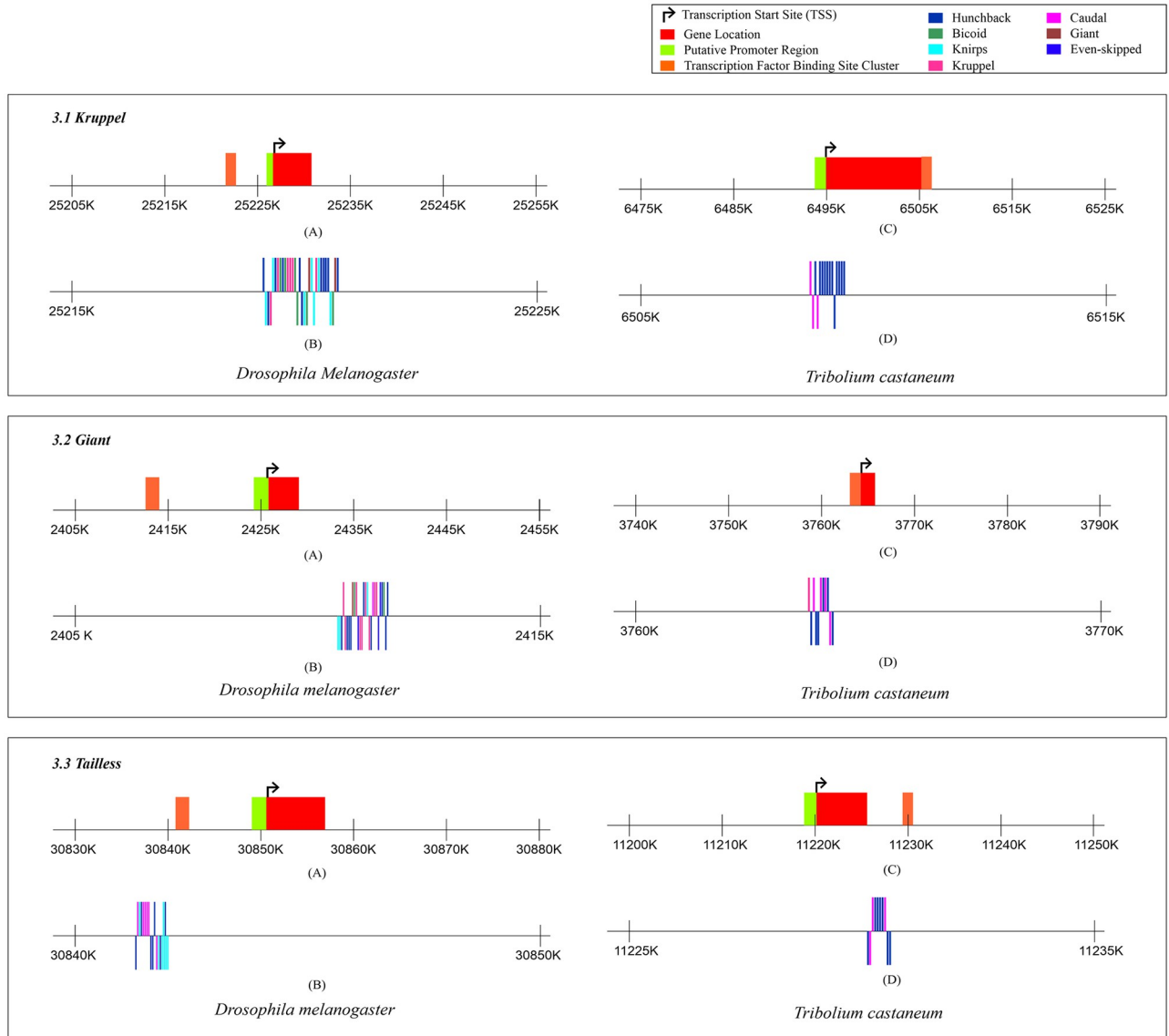
Fig 3. Figure showing the pCRM details of *Drosophila melanogaster* and *Tribolium castaneum* for different transcription factors in *Kruppel*, *Giant*, and *tailless*. 3.1 shows the predicted transcription factor binding site clusters in *Kruppel* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Kruppel* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Kruppel* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Kruppel* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Kruppel* for *T. castaneum*, 3.2 shows the predicted transcription factor binding site clusters in *Giant* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Giant* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Giant* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Giant* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Giant* for *T. castaneum* 3.3 shows the predicted transcription factor binding site clusters in *Tailless* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Tailless* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Tailless* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Tailless* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Tailless* for *T. castaneum*.
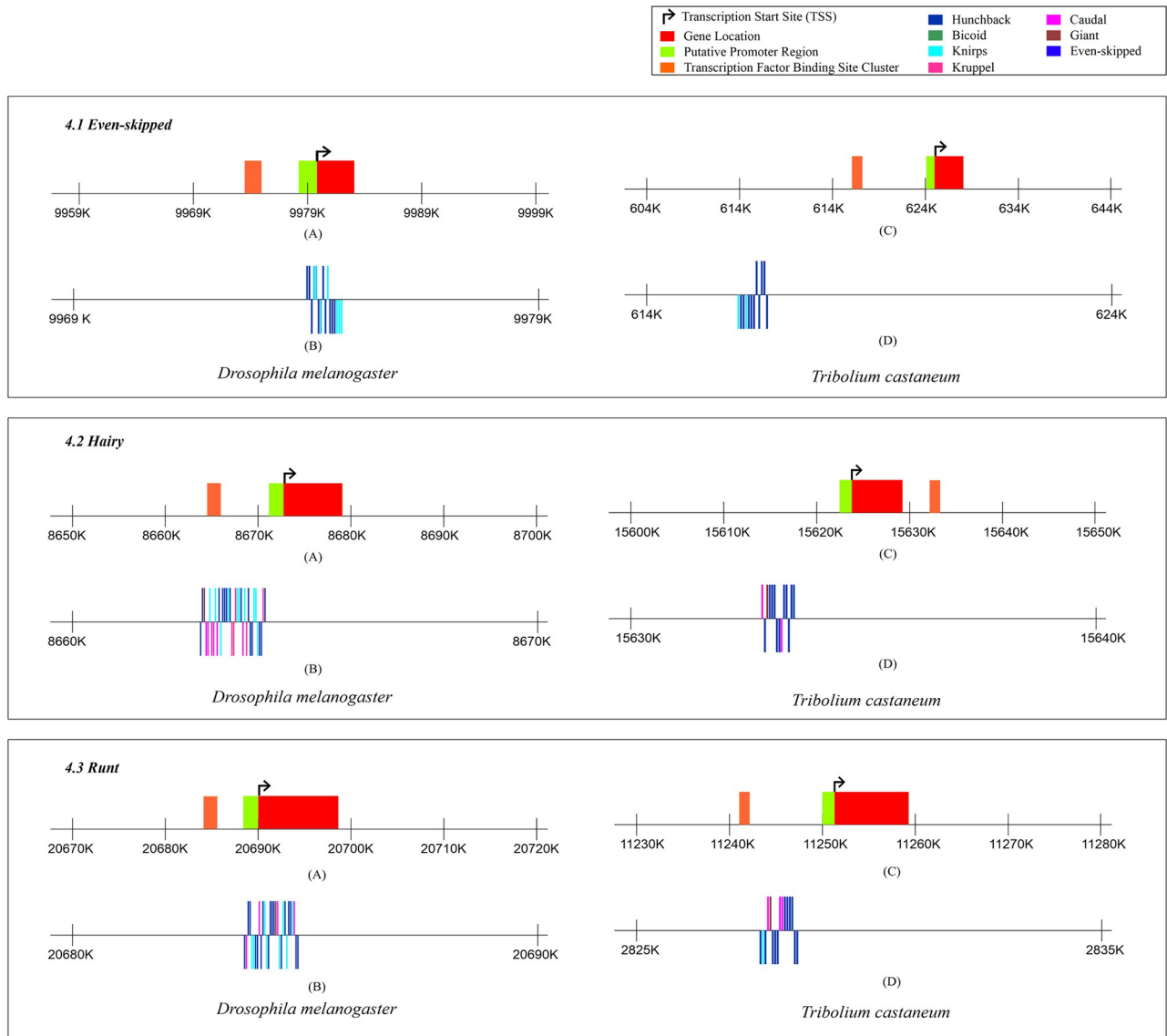
Fig 4. Figure showing the pCRM for *Drosophila melanogaster* and *Tribolium castaneum* for different transcription factors in *Even-skipped*, *Hairy*, and *Runt*. 4.1 shows the predicted transcription factor binding site clusters in *Even-skipped* for *D. melanogaster* and *T. castaneum*, A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Even-skipped* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Even-skipped* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Even-skipped* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Even-skipped* for *T. castaneum*. 4.2 shows the predicted transcription factor binding site clusters in *Hairy* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Hairy* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Hairy* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Hairy* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Hairy* for *T. castaneum*. 4.3 shows the predicted transcription factor binding site clusters in *Runt* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Runt* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Runt* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Runt* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Runt* for *T. castaneum*.
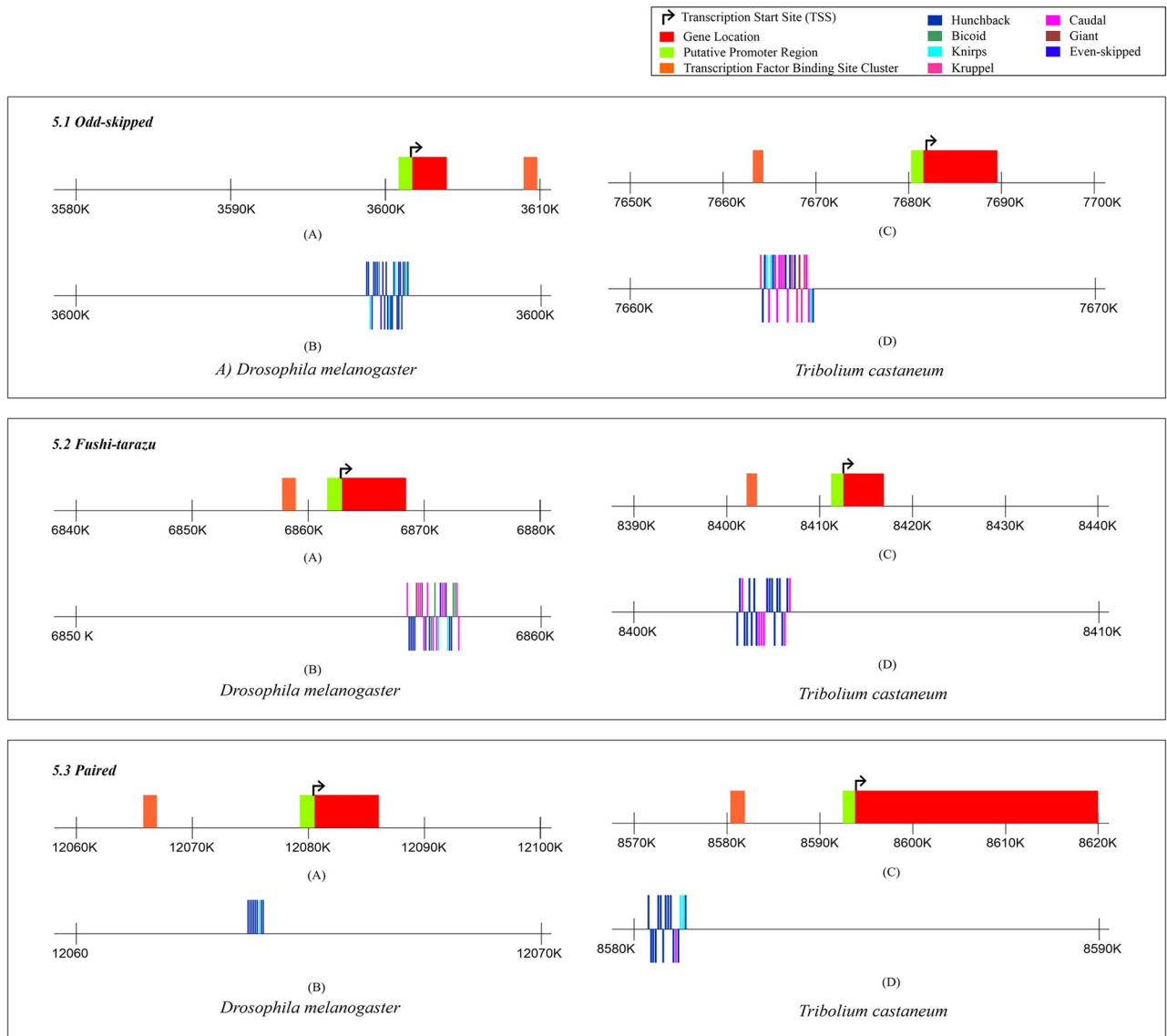
Fig 5. Figure showing the pCRM in *Drosophila melanogaster* and *Tribolium castaneum* for different transcription factors in *Odd-skipped*, *Fushi-tarazu*, and *Paired* genes. 5.1 shows the predicted transcription factor binding site clusters in *Odd-skipped* for *D. melanogaster* and *T. castaneum*, A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Odd-skipped* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Odd-skipped* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Odd-skipped* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Odd-skipped* for *T. castaneum*. 5.2 shows the predicted transcription factor binding site clusters in *Fushi-Tarazu* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Fushi-Tarazu* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Fushi-Tarazu* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Fushi-Tarazu* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Fushi-Tarazu* for *T. castaneum*. 5.3 shows the predicted transcription factor binding site clusters in *Paired* for *D. melanogaster* and *T. castaneum*. A) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Paired* gene in *D. melanogaster*, B) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Paired* for *D. melanogaster*, C) Locations of the gene, TSS and the TFBS cluster as predicted by the MCAST software for the *Paired* gene in *T. castaneum*, D) Binding sites for different transcription factors on the predicted TFBS cluster in the gene *Paired* for *T. castaneum*.

Table 3. Location of the transcription start sites and putative cis-regulatory modules and CRM's size as predicted by the MCAST software on the A-P patterning genes of *Drosophila melanogaster*. Here–and + represents the upstream and downstream location of predicted CRM respectively.

S. No.	Gene Name	Transcription Start Site	Predicted CRM's location (motif score of pCRM)	Predicted CRM's size	+,- location of pCRM
1	*Hunchback*	8690979	8684092–8684568 (64.46)	476 bp	-
2	*Knirps*	20692326	20694580–20695169 (84.08)	589 bp	-
3	*Caudal*	20770700	20777,675–20778,195 (28.18)	520 bp	+
4	*Kruppel*	25226609	25222746–25223355 (59.56)	609 bp	+
5	*Giant*	2427112	2412138–2412768 (44.80)	630 bp	-
6	*Tailless*	30852314	30842160–30842579 (44.68)	419 bp	-
7	*Even-skipped*	9979430	9974834–9975399 (42.35)	565 bp	-
8	*Odd-skipped*	3604995	3609367–3609952 (59.35)	585 bp	+
9	*Runt*	20694432	20685463–20686280 (73.21)	817 bp	-
10	*Hairy*	8675953	8665577–8666104 (85.88)	527 bp	-
11	*Fushi-tarazu*	6864324	6858324–6858811 (45.67)	487 bp	-
12	*Paired*	12082992	12067972–12068189 (18.52)	217 bp	-

https://doi.org/10.1371/journal.pone.0290035.t003

These interactions were evaluated using the STRING database. As predicted by the software, the average local clustering coefficient is 0.857 for *D. melanogaster and* 0.837 for *T. castaneum*. The PPI enrichment p-value is $< 1.0e\text{-}16$ for *D. melanogaster* and 3.28e-12 for *T. castaneum*. The interaction between the TFs in both *D. melanogaster* and *T. castaneum* shows more enrichment, which suggested that these TFs proteins are biologically connected as a group.

In *D. melanogaster*, the software predicted that the transcription factor bicoid shows co-expression with Hunchback. Hunchback shows protein homology with Kruppel. There was no co-expression of proteins detected in the *T. castaneum* and protein homology was identified between Hunchback and Kruppel as in *D. melanogaster*. The software depicted known experimentally determined interactions between all proteins except for Caudal, Knirps, and Even-skipped in both *D. melanogaster* and *T. castaneum*. No interaction between Caudal and Even-skipped proteins were predicted in both insects.

Discussion

The determination of the early body design in *D. melanogaster* is done by the action of cis-regulatory elements in the genome that regulate gene expression during development.

Table 4. Location of the transcription start sites and putative cis-regulatory modules and predicted CRM's size as predicted by the MCAST software on the A-P patterning genes of *Tribolium castaneum*.

S.No.	Gene Name	Transcription Start Site	Predicted CRM'S location (motif score of pCRM)	Predicted CRM'S size	+,- location of pCRM
1	*Hunchback*	8102130	8092849–8093278 (11.55)	429 bp	-
2	*Knirps*	4721907	4697483–4698242 (23.50)	759 bp	-
3	*Caudal*	2828217	2832700–2833237 (12.13)	537 bp	+
4	*Kruppel*	6495152	6502223–6502912 (23.50)	689 bp	+
5	*Giant*	3766076	3765042–3765574 (19.20)	532 bp	-
6	*Tailless*	11221703	11230486–11230843 (15.45)	357 bp	+
7	*Even-skipped*	624731	617466–617724 (19.58)	258 bp	-
8	*Odd-skipped*	7673308	7667718–7668180 (46.94)	462 bp	-
9	*Runt*	11253714	11243118–11243747 (18.58)	629 bp	-
10	*Hairy*	15627964	15635988–15636522 (26.15)	534 bp	+
11	*Fushi-tarazu*	8414663	8404780–8405724 (20.87)	944 bp	-
12	*Paired*	8598708	8580847–8581470 (28.38)	623 bp	-

https://doi.org/10.1371/journal.pone.0290035.t004

Table 5. Number of Transcription Factor binding sites on different A-P patterning genes in *Drosophila melanogaster* and *Tribolium castaneum* as predicted by the MCAST software. Here, *Drosophila melanogaster* and *Tribolium castaneum* are abbreviated as *Dm* and *Tc* respectively.

S.No.	Name of the A-P patterning gene	Bicoid		Hunchback		Caudal		Giant		Knirps		Kruppel		Even-skipped	
		Dm	*Tc*	*Dm*	*Tc*	*Dm*	*Tc*	*Dm*	*Tc*	*Dm*	*Tc*	*Dm*	*Tc*	*Dm*	*Tc*
1	*Hunchback*	6	-	10	5	-	4	3	-	4	-	1	-	-	-
2	*Knirps*	4	-	17	9	-	1	-	-	3	-	3	-	-	-
3	*Caudal*	8	-	9	8	6	3	-	-	-	-	-	-	-	-
4	*Kruppel*	6	-	11	11	-	3	2	-	7	-	6	-	-	-
5	*Giant*	2	-	9	5	1	5	1	-	3	-	9	1	3	-
6	*Tailless*	-	-	6	6	6	3	-	-	6	-	-	-	-	-
7	*Even-skipped*	-	-	9	10	-	-	-	-	7	3	-	-	-	-
8	*Hairy*	-	-	14	11	5	2	1	1	10	-	6	-	-	-
9	*Runt*	-	-	17	11	4	3	1	1	8	1	-	-	-	-
10	*Odd-skipped*	2	-	17	8	-	7	-	1	4	3	-	6	3	-
11	*Fushi-tarazu*	4	-	8	13	11	6	2	-	2	-	-	-	2	3
12	*Paired*	-	-	8	13	-	1	-	-	1	2	-	-	-	-

https://doi.org/10.1371/journal.pone.0290035.t005

Transcription factors (TFs) bind to these regions and regulate target gene expression [34]. Translation of localised maternal mRNAs during oogenesis constructs the initial TF gradients in the embryo at the top of the A-P patterning cascade [35, 36]. These maternal TFs then bind to gap gene specific embryonic regulatory regions, thereby regulating gap TF expression patterns during early A-P specification [9, 37]. The interaction between the different TFs was predicted using the STRING database. Based on computational modelling, it seems that these TFs are interconnected and have a role in coordinating the insect development. Previous studies suggested that the self-regulation pathway and the activities of other maternal-effect genes and gap genes control the expression of gap genes involved in the establishment of the A-P axis. [21]. In addition to the regulation of gap genes, these genes act as TFs for the regulation of genes involved in the A-P axis formation cascade which are pair-rule genes, segment polarity genes, and Homeotic selector genes [21, 38]. Consistent with the interaction identified by STRING software in the present study, previous studies on Bicoid and Hunchback have

Table 6. Number of the conserved, transition and transversional sites along with the transition/transversion ratio in the predicted pCRMs as depicted by the MEGA XI software.

S.NO.	Name of Gene	Number of conserved sites	Number of transition sites	Number of transversional sites	Transition/Transversional ratio
1	Hunchback	216	71	79	0.9
2	Knirps	267	134	178	0.8
3	Caudal	214	125	168	0.7
4	Kruppel	276	160	167	1.0
5	Giant	234	133	162	0.8
6	Tailless	134	80	111	0.7
7	Even-skipped	128	53	74	0.7
8	Odd-skipped	221	102	131	0.8
9	Runt	268	153	200	0.8
10	Hairy	201	133	184	0.7
11	Fushi-tarazu	271	78	137	0.6
12	Paired	84	31	102	0.3

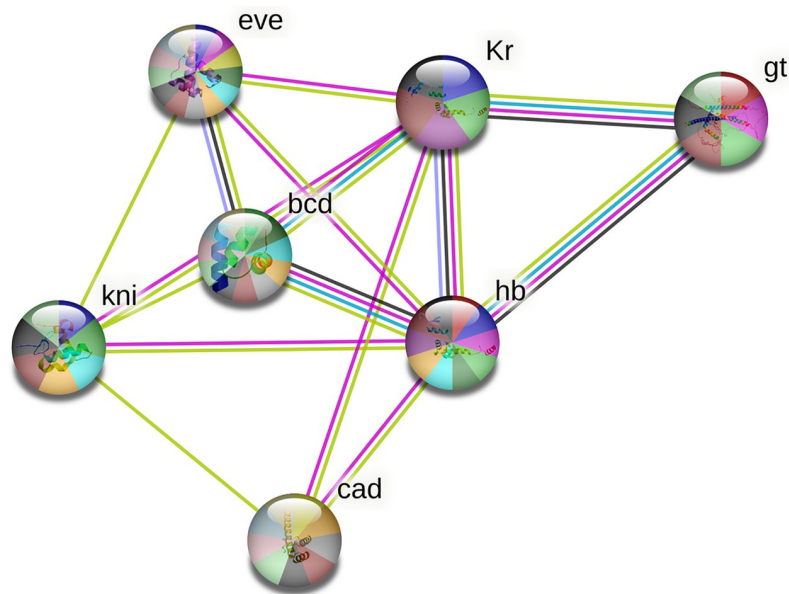https://doi.org/10.1371/journal.pone.0290035.t006

Fig 6. Protein-protein interactions between the different TFs which bind to the enhancers in the A-P patterning genes of *D. melanogaster*. Here in this figure, the pink colour represents the experimentally determined interactions, Sky blue shows the interactions curated from databases between the proteins. The blue colour node represents the gene co-occurrence. The black colour node depicts the co-expression of protein and the violet colour represents the protein homology. The yellow colour node represents the interactions that are predicted through text mining.

https://doi.org/10.1371/journal.pone.0290035.g006

revealed that these proteins have compatible binding with each other [39]. The Kruppel protein interacts with the Hunchback protein in order to repress the latter's expression and simultaneously regulate its own expression, as evidenced by previous research [40]. The findings of our study indicate that the proteins Hunchback and Kruppel exhibit homology, implying a shared evolutionary ancestry between these proteins. Gap TFs also control the downstream cascade of A-P Patterning genes [41, 42]. The binding of TFs to particular clusters of activator and repressor binding sites inside embryonic CRMs tightly controls gene expression patterns at each phase in the cascade. Individual CRMs have distinct molecular characteristics that influence transcriptional output. When a TF attaches to a CRM, it might behave as an activator or a repressor, depending on the situation [43, 44]. Hundreds of cis-regulatory motif sequences have been identified across all model species using both experimental [45–48] and bioinformatic [49–52] approaches, as well as the discovery of the related transcription factors binding to them [53]. The cluster of these transcription factor binding sites can be recognized as putative enhancers or CRM which help in the process of transcription. Many key regulators of early development have been identified courtesy of sophisticated genetic screening and the molecular biology and biochemistry of these factors, as well as their target sequences, have gained considerable interest in *Drosophila* [54, 55]. In-silico identification of CRM in early development in *D. melanogaster* and *D. pseudoobscura* were predicted and in-vivo testing was done on pCRMs [1, 56]. *Tribolium* is a good illustration of short germ embryogenesis in insects since it represents the ancestral kind of embryogenesis. In comparison to *Drosophila*, the blastoderm phase determines only the cephalic and thoracic segments, but not the abdominal segments. While the Bicoid gradient has been used to study pattern development in *Drosophila*, it is considered that this is not the case in *Tribolium* as *Bicoid* is not present in the insect. The alteration from *Caudal* activation of the *Hunchback* gap region to direct activation by *Bicoid* was an evolutionary shift from short to long germ embryogenesis [38, 57]. The
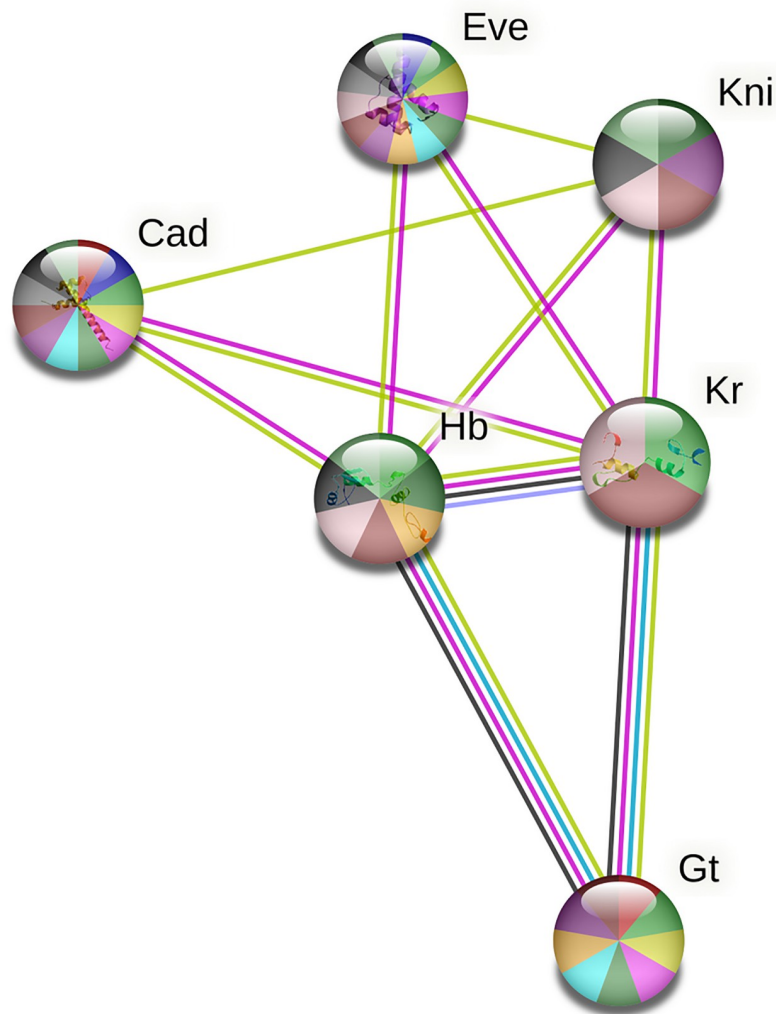
Fig 7. Protein-protein interactions between the different TFs which bind to the enhancers in the A-P patterning genes of *T. castaneum*. Here in this figure, the pink colour represents the experimentally determined interactions, Sky blue shows the interactions curated from databases between the proteins. The blue colour node represents the gene co-occurrence. The black colour node depicts the co-expression of protein and the violet colour represents the protein homology. The yellow colour node represents the interactions that are predicted through text mining.

https://doi.org/10.1371/journal.pone.0290035.g007

previous studies on A-P patterning comparison have been done on different species of *Drosophila* but there is no study which compares A-P patterning in different insect species belonging to different insect order [58]. Hence, the present study was performed to check whether the insects belonging to different orders, have different developmental patterns but have similar genes, and are controlled by similar cis-regulatory modules or not.

In *D. melanogaster*, the zygotic *Hunchback* gene is activated by the synergetic interaction between the Hunchback and Bicoid transcription factors [59–61], the result predicted by the present study also depicted a similar result. The cluster predicted by MCAST has the maximum number of Hunchback transcription factor sites in both *D. melanogaster* and *T. castaneum*. As mentioned above, in *Tribolium* the *bicoid* gene is absent therefore *Tribolium* has additional *caudal* sites which help in *Hunchback* gene activation and expression. In the case of *Knirps*, the maximum number of transcription factors binding sites again are of *Hunchback* in

addition to other factors in both *D. melanogaster* and *T. castaneum*. Previous *Drosophila* studies suggest that the *Hunchback* acts as a repressor for *Knirps*, as binding of the *Hunchback* suppresses the *Knirps* in the anterior half of the embryo [62, 63] *Caudal* gene is one of the most studied gap genes in *D. melanogaster* and *T. castaneum*, which help in the activation of gap genes and pair-rule genes in insects. *Caudal* is known to be a downstream core promoter element [39] and our results also depict the same. Both in *D. melanogaster* and *T. castaneum* the cluster is found to be downstream to the transcription start site. Both clusters have TFBS for Hunchback and Caudal, while *Drosophila* has additional sites for Bicoid also.

The previous studies of the Kruppel gene suggest that the *Kruppel* has TFBS for Hunchback, Bicoid, Giant, and Knirps [62, 64]. *Hunchback* and *Bicoid* are known as activators of the *Kruppel* gene, while *Knirps* and *Giant are* known to be their repressors [62]. The present study result has also predicted similar binding sites in *D. melanogaster* as in evident in previous studies while *T. castaneum* has TFBS clusters for Hunchback and Caudal only. For the expression of *Giant* gene in *Drosophila*, Hunchback functions as a concentration-dependent repressor of *Giant*, suppressing its most anterior expression [65–67]. Giant will not be transcribed in the anterior domain without the presence of *Bicoid* [65]. Activation in the posterior domain necessitates the combined actions of the *Caudal* and *Bicoid* [65, 68]. Similar TFBS was found in *D. melanogaster* and *T. castaneum*. In addition to these predicted pCRMs, binding sites for Kruppel and Knirps were predicted in *D. melanogaster* while only one Kruppel site was predicted in *T. castaneum*. These results suggest that the gap genes transcription in *D. melanogaster* is mainly controlled by Hunchback, Bicoid, and Caudal proteins while in *T. castaneum*, Hunchback and Caudal are majorly the transcription factors for the gap genes.

Gap genes with maternal gradients act across shorter distances and overlap at their borders, resulting in the seven-stripe expression patterns of the pair-rule genes [69, 70] *Even-skipped* is one of the most extensively studied pair-rule gene in *D. melanogaster*. Previous studies documented the individual *Even-skipped* stripe responds to the different gradients and combinations of gap gene transcription factors [43, 71–73]. The present study predicts the gene has TFBS clusters for Hunchback and Knirps only in both *D. melanogaster* and *T. castaneum*, this combination in previous studies was found to influence the stripe 3+7 enhancer in *Drosophila melanogaster* [74]. In the earlier *D. melanogaster* study, it was found that the gap genes mainly *Hunchback*, *Giant*, *Kruppel*, and *Knirps* act as a repressor for the *Runt*, which is a primary pair-rule gene [75]. Similar TFBS were predicted in the present study, with the exception that no Kruppel binding site was found in both *D. melanogaster* and *T. castaneum*. In addition to these known TFs, additional binding sites for Caudal were also found in both *D. melanogaster* and *T. castaneum*. The gap genes mainly, *Hunchback*, *Knirps*, and *Kruppel* are known to influence the expression of pair-rule genes [76, 77]. The *hairy* gene also has similar TFBS, which is alike to the cluster predicted in *D. melanogaster*. The cluster have TFBS for Hunchback, Knirps, and Kruppel in addition to these, binding sites for Caudal and one Giant transcription factor were also predicted. In contrast to *D. melanogaster*, hairy gene in *T. castaneum* showed binding sites for Hunchback, Caudal and Giant site. The *Hairy* functions in the trunk and head segmentation in *D. melanogaster*. A previous study on *Hairy* gene in *T. castaneum* suggests that the gene functions only during trunk segmentation in *T. castaneum* and is non-functional during the head segmentation pathway [78]. The *Fushi-tarazu*, *Odd-skipped*, and *Paired* are regarded as secondary pair rule genes in *D. melanogaster* [3, 79]. These genes are known to have Hunchback Kruppel, Knirps and Giant as repressor factors in different combinations [80]. The MCAST analysis revealed binding sites for Hunchback, Caudal and Even-skipped factors in the secondary pair rule genes for both *D. melanogaster* and *T. castaneum*. In addition, *Drosophila Fushi-tarazu* gene showed binding sites for bicoid and giant also. As far as the *Odd skipped* gene is concerned, both *D. melanogaster* and *T. castaneum* have TFBS for Hunchback and Knirps. The gene in *D.*

melanogaster also showed binding sites for Bicoid and Even-skipped. The *Odd skipped* gene in *T. castaneum* showed binding sites for Kruppel and Caudal also. *Paired* being a secondary pair-rule gene, is controlled by the primary pair-rule genes in *D. melanogaster* and *T. castaneum* [81]. Most of the Predicted Cis-regulatory elements of the pair rule genes have TFBS for the Hunchback, Bicoid, Caudal, Knirps, and Kruppel in *D. melanogaster*. However, in *T. castaneum*, most pCRMs have TFBS for hunchback, knirps, caudal, and kruppel. The size of the predicted Cis-regulatory elements is between 200 bp to 850 bp for *D. melanogaster* and 240 bp– 950 bp for *T. castaneum*. The results of the MCAST analysis in the present study suggest that most of the transcription factors which control the A-P patterning cascade are conserved in *D. melanogaster* and *T. castaneum* with the exception that there are no binding sites for Bicoid in *T. castaneum*.

Transitions (Ti) are referred to as pyrimidine- or purine-based A-G or C-T switching. A transversion (Tv) is the exchange of two-ring purine nucleobases for one-ring pyrimidine bases. There are four conversion possibilities: A-C, A-T, C-G, and G-T. From the last decade, the Ti/Tv ratio has been employed as a significant metric for the reconstruction of phylogenetic trees and the calculation of divergence. Even the High-throughput sequencing studies employ the Ti/Tv ratio as a quality control measure. Assuming that there are two potential transitions and four possible transversions, the Ti/Tv ratio, which divides the number of transition SNPs by the number of transversion SNPs, should equal 0.5 if replacement variations occur at random. Nonetheless, a transversion is considered a more important change than a transition since it requires more energy than replacement without affecting the ring structure. Hence, the transition and transversion ratio is frequently more than 0.5 in actual sequencing data [82–85]. Studies also suggest that the Tv's have more significant effects on regulatory DNA, such as TF binding motif studies and allele-specific TF binding [86]. Keeping in mind, the importance of Ttransitions and transversion, the present study also evaluated the number of Ti, Tv sites, and Ti/Tv ratio in the predicted enhancers of A-P patterning genes. The Tv sites were found to be more than the Ti sites and the average Ti/Tv ratio was 0.75 as given in Table 5. As the Tv is most likely to affect the amino acid sequence than the Ti, the more Tv can be indicative of a large number of variations, which will ultimately affect gene expression [81–85].

Conclusion

This study marks the first-ever attempt to conduct an integrated examination of the location, size, and composition of clusters of transcription factor binding sites (TFBS) within the cis-regulatory elements of multiple anterior-posterior (A-P) patterning genes that exhibit orthology to gene sequences found in *Drosophila*. The present investigation has revealed that comparable transcription factors (TFs) could regulate the expression of anterior-posterior (A-P) patterning genes in Diptera and Coleoptera taxa of insects. The majority of transcription factors (TFs) were observed to be situated upstream of the transcription start site (TSS), although a subset were also identified downstream of the TSS. In both *Drosophila melanogaster* and *Tribolium castaneum*, the Hunchback transcription factor binding site (TFBS) exhibited the highest frequency among all identified TFBS. The present study contributes to the advancement of our understanding regarding the evolutionary patterns of genes and cis-regulatory elements in two distinct orders of insects. Subsequent validation of these findings may be achieved through in-vitro and in-vivo experimentation.

Acknowledgments

Author Contributions

Conceptualization: Anshika Moudgil.

Data curation: Anshika Moudgil.

Formal analysis: Tejinder Kaur.

Methodology: Anshika Moudgil.

Supervision: Tejinder Kaur.

Writing – original draft: Anshika Moudgil.

Writing – review & editing: Ranbir Chander Sobti, Tejinder Kaur.

References

1. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci USA. 2002; 99: 757–762. https://doi.org/10.1073/pnas.231608898 PMID: 11805330

2. Davidson EH. Genomic regulatory systems: development and evolution. San Diego: Academic Press; 2001.

3. Dillon N, Sabbattini P. Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. Bioessays. 2000; 22: 657–665. https://doi.org/10.1002/1521-1878(200007)22:7<657::AID-BIES8>3.0.CO;2-2 PMID: 10878578

4. Davidson E. The regulatory genome - 1st edition. 2006 [cited 16 May 2023]. https://www.elsevier.com/books/the-regulatory-genome/davidson/978-0-12-088563-3

5. Ni P, Wilson D, Su Z. A map of cis-regulatory modules and constituent transcription factor binding sites in 80% of the mouse genome. BMC Genomics. 2022; 23: 714. https://doi.org/10.1186/s12864-022-08933-7 PMID: 36261804

6. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003; 424: 147–151. https://doi.org/10.1038/nature01763 PMID: 12853946

7. Ingham PW. The molecular genetics of embryonic pattern formation in *Drosophila*. Nature. 1988; 335: 25–34. https://doi.org/10.1038/335025a0 PMID: 2901040

8. Driever W, Thoma G, Nüsslein-Volhard C. Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the bicoid morphogen. Nature. 1989; 340: 363–367. https://doi.org/10.1038/340363a0 PMID: 2502714

9. Struhl G, Struhl K, Macdonald PM. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. Cell. 1989; 57: 1259–1273. https://doi.org/10.1016/0092-8674(89)90062-7 PMID: 2567637

10. Jaeger J. The gap gene network. Cell Mol Life Sci. 2011; 68: 243–274. https://doi.org/10.1007/s00018-010-0536-y PMID: 20927566

11. Johnston DS, Nüsslein-Volhard C. The origin of pattern and polarity in the *Drosophila* embryo. Cell. 1992; 68: 201–219. https://doi.org/10.1016/0092-8674(92)90466-P PMID: 1733499

12. Reinitz J, Sharp DH. Mechanism of *eve* stripe formation. Mechanisms of Development. 1995; 49: 133–158. https://doi.org/10.1016/0925-4773(94)00310-J PMID: 7748785

13. Small S, Kraut R, Hoey T, Warrior R, Levine M. Transcriptional regulation of a pair-rule stripe in *Drosophila*. Genes Dev. 1991; 5: 827–839. https://doi.org/10.1101/gad.5.5.827 PMID: 2026328

14. Rivera-Pomar R, Jäckle H. From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. Trends in Genetics. 1996; 12: 478–483. https://doi.org/10.1016/0168-9525(96)10044-5 PMID: 8973159

15. Langeland JA, Carroll SB. Conservation of regulatory elements controlling *hairy* pair-rule stripe formation. Development. 1993; 117: 585–596. https://doi.org/10.1242/dev.117.2.585 PMID: 8330529

16. Lukowitz W, Schröder C, Glaser G, Hülskamp M, Tautz D. Regulatory and coding regions of the segmentation gene *hunchback* are functionally conserved between *Drosophila virilis* and *Drosophila melanogaster*. Mechanisms of Development. 1994; 45: 105–115. https://doi.org/10.1016/0925-4773(94)90024-8 PMID: 8199047

17. Ludwig MZ, Patel NH, Kreitman M. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development. 1998; 125: 949–958. https://doi.org/10.1242/dev.125.5.949 PMID: 9449677

18. Papatsenko DA, Makeev VJ, Lifanov AP, Régnier M, Nazina AG, Desplan C. Extraction of Functional Binding Sites from Unique Regulatory Regions: The *Drosophila* Early Developmental Enhancers. Genome Res. 2002; 12: 470–481. https://doi.org/10.1101/gr.212502 PMID: 11875036

19. Berg OG, Von Hippel PH. Selection of DNA binding sites by regulatory proteins. Journal of Molecular Biology. 1987; 193: 723–743. https://doi.org/10.1016/0022-2836(87)90354-8 PMID: 3612791

20. Fickett JW. Finding genes by computer: the state of the art. Trends in Genetics. 1996; 12: 316–320. https://doi.org/10.1016/0168-9525(96)10038-x PMID: 8783942

21. Kimelman D, Martin BL. Anterior-posterior patterning in early development: three strategies: Anterior-posterior patterning in early development. WIREs Dev Biol. 2012; 1: 253–266. https://doi.org/10.1002/wdev.25 PMID: 23801439

22. Sobti RC, editor. Advances in animal experimentation and modeling: understanding life phenomena. London, United Kingdom: Elsevier, Academic Press, an Imprint of Elsevier; 2022.

23. Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, et al. FlyBase: updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Research. 2021; 49: D899–D907. https://doi.org/10.1093/nar/gkaa1026 PMID: 33219682

24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

25. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2019; gkz1001. https://doi.org/10.1093/nar/gkz1001 PMID: 31701148

26. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43: W39–W49. https://doi.org/10.1093/nar/gkv416 PMID: 25953851

27. Bailey TL, Noble WS. Searching for statistically significant regulatory modules. Bioinformatics. 2003; 19: ii16–ii25. https://doi.org/10.1093/bioinformatics/btg1054 PMID: 14534166

28. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res. 2021; 31: 159–169. https://doi.org/10.1101/gr.266932.120 PMID: 33239395

29. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Research. 2022; 50: D988–D995. https://doi.org/10.1093/nar/gkab1049 PMID: 34791404

30. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 1994; 22: 4673–4680. https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417

31. Hall TA. BIOEDIT: A USER-FRIENDLY BIOLOGICAL SEQUENCE ALIGNMENT EDITOR AND ANALYSIS PROGRAM FOR WINDOWS 95/98/ NT. 1999.

32. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Battistuzzi FU, editor. Molecular Biology and Evolution. 2021; 38: 3022–3027. https://doi.org/10.1093/molbev/msab120 PMID: 33892491

33. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research. 2021; 49: D605–D612. https://doi.org/10.1093/nar/gkaa1074 PMID: 33237311

34. Howard ML, Davidson EH. cis-Regulatory control circuits in development. Developmental Biology. 2004; 271: 109–118. https://doi.org/10.1016/j.ydbio.2004.03.031 PMID: 15196954

35. Berleth T, Burri M, Thoma G, Bopp D, Richstein S, Frigerio G, et al. The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. The EMBO Journal. 1988; 7: 1749–1756. https://doi.org/10.1002/j.1460-2075.1988.tb03004.x PMID: 2901954

36. Steward R, Zusman SB, Huang LH, Schedl P. The dorsal protein is distributed in a gradient in early *Drosophila* embryos. Cell. 1988; 55: 487–495. https://doi.org/10.1016/0092-8674(88)90035-9 PMID: 2460244

37. Driever W, Nüsslein-Volhard C. The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. Cell. 1988; 54: 95–104. https://doi.org/10.1016/0092-8674(88)90183-3 PMID: 3383245

38. Rudolf H, Zellner C, El-Sherif E. Speeding up anterior-posterior patterning of insects by differential initialization of the gap gene cascade. Developmental Biology. 2020; 460: 20–31. https://doi.org/10.1016/j.ydbio.2019.04.015 PMID: 31075221

39. Juven-Gershon T, Hsu J-Y, Kadonaga JT. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. Genes Dev. 2008; 22: 2823–2830. https://doi.org/10.1101/gad.1698108 PMID: 18923080

40. Treisman J, Desplan C. The products of the *Drosophila* gap genes *hunchback* and *Krüppel* bind to the hunchback promoters. Nature. 1989; 341: 335–337. https://doi.org/10.1038/341335a0 PMID: 2797150

41. Qian S, Capovilla M, Pirrotta V. The bx region enhancer, a distant cis-control element of the *Drosophila Ubx* gene and its regulation by *hunchback* and other segmentation genes. The EMBO Journal. 1991; 10: 1415–1425. https://doi.org/10.1002/j.1460-2075.1991.tb07662.x PMID: 1902784

42. Štanojević D, Hoey T, Levine M. Sequence-specific DNA-binding activities of the gap proteins encoded by *hunchback* and *Krüppel* in *Drosophila*. Nature. 1989; 341: 331–335. https://doi.org/10.1038/341331a0 PMID: 2507923

43. Tony Ip Y, Kraut R, Levine M, Rushlow CA. The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in drosophila. Cell. 1991; 64: 439–446. https://doi.org/10.1016/0092-8674(91)90651-e PMID: 1988156

44. Small S, Blair A, Levine M. Regulation of Two Pair-Rule Stripes by a Single Enhancer in the *Drosophila* Embryo. Developmental Biology. 1996; 175: 314–324. https://doi.org/10.1006/dbio.1996.0117 PMID: 8626035

45. Agostini F, Cirillo D, Ponti RD, Tartaglia GG. SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. BMC Genomics. 2014; 15: 925. https://doi.org/10.1186/1471-2164-15-925 PMID: 25341390

46. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research. 2006; 34: W369–W373. https://doi.org/10.1093/nar/gkl198 PMID: 16845028

47. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 2010; 20: 861–873. https://doi.org/10.1101/gr.100552.109 PMID: 20378718

48. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Research. 2008; 36: 5221–5231. https://doi.org/10.1093/nar/gkn488 PMID: 18684996

49. Luehr S, Hartmann H, Soding J. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. Nucleic Acids Research. 2012; 40: W104–W109. https://doi.org/10.1093/nar/gks602 PMID: 22693218

50. Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003; 5: 201. https://doi.org/10.1186/gb-2003-5-1-201 PMID: 14709165

51. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Reports. 2014; 8: 2015–2030. https://doi.org/10.1016/j.celrep.2014.08.019 PMID: 25220462

52. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005; 23: 137–144. https://doi.org/10.1038/nbt1053 PMID: 15637633

53. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014; 158: 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009 PMID: 25215497

54. Kapil S, Sobti RC, Kaur T. Prediction and analysis of cis-regulatory elements in Dorsal and Ventral patterning genes of *Tribolium castaneum* and its comparison with *Drosophila melanogaster*. Mol Cell Biochem. 2023 [cited 9 Apr 2023]. https://doi.org/10.1007/s11010-023-04712-4 PMID: 37004638

55. Lewis EB. A gene complex controlling segmentation in *Drosophila*. Nature. 1978; 276: 565–570. https://doi.org/10.1038/276565a0 PMID: 103000

56. Nüsslein-Volhard C, Wieschaus E. Mutations affecting segment number and polarity in *Drosophila*. Nature. 1980; 287: 795–801. https://doi.org/10.1038/287795a0 PMID: 6776413

57. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol. 2004; 5: R61. https://doi.org/10.1186/gb-2004-5-9-r61 PMID: 15345045

58. Wolff C, Schröder R, Schulz C, Tautz D, Klingler M. Regulation of the *Tribolium* homologues of *caudal* and *hunchback* in *Drosophila*: evidence for maternal gradient systems in a short germ embryo. Development. 1998; 125: 3645–3654. https://doi.org/10.1242/dev.125.18.3645 PMID: 9716530

59. Wunderlich Z, Fowlkes CC, Eckenrode KB, Bragdon MDJ, Abiri A, DePace AH. Quantitative comparison of the anterior-posterior patterning system in the embryos of five *Drosophila* species. G3 Genes|Genomes|Genetics. 2019; 9: 2171–2182. https://doi.org/10.1534/g3.118.200953 PMID: 31048401

60. Simpson-Brose M, Treisman J, Desplan C. Synergy between the *hunchback* and *bicoid* morphogens is required for anterior patterning in *Drosophila*. Cell. 1994; 78: 855–865. https://doi.org/10.1016/S0092-8674(94)90622-X PMID: 8087852

61. Sauer F, Rivera-Pomar R, Hoch M, Jäckle H. Gene regulation in the *Drosophila* embryo. Phil Trans R Soc Lond B. 1996; 351: 579–587. https://doi.org/10.1098/rstb.1996.0057 PMID: 8735281

62. Gilbert SF. The origins of anterior-posterior polarity. Developmental Biology 6th edition. 2000 [cited 12 Jun 2023]. https://www.ncbi.nlm.nih.gov/books/NBK10039/

63. Pankratz MJ, Busch M, Hoch M, Seifert E, Jäckle H. Spatial Control of the Gap Gene *knirps* in the *Drosophila* Embryo by Posterior Morphogen System. Science. 1992; 255: 986–989. https://doi.org/10.1126/science.1546296 PMID: 1546296

64. Hoch M, Gerwin N, Taubert H, Jäckle H. Competition for Overlapping Sites in the Regulatory Region of the *Drosophila* Gene *Krüppel*. Science. 1992; 256: 94–97. https://doi.org/10.1126/science.1348871 PMID: 1348871

65. Eldon ED, Pirrotta V. Interactions of the *Drosophila* gap gene giant with maternal and zygotic pattern-forming genes. Development. 1991; 111: 367–378. https://doi.org/10.1242/dev.111.2.367 PMID: 1716553

66. Kraut R, Levine M. Mutually repressive interactions between the gap genes *giant* and *Kruppel* define middle body regions of the *Drosophila* embryo. Development. 1991; 111: 611–621. https://doi.org/10.1242/dev.111.2.611 PMID: 1893878

67. Struhl G, Johnston P, Lawrence PA. Control of *Drosophila* body pattern by the *hunchback* morphogen gradient. Cell. 1992; 69: 237–249. https://doi.org/10.1016/0092-8674(92)90405-2 PMID: 1568245

68. Rivera-Pomar R, Lu X, Perrimon N, Taubert H, Jäckle H. Activation of posterior gap gene expression in the *Drosophila* blastoderm. Nature. 1995; 376: 253–256. https://doi.org/10.1038/376253a0 PMID: 7617036

69. Frasch M, Hoey T, Rushlow C, Doyle H, Levine M. Characterization and localization of the even-skipped protein of *Drosophila*. The EMBO Journal. 1987; 6: 749–759. https://doi.org/10.1002/j.1460-2075.1987.tb04817.x PMID: 2884106

70. Macdonald PM, Ingham P, Struhl G. Isolation, structure, and expression of *even-skipped*: A second pair-rule gene of *Drosophila* containing a homeo box. Cell. 1986; 47: 721–734. https://doi.org/10.1016/0092-8674(86)90515-5 PMID: 2877745

71. Fujioka M, Emi-Sarker Y, Yusibova GL, Goto T, Jaynes JB. Analysis of an *even-skipped* rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients*. Development. 1999; 126: 2527–2538. https://doi.org/10.1242/dev.126.11.2527 PMID: 10226011

72. Goto T, Macdonald P, Maniatis T. Early and late periodic patterns of *even skipped* expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell. 1989; 57: 413–422. https://doi.org/10.1016/0092-8674(89)90916-1 PMID: 2720776

73. Small S, Blair A, Levine M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. The EMBO Journal. 1992; 11: 4047–4057. https://doi.org/10.1002/j.1460-2075.1992.tb05498.x PMID: 1327756

74. Struffi P, Corado M, Kaplan L, Yu D, Rushlow C, Small S. Combinatorial activation and concentration-dependent repression of the *Drosophila even skipped* stripe 3+7 enhancer. Development. 2011; 138: 4291–4299. https://doi.org/10.1242/dev.065987 PMID: 21865322

75. Klinger M, Gergen JP. Regulation of runt transcription by *Drosophila* segmentation genes. Mechanisms of Development. 1993; 43: 3–19. https://doi.org/10.1016/0925-4773(93)90019-T PMID: 8240970

76. Carroll SB, Laymon RA, McCutcheon MA, Riley PD, Scott MP. The localization and regulation of Antennapedia protein expression in *Drosophila* embryos. Cell. 1986; 47: 113–122. https://doi.org/10.1016/0092-8674(86)90372-7 PMID: 3093083

77. Frasch M, Levine M. Complementary patterns of *even-skipped and fushi tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. Genes Dev. 1987; 1: 981–995. https://doi.org/10.1101/gad.1.9.981 PMID: 2892761

78. Aranda M, Marques-Souza H, Bayer T, Tautz D. The role of the segmentation gene hairy in *Tribolium*. Dev Genes Evol. 2008; 218: 465–477. https://doi.org/10.1007/s00427-008-0240-1 PMID: 18679713

79. Nasiadka A, Dietrich BH, Krause HM. Anterior-posterior patterning in the *Drosophila* embryo. Advances in Developmental Biology and Biochemistry. Elsevier; 2002. pp. 155–204. https://doi.org/10.1016/S1569-1799(02)12027-2

80. Schroeder MD, Greer C, Gaul U. How to make stripes: deciphering the transition from non-periodic to periodic patterns in *Drosophila* segmentation. Development. 2011; 138: 3067–3078. https://doi.org/10.1242/dev.062141 PMID: 21693522

81. Choe CP, Miller SC, Brown SJ. A pair-rule gene circuit defines segments sequentially in the short-germ insect *Tribolium castaneum*. Proc Natl Acad Sci USA. 2006; 103: 6560–6564. https://doi.org/10.1073/pnas.0510440103 PMID: 16611732

82. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061–1073. https://doi.org/10.1038/nature09534 PMID: 20981092

83. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Briefings in Bioinformatics. 2014; 15: 879–889. https://doi.org/10.1093/bib/bbt069 PMID: 24067931

84. Wang GT, Peng B, Leal SM. Variant Association Tools for Quality Control and Analysis of Large-Scale Sequence and Genotyping Array Data. The American Journal of Human Genetics. 2014; 94: 770–783. https://doi.org/10.1016/j.ajhg.2014.04.004 PMID: 24791902

85. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015; 31: 318–323. https://doi.org/10.1093/bioinformatics/btu668 PMID: 25297068

86. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, et al. Transversions have larger regulatory effects than transitions. BMC Genomics. 2017; 18: 394. https://doi.org/10.1186/s12864-017-3785-4 PMID: 28525990