microPublication BIOLOGY

2/5/2025 - *Open Access*

In silico identification and analysis of paralogs encoding enzymes of carbohydrate metabolism in *Drosophila melanogaster*

Rossana Zaru[1], Steven J Marygold[1§]

[1]FlyBase, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, U.K.

[§]To whom correspondence should be addressed: sjm41@cam.ac.uk

Abstract

The identification and characterization of gene paralogs is crucial to understand the functional contribution of individual genes/proteins to biological pathways. Here, we have identified 51 genes belonging to fifteen paralogous groups encoding enzymes involved in carbohydrate metabolism in *Drosophila melanogaster*. Strikingly, most paralogous groups comprise a single ‘canonical' enzyme that is expressed ubiquitously and one or more variants expressed predominantly in the testis. Most of these testis-specific forms are predicted to be catalytically inactive, suggesting they may have adopted regulatory roles. This work will aid the planning and interpretation of experimental studies of several Drosophila metabolic pathways, including glycolysis, gluconeogenesis and the pentose phosphate pathway.

Enzyme name	Human symbol	Metabolic pathway	Drosophila enzymes				
			Symbol	CG#	Iden	Active?	TSI
hexokinase	HK2	GLYCOLYSIS	**Hex-A**	CG3001	47%	Yes (E)	-1.43
			Hex-C	CG8094	42%	Yes (E)	-0.46
			Hex-T*	CG32849	43%	Yes (P)	5.07
			Hexl*	CG33102	31%	No (P)	4.97
aldolase	ALDOA	GLYCOLYSIS / GLUCONEOGENESIS	**Aldo***	CG6058	69%	Yes (E)	-0.29
			Aldol*	CG5432	57%	No (P)	4.75
glyceraldehyde-3-phosphate dehydrogenase	GAPDH	GLYCOLYSIS / GLUCONEOGENESIS	**Gapdh1**	CG12055	76%	Yes (E)	-0.65
			Gapdh2	CG8893	76%	Yes (E)	-1.09
			Gapdh3*	CG9010	66%	Yes (P)	5.07
phosphoglycerate kinase	PGK1	GLYCOLYSIS / GLUCONEOGENESIS	**Pgk**	CG3127	70%	Yes (P)	-0.92
			Pgk2*	CG9961	60%	Yes (P)	5.06
phosphoglycerate mutase	PGAM2	GLYCOLYSIS / GLUCONEOGENESIS	**Pgam1**	CG1721	68%	Yes (P)	-0.91
			Pgam2	CG17645	63%	Yes (P)	5.16
			Pgaml*	CG7059	39%	No (P)	0.75
pyruvate kinase	PKM	GLYCOLYSIS	**Pyk**	CG7070	63%	Yes (E)	-0.91
			Pykl1*	CG7069	48%	No (P)	4.44
			Pykl2*	CG2964	43%	No (P)	5.13
			Pykl3*	CG7362	36%	No (P)	4.63
			Pykl4*	CG12229	21%	No (P)	5.02
			Pykl5*	CG11249	19%	No (P)	5.09
malate dehydrogenase (mitochondrial)	MDH2	GLUCONEOGENESIS	**Mdh2**	CG7998	59%	Yes (E)	-0.85
			Mdh2b*	CG10749	51%	Yes (P)	5.00
			Mdh2c*	CG10748	44%	Yes (P)	4.13
glucose-6-phosphate dehydrogenase	G6PD	PENTOSE PHOSPHATE	**G6pd**	CG12529	64%	Yes (E)	2.75
			G6pdl*	CG7140	40%	No (P)	5.03
transketolase	TKT	PENTOSE PHOSPHATE	**Tkt**	CG8036	58%	Yes (P)	0.76
			Tktl*	CG5103	53%	No (P)	5.09
galactose mutarotase	GALM	GALACTOSE METABOLISM	**Galm1**	CG32444	40%	Yes (P)	1.88
			Galm2	CG10467	46%	Yes (P)	-0.45
			Galml1*	CG32445	41%	No (P)	5.02
			Galml2*	CG10996	37%	No (P)	4.36
			Galml3*	CG4988	40%	No (P)	5.18
trehalase	TREH	TREHALOSE METABOLISM	**Treh**	CG9364	42%	Yes (E)	-0.71
			Trehl*	CG6262	33%	No (P)	4.98
trehalose phosphatase	N/A	TREHALOSE METABOLISM	**Tpp**	CG5171	N/A	Yes (E)	-0.30
			Tppl	CG5177	N/A	No (E)	-0.38
ketohexokinase	KHK	FRUCTOSE METABOLISM	**Khk**	CG7328	27%	Yes (P)	-0.03
			Khk2*	CG7551	26%	Yes (P)	5.09
			Khkl1*	CG7335	27%	No (P)	5.15
			Khkl2*	CG12289	27%	No (P)	5.02
sorbitol dehydrogenase	SORD	FRUCTOSE METABOLISM	**Sord1**	CG1982	55%	Yes (E)	-0.60
			Sord2	CG4649	55%	Yes (P)	0.36
			Sordl*	CG4836	29%	No (P)	5.01
aldose reductase	AKR1B1	FRUCTOSE METABOLISM	**Ar1**	CG6084	57%	Yes (E)	-0.36
			Ar2	CG10638	45%	Yes (P)	0.07
			Ar3*	CG10863	46%	Yes (P)	0.24
			Ar4*	CG9436	44%	Yes (P)	-0.97
			Ar5*	CG2767	42%	Yes (P)	-1.27
			Ar6*	CG12766	48%	Yes (P)	-0.22
			Ar7*	CG6083	49%	Yes (P)	4.69
			ARY	CG40064	37%	Yes (P)	2.81

Figure 1. Paralogs encoding enzymes of simple carbohydrate metabolism in Drosophila.:

Human symbol: HGNC symbol for the human gene encoding the enzyme most similar to the Drosophila gene(s); Metabolic pathway: major pathway(s) in which the enzyme acts; Symbol: current or new (*) symbol for the Drosophila gene in FlyBase - canonical genes are in bold type; CG#: gene model annotation number in FlyBase; Iden: % amino acid identity between the Drosophila protein and the given human enzyme; Active?: an assessment of whether the enzyme is catalytically active or not, based either on experimental data (E) or predicted from sequence analyses (P) - catalytically inactive pseudoenzymes are indicated in red; TSI: testis-specificity index, which ranges from -2.52 (underrepresented in testis) to 5.2 (very high testis bias) - testis-enriched genes are indicated in blue.

Description

The biochemical pathways governing the metabolism of simple carbohydrates are highly conserved and critical to life (reviewed by Chandel 2021). For example, glycolysis is a series of enzymatic reactions that catabolize glucose to generate ATP, NADH and pyruvate, the latter of which can be used to fuel energy production or anabolic pathways. Gluconeogenesis is the process of regenerating glucose molecules from non-carbohydrate precursors. This pathway uses many of the same enzymes as glycolysis together with additional enzymes that bypass irreversible glycolytic reactions. The pentose phosphate pathway is an alternative way to oxidize glucose that generates NADPH and ribose-5-phosphate, which are used in many important biosynthetic reactions. Additional pathways metabolize other simple sugars, such as fructose, galactose and trehalose.

We are currently reviewing and annotating the carbohydrate metabolic pathways of *Drosophila melanogaster* (hereafter, Drosophila). In so doing, we have identified fifteen instances of duplicate genes (paralogs) encoding a total of 51 enzymes predicted to metabolize simple sugars (Table 1). Most paralogous groups comprise two or three members, though there are five galactose mutarotase genes, six pyruvate kinase genes (Heidarian et al. 2023) and eight paralogs encoding aldose reductase.

Each group has at least one ‘canonical' gene, defined as a paralog with widespread expression in somatic tissues and whose encoded protein is known/predicted to be catalytically active (shown in bold text in Table 1; also see Extended Data Files). Most of these 23 genes have been identified previously and are named using standard enzyme nomenclature, including a distinguishing alpha-numerical suffix where necessary. Five groups contain multiple canonical genes. In four of these cases, the paralogs exhibit some qualitative and/or quantitative differences in their expression: *Hex-A* is expressed in most tissues throughout development whereas *Hex-C* has peak expression in the fat body and digestive system of adults (Duvernell and Eanes 2000); *Galm1* expression is high in most adult tissues, whereas *Galm2* is generally expressed at lower levels and mainly in the digestive system of adults (Öztürk-Çolak et al. 2024); *Sord1* expression is higher than *Sord2* across all tissues (Luque et al. 1998); and *Ar2* is expressed at distinctly lower levels compared to *Ar1*, *Ar3*, *Ar4* and *Ar5* (Öztürk-Çolak et al. 2024). In contrast, the two *Gapdh* paralogs show near identical expression profiles across different tissues and developmental stages (Öztürk-Çolak et al. 2024).

The 28 ‘non-canonical' paralogs are characterized by their sequence divergence from the canonical gene(s), predicted lack of catalytic activity in their encoded protein, and/or tissue-restricted expression pattern (Table 1; also see Extended Data Files). Most are identified, analyzed and named here for the first time. 18 of these genes are known or predicted to encode catalytically inactive proteins owing to loss of critical residues in active sites and/or substrate-binding sites (Yoshida et al. 2016; Duvernell et al. 2000; see Extended Data File 1). Such genes are therefore considered to encode pseudoenzymes and are referred to using the standard enzyme nomenclature with a ‘like' suffix, adding a distinguishing numerical suffix where necessary. For example, the *Transketolase* (*Tkt*) gene encodes the canonical enzyme, whereas *Transketolase like* (*Tktl*) is a divergent paralog predicted to encode a pseudoenzyme. Strikingly, the expression of almost all these non-canonical paralogs is restricted to the testis. This is evident by inspecting their ‘testis specificity index', where higher scores indicate testis-restricted expression (Vedelek et al. 2018; Table 1). It is possible that these duplicates have diverged to meet distinct metabolic demands of the testis environment, functioning either as tissue-specific enzymes or as non-catalytic modulators of canonical enzymes. Significantly, several of the Drosophila testis-specific paralogs have human counterparts with testis-enriched expression, including *HK1*, *GAPDHS*, *PGK2*, *TKTL1*, *TKTL2* and AKR1E2 (UniProt Consortium 2023).

The paralogs mentioned herein arose in the Drosophila genome through different evolutionary means (see Extended Data File 1). Several paralogs have been classed as ‘retrogenes', meaning that they originated via retrotransposition from a parental gene copy, namely: *Gapdh1*, *Pgam2*, *Mdh2b* or *Mdh2c*, *G6pdl*, *Galm2*, *Galml2* and *Ar5* (Currie and Sullivan 1994; Dai et al. 2006; Bai et al. 2007; Langille and Clark 2007; Pan and Zhang 2009). Other gene pairs are directly adjacent in the genome and therefore probably arose by a tandem duplication event. These are: *Hex-T*/*Hexl*; *Pgk*/*Pgk2*; *Pyk*/*Pykl1*; *Mdh2b*/*Mdh2c*; *Galm1*/*Galml1*; *Tpp*/*Tppl*; *Khk*/*Khkl1*; *Khk2*/*Khkl2*; *Ar1*/*Ar7* and *Ar3*/*Ar6* (Duvernell and Eanes 2000; Heidarian et al. 2023; Öztürk-Çolak et al. 2024). The remaining paralogs likely derived from other types of gene duplication event (Reams and Roth 2015).

In summary, we find that many genes encoding enzymes governing carbohydrate metabolism are duplicated in Drosophila. Several of these paralogs display tissue-specific, notably testis-specific, expression, and a third are predicted to encode pseudoenzymes. This information is critical to experimental studies of these genes and, as such, will be integrated into FlyBase and allied databases in the form of updated gene nomenclature and Gene Ontology annotations.

Methods

Drosophila genes encoding enzymes involved in carbohydrate metabolic pathways were identified using an integrative approach combining Gene Ontology (GO) annotation data at FlyBase (Öztürk-Çolak et al. 2024) with computed pathway resources at FlyCyc (Caspi et al. 2020), Reactome (Milacic et al. 2024) and KEGG (Kanehisa et al. 2024). Fly gene and protein data were obtained from FlyBase (http://flybase.org) version FB2024_06. Paralogous genes, human-Drosophila orthologs and amino acid identities were determined using the implementation of DIOPT (Hu et al. 2011) within FlyBase. The testis-specificity index (TSI) for each gene was originally computed from modENCODE RNAseq data by Vedelek et al. (2018) and was retrieved from the TSI implementation in FlyBase. High-throughput expression data shown in Extended Data File 2 were derived from the implementation of FlyAtlas 2 (Leader et al. 2018) within FlyBase. The potential catalytic activity of uncharacterized Drosophila enzymes was assessed by using the UniProt ‘Align' tool to align their protein sequences with those of characterized (usually human) orthologs and manually inspecting conservation of critical residues at the annotated active site and substrate/cofactor binding sites (Zaru et al. 2023).

Acknowledgements: We thank Jason Tennessen for comments on the manuscript.

Extended Data

Description: High-throughput expression data (from FlyAtlas 2) for the Drosophila paralogs listed in Table 1.. Resource Type: Dataset. File: Extended Data Table 2.xlsx. DOI: 10.22002/dd86q-snk22

Description: Additional data about the Drosophila paralogs listed in Table 1.. Resource Type: Dataset. File: Extended Data Table 1-v2.xlsx. DOI: 10.22002/9exqh-dcv34

References

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol 8(1): R11. PubMed ID: 17233920

Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al., Karp PD. 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res 48(D1): D445-D453. PubMed ID: 31586394

Chandel NS. 2021. Carbohydrate Metabolism. Cold Spring Harb Perspect Biol 13(1). PubMed ID: 33397651

Currie PD, Sullivan DT. 1994. Structure, expression and duplication of genes which encode phosphoglyceromutase of Drosophila melanogaster. Genetics 138(2): 352-63. PubMed ID: 7828819

Dai H, Yoshimatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene 385: 96-102. PubMed ID: 17101240

Duvernell DD, Eanes WF. 2000. Contrasting molecular population genetics of four hexokinases in Drosophila melanogaster, D. simulans and D. yakuba. Genetics 156(3): 1191-201. PubMed ID: 11063694

Heidarian Y, Tourigny JP, Fasteen TD, Mahmoudzadeh NH, Hurlburt AJ, Nemkov T, et al., Tennessen JM. 2023. Metabolomic analysis of Drosophila melanogaster larvae lacking pyruvate kinase. G3 (Bethesda) 14(1). PubMed ID: 37792629

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12: 357. PubMed ID: 21880147

Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. 2024. KEGG: biological systems database as a model of the real world. Nucleic Acids Res. PubMed ID: 39417505

Langille MG, Clark DV. 2007. Parent genes of retrotransposition-generated gene duplicates in Drosophila melanogaster have distinct expression profiles. Genomics 90(3): 334-43. PubMed ID: 17628393

Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. 2018. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic Acids Res 46(D1): D809-D815. PubMed ID: 29069479

Luque T, Hjelmqvist L, Marfany G, Danielsson O, El-Ahmad M, Persson B, Jörnvall H, González-Duarte R. 1998. Sorbitol dehydrogenase of Drosophila. Gene, protein, and expression data show a two-gene system. J Biol Chem 273(51): 34293-301. PubMed ID: 9852094

Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al., D'Eustachio P. 2024. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res 52(D1): D672-D678. PubMed ID: 37941124

Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al., FlyBase Consortium. 2024. FlyBase: updates to the Drosophila genes and genomes database. Genetics 227(1). PubMed ID: 38301657

Pan D, Zhang L. 2009. Burst of young retrogenes and independent retrogene formation in mammals. PLoS One 4(3): e5040. PubMed ID: 19325906

Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. Cold Spring Harb Perspect Biol 7(2): a016592. PubMed ID: 25646380

UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51(D1): D523-D531. PubMed ID: 36408920

Vedelek V, Bodai L, Grézal G, Kovács B, Boros IM, Laurinyecz B, Sinka R. 2018. Analysis of Drosophila melanogaster testis transcriptome. BMC Genomics 19(1): 697. PubMed ID: 30249207

Yoshida M, Matsuda H, Kubo H, Nishimura T. 2016. Molecular characterization of Tps1 and Treh genes in Drosophila and their role in body water homeostasis. Sci Rep 6: 30582. PubMed ID: 27469628

Zaru R, Orchard S, UniProt Consortium. 2023. UniProt Tools: BLAST, Align, Peptide Search, and ID Mapping. Curr Protoc 3(3): e697. PubMed ID: 36943033

Funding: Supported by National Institute of Diabetes and Digestive and Kidney Diseases (United States) 1R01DK136945-01 to Jason Tennessen, Angelo D'Alessandro, Steven Marygold, Norbert Perrimon.

Author Contributions: Rossana Zaru: data curation, methodology, writing - review editing, formal analysis. Steven J Marygold: conceptualization, writing - original draft, methodology, investigation, data curation, formal analysis, funding acquisition.

Reviewed By: Anonymous

History: Received November 19, 2024 **Revision Received** January 28, 2025 **Accepted** February 2, 2025 **Published Online** February 5, 2025

Citation: Zaru R, Marygold SJ. 2025. *In silico* identification and analysis of paralogs encoding enzymes of carbohydrate metabolism in *Drosophila melanogaster*. microPublication Biology. 10.17912/micropub.biology.001425