BTCH90009

Assignment 2: Generating hypothesis through genomics and bioinformatics

Name: Jiayu Wang

Student ID: 1039580

Word Count: 1186

Introduction

Drosophila melanogaster is a member of *Drosophila* family and is commonly used as a model organism for genetic and genomic studies [1]. *Drosophila melanogaster* is a typical holometabolous insect species which means it needs to experience a complete metamorphosis as shown in **Fig 1**, which needs to be considered when studying it [2]. Recently, the government is aiming to eradicate *Drosophila melanogaster* since it has become a harmful pest. To achieve this, three chemical inhibitors have been developed that respectively target proteins encoded by gene CG5939, CG2302, and CG3757. These inhibitors have been proven effective *in vitro*, hence this research aims to study the potential effectiveness, specificity and accessibility of the inhibitors *in vivo* and nominate the best one for global deployment.

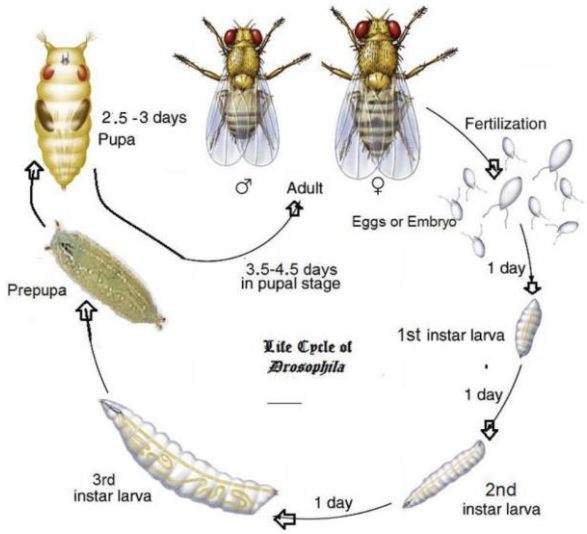

Fig 1. Metamorphosis cycle of *Drosophila* [2]

Methods

FlyBase

Flybase is a comprehensive genetic and genomic database specifically for model organism *Drosophila melanogaster* that includes many different tools and comprises millions of references [3]. We used FlyBase and its G Browser tool to extract information about functions, expression and homology of the candidate genes. We also used several key tool links on FlyBase webpage, like DIOPT to perform encoded-protein alignment and ModENCODE to find expression information[4, 5].

UniprotKB

UniprotKB is a comprehensive hub of protein information collection [6]. We used Uniprot to browse protein-level annotation about these genes.

NCBI-BLAST and accessed databases

NCBI-BLAST is a protein alignment tool that was used to find protein similarity between orthologs in this research [7]. We chose non-redundant (NR) database to have a more comprehensive alignment search [8].

Results

Specificity of the three candidate genes

(i) Similarity between the three candidate genes and their human orthologs

As shown in **Table 1**, gene CG2302 shows a higher homology with its human orthologs while CG3757 shows the lowest homology and has only one ortholog. Similarities between these three genes and their human orthologs are all relatively low, which means the inhibitors against the encoded proteins of these three genes are less possible to affect humans. Overall, gene CG3757 shows the best specificity here.

Table 1. Information about three candidate genes and their human orthologs

Gene	Number of human ortholog(s)	The highest score among human orthologs	encoded-protein alignment against the highest-scored human orthologs by DIOPT (Identity, Similarity)
CG5939	20	3 of 15	36%, 60%
CG2302	21	9 of 15	32%, 43%
CG3757	1	1 of 15	21%, 34%

(ii) Similarity between the three candidate genes and their homologs in *Drosophila* species

There are 13 *Drosophila* species in total, among which the two closest related species to *Drosophila melanogaster* are *Drosophila simulans* and *Drosophila sechellia.* As shown in **Table 2**, all three candidate genes have orthologs nearly all other *Drosophila* species with close to 100% identities and high BLAST scores, though **Table 2** only exhibits alignment results against the two closest related species. This indicates a potential risk of off-targeting of the inhibitors to the other *Drosophila* species.

Table 2. Information about three candidate genes and their orthologs in *Drosophila* species

Gene	Number of *Drosophila* species that have ortholog(s)	Encoded-protein alignment against the highest-scored orthologs of the two closest related species (Identity, Scores) by BLASTP	
		Drosophila simulans	***Drosophila sechellia***
CG5939	11	99.68%, 1217	99.77%, 1707
CG2302	10	99.75%, 1658	99.37%, 1647
CG3757	11	99.45%, 1115	99.42%, 714

(iii) Similarity between the three candidate genes and their paralogs

As shown in **Table 3**, all these three candidate genes have paralogs in genomes. The highest homologous score among these paralogs is that of gene CG2302 at 8 of 10. For the other two genes, the similarities between them and their paralogs are not significant. Besides, similarities between the encoded protein of the three candidate genes and their corresponding paralogs are all not significant with identities at 26% - 57%. This indicates less possibility of these paralogs making up for the inhibition by gene redundancy.

Table 3. Information about three candidate genes and their paralogs

Gene	Number of paralog(s)	The highest score among paralog(s)	encoded-protein alignment against the highest-scored human orthologs by DIOPT (Gene symbol, Identity, Similarity)
CG5939	13	3 of 10	*Zip*, 38%, 60%
CG2302	22	8 of 10	*nAChRα3*, 44%, 54%; *nAChRα4*, 57%, 70%; *nAChRβ2*, 54%, 70%
CG3757	16	3 of 10	*yellow-b*, 37%, 59%; *yellow-c*, 40%, 60%; *yellow-d2*, 28%, 48%; *yellow-d*, 29%, 47%; *yellow-e2*, 28%, 47%; *yellow-e3*, 30%, 45%; *yellow-e*, 26%, 42%;

Effectiveness of the three candidate genes

The lethality of alleles, phenotypes, molecular function, biological process and cellular component are presented below in **Table 4**. The molecular function of gene CG5935, unlike that of the other two genes, is not based on experimental evidence but based on only prediction. Remarkably, missing and mutation of gene CG5935 is lethal, which indicates a more fundamental function for the survival of *Drosophila melanogaster* compared with the other two genes. Gene CG2303 is part of the acetylcholine-gated channel complex and is related to cation transport, though mutation on this gene is not necessarily lethal. Gene CG3757 is related to body pigmentation and might cause both genders fertile if mutant. Overall, we can rank the essentiality of these three candidate genes from high to low as CG5935, CG2303, and then CG3757.

Table 4. Gene ontology of the three candidate genes

Genes	Lethality of alleles	Other phenotypes except for lethality	Ontology		
			Molecular function	**Biological process**	**Cellular Component**
CG5935	Lethal/ Partially lethal/ Recessively lethal	Abnormal flight/neuro-physiology	Motor actitivity (predicted)	Mesoderm development, Myofibril assembly	M band, striated muscle myosin thick filament
CG2303	Variable	Abnormal sleep	acetylcholine receptor activity, acetylcholine-gated cation-selective channel activity	cation transport	Part of acetylcholine-gated channel complex
CG3757	Variable	fertile, abnormal behavior and body color	NOT enables dopachrome isomerase activity	pigmentation, male courtship and mating behavior, veined wing extension	Cell hair, cytoplasm, extracellular region

Accessibility of the three candidate genes

(i) Anatomical location of expression

As shown in **Fig.2**, gene CG5935 shows a high expression in adult heads and the circulatory system in adult hearts and relatively high expression adult digestive system and accessory gland of male adults. Gene CG2303 shows low to moderate expression in adult heads and the nervous system in lamina monopolar neuron L2. Gene CG3757 shows very low to moderate expression in the sensory system, digestive system, and integumentary system. Overall, all three genes have easily accessed expression locations while gene CG5935 has the highest expression level.

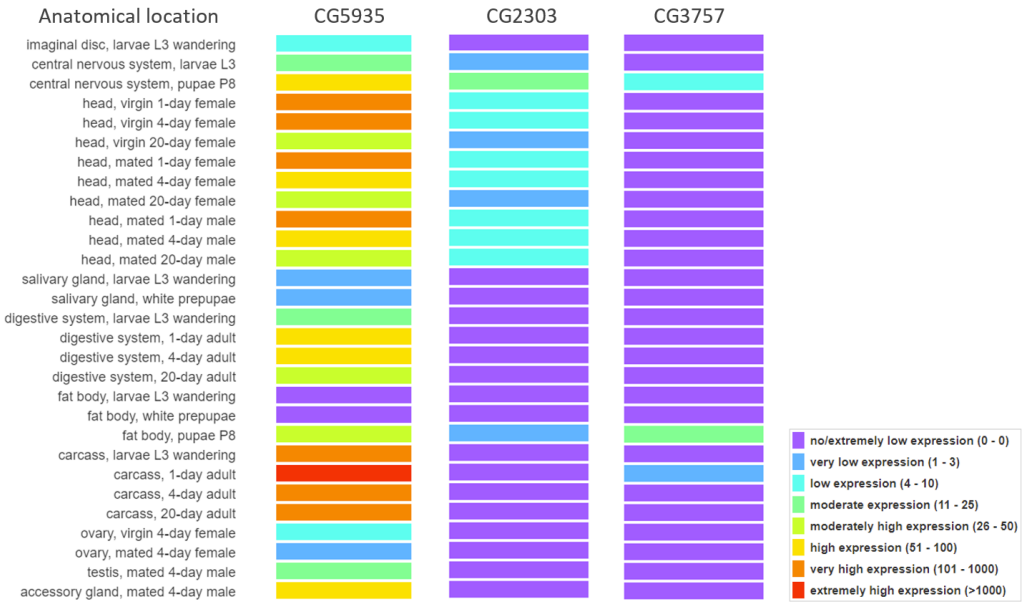

Fig.2 Heatmap of the expression level of the three candidate genes against different anatomical locations

(ii) Life stages of expression

As shown in **Fig. 3**, gene CG5935 shows relatively significant expression nearly throughout the entire life except for the first few hours in embryo stages. Gene CG2303 has a very low expression only during late embryo, early larva, entire pupa, early female adult and entire male adult stages. Gene 3757 shows moderate to moderately high expression during late embryo, late larva, late pupa stages. Overall, gene CG5935 shows a much longer duration of expression compared with the other two genes. Also, for gene CG2302 and CG3757, expression during the pupa stage might be less accessible due to the hard shell.

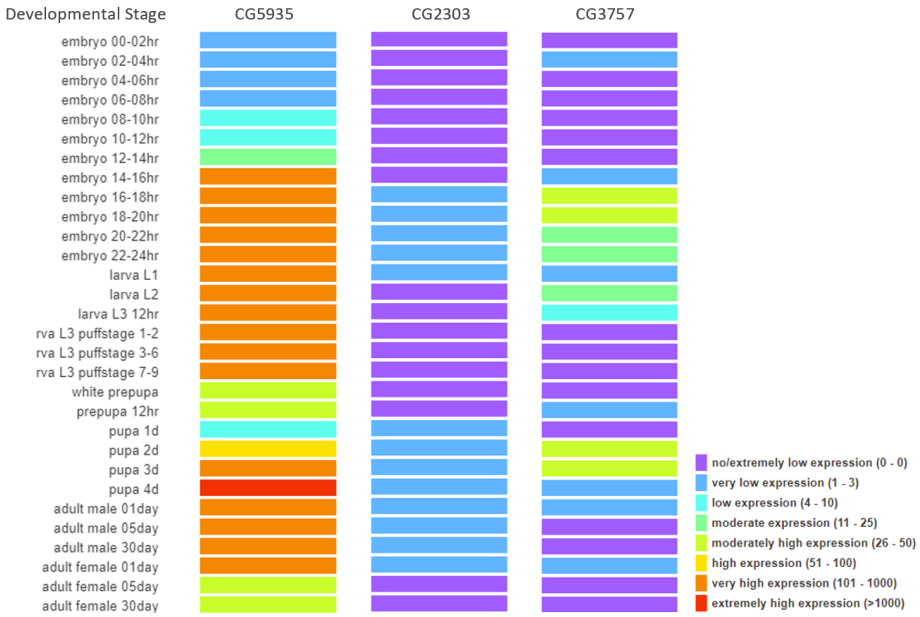

Fig.3 Heatmap of the expression level of the three candidate genes throughout different life stages

Discussions

All three candidate genes have orthologs in most of the other *Drosophila* species with high similarity among the encoded proteins, which indicates a poor inter-species specificity. Gene CG5939 performs a fundamental function in motor activity and has lethal alleles and hence shows better effectiveness than the other genes. Besides, it shows a nearly life-span high-level expression in the important and easily accessed tissues, along with relatively low similarity with paralogs and human orthologs. This indicates high accessibility and acceptable specificity against humans. The main reason for rejecting gene 2303 and 3757 is poor effectiveness. In aggregate, CG5939 is the nominated gene in this research.

Besides, though the genomic information of *Drosophila melanogaster* as a model organism is quite exhaustive, genomic data regarding the three genes in *Drosophila melanogaster* and other *Drosophila* species are still insufficient. This may cause inaccurate judgment when evaluating the effect of the inhibitors on these species.

Appraisal

A good reference genome is fundamental for genomic studies. In this research, with the reliable reference genome of *Drosophila melanogaster*, we can get to know the information about the functions, homology and expression of these genes. We can have access to the different encoded protein sequences of a given gene to then perform peptide alignment for further study of the similarity and homology. We can use alleles of these genes and the corresponding phenotypes to infer their lethality and essentiality. We can also utilize the transcriptome information to know which segment of the gene expresses during which life stages. In contrast, the orthologs of the three candidate genes in other *Drosophila* species lack annotations and are hence difficult to study.

Conclusions

In summary, gene CG5939 is the most suitable target gene to develop and deploy the chemical inhibitor due to its high effectiveness and accessibility, though it may off-target and influence other *Drosophila* species.

Reference

1. Markow TA: **The secret lives of Drosophila flies.** *eLife* 2015, **4:**e06793.
2. Allocca M, Zola S, Bellosta P: **The Fruit Fly, Drosophila melanogaster: The Making of a Model (Part I).** In; 2018
3. Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ, the FlyBase c: **FlyBase: improvements to the bibliography.** *Nucleic Acids Research* 2013, **41:**D751-D757.
4. The UniProt C: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Research* 2019, **47:**D506-D515.
5. Muers M: **The modENCODE guide to the genome.** *Nature Reviews Genetics* 2011, **12:**80-80.
6. The UniProt C: **UniProt: the universal protein knowledgebase in 2021.** *Nucleic Acids Research* 2021, **49:**D480-D489.
7. **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2018, **46:**D8-d13.
8. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* 2007, **35:**D61-D65.