

Seeking the Common Ground: Regulation of Retrotransposed Genes in the *Drosophila* Genus

Lan Hu¹, Daniel Segrè², Temple F. Smith³

1 Introduction

Gene duplication via retrotransposition [4] creates the new copy by reverse transcription of a spliced mRNA, typically resulting in a single-exon copy of a multi-exon ancestral gene. Furthermore, the new copy is deposited in a novel chromosomal environment with a different set of gene neighbors. Potential genetic regulation changes are particularly acute in retrotransposition, because the original regulatory region, lost during the gene transition, is unlikely to be replaced by a compatible and useful transcriptional signal. Thus most retrotransposed copies became pseudogenes.

Thirty-four functional retrotransposed genes have been reported in the *Drosophila* genus [2, 3]. The question is how these genes are regulated or even actively used in which the original regulatory signals are either missing or incompatible with the new position. In this study, we focused on a subset of the retrotransposed genes that are lineage supported in the 12 *Drosophila* species. By seeking the most informative ordered conserved regulatory elements among the orthologs of the retrotransposed genes in available species, we are able to identify the regulatory elements by which a retrotransposed gene could be regulated and further shed light on the functions of unannotated retrotransposed genes.

2 Method

For each gene, its functional orthologs from 12 species constitute an ortholog set. We focused on the upstream 2Kb region of each sequence in an ortholog set for identifying potential regulatory elements. A collection of Transcription Factor Binding Sites (TFBS) in *D. melanogaster* (which is well studied) is compiled from various databases and literature.

A general computational framework was developed to efficiently identify the most informative ordered conserved regulatory elements simultaneously in 12 species. In the first phase, the TFBS's are identified in individual sequences, allowing degeneracy of a TFBS. The second phase chains the TFBS's that are conserved in up to 12 species, enabling loss of a TFBS in some species, to construct the most informative ordered conserved regulatory elements.

The specific chromosomal environment of retrotransposed genes was also taken into account. The gene *l(1)10Bb*, retrotransposed in *D. pseudoobscura* and *D. persimilis* lineage which branched off about 25 million years ago. At the 'ancestral' synteny location, we have identified a potential pseudogene with multi-exons as the likely ancestor. The upstream regions of this pseudo copy were included into the analysis as well.

3 Results

	(A)	(B)
	Orthologs of <i>l(1)10Bb</i> from 12 Species (2 were retrotransposed)	Orthologs of <i>l(1)10Bb</i> from 10 Species + 2 pseudogenes
the Most Informative Ordered Conserved Regulatory Elements	<u>[zeste, tin, tin, br, br]</u>	<u>[zeste, ac, tin, br, br]</u>

Table 1: Comparison of the most informative ordered conserved regulatory elements. The underlined TFBS's are common ordered sites between (A) and (B). The brackets indicate the order of TFBS's.

To test the statistical significance of this particular ordered combination of 4 TFBS's in Table 1, we compiled a control of 2074 genes that are annotated and located internal of synteny blocks [1]. The 2

¹ Graduate Program of Bioinformatics, Boston University, Boston, MA, USA. E-mail: hulan@bu.edu

² Graduate Program of Bioinformatics, Boston University, Boston, MA, USA. E-mail: dsegre@bu.edu

³ BioMolecular Engineering Research Center, Boston University, Boston, MA, USA. E-mail: tsmith@darwin.bu.edu

pseudo genes in (B) of Table 1 were replaced in turn by orthologs of control genes in *D. pseudoobscura* and *D. persimilis*. For the combination of sequences in which the above ordered 4 TFBS's were found, the analysis of the functions of those genes was conducted. Results are shown in Table 2.

	Have Ordered Combination of [zeste, tin, br, br]	Do Not Have the Ordered Combination
Involved in Gamete Generating Processes [†]	10	62
Not Involved in the Gamete generating Processes	73	1991

Table 2: The distribution of 2074 tested genes. Fisher's exact test p-value=0.0004.

[†]: The corresponding GO terms are:

GO:0007292 female gamete generation	GO:0007314 oocyte anterior/posterior axis determination
GO:0016325 oocyte microtubule cytoskeleton organization	GO:0030720 oocyte localization during oogenesis
GO:0016321 female meiosis chromosome segregation	GO:0035211 spermathecum morphogenesis
GO:0007283 spermatogenesis	GO:0007286 spermatid development

4 Conclusion

Identifying the common regulatory elements for orthologs in general is difficult, even for genes that remain in the same synteny environment across 12 *Drosophila* species. It is particularly acute in retrotransposed genes where sudden, drastic changes happened. More computational tools yet need to be developed to answer these questions.

By seeking the common ground, i.e. the most informative ordered conserved regulatory elements, in retrotransposed genes, our framework was able to identify significant common TFBS's. Furthermore, in the example of *l(1)10Bb*, the found conserved regulatory elements were confirmed in the pseudogenes. This has provided us a fuller picture of how regulation of retrotransposed copy could have been evolved. The co-existence of the ancestral and retrotransposed copies could provide time for independent regulatory elements evolution for the retrotransposed copy. The enrichment of GO terms correlating to the found ordered conserved regulatory elements was consistent with the known functions of homologs of *l(1)10Bb* in other organisms.

References

- [1] Bhutkar, A. et al. 2006. Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations. *Genome Informatics* 17(2): 152-161.
- [2] Bhutkar, A. et al. 2007. Genome-scale analysis of positionally relocated genes. *Genome Research* 17(12).
- [3] *Drosophila* 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203-218.
- [4] Soares, M. B. et al. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Molecular and Cellular Biology* 5:2090-2103.