Modeling gene expression evolution with EvoGeneX uncovers differences in evolution of species, organs and sexes

Soumitra Pal[1], Brian Oliver[2,*], and Teresa M. Przytycka[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA
[2]Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, 50 South Drive, Bethesda, MD 20892, USA

Abstract

While DNA sequence evolution has been well studied, the expression of genes is also subject to evolution. Yet the evolution of gene expression is currently not well understood. In recent years, new tissue/organ specific gene expression datasets spanning several organisms across the tree of life, have become available providing the opportunity to study gene expression evolution in more detail. However, while a theoretical model to study evolution of continuous traits exist, in practice computational methods often cannot distinguish, with confidence, between alternative evolutionary scenarios. This lack of power has been attributed to the modest number of species with available expression data.

To solve this challenge, we introduce EvoGeneX, a computationally efficient method to uncover the mode of gene expression evolution based on the Ornstein-Uhlenbeck process. Importantly, EvoGeneX in addition to modelling expression variations between species, models within species variation. To estimate the within species variation, EvoGeneX formally incorporates the data from biological replicates as a part of the mathematical model. We show that by modelling the within species diversity EvoGeneX significantly outperforms the currently available computational method. In addition, to facilitate comparative analysis of gene expression evolution, we introduce a new approach to measure the dynamics of evolutionary divergence of a group of genes.

We used EvoGeneX to analyse the evolution of expression across different organs, species and sexes of the *Drosophila* genus. Our analysis revealed differences in the evolutionary dynamics of male and female gonads, and uncovered examples of adaptive evolution of genes expressed in the head and in the thorax.

1 Introduction

Studies of species evolution typically focus on the evolution of species' genetic code. However, in complex multi-cellular organisms all cells utilize the same genetic information, yet they show remarkable phenotypic differences arising from distinct transcriptional programs executed in different tissues/organs. These divergent transcriptional programs underline tissue/organ specific evolutionary adaptations and are subject to tissue specific constraints. Evolutionary analysis of gene expression can thus shed light on the evolutionary processes in ways that cannot be achieved by the analyses of sequence alone.

The understanding of the interplay between species evolution and tissue specific constraints is still limited. The results of initial analyses of gene expression evolution were controversial. For example, early studies of the evolution of primate gene expression suggested that expression evolution is largely consistent with the neutral theory of evolution (Khaitovich et al. 2004, 2005) while subsequent analyses uncovered signature of positive selection in brain (Khaitovich et al. 2006b) and testis (Khaitovich et al. 2006a). However, at the same time, other studies suggested that in both, primates and model organisms, stabilizing selection is likely to be the dominant mode of gene expression evolution (Gilad et al. 2006; Rifkin et al. 2003).

In the past decade, several gene expression datasets encompassing a larger number of species have been collected and analysed [mammals: (Brawand et al. 2011; J. Chen et al. 2019), vertebrates: (Chan et al. 2009)]. Recently, we collected a large dataset of gene expression focusing on the *Drosophila* phylogeny (Yang et al. 2018). Analyses of several of these datasets revealed that the samples clustered generally by tissues rather than by species (or study), suggesting strong evolutionary constraints on tissue-specific gene expression. Yet, there are some sets of genes that cluster by species (Breschi et al. 2016). These studies indicate that the interplay between tissue and species evolution is complex and gene-dependent.

Following the pioneering work of Felsenstein (1973), neutral evolution of continuous traits, such as gene expression, is formally modelled by Brownian-motion (BM). Along a similar line of thought, Lande (1976) pioneered the use of the Ornstein-Uhlenbeck (OU) process to model evolutionary constraints and adaptive evolution of continuous traits. The OU process is stochastic and extends the BM model by adding an "attraction force" towards an optimum value. Combining the OU process on gene expression with the information about the underlining evolutionary tree (obtained, for example, from sequence analysis) allows modeling the differences in evolution along individual branches and thus helps uncover branch specific adaptation (Hansen 1997; Butler and King 2004).

The newer gene expression datasets that span many organisms across the tree of life and include expression data on multiple tissues/organs have challenged us to put these theoretical models into practical use. Several recent studies used such stochastic models to study evolution of gene expression (Bedford and Hartl 2009; Kalinka et al. 2010; Nourmohammad et al. 2017; Brawand et al. 2011; J. Chen et al. 2019). Unfortunately, these formal stochastic methods for modelling continuously varying phenotypic traits typically ignore the within-species variations. Yet, the gene expression differs even between genetically identical multi-cellular organisms, including significant individual to individual variation in *Drosophila* (H. Lee et al. 2018; Lin et al. 2016). It has been long appreciated that ignoring these variations might bias the results in comparative data analysis (Joseph Felsenstein et al. 2008; Ives et al. 2007). In support of this point, Rohlfs et al. (2014) built a model that ignored phylogenetic tree but included the within species variation where gene expression levels were normally distributed across species and individuals and showed that the current methods cannot distinguish data generated by such a model from constrained evolution. Within species variation was considered in the initial analysis of mammalian gene expression (Brawand et al. 2011) but without attention to computational efficiency of the approach or a demonstration that it impacted the results. More importantly, this study considered only constrained and adaptive evolution models ignoring the possibility of neutral evolution. A subsequent analysis of extended mammalian phylogeny showed that neutral evolution cannot be rejected for a large fraction of mammalian genes (J. Chen et al. 2019). However, this last study did not consider the within species variations which might have impacted the results. In addition, studies on the OU models that do not account for within species variations, demonstrated this more complex model to be often incorrectly favored over simpler models such as the BM (Cooper et al. 2016).

Given these conflicting approaches and having in mind the need to facilitate future analyses of gene expression evolution, we developed a rigorous and computationally efficient approach, EvoGeneX, to model gene expression evolution for a given set of species under the assumption that the expression data for each species includes biological replicates. We show that by modelling the within species diversity EvoGeneX outperforms the currently available computational method OUCH (Hansen 1997) that does not use replicates. EvoGeneX has not only increased precision in detecting constrained evolution but was also able to uncover examples of adaptive evolution with relatively small FDR – a task that has been difficult to achieve with the previous method (J. Chen et al. 2019). At the same time, thanks to our improved Maximum Likelihood estimation, EvoGeneX is very efficient despite the increased size of analysed data and increased number of parameters to be optimised.

We applied EvoGeneX to analyse our new expression dataset encompassing 5 different organs, from carefully selected representatives of the *Drosophila* genus (both male and female) where the expression data for each species, organ and sex was measured in 4 biological replicates (Yang et al. 2018). The genus *Drosophila* is particularly well suited for studying gene expression evolution. The last common ancestor of the genus is assumed to date to the Cretaceous period about 112 ± 28 million years ago (Wheat and Wahlberg 2013). The *Drosophila* species occupy diverse geographic locations and ecological niches (Morales-Hojas and Vieira 2012). Compared to the previous studies in mammals (Brawand et al. 2011; J. Chen et al. 2019) and vertebrates (Chan et al. 2009), the morphology of the *Drosophila* species is similar while at the same time the evolutionary distances are significantly larger relative to mammals or even vertebrates typically included in such studies. This makes *Drosophila* an ideal phylogeny to study the interplay among different modes of gene expression evolution.

Our analysis demonstrated that, in *Drosophila*, constrained evolution is more abundant than neutral evolution, however, neutral evolution cannot be rejected in a very large fraction of the genes. In addition, we found that many of the genes that are subject to constrained evolution are common to all organs and both sexes. To gain further understanding of variations in evolutionary dynamics between organs and sexes, we introduced an approach based on Michaelis-Menten kinetics. Our approach revealed striking differences in evolutionary dynamics of gene expression in male and female gonads. Finally, EvoGeneX revealed compelling examples of adaptive evolution.

2 Results and Discussion

2.1 EvoGeneX: A new method to model gene expression evolution

Given the gene expression data and a species evolutionary tree, our goal is to uncover the most likely evolutionary scenario of gene expression. The three basic modes of evolution considered in this study are: (i) neutral evolution, (ii) constrained evolution (purifying/stabilising selection) where the evolution of gene expression is biased against divergence from some optimum value, and (iii) adaptive evolution where the expression in different groups of species is biased towards different optimum values. With respect to the adaptive evolution, in this study we will focus on a special

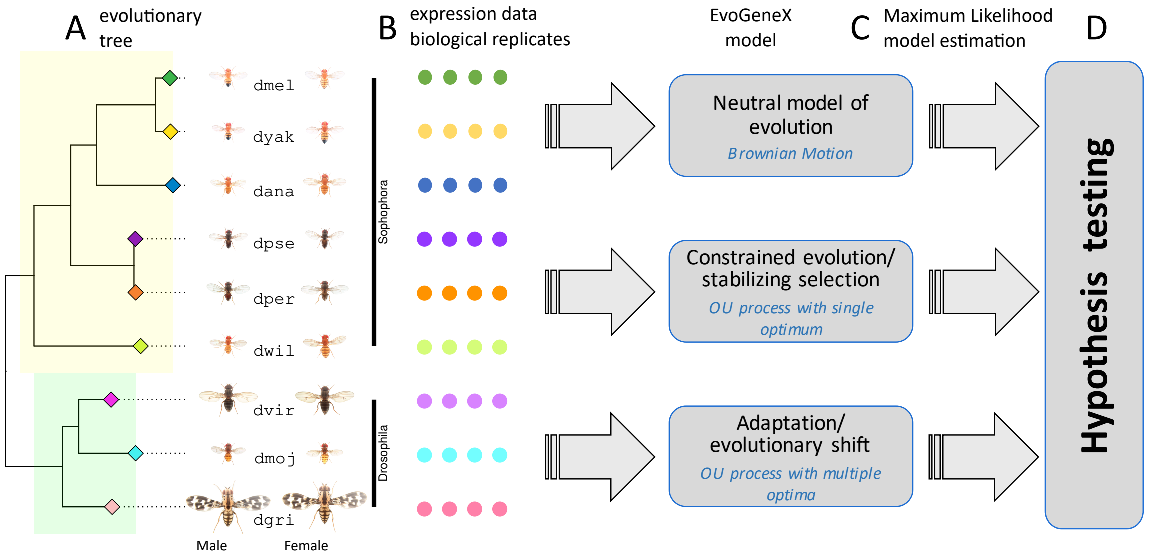

A
evolutionary tree
B
expression data
biological replicates
EvoGeneX model
C
Maximum Likelihood model estimation
D
dmel
dyak
dana
dpse
dper
dwil
dvir
dmoj
dgri
Male
Female
Sophophora
Drosophila
Neutral model of evolution
Brownian Motion
Constrained evolution/ stabilizing selection
OU process with single optimum
Adaptation/ evolutionary shift
OU process with multiple optima
Hypothesis testing

Figure 1: **Workflow diagram.** EvoGeneX takes as the input evolutionary tree (including evolutionary distances) (A) and normalised expression values across biological replicates and species (B). It models evolutionary scenarios as stochastic processes as described in the text and uses Maximum Likelihood approach to fit the parameters of the model to the data (C). For the adaptive model, the proposed adaptation regimes (here *Sophophora* and *Drosophila* subgroups) are required as part of the input. Finally, model selection is preformed using hypothesis testing (D).

case where with two groups of species defined by two sub-trees of the evolutionary tree. However the method is general and can also be used when the groups of species are defined by, say, common habitat rather than an evolutionary relation.

Following standard practice, we assume that the evolution of gene expression follows a stochastic process. In particular, we assume that neutral evolution follows Brownian Motion (BM) – a special case of the more general mean-reverting Ornstein-Uhlenbeck (OU) process – the broadly accepted model for the evolution of continuous traits such as gene expression (Hansen 1997; Butler and King 2004; Bedford and Hartl 2009). This model assumes that gene expression follows a stochastic process that is “attracted” towards some optimum value. The strength of the bias is modelled as a constant α. The optimum value to which the process is attracted is allowed to change over the evolutionary time, reflecting changes in environmental or other constraints acting on the trait.

Assuming neutral evolution as a null model, our goal is to use a hypothesis testing framework to uncover alternative modes of evolution. Towards this end we developed EvoGeneX, a mathematical framework to estimate the parameters of the process given the data. To account for the species variability, we assume that for each species i, there are several observations reflecting within-species variability of the trait of interest (here expression of a specific gene). Denoting the value of the kth observation of the trait for species i by $y_{i,k}$ we set:

$$Y_{i,k}(T_i) = y_{i,k} \tag{1}$$

where T_i is the evolutionary time from the least common ancestor of all species in this tree to species i. Furthermore, for $t \leq T_i$ we define:

$$Y_{i,k}(t) = X_i(t) + \varepsilon_{i,k} \tag{2}$$

Table 1: Summary of performances of EvoGeneX and OUCH on simulated datasets

		precision		recall		specificity	
threshold	**FDR**	OUCH	EvoGeneX	OUCH	EvoGeneX	OUCH	EvoGeneX
0.00	**0.01**	0.5916	0.7646	0.4082	0.0216	0.7183	0.9933
	0.05	0.5779	0.6799	0.5789	0.0799	0.5772	0.9624
0.55	**0.01**	0.6770	0.8161	0.2894	0.0418	0.8619	0.9906
	0.05	0.6742	0.7397	0.4480	0.1526	0.7835	0.9463

where $X_i(t)$ follows the OU process:

$$dX_i(t) = \alpha_i \left[\beta_i(t) - X(t)\right] dt + \sigma dB_i(t). \tag{3}$$

In equation (3) the term $\sigma dB_i(t)$ models the increments of standard Brownian motion (BM), $\beta_i(t)$ is the optimum trait value for species i at time t, and α_i is the strength of the attraction towards the optimum value. In addition, $\beta_i(t)$ is assumed to change in discrete steps corresponding to the speciation events (internal nodes of the evolutionary tree) (Hansen 1997; Butler and King 2004; J. Chen et al. 2019). Finally, $\varepsilon_{i,k} \sim N(0, \gamma\sigma^2)$ is an identically distributed Gaussian variable with mean 0 and variance $\gamma\sigma^2$ that models the within species variability. We assume that within species variance is smaller than evolutionary variance, hence the factor γ is assumed to be less than 1.

Given this general framework, we use hypothesis testing (Figure 1D) to differentiate among the following evolutionary models: (i) **neutral evolution:** $\alpha = 0$, (ii) **constrained evolution (stabilising selection):** $\alpha > 0$ and one common optimum $\beta_i(t) = \theta_0$ for all i and t, and (iii) **adaptive evolution (evolutionary shift):** $\alpha > 0$ and two different optima θ_0, θ_1 representing two regimes with optimal β values θ_1 in a specific subtree and θ_0 in the rest of the evolutionary tree. In all cases appropriate θ values together with α, σ, γ must be estimated from the data (i.e. values $y_{i,k}$). Towards this end, in the Methods section, we describe our efficient method to compute Maximum Likelihood estimate of the parameters of the model.

2.2 EvoGeneX outperforms the previous leading approach

First, we use simulations to test whether EvoGeneX preforms better than the previous approach, OUCH (Hansen 1997; Butler and King 2004), that does not account for within species variation. We simulated 2000 genes with one optimum expression value θ_0 and different values of the parameters σ^2, γ and α where 1000 are constrained and the rest are neutrally evolving (Supplementary Section S5).

To each of these 2000 simulated gene expressions we applied EvoGeneX and OUCH, to predict the mode of evolution (constrained or neutral). We used two P-value thresholds: 0.01 and 0.05 both adjusted for multiple testing using Benjamini-Hochberg correction. In Table 1 we report summary results for all simulations (threshold 0.00) and, separately, summary of the results where at least one of the methods achieved precision at least 0.55 thus excluding the instances where constrained evolution was hard to distinguish from neutral evolution by either method.

Overall, EvoGeneX has consistently higher precision and specificity at the expense of recall. Thus, EvoGeneX is more conservative in predicting constrained evolution and is less likely to predict a more complex model for data simulated using the neutral model, correcting an important issue of the OUCH approach. In addition, despite the need of estimating an additional parameter and considering larger input data, with more complex relations between them, EvoGeneX is only about 2 fold slower than OUCH to infer from 4 replicates (Supplementary Section S9).

2.3 Stabilising selection and neutral gene expression evolution in *Drosophila*

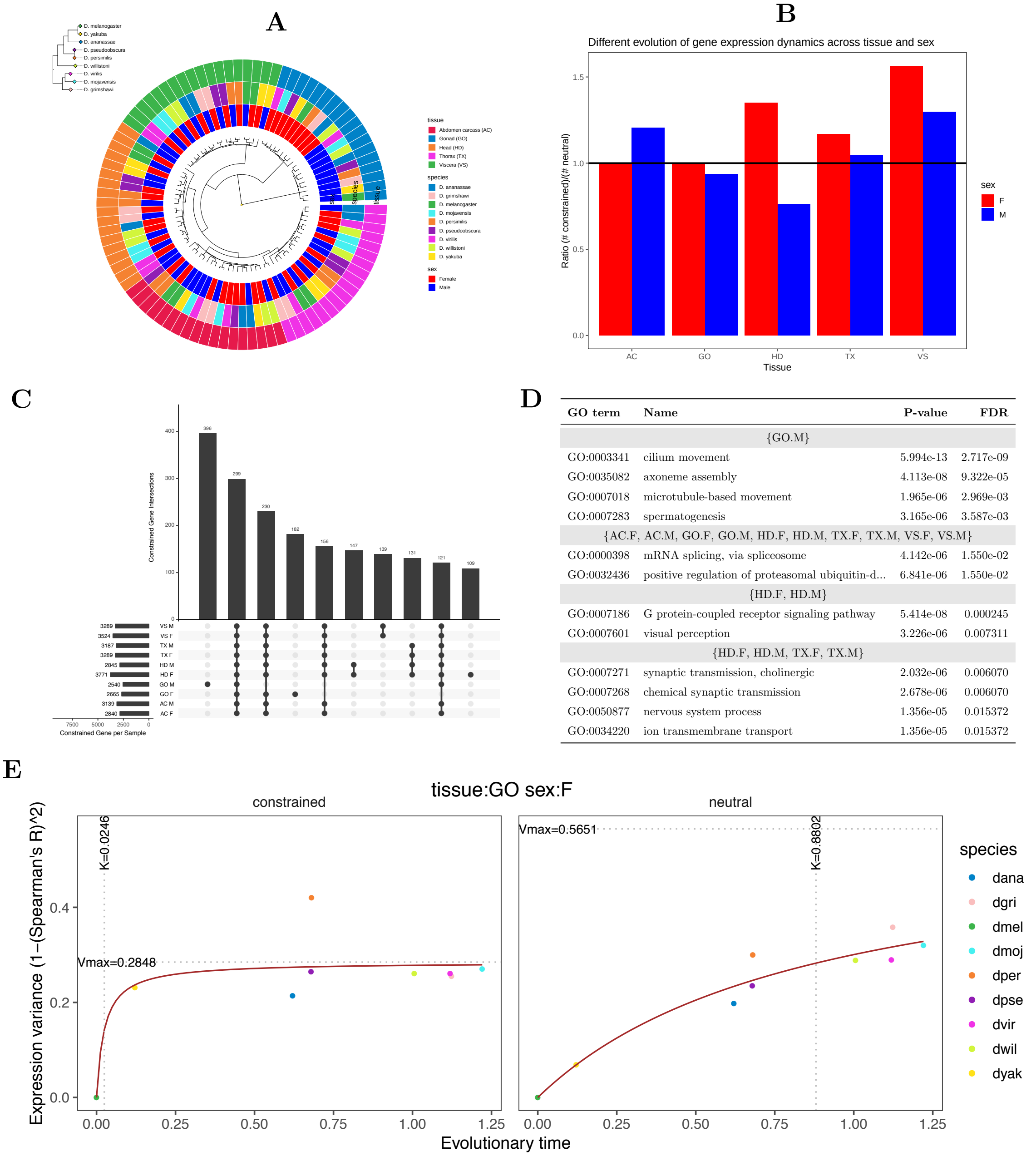
Previous studies in mammals and some other vertebrates report, that when using Pearson's or Spearman's correlation as a similarity measure, the expression data from different organs and species predominately clusters by organs (Brawand et al. 2011; Chan et al. 2009; J. Chen et al. 2019; Sudmant et al. 2015). Given large evolutionary distances within the *Drosophila* genus, we first asked if this observation also holds for *Drosophila*. We used previously collected expression data (Yang et al. 2018) from 9 *Drosophila* species (Figure 1): *D. melanogaster* (dmel), *D. yakuba* (dyak), *D. ananassae* (dana), *D. pseudoobscura* (dpse), *D. persimilis* (dper), *D. willistoni* (dwil), *D. virilis* (dvir), *D. mojavensis* (dmoj) and *D. grimshawi* (dgri) across both sexes and five organs: head (HD), gonads (GO), thorax (TX), viscera (VS) and abdomen carcass (AC) (see Supplementary Figure S1 for an illustrative cartoon of the tissues). Indeed, using $1-S(X,Y)$ where $S(X,Y)$ is the Spearman's correlation between replicate-averaged genes expression, we observe that the expression data predominantly clusters by tissue (Figure 2A, Supplementary Figure S2 for PCA analysis). However, when we used log gene expression values and Euclidean distances then related tissues of the same species often clustered first (Supplementary Figure S3). Specifically, with this similarity measure, head and thorax typically cluster together. This reflects the fact that in fly these two organs are related. Thus both tissue-specific and species-specific trends can be observed, depending on the weight given to genes in the tails of the expression distributions.

Next, we used EvoGeneX to identify genes that are subject to constrained evolution and those for which the null hypothesis of neutral evolution could not be rejected. To focus on the genes that are relevant for a given tissue and sex, we included in this analysis only those genes that have normalised read counts of more than one in all 4 replicates and all species. Independent of organ and sex, and despite the stringency of our model, the hypothesis of neutral evolution could be rejected for more than 50% of the expressed genes (Figure 2C) . In what follows we refer to a gene as "constrained" if the null hypothesis was rejected for that gene, "neutral" otherwise. Note that, technically, this neutral group also contains weakly constrained genes which could not be confidently distinguished from the neutrally evolving group.

Subsequently, we asked if there are genes that are shared by constrained groups in more than one organ/sex. Perhaps not surprisingly, the biggest set of genes jointly constrained was in the intersection of all the organs (Figure 2D for overlaps of more than 100 genes and Supplementary Figure S4 for all overlaps). Gene Ontology Enrichment Analysis (GOEA) revealed that the most significant GOEA term for biological process for this gene set is mRNA splicing, via spliceosome. Alternative splicing is a highly conserved mechanism common to eukaryotes which explains the constraints that are put on its expression divergence in all tissues. Interestingly, genes constrained in male and female samples of head only were enriched in the GO terms for visual perception. Vision is an important evolutionary trait and, thus, many associated genes are indeed expected to be a subject to purifying selection. However, differences in visual perception might also contribute to adaptation and the trade-off between the two modes of evolution of the visual system warrants further investigation. More details on the GOEA of the overlapping constrained gene groups can be found in Supplementary Table S1.

2.4 Differences in evolutionary dynamics across sexes and organs

Next we wanted to compare evolutionary dynamics between sexes and organs. To do this, we considered the relation between the evolutionary distances and expression divergence, measured as $1-S(dmel,X)$ where $S(dmel,X)$ is the square of Spearman's correlation between replicate-averaged expression in *D. melanogaster* and species X. Subsequently, we fitted the Michaelis-Menten curve



GO term	Name	P-value	FDR
{GO.M}			
GO:0003341	cilium movement	5.994e-13	2.717e-09
GO:0035082	axoneme assembly	4.113e-08	9.322e-05
GO:0007018	microtubule-based movement	1.965e-06	2.969e-03
GO:0007283	spermatogenesis	3.165e-06	3.587e-03
{AC.F, AC.M, GO.F, GO.M, HD.F, HD.M, TX.F, TX.M, VS.F, VS.M}			
GO:0000398	mRNA splicing, via spliceosome	4.142e-06	1.550e-02
GO:0032436	positive regulation of proteasomal ubiquitin-d...	6.841e-06	1.550e-02
{HD.F, HD.M}			
GO:0007186	G protein-coupled receptor signaling pathway	5.414e-08	0.000245
GO:0007601	visual perception	3.226e-06	0.007311
{HD.F, HD.M, TX.F, TX.M}			
GO:0007271	synaptic transmission, cholinergic	2.032e-06	0.006070
GO:0007268	chemical synaptic transmission	2.678e-06	0.006070
GO:0050877	nervous system process	1.356e-05	0.015372
GO:0034220	ion transmembrane transport	1.356e-05	0.015372

Figure 2: **Constrained and neutral gene evolution in *Drosophila* genus.** A) Hierarchical clustering of gene expression data using Spearman's correlation. The outside circles correspond to tissue, the middle circles to species and the inside circles to sex. The main organizing factor is cluster first (except for gonads which cluster first by sex). B) The ratio of number of constrained genes to number of neutral genes in each organ-sex pairs. C) Venn diagram for overlaps between constrained genes across organs, D) GO enrichment analysis of the overlaps. Only terms enriched with FDR at most 0.05 are shown. For male gonads M.GO additional mostly related terms are listed in supplement E) Modeling evolutionary dynamics in female gonads with the Michaelis-Menten curve. Constrained genes are characterized by small K values while V_{max} estimates equilibrium divergence.

$f(x) = xV_{max}/(K + x)$ for constrained and neutrally evolving gene sets separately (Figure 3E, Table 2). Here V_{max} is the maximum (putative equilibrium value) of $1 - S(dmel, X)$ and K is

Table 2: Summary of V_{max} and K of constrained and neutral genes for different tissues and sexes

	V_{max}				K			
group	constrained		neutral		constrained		neutral	
sex	F	M	F	M	F	M	F	M
AC	0.301	0.272	0.345	0.387	0.050	0.049	0.200	0.205
GO	**0.285**	**0.471**	**0.565**	**0.761**	0.025	0.008	**0.880**	0.448
HD	0.243	0.265	0.428	0.472	0.035	0.027	0.518	0.820
TX	0.275	0.291	0.404	0.546	0.052	0.068	0.343	0.810
VS	0.263	0.278	0.332	0.384	0.048	0.054	0.293	0.410

the value of x where $f(x) = V_{max}/2$. Thus V_{max} estimates the constraints on the divergence of the given group of genes while K measures the 'speed' of evolutionary dynamics: K close to zero corresponds to the situation where the value V_{max} corresponding to the equilibrium divergence is immediately achieved. In contrast, K can be seen seen as a measure of the speed of the divergence of Spearman's correlation relative to its estimated maximum value V_{max} where larger K correspond to slower relative divergence to the projected equilibrium at V_{max}.

Table 2 shows the values of V_{max} and k for the two gene groups for all combinations of tissues and sexes. As expected, V_{max} values for constrained groups of genes are systematically smaller than the V_{max} values of neutrally evolving genes for the same tissue/sex, although in the case of abdomen carcass (AC) the difference is small (consistent with small K values for neutral AC groups). Strikingly, male gonads are far less constrained (larger V_{max}) than any other organ in any sex. This is true not only for the constrained group, but also the V_{max} for neutrally evolving male gonads is the highest. The high V_{max} value for neutrally evolving genes in female gonad is surprising since, unlike male, female gonads don't stand out as being less constrained than other tissues in the constrained groups. Indeed, the large K value for neutrally evolving female gonads genes suggest that the changes (relative to the V_{max} value) are slow. These results show that the expression evolution of male and female gonads follow drastically different dynamics. Moreover, these results do not depend on the species taken as reference (*D. melanogaster* here) as we observe similar results when *D. pseudoobscura* or *D. virilis* is taken as reference (Supplementary Figure S5 and Table S2).

2.5 EvoGeneX reveals examples of adaptive evolution

Differences in the habitat, evolutionary bottle necks, or other factors might lead to evolutionary shifts of optimal expression values. As a result, the optimum value of a gene expression might be different in different branches of the evolutionary tree. Statistically, detection of such evolutionary adaptation is very challenging and current methods have very limited statistical power that has been attributed to the relatively small sizes of evolutionary trees for which the tissue specific gene expression data is available (J. Chen et al. 2019). We reasoned that by utilizing replicates for the estimation of within species variability, EvoGeneX might have higher power for detecting such adaptive evolution. We used a hypothesis testing framework, where we first tested if both, the neutral BM model and constrained model are rejected in favor of adaptive evolution. We focused on the two scenarios below.

First, we tested for adaptive evolution two subtrees: *Sophophora* and *Drosophila* (Figure 1). The species in these subtrees differ in many behavior factors including diet: *Sophophora* species feed on fruits while *Drosophila* species feed on vegetable diet (Markow 2015). In addition, presence of multiple species in each subgroup allows for assessing FDR using a permutation test (J. Chen

Table 3: Summary of results on adaptation test on selected two-regime scenarios

regime	Sophophora/Drosophila						dgri/others	
statistics	#genes		FDR		#genes		overlap	P-value
sex	F	M	F	M	F	M		
AC	43	33	0.462093	0.507576	28	33	17	2.706748e-36
GO	36	58	0.422222	0.461724	53	20	4	6.227412e-06
HD	41	83	0.384634	**0.188554**	76	81	51	**1.845591e-92**
TX	38	47	0.481579	0.335532	**127**	**150**	72	**1.654622e-98**
VS	38	48	0.503158	0.372292	47	51	28	2.242926e-53

et al. 2019). In this case the most striking signature of adaption was displayed by male heads where EvoGeneX discovered 83 adaptive genes with the relatively low FDR of less than 0.2 (Table 3). This should be contrasted with the FDR of 0.56 obtained for the same test using OUCH approach. We note that there are substantial differences in wiring and gross anatomy between male and female fly brains (Cachero et al. 2010). In particular, male heads express male-specific isoform of the *fruitless* gene that is known to regulate male courtship behavior (Kimura et al. 2005; G. Lee et al. 2000). The set of adaptive genes uncovered by EvoGeneX can provide additional cues for male specific adaptation of fly brain and warrants further investigation.

The second very interesting case is the evolution of *D. grimshawi*, a unique Hawaiian species which exhibits stunning phenotypic differences relative to the other species in this study (Figure 1) such as very limited ability to fly and very unusual diet. We tested the two regime scenario: *D. grimshawi* vs others. Since in this case we could not apply permutation test as only one species is under a regime, we tested for the statistical significance of the overlap between the putative male and female adaptive genes. Given that the measurements of male and female expression are completely independent, a statistically significant overlap provides a strong support for an adaptive evolution of these genes. We found significant overlap for all tissues with the most spectacular overlap for the thorax genes (Fisher's test P-value 1.845591e-92, Table 3) closely followed by genes active in head which is consistent with the pronounced differences in the body and the expected adaptation to the highly specific diet and environment of this Hawaiian species.

3 Conclusions

Studies of the evolution of gene expression can profoundly contribute to the understanding of the molecular system underlying phonetic traits. However these studies have been hampered by limited statistical power of the available methods. For example, it has been recognised that the power of the existing approach to detect adaptive evolution (evolutionary shift) is low and suggested that a larger evolutionary tree is required to preform such analysis (J. Chen et al. 2019). We argue, that rather than increasing the size of the tree, which might be difficult, one can utilize biological replicates that are often readily available. Indeed, scientists increasingly collect their data in multiple replicates. This data provides very valuable information about the within species variation that could be used to boost the performance of the method without increasing the number of species under study. Addressing these needs EvoGeneX formally includes within species variation estimated using biological replicates ensuring increased precision relative to the previous leading method.

We used EvoGeneX to provide the first analysis of gene expression evolution across different organs, species, and sexes of the *Drosophila* genus. Our analysis uncovered interesting but differential evolutionary dynamics of male and female gonads. It has also been able to detect, with high confidence, examples of adaptive evolution in the context of the very unique Hawaiian fly – *D. grimshawi* As an species unique to Hawaiian islands, *D. grimshawi* experienced evolution bottleneck. It also had to adopt to unique environment and diet. In addition, in the context of *Sophophora* and *Drosophila* subgroups, we found a signature of adaption in gene expression in male heads that we plan to investigate in follow-up studies. While these two subgroups differ in their diet the fact that the adaptive genes were found only in males suggest that they might be related to the difference in sexual relations such as courtship, response to pheromones or other aspects of sexual dimorphism of gene expression in head.

Overall, our results suggest that, by taking full advantage of the existing gene expression data, EvoGeneX provides a new and powerful tool to study gene expression evolution.

4 Materials and Methods

Data and data / software availability The expression data (Yang et al. 2018) was obtained from the NCBI GEO database under accession numbers GSE99574 and GSE80124, and the phylogenetic tree was obtained from Z.-X. Chen et al. (2014) (Supplementary Section S1). The source code for EvoGeneX is available at the NCBI public Github repository: `https://github.com/ncbi/EvoGeneX`. A software pipeline to analyze the data using EvoGeneX was built using JUDI (Pal and Przytycka 2019).

EvoGeneX model and its parameters EvoGeneX takes as its input a rooted evolutionary tree and the values of quantitative characters $y_{i,k}$ for all terminal taxa i and biological replicates k. Two sets of random variables, $X_i(t)$ at the taxa level and $Y_{i,k}$ at the replicate level, are used to model the evolution of trait value across time such that the observed trait value at time T_i, the evolutionary time from the least common ancestor of all species in this tree to species i, is $Y_{i,k}(T_i) = y_{i,k}$. The two sets of random variables are governed by equations (2) and (3).

EvoGeneX further assumes that the optimum value $\beta_i(t)$ of "attraction" in (3) changes at speciation events only and remains constant along individual edges of the phylogenetic tree. The history of the ith lineage consists of a number, $\kappa(i)$, of sequential branch segments demarcated by speciation events $0 = t_i^0 < t_i^1 < t_i^2 < \ldots < t_i^{\kappa(i)} = T_i$. Let all $t_i^{\tau-1} \leq t \leq t_i^{\tau}$ represent a *selective regime* where the evolution "attracts" towards a fixed optimum value β_i^{τ} of $\beta_i(t)$. EvoGeneX further simplifies the model by letting a small number, R, of distinct optimum values θ_r, $r = 1, \ldots, R$, each corresponding to one selective regime. In fact one of the most interesting cases corresponds to the model with two optima where one branch of the tree follows a regime of optimum values θ_1 and the rest of the tree θ_0 (J. Chen et al. 2019; Brawand et al. 2011).

Let the binary variable $\beta_{i,r}^{\tau}$ indicate if the τth branch on lineage i has operated in rth regime. Then we have $\beta_i^{\tau} = \sum_{r=1}^{R} \beta_{i,r}^{\tau} \theta_r$. Since each branch is associated with exactly one optimum, for each i, τ there is exactly one r such that $\beta_{i,r}^{\tau} = 1$ and $\beta_{i,r'}^{\tau} = 0$ for all $r \neq r'$. Further, self-consistency requires that $\beta_{i,r}^{\tau} = \beta_{j,r}^{\eta}$ whenever lineage i and j share the branch ending in epoch $t_i^{\tau} = t_j^{\eta}$.

Thus, for a given tree, $y_{i,k}$s and $\beta_{i,r}^{\tau}$s, EvoGeneX estimates the parameters $\alpha, \sigma, \gamma, \theta_0, \theta_1, \ldots, \theta_R$.

Inference of EvoGeneX parameters using Maximum Likelihood estimates In the following, it will be convenient to make use of matrix notation. Accordingly, we collect our random variables, $X_i(t)$ for the trait values at the taxa, and $Y_{i,k}(t)$ for the replicated trait values, in vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$, respectively, and our observed quantitative data in vector $\mathbf{y}$ with components $y_{i+(k-1)N} = Y_{i,k}(T_i)$, the observed value of replicate k of taxa i at the evolutionary time T_i. The expected value of random variable $\mathbf{y}(t)$ at the taxa can be shown to be (Supplementary Section S4)

$$\mathrm{E}[\mathbf{y}(T) \mid \mathbf{x}(0) = \theta_0 \mathbb{1}] = \mathbf{W}\boldsymbol{\theta} \tag{4}$$

where $\mathbb{1}$ is a vector of all 1s, column vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_R)^T$ and the weight matrix W is dependent only on α among the parameters and has entries

$$W_{i+(k-1)N,0} = e^{-\alpha T_i}, \quad W_{i+(k-1)N,r} = \sum_{\tau=1}^{\kappa(i)} \left(e^{-\alpha(T_i - t_i^{\tau})} - e^{-\alpha(T_i - t_i^{\tau-1})} \right) \beta_{i,r}^{\tau}$$

for $i = 1, \ldots, N$, $k = 1, \ldots, M$ and $r = 1, \ldots, R$. Similarly, let an $MN \times MN$ matrix $\mathbf{V}$ give the covariance between species i, replicate k and species j, replicate l by the entry (Supplementary Section S4)

$$\begin{aligned} v_{(i,k),(j,l)} &= v_{i+(k-1)N, j+(l-1)N} = \mathrm{Cov}[Y_{i,k}(t_i), Y_{j,l}(t_j) \mid X_i(0) = X_j(0) = \theta_0] \\ &= \frac{\sigma^2}{2\alpha} e^{-\alpha(t_i + t_j - 2s_{i,j})}(1 - e^{-2\alpha s_{i,j}}) + \begin{cases} \gamma\sigma^2 & \text{if } (i = j) \text{ and } (k = l) \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

It is known that $\mathbf{y}$ is a multi-variate Gaussian $\sim \mathcal{N}(\mathbf{W}\boldsymbol{\theta}, \mathbf{V})$ with mean and co-variance given by equations (4) and (5) (Hansen and Martins 1996). Thus, the likelihood of the parameters α, σ, γ, and $\boldsymbol{\theta}$, given the data $\mathbf{y}$ is

$$\mathcal{L}(\alpha, \sigma, \gamma, \boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{NM} \det \mathbf{V}}} \exp\left[-\frac{(\mathbf{y} - \mathbf{W}\boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})}{2} \right]$$

As maximizing $\mathcal{L}$ is equivalent to minimizing $U = -2\log\mathcal{L}$, we seek to minimize

$$U(\alpha, \sigma, \gamma, \boldsymbol{\theta} \mid \mathbf{y}) = NM \log(2\pi) + \log\det \mathbf{V} + (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})$$

However, it can be noted that $\mathbf{V}$ has a nice structure and can be expressed as $\sigma^2(\tilde{\mathbf{V}} + \gamma\mathbb{I})$ where $\tilde{\mathbf{V}}$ is dependent on α only among all the parameters and $\mathbb{I}$ is an identity matrix of size MN. The elements of $\tilde{\mathbf{V}}$ are given by $\tilde{V}_{(i,k),(j,l)} = \frac{1}{2\alpha} e^{-\alpha(t_i + t_j - 2s_{i,j})}(1 - e^{-2\alpha s_{i,j}})$. Thus, U can be expressed as

$$U(\alpha, \sigma, \gamma, \boldsymbol{\theta} \mid \mathbf{y}) = NM \log(2\pi\sigma^2) + \log\det(\tilde{\mathbf{V}} + \gamma\mathbb{I}) + \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})^T (\tilde{\mathbf{V}} + \gamma\mathbb{I})^{-1} (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})$$

whose minimum can be estimated using any off-the-self nonlinear optimization solver.

However, we improve efficiency by utilizing Karush-Kuhn-Tucker conditions. By setting the partial derivatives of U with respect to σ and $\boldsymbol{\theta}$ to 0 at an optimal solution $(\hat{\alpha}, \hat{\sigma}, \hat{\gamma}, \hat{\boldsymbol{\theta}})$, we get

$$\hat{\sigma}^2 = \frac{1}{NM} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\theta}})^T (\tilde{\mathbf{V}} + \hat{\gamma}\mathbb{I})^{-1} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\theta}}), \text{ and, } \hat{\boldsymbol{\theta}} = \left(\mathbf{W}^T (\tilde{\mathbf{V}} + \hat{\gamma}\mathbb{I})^{-1} \mathbf{W} \right)^{-1} \mathbf{W}^T (\tilde{\mathbf{V}} + \hat{\gamma}\mathbb{I})^{-1} \mathbf{y}$$

Thus, instead of minimizing function U of four parameters α, σ, γ and $\boldsymbol{\theta}$, it is enough to minimize a new function of two parameters, α and γ, $\tilde{U}(\alpha, \gamma) = NM \left[1 + \log 2\pi\hat{\sigma}^2(\alpha, \gamma) \right] + \log\det(\tilde{\mathbf{V}} + \gamma\mathbb{I})$ where the following two intermediate functions

$$\begin{aligned} \hat{\boldsymbol{\theta}}(\alpha, \gamma) &= \left(\mathbf{W}^T (\tilde{\mathbf{V}} + \gamma\mathbb{I})^{-1} \mathbf{W} \right)^{-1} \mathbf{W}^T (\tilde{\mathbf{V}} + \gamma\mathbb{I})^{-1} \mathbf{y} \\ \hat{\sigma}^2(\alpha, \gamma) &= \frac{1}{NM} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\theta}}(\alpha, \gamma))^T (\tilde{\mathbf{V}} + \gamma\mathbb{I})^{-1} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\theta}}(\alpha, \gamma)) \end{aligned} \tag{6}$$

give the values of the remaining two parameters σ and $\boldsymbol{\theta}$ at the optimal solution.

Maximum Likelihood estimates for Brownian model In order to compare EvoGeneX model of evolution with the BM model, we need to compute maximum likelihood estimates for the Brownian model (BM) accounting for the within species variation. BM is a simplified model in comparison to EvoGeneX: there is no "attracting" optimal values and hence there is no α parameter and θ has only one value, θ_0, to be estimated. Using a procedure similar to the previous section, we estimate σ, γ, θ_0 from the given data (Supplementary Section S4).

Computing statistical significance We use statistical hypothesis testing to decide which of the three different modes of evolution the trait has undergone: i) neutral, ii) constrained and ii) adaptive (Section 2.1). For this purpose we use likelihood ratio test (Supplementary Section S4).

Acknowledgement

The authors would like to thank Haiwang Yang and Justin Fear, for their helpful comments. This research was supported by the Intramural Research Program of the National Library of Medicine and the National Institute of Diabetes and Digestive and Kidney Diseases at the National Institutes of Health, USA.

References

Bedford, Trevor and Daniel L. Hartl (2009). "Optimization of Gene Expression by Natural Selection". en. In: *Proceedings of the National Academy of Sciences* 106.4, pp. 1133–1138. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0812009106.

Brawand, David et al. (2011). "The Evolution of Gene Expression Levels in Mammalian Organs". en. In: *Nature* 478.7369, pp. 343–348. ISSN: 1476-4687. DOI: 10.1038/nature10532.

Breschi, Alessandra et al. (2016). "Gene-Specific Patterns of Expression Variation across Organs and Species". In: *Genome Biology* 17.1, p. 151. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1008-y.

Butler, Marguerite A. and Aaron A. King (2004). "Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution." In: *The American Naturalist* 164.6, pp. 683–695. ISSN: 0003-0147. DOI: 10.1086/426002.

Cachero, Sebastian et al. (2010). "Sexual Dimorphism in the Fly Brain". In: *Current Biology* 20.18, pp. 1589–1601. ISSN: 0960-9822. DOI: 10.1016/j.cub.2010.07.045.

Chan, Esther T. et al. (2009). "Conservation of Core Gene Expression in Vertebrate Tissues". In: *Journal of Biology* 8.3, p. 33. ISSN: 1475-4924. DOI: 10.1186/jbiol130.

Chen, Jenny et al. (2019). "A Quantitative Framework for Characterizing the Evolutionary History of Mammalian Gene Expression". en. In: *Genome Research* 29.1, pp. 53–63. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.237636.118.

Chen, Zhen-Xia et al. (2014). "Comparative Validation of the D. Melanogaster modENCODE Transcriptome Annotation". en. In: *Genome Research* 24.7, pp. 1209–1223. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.159384.113.

Cooper, Natalie et al. (2016). "A Cautionary Note on the Use of Ornstein Uhlenbeck Models in Macroevolutionary Studies". en. In: *Biological Journal of the Linnean Society* 118.1, pp. 64–77. ISSN: 1095-8312. DOI: 10.1111/bij.12701.

Felsenstein, J (1973). "Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characters." In: *American Journal of Human Genetics* 25.5, pp. 471–492. ISSN: 0002-9297.

Felsenstein, Joseph, Associate Editor: Sarah Perin Otto, and Editor: Michael C. Whitlock (2008). "Comparative Methods with Sampling Error and Within-Species Variation: Contrasts Revisited and Revised". In: *The American Naturalist* 171.6, pp. 713–725. ISSN: 0003-0147. DOI: 10.1086/587525.

Gilad, Yoav, Alicia Oshlack, and Scott A. Rifkin (2006). "Natural Selection on Gene Expression". In: *Trends in Genetics* 22.8, pp. 456–461. ISSN: 0168-9525. DOI: 10.1016/j.tig.2006.06.002.

Hansen, Thomas F. (1997). "Stabilizing Selection and the Comparative Analysis of Adaptation". In: *Evolution* 51.5, pp. 1341–1351. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.1997.tb01457.x.

Hansen, Thomas F. and Emília P. Martins (1996). "Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data". en. In: *Evolution* 50.4, pp. 1404–1417. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.1996.tb03914.x.

Ives, Anthony R., Peter E. Midford, and Theodore Garland (2007). "Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods". en. In: *Systematic Biology* 56.2, pp. 252–270. ISSN: 1063-5157. DOI: 10.1080/10635150701313830.

Kalinka, Alex T. et al. (2010). "Gene Expression Divergence Recapitulates the Developmental Hourglass Model". en. In: *Nature* 468.7325, pp. 811–814. ISSN: 1476-4687. DOI: 10.1038/nature09634.

Khaitovich, Philipp, Wolfgang Enard, Michael Lachmann, and Svante Pääbo (2006a). "Evolution of Primate Gene Expression". en. In: *Nature Reviews Genetics* 7.9, pp. 693–702. ISSN: 1471-0064. DOI: 10.1038/nrg1940.

Khaitovich, Philipp, Svante Pääbo, and Gunter Weiss (2005). "Toward a Neutral Evolutionary Model of Gene Expression". en. In: *Genetics* 170.2, pp. 929–939. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.104.037135.

Khaitovich, Philipp et al. (2004). "A Neutral Model of Transcriptome Evolution". en. In: *PLOS Biology* 2.5, e132. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020132.

Khaitovich, Philipp et al. (2006b). "Positive Selection on Gene Expression in the Human Brain". In: *Current Biology* 16.10, R356–R358. ISSN: 0960-9822. DOI: 10.1016/j.cub.2006.03.082.

Kimura, Ken-Ichi, Manabu Ote, Tatsunori Tazawa, and Daisuke Yamamoto (2005). "Fruitless Specifies Sexually Dimorphic Neural Circuitry in the Drosophila Brain". en. In: *Nature* 438.7065, pp. 229–233. ISSN: 1476-4687. DOI: 10.1038/nature04229.

Lande, Russell (1976). "Natural Selection and Random Genetic Drift in Phenotypic Evolution". en. In: *Evolution* 30.2, pp. 314–334. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.1976.tb00911.x.

Lee, Gyunghee et al. (2000). "Spatial, Temporal, and Sexually Dimorphic Expression Patterns of the Fruitless Gene in the Drosophila Central Nervous System". en. In: *Journal of Neurobiology* 43.4, pp. 404–426. ISSN: 1097-4695. DOI: 10.1002/1097-4695(20000615)43:4<404::AID-NEU8>3.0.CO;2-D.

Lee, Hangnoh et al. (2018). "Dosage-Dependent Expression Variation Suppressed on the Drosophila Male X Chromosome". en. In: *G3: Genes, Genomes, Genetics* 8.2, pp. 587–598. ISSN: 2160-1836. DOI: 10.1534/g3.117.300400.

Lin, Yanzhu, Zhen-Xia Chen, Brian Oliver, and Susan T. Harbison (2016). "Microenvironmental Gene Expression Plasticity Among Individual Drosophila Melanogaster". en. In: *G3: Genes, Genomes, Genetics* 6.12, pp. 4197–4210. ISSN: 2160-1836. DOI: 10.1534/g3.116.035444.

Markow, Therese Ann (2015). "The Secret Lives of Drosophila Flies". In: *eLife* 4, e06793. ISSN: 2050-084X. DOI: 10.7554/eLife.06793.

Morales-Hojas, Ramiro and Jorge Vieira (2012). "Phylogenetic Patterns of Geographical and Ecological Diversification in the Subgenus Drosophila". en. In: *PLOS ONE* 7.11, e49552. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0049552.

Nourmohammad, Armita et al. (2017). “Adaptive Evolution of Gene Expression in Drosophila”. English. In: *Cell Reports* 20.6, pp. 1385–1395. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.07.033.

Pal, Soumitra and Teresa M. Przytycka (2019). “Bioinformatics Pipeline Using JUDI: Just Do It”. en. In: *bioRxiv*, p. 611764. DOI: 10.1101/611764.

Rifkin, Scott A., Junhyong Kim, and Kevin P. White (2003). “Evolution of Gene Expression in the Drosophila Melanogaster Subgroup”. en. In: *Nature Genetics* 33.2, pp. 138–144. ISSN: 1546-1718. DOI: 10.1038/ng1086.

Rohlfs, Rori V., Patrick Harrigan, and Rasmus Nielsen (2014). “Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation”. en. In: *Molecular Biology and Evolution* 31.1, pp. 201–211. ISSN: 0737-4038. DOI: 10.1093/molbev/mst190.

Sudmant, Peter H., Maria S. Alexis, and Christopher B. Burge (2015). “Meta-Analysis of RNA-Seq Expression Data across Species, Tissues and Studies”. In: *Genome Biology* 16.1, p. 287. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0853-4.

Wheat, Christopher W. and Niklas Wahlberg (2013). “Critiquing Blind Dating: The Dangers of over-Confident Date Estimates in Comparative Genomics”. In: *Trends in Ecology & Evolution* 28.11, pp. 636–642. ISSN: 0169-5347. DOI: 10.1016/j.tree.2013.07.007.

Yang, Haiwang et al. (2018). “Re-Annotation of Eight Drosophila Genomes”. en. In: *Life Science Alliance* 1.6, e201800156. ISSN: 2575-1077. DOI: 10.26508/lsa.201800156.