Adaptive Constraint and its Natural Rescuers: Insights from *Drosophila*

Aldehyde Dehydrogenase

by

Mahul Chakraborty

Submitted in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Supervised by Professor James D. Fry

Department of Biology

Arts, Sciences and Engineering

School of Arts and Sciences

University of Rochester

Rochester, New York

2013
Dedication

This thesis is dedicated to all my teachers, without whom I would not have come this far.
Biographical Sketch

The author was born in Kolkata, India, on August 15, 1983. He attended Bidhannagar College at the University of Calcutta (India) and graduated with the Bachelor of Science degree in Zoology in 2005. He received his Master of Science degree in Biotechnology from Indian Institute of Technology, Roorke (India) in 2007. He came to the University of Rochester, Rochester, NY (USA) in the fall of 2007 and began graduate studies in the Department of Biology. He received a second Master of Science degree in Biology from University of Rochester in 2009. He has worked as teaching assistant at Department of Biology from 2008 to 2013. He pursued his research on genetics of adaptation under the guidance of Professor James D. Fry.

List of publication

Acknowledgements

I owe my deepest gratitude to my advisor, Dr. James Fry, for providing a great research ambience and for his guidance, support, and encouragement. I would like to thank the members of my Thesis Advisory Committee, Dr. Allen Orr, Dr. Gloria Culver, and Dr. Walter Eanes for their expert advice during my graduate career. I would like to thank Dr. David Lambert for his technical advice. Thanks to my colleague and friend, Jing Zhu – as well as past members of Fry lab: Justin Budnik, Rebecca Flannagan, Brigid Hedlund, and Oscar Yang for their assistance with various experiments and many stimulating discussions. Thanks to Kathy Donlon for her support inside and outside the lab. I would also like to thank members of the PERL group – Jeff, Rob, Anthony, and LeAnne – for the brainstorming sessions. I express my gratitude to the Biology Department at University of Rochester and its staff, especially Cynthia Landry, Deborah Lawrence, and Rose Covello. I thank my friends – Krishanu, Nilotpal, Wreetabrata, Swagata, Rashmi, Supriya, Arnab, Rahul, Amrita, Simantini, Itender and many others for their valued friendship. I thank Joy for his constant encouragement and infectious optimism. This thesis would not have been possible without the continual support and motivation from two most important people of my life, Rita and Aniruddha Chakraborty. I thank Gitangshu Kar and Rama Kar for their love and care. Finally, I would like to thank Dr. Lucien Rouse, Amanda, and Sarah for ensuring my unhindered bipedalism.
Abstract
Adaptation involves trait optimization: mutations optimizing a phenotype spread in a population, replacing their ancestral counterparts. Mutations that improve turnover rate of a promiscuous enzyme for one substrate often reduce turnover rate of the enzyme for other substrates. Such tradeoff has been predicted to impede spread of potentially beneficial mutations and thus constrain adaptation. Protein structural constraints acting as adaptive constraint are likely to be common, but the only convincing example of such adaptive constraint is based on an engineered bacterial enzyme. Using *D. melanogaster* aldehyde dehydrogenase enzyme gene (*DmAldh*), I investigated the hypothesis that enzyme structural constraint hinders spread of adaptive mutation in natural populations. A *DmAldh* replacement polymorphism causes substitution of a conserved residue Leu479 by Phe479, which increases turnover rate of the enzyme for acetaldehyde, a toxic derivative of dietary ethanol, and increases resistance of the flies to acetaldehyde and ethanol. Nonetheless, frequency of the Phe allele remains low in temperate region *D. melanogaster* populations, where increased resistance to ethanol is adaptive. Further investigation revealed that Leu→Phe substitution compromises protection of DmALDH from toxic aldehydes generated by lipid peroxidation within mitochondria, because the Phe enzyme detoxifies these aldehydes more slowly than the Leu enzyme. The fitness defect of the Phe allele in absence of ethanol is the likely reason of why Phe allele does not spread in ethanol resistant populations. Furthermore, an intronic eSNP and a set of deletions within *DmAldh* 3’ UTR were found to increase expression of *DmAldh*. The eSNP appears to be adaptive in populations where increased resistance to ethanol is adaptive, likely because it increases
ethanol resistance. The Phe allele is in linkage disequilibrium with the eSNP, suggesting compensatory interaction between the two. Additionally, Aldh duplicates with amino acid substitutions showing similar structural, and perhaps functional, effects as Leu → Phe substitution have originated independently in *Drosophila* and *obscura* groups. Evidence based on amino acid and expression changes suggests that duplicates have specialized on an enzyme function that may not have been possible in ancestral ALDH. The expression polymorphism and duplications are inferred as potential mechanisms which can rescue adaptive constraint involving *Aldh*. 
Contributors and Funding Sources

The work was supervised by a dissertation committee consisting of Professors James D. Fry, H. Allen Orr, and Gloria Culver of Department of Biology, University of Rochester and Professor Walter F. Eanes of Department of Ecology and Evolution, Stony Brook University. Some of the DNA sequence data analyzed for Chapter 2 were provided by Professor James D. Fry. All experiments were performed by the author. Data from experiments described in Chapter 2 and Chapter 3 were jointly analyzed by the author and Professor James D. Fry. The Department of Biology server, Augusta, was used to run computationally demanding programs. Confocal images presented in Introduction were obtained from Department of Biology Imaging Facility. The purification of recombinant enzymes described in Chapter 2 was performed in the laboratory of Professor Gloria Culver. Funding for the research was provided by National Science Foundation and National Institute of Health through grants awarded to Professor James D. Fry.

Chapter 2 is derived from the paper titled “Choose your poison: opposing selection pressures on a detoxifying enzyme constrain adaptation to dietary ethanol in natural populations of Drosophila”, the manuscript of which is under preparation.

Chapter 3 is derived from a paper titled “Parallel functional changes in independent testes-specific duplicates of Aldehyde Dehydrogenase in Drosophila”, the manuscript of which is under preparation.
Chapter 4 is derived from a paper titled “Adaptive cis-regulatory variant of *D. melanogaster* aldehyde dehydrogenase in ethanol adapted natural populations”, manuscript of which is under preparation.
# Table of Contents

Dedication ii  
Biographical Sketch iii  
Acknowledgements iv  
Abstract v  
Contributors and Funding Sources vii  
Table of Contents ix  
List of Tables xi  
List of Figures xii  
Chapter 1 Introduction 1  
   References 16  
Chapter 2 Choose your poison: opposing selection pressures on a detoxifying enzyme constrain adaptation to dietary ethanol in natural populations of *Drosophila* 22  
   Abstract 23  
   Materials and Methods 34  
   Supplementary results 44  
   References 48
Chapter 3  Adaptive cis-regulatory variant of *D. melanogaster* aldehyde dehydrogenase in ethanol adapted natural populations

Abstract 52

Introduction 54

Results 57

Discussion 68

Materials and Methods 733

References 78

Chapter 4  Parallel functional changes in independent testes-specific duplicates of Aldehyde Dehydrogenase in *Drosophila* 82

Abstract 83

Introduction 85

Results 87

Discussion 101

Materials and methods 107

References 112

Chapter 5  Conclusion 116

References 123
List of Tables

Table

1. The genomic locations and frequencies of the potential *DmAldh* eSNPs in
   DPGP and DGRP populations .................................................................61
2. Relative expression level of *Aldh* and *Aldh-dup* in different species. ...............90
3. dN/dS for different site classes (Zhang, et al. 2005) of ALDH and ALDH-dup
   in *Obscura* group. ......................................................................................92
4. Diameter of the SEC mouth in DpseALDH, DpseALDH-dup, DmojALDH,
   and their mutants created *in silico*. ..........................................................98
5. *Aldh* genes and the source of their corresponding RNAseq data......................108
List of Figures

Figure

1-1. Co-localization pattern of mito-GFP and DmALDH-RFP (Larval gut epithelial cell); Nucleus is stained with DAPI stain (blue). ...........................................8

1-2. A schematic diagram showing the mechanism by which reactive aldehydes are produced within mitochondria. .................................................................9

1-3. A possible mechanism by which cis-regulatory variant increasing enzyme concentration can compensate for reduced turnover rate of an enzyme variant. ..............................................................................................12

2-1. Worldwide distribution of frequency of the Phe allele..............................................26

2-2. a) Longitudinal section through the substrate entry channel (SEC) of DmALDH\textsuperscript{Phe} superposed on DmALDH\textsuperscript{Leu} .........................................................................................................28

2-3. a) Comparison between the resistance to acetaldehyde stress of Leu and Phe homozygous insert lines ...........................................................................30

2-4. a) Comparison between the survival of Leu and Phe homozygote transformant lines in presence of hyperoxia........................................................................32

2-s1. Turnover rates of DmALDH\textsuperscript{Leu} and DmALDH\textsuperscript{Phe} for trans-2-Hexenal and trans-2-Octenal................................................................................44

2-s2. Estimated selection co-efficient of the Phe allele from the long term competition experiment. ..................................................................................................45

2-s3. Turnover rates of extracts of Leu and Phe homozygotes ......................................46
2-s4. Expression level of *DmAldh* relative to *β-Tub56D* in Leu and Phe insert homozygotes. .................................................................47

3-1. a) Relative expression level of *DmAldh* (*DmAldh*/βtub56D) in two African and two European populations..................................................58

3-2. a) Association probabilities of the 1450197 SNPs used in association mapping..................................................................................60

3-3. a) Distribution of derived SNP4809T (black) and SNP4809G (grey) in four different populations located at different latitudes.........................63

3-4. Distribution of composite likelihood ratio (CLR) for selective sweep across the *DmAldh* genomic region calculated using 100 grids. ...............64

3-5. a) Schematic diagrams of different types of deletions present in 3’ UTR of *DmAldh*..........................................................................................67

4-1. A) Phylogenetic relationships between *Aldh* and *Aldh-dup* genes. .................88

4-2. A) Front view of the entrance/mouth of the substrate entry tunnel of DpseALDH. .........................................................................................95

4-3. Normalized diameter of the channel mouth in ALDH and duplicated ALDH in different *Drosophila* species......................................................100

4-4. Acetaldehyde dehydrogenase activity of extracts from abdomens of *D. pseudoobscura* and *D. melanogaster*. .......................................................102
Chapter 1

Introduction
Adaptive constraint

Natural selection is an optimizing force: mutations that optimize a phenotype spread in a population and supplant their predecessors. However, adaptive mutations often have deleterious side effects (Andersson, 2003), which undermine their overall beneficial effect on fitness. Opposing effects of the mutations on two fitness related traits can maintain the mutations as polymorphisms, thereby impeding adaptation (Tian et al., 2003; Yang et al., 2003; Luong and Polak, 2007; Todesco et al., 2010). Hence, such conflicting effects of a mutation on fitness components has been referred to as adaptive constraint (Etterson and Shaw, 2001). These adaptive constraints are generally inferred based on statistical methods (Schluter, 1996; Etterson and Shaw, 2001); their molecular basis are largely unknown.

Phenotypic effects of mutations often entail changes in protein structure and function (Hoekstra and Coyne, 2007). Therefore, protein structural constraints are considered good candidates for molecular basis of adaptive constraints, partly due to the different types of physical and thermodynamic constraints that limit number of functional forms for a protein structure (Kim et al., 2007; Feldman et al., 2012; Harms and Thornton, 2013). Nonetheless, whether such adaptive constraints are common in natural populations remains unclear.

Protein structural constraint as adaptive constraint

Initial evidence supporting the role of protein structures in maintenance of genetic variation came nearly 50 Years ago, when Gillespie and Kojima (G & K) pointed out that
naturally occurring functional variants of an enzyme was correlated with substrate specificity of the enzyme (Gillespie and Kojima, 1968). It was hypothesized that the promiscuous enzymes (Type II) were more polymorphic than the single substrate enzymes because each of the variants was specialized on each substrate the enzyme encountered (Gillespie and Langley, 1974; Lunzer et al., 2002).

For next 40 years, G & K’s hypothesis remained untested, in part because the generality of their observation was questioned (Sole-Cava and Thorpe, 1989; Ward et al., 1992), in part because other explanation of their observation seemed equally plausible. For example, type II enzymes could be more polymorphic simply because purifying selection acting on these enzymes is weaker than selection acting on the type I enzymes (Singh, 1976). Furthermore, it was argued that enzyme variants showing difference in turnover rates towards different substrates under experimental conditions had little or no biological relevance (Kacser and Burns, 1973; Middleton and Kacser, 1983). However, evidence collected in the last 10 years from directed protein evolution and bacterial artificial evolution experiments seem to support G & K’s hypothesis (see below).

G & K’s hypothesis can be divided into two parts: the first part postulates that an enzyme cannot be optimized for all substrates; the second part holds that this property contributes to maintenance of enzyme polymorphism. The first part is relatively easy to intuit: if a mutation improves catalytic efficiency (turnover rate or Michaelis constant, or both) for one of the substrates of the enzyme, it is likely to reduce the catalytic efficiency of the enzyme for other substrate(s) (Lunzer et al., 2005; DePristo, 2007; Romero and Arnold, 2009). Nonetheless, the notion of functional tradeoff due to structural constraint in
enzymes has been subject of much debate (Khersonsky and Tawfik, 2010). While it has been shown using directed protein evolution that improvement of enzyme catalytic efficiency for one substrate may not cause substantial changes in its catalytic efficiencies for other substrates (Aharoni et al., 2005), efforts in optimizing catalytic efficiency of other enzymes for multiple substrates in similar, but separate, experiments were unsuccessful (Fasan et al., 2008; Tokuriki et al., 2012). Aside from these experiments, which focused solely on enzyme function and not their effects on organismal fitness, another experiment has demonstrated that tradeoff in enzyme substrate (or co-enzyme) specificity leads to fitness tradeoff across environments imposing different artificial selective pressures (Miller et al., 2006). Although these studies underscore G & K’s hypothesis, they involve engineering enzymes to obtain unnatural functional changes. Hence, whether mutations in natural populations have similar functional effects at molecular and organismal level remains an open question.

**Maintenance of variation: population genetic theories**

Various forms of balancing selection can maintain enzyme variants showing tradeoff between enzyme functions. Perhaps the most prominent among these is the one involving overdominance. For a gene with two alleles, complete overdominance or heterozygote superiority occurs when fitness of the heterozygote is higher than that of either homozygote. The sickle cell anemia is a famous example of maintenance of polymorphism via overdominance: the heterozygotes, who possesses one copy of normal hemoglobin and one copy of sickle cell hemoglobin (HbS), are protected from both malaria and sickle cell anemia (Kwiatkowski, 2005). In contrast, the homozygotes for normal
hemoglobin allele or the Hb\textsuperscript{S} allele are predisposed to the risk of malaria and anemia, respectively. While a theoretical prediction argue in favor of widespread contribution of heterozygote advantage in maintenance of polymorphism (Sellis et al., 2011), lack of examples opposes such claim (Hedrick, 2012).

Apart from complete overdominance, other forms of balancing selection – especially variable selection across time or space, and frequency dependent selection – can also maintain enzyme polymorphism (Ayala, 1974; Rainey et al., 2000; Hedrick, 2006). For example, Levene (1953) first showed theoretically that spatially varying selection can maintain stable polymorphism without overdominance. According to the model, when strength of selection varies across subpopulations separated in space, stable polymorphism can be maintained if weighted harmonic mean fitness of the homozygotes among all subpopulations is less than the standardized mean fitness of the heterozygotes. The model is often criticized for two of its seemingly unrealistic assumptions: number of adult individuals has to remain same in every subpopulation (soft selection) and members of all subpopulation come together every generation and mate randomly. An alternative model, also known as the SAS-CFF model, holds that spatially variable selection pressure can maintain polymorphism when fitness of the heterozygotes, across different environmental conditions, is higher than the arithmetic mean of the fitness of the homozygotes (Gillespie, 1978; Hedrick, 1986). This model, however, requires reversal of dominance: fitness of the heterozygote has to be closer to the fittest genotype under a given environmental condition. In terms of enzyme function, this may result when relative contribution of each enzyme variant towards each biological function of the enzyme changes with environmental
conditions, so that the enzyme variant beneficial for a function will contribute more towards the function than the less advantageous allele (Gillespie, 1976).

Heterogeneity in selection pressure is generally inferred from correlation between varying ecological or environmental factors and frequency of different alleles (Powell, 1971; Johannesson and Tatarenkov, 1997; Powers and Schulte, 1998; Schmidt and Rand, 2001; Salvi et al., 2009; Lee and Mitchell-Olds, 2012). Because experimental data on molecular and functional bases of inferred fitness tradeoffs in these studies is mostly lacking, functional tradeoff cannot be established as the cause of the correlations. Consequently, the number of polymorphism in natural populations maintained due to functional tradeoff is considered low by some authors (Hedrick, 2006; Van Dooren, 2006).

To demonstrate that naturally occurring variants are maintained due to negative pleiotropic effect of adaptive mutations, the molecular and phenotypic bases of fitness tradeoff under different environmental conditions need to be elucidated – a task that is nearly impossible when different functions of a polymorphic gene are unknown and genetic manipulations are not feasible. For example, persistence of many disease variants in human populations suggest that their maintenance may not be solely dependent on mutational inputs or stochastic forces; instead, deletions may be deleterious pleiotropic effects of mutations that are beneficial for some other functions (Carter and Nguyen, 2011). Nonetheless, it is difficult to test experimentally if change in fitness effects of disease variants under different selective regimes is conducive to maintenance of disease polymorphisms. The best possible solution is then to look for such polymorphisms in
natural populations of model organisms and study their phenotypic and fitness effects under relevant environmental conditions.

**Drosophila Aldehyde Dehydrogenase**

The enzyme aldehyde dehydrogenase (DmALDH) of *Drosophila melanogaster* falls into the category of Type II enzymes as described by G & K. The enzyme can oxidize a wide array of aldehydes, albeit with different catalytic efficiency (Rothacker and Ilg, 2008). Acetaldehyde, formed by enzymatic oxidation of dietary ethanol, is the first identified biological substrate of the enzyme. In absence of functional DmALDH enzyme, flies are killed by an ethanol concentration that is easily tolerated by wild type flies, presumably due to accumulation of the highly toxic acetaldehyde (Fry and Saweikis, 2006).

Ethanol resistance of *D. melanogaster* varies across latitudes in all major continents: flies collected from higher latitudes show more resistance to ethanol than their counterparts from lower latitudes (David and Bocquet, 1975; Cohan and Graf, 1985; Chakir et al., 1993; Parkash et al., 1999). Existence of such parallel clines in different continents suggest that increased resistance to ethanol is under selection in temperate region populations of *D. melanogaster*. Because DmALDH plays an important role in ethanol resistance, any replacement polymorphism that increases acetaldehyde turnover rate of the enzyme should be beneficial in ethanol resistant *D. melanogaster* populations. Consistent with this hypothesis, a replacement polymorphism at *DmAldh* was discovered that is associated with increased acetaldehyde turnover rate (Fry et al., 2008). The derived allele, which replaces a leucine residue at 479\(^{th}\) position of the enzyme by a phenylalanine
Figure 1-1. Co-localization pattern of mito-GFP (green) and DmALDH-RFP (red) in larval gut epithelial cell. Signals for mito-GFP (left) and DmALDH-RFP (center) match almost entirely (right), indicating mitochondrial localization of DmALDH. Nucleus is stained with DAPI stain (blue).
Figure 1-2. A schematic diagram showing the mechanism by which reactive aldehydes are produced within mitochondria. The superoxide anion produced at the end of oxidative phosphorylation is broken down into hydroxyl radical (OH•), which react with the polyunsaturated membrane lipids to produce lipid peroxides. The lipid peroxides thus produced give rise to toxic aldehydes.
residue, is more frequent in ethanol resistant temperate fly populations than ethanol sensitive ancestral African and other tropical *D. melanogaster* populations. Fry et al. (2008) showed that the Phe allele is likely beneficial for ethanol resistance, but its frequency does not increase beyond 20% in ethanol resistant temperate fly populations. Because DmALDH is a type II enzyme, it was hypothesized that DmALDH may have other biological substrates and Phe allele perhaps affects catalytic efficiency of DmALDH for those substrates (Fry et al., 2008). This hypothesis is supported by the observation that Leu, the residue present in the ancestral DmALDH, is present in all other sequenced *Drosophila* species, which are not as resistant to ethanol as *D. melanogaster* (Mercot et al., 1994). Supporting the above hypothesis I found that DmALDH protects from reactive aldehydes produced from lipid peroxidation during aerobic respiration, and is localized in mitochondria, the main site of lipid peroxidation (Fig 1-1, Fig 1-2) (Chakraborty and Fry, 2011). Therefore it was hypothesized that although DmALDH<sup>Phe</sup> is advantageous in ethanol resistant populations because it increases acetaldehyde turnover rate, fitness advantage of the Phe allele through increased acetaldehyde turnover rate is undermined by fitness defects resulting from its reduced turnover rate for large, reactive aldehydes produced by lipid peroxidation. Results from investigation of these hypotheses comprise the contents of Chapter 2.

If functional tradeoff shown by the Phe allele undermines its adaptive significance, mechanisms that compensate for the fitness disadvantage are likely to evolve. Cis-regulatory polymorphism can accomplish this by altering enzyme concentration. On the other hand, gene duplication can compensate by making a new copy of *Aldh* that is freed
from the constraint of carrying out the ancestral function at an optimum level. These possibilities are further discussed in the following sections.

**Rescuers of Adaptive constraint: gene expression changes**

Cis-regulatory mutations are often considered as alternative to replacement mutations for accomplishing a phenotypic change (Hoekstra and Coyne, 2007; Stern and Orgogozo, 2008). However, cis-regulatory mutations can also be associated with replacement mutations and modify the functional effects of the latter (Dimas et al., 2008; Olson-Manning et al., 2012; Rzezniczak et al., 2012). Cis-regulatory mutations improve enzyme functionality by increasing *in vivo* enzyme concentration, which elevates turnover rate of the enzyme for all of its substrates (Stam and Laurie, 1996; Eanes, 1999; Goto et al., 2005).

If such a cis-regulatory polymorphism co-occurs with an amino acid polymorphism at a gene, they can improve catalytic activity of the enzyme for a substrate independently, and jointly too. Therefore their adaptive significance could be based on their independent or joint contributions towards a phenotype that is dependent on the turnover rate of the enzyme for the particular substrate. Now, if the coding mutation shows functional tradeoff, the turnover rate of the enzyme for other substrate(s) will go down to suboptimal level (Fig 1-3). In this case, a cis-regulatory variant that increases the enzyme concentration can compensate for the reduced turnover rate of the enzyme variant for the particular substrate, and thus reduce the fitness cost of the enzyme variant.

If increased turnover rate of DmALDH for acetaldehyde provides increased
The Michaelis–Menten equation. Substrate turnover rate \( V \) is dependent on enzyme concentration \([E]_p\).

\[ V = V_{\text{max}} \frac{[S]}{[S] + K_M} \quad V_{\text{max}} = k_2[E]_p \]

The amino acid variant has a tradeoff. It reduces the turnover rate of the enzyme for a particular substrate to suboptimal level \( V_a \).

\[ V_a = k_2[E]_p \frac{[S]}{[S] + K_M} \quad V_a < V_{\text{opt}}, k_2 < k_2 \text{ opt} \text{ or } K_M > K_M \text{ opt} \]

C. Cis-regulatory variant increases enzyme concentration \( n \) times.

\[ V_{ca} = k_2[E]_p \frac{[S]}{[S] + K_M} \quad [E]_p \text{ }' = n[E]_p \quad (n>1) \]

D. Effective turnover rate of the enzyme is now closer to the optimum level.

\[ V_{\text{opt}} \geq V_{ca} = nV_a > V_a \]

Figure 1-3. A possible mechanism by which cis-regulatory variant increasing enzyme concentration can compensate for reduced turnover rate of an enzyme variant. \( V_a = \) turnover rate of the enzyme for a particular substrate in absence of the cis-regulatory variant; \( V_{ca} = \) effective turnover rate of the enzyme in presence of the cis-regulatory variant.
resistance to ethanol, then cis-regulatory mutations that increase DmALDH concentration should also be able to increase ethanol resistance in flies. Hence, such cis-regulatory mutations should be under positive selection in ethanol resistant temperate populations. Because activity of DmALDH varies within $DmAldh^{Leu}/DmAldh^{Leu}$ homozygotes (Fry et al., 2008), it was hypothesized that apart from the amino acid variant Phe, nonreplacement polymorphisms might also contribute to determining DmALDH activity. Accordingly, I investigated if any adaptive cis-regulatory polymorphism exists in the natural populations of $D. melanogaster$ that may offset the fitness cost of the Phe allele. Chapter 3 contains the details of this investigation.

**Rescuers of Adaptive constraint: Gene duplication**

Conserved proteins are constrained from evolving new functions because their structures are optimized for the function they are conserved for. Duplication of the gene encoding such a protein, however, offers the freedom to optimize the same protein for a new function, without incurring any fitness cost (Prince and Pickett, 2002; Hahn, 2009). The optimization of the new gene copy for the new function is accomplished by changing amino acid residues under positive selection (Hughes, 1994; Lynch and Conery, 2000). If the new function is tissue, developmental stage, or sex specific, then the gene copy undergoing optimization for the new function may acquire expression pattern that is limited in space, time, or sex, respectively (Force et al., 1999; Innan and Kondrashov, 2010). Although gene duplication is considered as a principal mechanism by which adaptive constraints due to protein structural constraint are rescued, its actual contribution to adaptation is unclear (Carroll, 2005; Hoekstra and Coyne, 2007).
In this context, it has been noted that a large number of duplicates in animal genomes, originating from ubiquitously expressed genes, acquire testis biased gene expression (Vinckenbosch et al., 2006; Gallach et al., 2010; Baker et al., 2012). One hypothesis suggests that the ancestral genes are constrained from optimizing on testis specific function, whereas the new copy, freed from the constraint, can accomplish that (Gallach and Betran, 2011). Supporting this hypothesis, many of these duplicates have been found to have accumulated amino acid changes under positive selection, presumably to become more efficient for the testis specific function (Gallach and Betran, 2011). However, it remains untested if the amino acid changes associated with ancestral to duplicate functional shifts were truly under structural constraint in the ancestral protein.

Detecting constraint in the ancestral copy entails characterization of functional changes the duplicate has undergone, and identifying the factor(s) that constrained the ancestral gene from undergoing those functional changes. As mentioned previously, structural limitations can constrain protein evolution (see the section “protein structural constraint as adaptive constraint”). Hence, if we know apriori how structural constraint limits a particular protein’s ability to undergo functional shift, all we need to show is that the amino acid changes in the duplicate are relevant to the known structural constraint.

The DmALDH Leu-Phe polymorphism indicates that ancestral, conserved substrate specificity of DmALDH may not be altered without functional tradeoff. Because the aldehyde dehydrogenase in Drosophila is highly conserved, this prediction is likely to hold for ALDH of other Drosophila species as well. If structural constraint in ALDH opposes alteration of ancestral substrate specificity, studying substitutions in Aldh
duplicates that are likely to alter substrate specificity of the ancestral enzyme can help us understand if duplication of Aldh rescues from adaptive constraint.

In two DmAldh paralogs of independent origin, one in Drosophila subgenus and another in Obscura species group, a conserved residue – Thr144 – has been substituted by bulkier apolar amino acid residues. The substitutions are predicted to affect enzyme function in similar manner as Leu479 → Phe479 substitution in DmALDH. The duplicates in both cases have acquired testis biased expression, indicating that the functional shifts are relevant to testis specific roles of the enzyme. I examined the prediction that evolution of the duplicates allowed functional shift that the ancestral enzyme was constrained from undergoing. Details of this study comprise the contents of Chapter 4.
References


Levene, H. 1953. Genetic equilibrium when more than one ecological niche is available. American Naturalist 87, 331-333.


Chapter 2

Choose your poison: opposing selection pressures on a detoxifying enzyme constrain adaptation to dietary ethanol in natural populations of *Drosophila*
Abstract

There are two prevailing views of adaptation: the sweep model, in which selection quickly drives a formerly rare allele to fixation, and the infinitesimal model, in which selection causes subtle allele frequency shifts at numerous loci with small phenotypic effects. A third, less often considered possibility is that adaptation involves large-effect alleles that create a novel benefit, but whose frequency increases are arrested by strong adverse pleiotropic effects. We report here an example of such a situation involving adaptation to naturally-occurring ethanol in D. melanogaster. Aldehyde dehydrogenases (class 2) are highly conserved detoxifying enzymes that appear to be universally present in mitochondria. Although they are capable of detoxifying many aldehydes, the function that accounts for their conservation, as well as their mitochondrial localization, appears to be detoxification of reactive, high-molecular weight aldehydes created when reactive oxygen species attack membrane lipids. In both Drosophila and humans, however, ALDH is also essential for detoxification of acetaldehyde derived from dietary ethanol. In highly ethanol-resistant D. melanogaster populations from the temperate zone, a novel ALDH allele, in which a Phenylalanine replaces a highly conserved Leucine, is consistently present in low to moderate frequencies, but is nearly absent in tropical populations. Here, we show that the Phe substitution causes a substrate specificity shift that makes it advantageous in flies exposed to high levels of ethanol or acetaldehyde, but strongly disadvantageous in the absence of ethanol. The polymorphism is apparently maintained in temperate zone populations by a trade-off between the benefit of the Phe
allele for ethanol detoxification, a trait at a high premium in the temperate zone, and the ancestral function of detoxification of endogenous mitochondrial aldehydes.
Although adaptation is often portrayed as a process in which one allele replaces another, allele frequency changes underlying adaptation to new environments are sometimes modest. For example, population-resequencing studies have shown that in humans and *Drosophila melanogaster*, two originally African species that have spread across the globe relatively recently, adaptation to new regions has taken place largely without fixation of alleles (Coop et al., 2009; Burke et al., 2010; Hancock et al., 2011). Attempts to explain such observations usually invoke an "infinitesimal" model, in which phenotypic change results from subtle shifts in allele frequency at many loci of very small effect (Pritchard and Di Rienzo, 2010). A less-often considered explanation is that some variants that promote adaptation to a new environmental challenge have large effects, but are prevented from fixation because they impair a still important ancestral function (Sellis et al., 2011). Here, we show that this is the case for a mutation in *D. melanogaster* that increases ethanol resistance, a trait more important to temperate populations of this species than ancestral African populations.

In temperate regions, *D. melanogaster* can regularly be found breeding in breweries and wineries, as well as in fermenting fruits with ethanol concentrations as high as 6% (McKenzie and McKechnie, 1979; McKechnie and Morgan, 1982). Although less is known about breeding sites of tropical *D. melanogaster*, temperate populations are consistently more resistant to ethanol poisoning than their tropical and subtropical counterparts, suggesting that selection for ethanol resistance is stronger in the temperate zone (David and Bocquet, 1975; Chakir et al., 1993). Although multiple genes likely contribute to the resistance difference, we focus here on *Aldh*, a gene encoding a
Figure 2-1. Worldwide distribution of frequency of the Phe allele of DmAldh. Frequency of the Phe allele is positively correlated with latitude in North America ($r = 0.641$) and Australia ($r = 0.84$). The Phe allele is absent in most African populations, rare in Southern USA and Northern Australia. Frequency of Phe reaches up to 20% in populations in Northern U.S.A., France, and Southern Australia.
mitochondrial aldehyde dehydrogenase (DmALDH) that, like its orthologue ALDH2 in humans, is essential for detoxification of acetaldehyde, the highly toxic initial breakdown product of ethanol (Fry and Saweikis, 2006). Most *D. melanogaster* have a leucine at the 479th position of DmALDH (466th in the mature enzyme), as do all other species of Diptera for which genome sequences are available. In about 10% of alleles in geographically diverse temperate populations, however, leucine is replaced by phenylalanine (Fig. 2-1). In a laboratory study, the phenylalanine variant appeared to be associated with higher ethanol resistance (Fry et al., 2008).

Aldehyde dehydrogenases of the ALDH2 and closely related ALDH1 groups are typically capable of oxidizing a wide range of aldehydes to their corresponding acids, but can vary considerably in their substrate preference (Sophos and Vasiliou, 2003). A reasonable hypothesis is that the Leu→Phe substitution creates a substrate specificity shift that increases the rate at which acetaldehyde can be detoxified, likely at the expense of detoxification of one or more ancestrally important aldehydes. Indeed, the corresponding position in vertebrates (459) has been identified as one of three critical positions that account for substrate specificity variation of ALDH1/2 enzymes by altering the size of the substrate entry channel (SEC), with the bulkier phenylalanine, as found in human ALDH2, constricting the channel and favoring binding of smaller aldehydes like acetaldehyde (Sobreira et al., 2011). When we aligned the model structures of the Phe466 and Leu466 forms of DmALDH using the crystal structure of human ALDH2 and sheep liver ALDH1 as references, we found that replacing Leu466 with Phe466 constricted the SEC proximal to the active site (Fig 2-2a,b), decreasing SEC volume from 535 to 456 Å³. To determine
Figure 2-2. a) Longitudinal section through the substrate entry channel (SEC) of DmALDH\textsuperscript{Phe} superposed on DmALDH\textsuperscript{Leu} (residues Phe479 and Leu479 have been colored in red and blue respectively). b) The Phe479 residue has been removed to show the extra space (arrowhead) available in DmALDH\textsuperscript{Leu} for substrate binding. The residue in yellow is the catalytic residue cys322. C) Comparison between the turnover rates of DmALDH\textsuperscript{Leu} and DmALDH\textsuperscript{Phe} for various substrates using purified recombinant enzymes. DmAldhphe catalyzes oxidation of acetaldehyde faster than DmALDH\textsuperscript{Leu}. However, substrates larger than acetaldehyde: butanal, hexanal, and benzaldehyde are metabolized by DmALDH\textsuperscript{Leu} increasingly faster than DmALDH\textsuperscript{Phe}. 
how the substitution affects substrate specificity, we over-expressed the Phe and Leu variants *in vitro*, purified the recombinant enzymes, and compared their turnover rates with saturating levels of a variety of aldehydes (Fig 2-2c; supplementary Fig 2-s1). As predicted, the Phe variant detoxifies acetaldehyde faster than the Leu variant, by about 20%. With larger aldehydes, however, the difference is dramatically reversed (Fig. 2-2c), presumably because the smaller SEC of the Phe variant restricts entry of the substrate.

The kinetic results suggest that the Phe466 variant should improve resistance to acetaldehyde and ethanol, which is rapidly converted to acetaldehyde *in vivo* by alcohol dehydrogenase. To test this prediction, we created two sets of transformant lines in an isogenic *Aldh*-null background, carrying either an *Aldh* genomic insert with the Leu variant, or the same insert in which a single nucleotide had been mutated to convert the Leu codon to Phe. Consistent with the predictions, flies with Phe466 survived significantly longer in the presence of toxic levels of acetaldehyde (Fig. 2-3A) and ethanol (Fig. 2-3B) than flies with Leu466.

While the above results likely explain why the Phe466 variant is found in ethanophilic temperate *D. melanogaster* populations, they do not explain why the frequency of the Phe allele is seldom more than 15% in these populations (Fig. 2-1). Notably, the frequency of Phe is similar in southern Australia, the northern U.S., and Europe, even though *D. melanogaster* has been present in Europe for thousands of years, compared to only ~200 in the Americas and Australia. This is not consistent with the hypothesis that the Phe allele is gradually increasing in frequency under weak positive selection, which would predict that its frequency should be highest in Europe (especially
Figure 2-3. a) Comparison between the resistance to acetaldehyde stress of Leu and Phe homozygous insert lines. Male and female Phe homozygotes survive at a significantly higher proportion than the male and female Leu homozygotes. b) Comparison between the resistance to ethanol stress of males and females homozygous for Leu and Phe insert. Male Phe homozygotes show significantly higher survival in presence of ethanol.
given this continent's relative isolation from the tropics, due to the barrier of the Sahara).

An alternative hypothesis, suggested by the conservation of the corresponding position in other Diptera, as well as the impaired detoxification of large aldehydes by the Phe variant, is that the Phe allele remains in low frequency because it is defective in an important ancestral function of DmA LDH. One such function is detoxification of reactive aldehydes generated by lipid peroxidation (Chakraborty and Fry, 2011), a process initiated when reactive oxygen species in mitochondria attack membrane lipids (Esterbauer et al., 1991). Such aldehydes covalently modify proteins and DNA, disrupt their normal functions, and lead to development of various pathological conditions (Berlett and Stadtman, 1997).

Although the spectrum of aldehydes generated by lipid peroxidation in Drosophila has not been characterized, most are likely to be considerably larger than acetaldehyde (Esterbauer et al., 1991), and thus would be expected to be detoxified more slowly by the Phe enzyme than the Leu form.

Consistent with the above hypothesis, the Phe lines survive markedly less well than the Leu lines under hyperoxia, a manipulation that increases ROS production, lipid peroxidation, and resulting protein damage due to carbonylation (Chakraborty and Fry, 2011) (Fig. 2-4a). To determine whether the Phe variant decreases fitness under normal ROS levels, we established four independent experimental populations from F1 hybrids between Phe and Leu transgenic lines and maintained them on ethanol-free medium. By the F10 generation, the frequency of the Phe allele had fallen from 0.5 to an average of 0.27 (range 0.18–0.32) (Fig. 2-4b), indicating a selective disadvantage of phe homozygotes
Figure 2-4. a) Comparison between the survival of Leu and Phe homozygote transformant lines in presence of hyperoxia. b) Frequency of the Phe allele in four replicate populations after being maintained for 10 generations on medium lacking ethanol, starting with equal frequencies.
relative to leu homozygotes of 7 to 27%, depending on the level of dominance assumed (Supplemental Fig 2-s2).

Our results show that the relatively small difference in allele frequencies between tropical and temperate populations (Fig. 2-1) cannot be explained by weak phenotypic effects of the polymorphism: relative to the Leu allele, the Phe allele increases survival in the presence of ethanol by on the order of 10%, while reducing fitness in the absence of ethanol by a similar amount. Importantly, the strength of these phenotypic effects could not have been inferred from patterns of sequence variation in populations. Although the Phe variant occurs on a single haplotype, indicating a single origin, the unbroken haplotype is relatively short (Fry et al., 2008), and the region does not stand out in screens for selective sweeps (Kolaczkowski et al., 2011; Fabian et al., 2012; Mackay et al., 2012).

Our results add to the evidence that alleles with considerable beneficial effects often remain segregating in populations because of deleterious side-effects (Carter and Nguyen, 2011). As many of these evolutionarily young polymorphisms bear molecular signatures elusively similar to those maintained by neutral processes (Thornton et al., 2007; Dudley et al., 2012), current statistical methods used to detect sites under selection are likely to miss them, underestimating the proportion of phenotypically important genetic variation in a population.
Materials and Methods

Allele frequency estimation (Fig. 2-1)

Illumina short reads of genomic DNA of the North American populations (Florida, Pennsylvania, and Maine) were obtained from EBI Short Read Archive, as described in Fabian et al (Fabian et al., 2012). Data for African populations and one French population were obtained from the assembled genome sequence made available by Drosophila Population Genomics Project (www.dpgp.org). Illumina short reads of 20 French isofemale lines were made available by Casey Bergman (Haddrill & Bergman, 2012). Genome sequences of 162 Raleigh lines were obtained from Drosophila Genetics Reference Panel (Mackay et al., 2012).

Paired end reads were aligned to the reference genome sequence of *DmAldh* by Bowtie2 (Langmead and Salzberg, 2012). The SAM files obtained from Bowtie2 were then converted to BAM files using SAMtools (Li et al., 2009) and used for calculating allele frequency.

For the Australian isofemale lines genotyped by restriction enzyme digestion (Fry et al., 2008), we estimated population allele frequencies from the frequencies of lines lacking the Phe allele, assuming that the founding wild caught females were mated by two males on average. Under this assumption, if $p$ is the population frequency of the Leu allele, the probability that a line contains no copies of the Phe allele is $p^6$. 
Protein structure modeling

Structural models of DmALDH<sup>Leu</sup> and DMALDH<sup>Phe</sup> were constructed by Populus (Offman et al., 2008) using structure of human mitochondrial aldehyde dehydrogenase (PDB id 1O04) and sheep liver class 1 aldehyde dehydrogenase (PDB id 1BXS) as the templates. The model structures of DmALDH<sup>Leu</sup> and DmALDH<sup>Phe</sup> were aligned in Pymol. We measured the volume of the substrate entry channel using FRED Receptor v 2.2.5 (Openeye Scientific, Santa Fe, NM) (Hawkins et al., 2010). First, potential substrate-binding spaces were searched within the enzyme with molecular probes. Next, using the prior knowledge of the active site from alignments between hALDH2 and DmALDH, a box is created containing the substrate entry channel. The volume of the channel was then computed using molecular probes.

In vitro expression of recombinant DmALDH

For cloning of DmAldh cDNA into an expression vector, we followed the protocol used by Rothacker and Ilg (Rothacker and Ilg, 2008). Accordingly, we subcloned DmAldh open reading frame sans the mitochondrial leader peptide into the expression vector pMALc2x (New England Biolabs, Ipswich, MA). We used the cDNA clone (GH22814) of DmAldh available from Drosophila Genomics Research Center (Bloomington, IN) as the source of the DmAldh<sup>Leu</sup> cDNA, and a clone gifted by Dr. Thomas Ilg as the source of the DmAldh<sup>Phe</sup> cDNA. The cDNAs for the two alleles, although taken from different sources, do not have any replacement difference other than the Leu-Phe position. Using a pair of
primers originally designed by Rothacker and Ilg (Rothacker and Ilg, 2008), we subcloned the *DmAldh* ORF into the BamHI-HindIII restriction sites, which were located downstream of the open reading frame of the maltose binding protein within the vector pMALc2X. The constructs were sequenced to ensure that no unwanted mutation has been introduced. Next, we transformed TB1 cells (New England Biolabs) with pMALc2x-*DmAldh*, and induced over-expression of the MBP-DmALDH following the manufacturer’s (New England Biolabs) protocol. To purify the protein, we ran the crude cellular extract through a 1ml MBPTrap-HP amylose column (GE Healthcare Life Sciences, PA, U.S.A.) connected to an automated FPLC machine. The purified protein was eluted using 50mM amylose solution, and stored at -80°C with 43% glycerol and 1mM DTT. Enzyme concentrations were measured using Lowry reagent (Sigma-Aldrich, St. Louis, MO).

*Enzyme kinetics assay*

Enzyme assays were carried out at 25°C in cuvettes of 1 cm path length in 1 ml volume of Na2HPO4/NaH2PO4 buffer containing 1mM DTT (pH = 8.5). The reduction rate of NAD+ (Ɛ<sub>340</sub> = 6220 M<sup>-1</sup>cm<sup>-1</sup>) was measured by Ultrospec spectrophotometer. All substrates were purchased from Sigma. Substrates were dissolved in DMSO. For acetaldehyde, butanal, hexanal, and benzaldehyde, we used 1mM ([S]>>K<sub>M</sub>) (Rothacker and Ilg, 2008) substrate in each reaction to estimate the turnover rates of the enzymes. Substrates trans- 2-Hexenal and trans-2-Octenal were used at a concentration of 1uM because they inhibited DmALDH at a concentration of 1mM. Each assay was performed
in three replicates and assays for both forms of the enzyme, DmALDH\textsuperscript{Leu} and DmALDH\textsuperscript{Phe}, for a specific substrate were always conducted in pairs

*Creation of the insert lines*

We created the insertion lines by inserting genomic fragment of *DmAldh* at a known genomic position on the 3\textsuperscript{rd} chromosome by \(\phi C31\) integrase mediated recombination (Bischof et al., 2007). The genomic fragment was amplified from the BAC clone BACR20B09 (Berkeley Drosophila Genome Project) (Hoskins et al., 2000). The BAC clone was extracted and purified using the BACMAX reagent following the manufacturer’s protocol (Epicentre Biotechnologies, Madison, WI). The genomic fragment containing *DmAldh* coding sequence along with 2Kb of both upstream and downstream sequence (total 8153 bp) was amplified using the forward primer – 5’ GAGGGAGGAAAAGGTAAG 3’ - and reverse primer – 5’ TTTAATTATCCTGCCACGCC 3’. As the BAC clone is derived from the genomic DNA of the reference strain, the amplified *Aldh* fragment had the Leu allele. The amplified fragment was cloned into the TOPO XL vector (Invitrogen). The Phe allele was then created by site directed mutagenesis. We used the Phusion DNA polymerase (New England Biolabs, Ipswich, MA) to amplify the TOPO XL- *DmAldh* construct using the reverse primer- 5’CCTGCGTGCCGGAACCTGATGTG 3’- and the forward, mutagenizing primer-5’GGTAAACACCTACAATGTCTTCGCTGCCCAGGC 3’. The genomic fragments containing the Leu and Phe alleles were excised from the TOPO XL-*DmAldh* construct by the restriction enzymes NotI and DpnI and inserted into the pattB vector (Bischof et al., 2007). The inserts were sequenced to confirm that they were free of any unwanted
mutations. The pattB-DmAldh constructs were injected into fly embryos containing attP receptor sites by Bestgene Inc (Chino hills, CA). We made the transformant flies homozygous for the inserts by crossing them with a 3rd chromosome balancer (TM3, Sb) stock. Next, we brought these insert carrying 3rd chromosome into the genetic background of a null DmAldh mutant strain [Line 17 in (Fry and Saweikis, 2006)]. We used three transformant lines derived from independent integrations for each allele for our experiments. In subsequent sections, we refer to these three transformant lines as sublines and each genotype, Leu or Phe, as lines.

**ALDH activity assay of the insert lines**

To compare the aldehyde dehydrogenase activity of the insert lines, we collected 0-2 day old male flies which were then aged for another 2 days in vials with medium. For each assay, 20 flies were ground in grinding buffer (0.25M sucrose, 5mM EDTA, 15 mM Triton X-100, 5 mM DTT) with 25ul of grinding buffer per fly. Protease inhibitor (Roche Applied Science) was added to the grinding buffer to prevent proteolytic degradation of the enzymes. After grinding, the extract was kept on ice for 15 minutes and then centrifuged at 16,300x g for 20 minutes. The supernatant was transferred to chilled 1.5ml tubes. Protein concentration of the supernatant was measured by Qubit fluorometer following manufacturer’s (Life technologies, Grand Island, NY) protocol. Enzyme activity was measured in the same way as for the purified recombinant proteins, except the assay buffer contained 0.02M pyrazole to inhibit aldehyde dehydrogenase activity of the alcohol
dehydrogenase enzyme (DmADH) present in the extracts. The difference between
DmALDH activities between the lines (Supplementary Fig 2-s3) were similar to that found
by purified recombinant proteins.

*Gene expression assay*

To rule out the possibility that observed difference in ALDH activity between the
Leu and Phe insert could arise from difference in expression level of *DmAldh* between the
two lines, we compared the amount of *DmAldh* transcript present in the two insert lines.
We quantified the relative abundance of *DmAldh* transcript with respect to the transcript
level of the βTub56D gene in six sublines by real time PCR. Total RNA from 2-4 day old
15 male flies for each subline was extracted by RNAeasy mini kit (Qiagen, Valencia, CA)
and cDNA was synthesized by iScript cDNA synthesis kit (Biorad, Hercules, CA).
Abundance of Aldh transcript was measured using the internal standard βTub56D as the
reference. Real time PCR reactions were performed in Applied Biosystems 7300 real time
PCR system (with 3 wells per subline) using TaqMan probes (Dm01809880_g1 for Aldh
and Dm02362299_u1 for βTub56D) following manufacturer’s protocol (Life
Technologies). Relative expression of Aldh (ΔCₜ) did not differ between lines
(supplementary Fig 2-s4).
**Oxidative stress assay**

Parental flies of six sublines were placed into vials with medium on day 0 at a density of 40-50 flies per vial. The vials were turned over after 6 days, and then cleared on day 10. Hyperoxia resistance was measured by placing 15 male flies emerging between day 10- day12 in vials with medium, with 10 vials per subline. Flies were allowed to recover from anesthesia for a day. One day later, the vials were placed inside an airtight plastic container with an inlet tube through which 100% oxygen entered from a tank. Oxygen concentration, as determined by an oxygen meter (Extech Instruments, Waltham, MA), was maintained in the range 90-100%. Flies were transferred to vials with fresh medium after the third day and survival of the flies was recorded after the fifth day. The assay was replicated a second time using flies derived from a different generation.

**Acetaldehyde resistance assay**

Parental flies were placed in vials with medium on day 0 at a density of 40-50 flies per vial. The vials were turned over after the 6th day. On day 11, newly emerged flies were discarded. On day 13, flies were transferred to vials containing medium. On day 15, flies were sorted into groups of 20 males or females and placed into new vials with medium. After 24 hours, flies from each collecting vial were transferred to assay vials. Each assay vial had 0.5g of cotton ball moistened with 2ml of 2.5% sucrose with 0.85% acetone at the bottom of the vial. Acetone inhibited alcohol dehydrogenase (DmADH) enzyme and thus removed its contribution to detoxification of acetaldehyde (Barbancho, 1992). A second
0.5g cotton ball was placed into the middle of each vial, which was then sealed by a cork. After 8 hours the corks were briefly removed, and 700ul 2.5% acetaldehyde solution was added to the top of the second cotton ball in assay vials containing male flies. Vials containing females received 3.5% acetaldehyde because 2.5% induced zero mortality in female flies. The assay was set up with 4-5 vials per sub-lines per sex. The number of dead flies was counted after the third day.

_Ethanol resistance assay_

Parental flies were set up on day 0 at a density of 40-50 flies per vial. The vials were turned over after sixth day to remove the parental flies. Vials were cleared on day 11 to discard any newly emerged flies. On day 13, flies were transferred to vials with medium. On day 15, 20 flies were placed into new vials with medium. On day 16, assay vials were prepared by inserting 0.5g cotton ball at the bottom of the vial and adding 2ml 5% sucrose solution to the cotton ball. Flies, either males, or females, were transferred to an assay vial and then a second cotton ball weighing 0.5g was pushed to the middle of the vial. Then, 800ul 25% ethanol was added to the top cotton ball of the vials containing the females; males received 800ul of 35% ethanol. After adding ethanol, the vials were securely corked to prevent loss of ethanol by evaporation. Number of dead flies was counted after the second day. Each assay was replicated 3-4 times using flies derived from different generations. Each assay consisted of 6-8 vials per sub-line per sex.
Long term competition experiments

200 males and 200 virgin females of each subline were collected. Males from a randomly chosen Leu or Phe subline were crossed to virgin females from a randomly chosen Phe or Leu subline to obtain the F1 Leu-Phe heterozygotes in the F1 generation. 1200 F1 Leu-Phe heterozygotes were collected from each of four types of crosses, and distributed over 20 vials with medium lacking ethanol to establish four populations of Leu-Phe heterozygotes. The adults were removed from the vials on the 6th day and flies emerging on the 14th day from each population were pooled and redistributed over 20 new vials. In 10th generation 25 males from each population were collected for genotyping.

Fly genotyping

Genomic DNA was extracted separately from individual flies. Part of the genomic sequence of \textit{DmAldh} insert containing the Leu-Phe polymorphic site was amplified using the forward primer - 5’ CCGATGTCCAGGATGATATG 3’ and reverse primer – 5’ CATATGTACTAGATAGAAATG 3’. The primer set does not amplify endogenous \textit{DmAldh} locus because the reverse primer binding site in the insert is split by an insertion (~50bp) present in the endogenous \textit{DmAldh}. Phe and Leu alleles were identified using the allele specific restriction cut sites of BbsI and MnII restriction enzymes respectively.
Statistical analyses

For the acetaldehyde resistance experiment statistical significance of the effect of line (Phe or Leu) was tested by nested ANOVA, with sublines (random effect) nested within lines (fixed effect). For the oxidative stress and ethanol resistance experiments, random effects of block (i.e., assay generation), as well as the line x block and subline x block interactions, were also included, and non-significant random effects were dropped from the analysis if doing so improved the fit of the model, as judged by a decrease in both AIC and BIC. Prior to the analysis, all survival data were arcsine square root transformed.
Supplementary results

Figure 2-s1. Turnover rates of DmALDH\textsuperscript{Leu} and DmALDH\textsuperscript{Phe} for trans-2-Hexenal and trans-2-Octenal, two of the known products of lipid peroxidation (Novotny et al., 1994). Error bars represent S.E.M.
Figure 2-s2. Estimated selection co-efficient of the Phe allele from the long term competition experiment (9 generations of selection). Selection co-efficients were estimated by simulation iterating over the full range of selection coefficient \([0,1]\), and selecting the value that, starting with initial allele frequency of 0.5, yielded the observed F10 allele frequencies. The Phe allele was assumed recessive, co-dominant, or dominant with respect to the Leu allele. Error bars represent S.E.M among the four replicate populations.
Figure 2-s3. Turnover rates of extracts of Leu and Phe homozygotes. Turnover rate of extract of Phe homozygotes for acetaldehyde is higher than Leu homozygotes (t-test, df = 4, p = 0.02); Turnover rates of Leu homozygote extracts are higher than Phe homozygotes for hexanal (t-test, df = 4, p = 0.01) and benzaldehyde (t-test, df = 4, p = 0.01). Turnover rates for butanal are not significantly different between the extracts of the two homozygotes (t-test, df = 4, p = 0.19). Error bars represent S.E.M.
Figure 2-s4. Expression level of DmAldh relative to β-Tub56D in Leu and Phe insert homozygotes. Difference between DmAldh expression levels in Leu and Phe are not statistically significant in males (t-test, df=4, p=0.16) or females (t-test, df = 4, p=0.27). Error bars represent S.E.M.
References


Chapter 3

Adaptive cis-regulatory variant of *D. melanogaster* Aldehyde Dehydrogenase in ethanol adapted natural populations
Abstract

Adaptive mutations are either replacement or cis-regulatory. The latter type is predicted to play a larger role in adaptation than the former. Yet, the number of known adaptive cis-regulatory variants are relatively few and biased towards atypical mutations, those with conspicuous phenotypic effects. To address these issues, I studied cis-regulatory variants of *D. melanogaster* aldehyde dehydrogenase enzyme gene (*DmAldh*). DmAldh detoxifies acetaldehyde formed during oxidation of dietary ethanol. Increased acetaldehyde turnover rate should lead to faster acetaldehyde detoxification, and hence increased resistance to ethanol. Cis-regulatory variants of *DmAldh* that increases transcript level and enzyme concentration should therefore be beneficial in populations where increased resistance to ethanol is adaptive. I found an intronic polymorphism (eSNP) associated with two-fold variation in transcript level of *DmAldh*. Frequency of the high expression variant (eSNP-T) is higher in European and North American temperate region populations than in the more ethanol sensitive ancestral African populations, consistent with beneficial role of the variant in ethanol metabolism. Consistent with its predicted adaptive significance, eSNP-T has undergone a selective sweep in a European population. In addition to the eSNP, a set of overlapping 3’ UTR deletions segregating in a North American population is also associated with increased *DmAldh* expression. The deletions are predicted to increase expression by disrupting binding of miRNA miR980 to *DmAldh* 3’ UTR. Consistent with this prediction, overexpression of miR980 in a background with full-length, ancestral Aldh 3’ UTR reduces DmAldh activity. Interestingly, eSNP-T is in
linkage disequilibrium with the Phe allele described in Chapter 2, suggesting that high expression may offset the fitness disadvantage of the Phe allele in absence of ethanol.
Introduction

Mutations can optimize adaptive phenotypes in two different ways: they either alter amino acid sequence or change cis-regulatory sequences. Most examples of adaptive mutations involve protein coding changes, which seem to suggest that adaptive protein coding changes are more common than regulatory changes (Hoekstra and Coyne, 2007). However, overrepresentation of protein coding changes among the known adaptive mutations could also be due to the ease of their detection; there is nothing analogous to the “genetic code” for cis-regulatory variants (Wray, 2007). Hence, detecting cis-regulatory mutations underlying adaptive phenotypic change requires extensive experimental work (Wittkopp and Kalay, 2012). Consequently, despite the prediction that regulatory variants contribute to adaptation more frequently than coding variants (Carroll, 2008; Stern and Orgogozo, 2008), examples of adaptive regulatory variants are relatively rare.

A common approach to identify adaptive cis-regulatory variants is to genetically map the variants underlying heritable phenotypic variation (Cheung and Spielman, 2009). Genetic variants discovered by this method are often biased towards variants with conspicuous phenotypic effects (Wray, 2007). One of the best examples of adaptive cis-regulatory variants in *D. melanogaster*, the DDT resistance alleles of *D. melanogaster Cyp6g1* gene, is a notable example of this kind of adaptive cis-regulatory change (Daborn et al., 2002). Most traits, in contrast, show continuous variation – subtle phenotypic changes resulting from contributions from many variants. Hence, typical adaptive cis-regulatory variants are less likely to have striking phenotypic effects (Mackay et al., 2009; Rockman, 2012; Wittkopp and Kalay, 2012). An alternative approach is often adopted that
can partially circumvent the bias of detecting large effect cis-regulatory changes. This approach identifies genes showing polymorphic expression pattern and then sequence variants associated with the expression variation are screened for signatures of positive selection (Fraser et al., 2010; Fraser et al., 2011). However, a potential disadvantage of this approach is that variants found by this method may not be linked to any known adaptive phenotype (Saminadin-Peter et al., 2012). This limitation can be overcome by studying cis-regulatory polymorphisms that are located within genes with known contributions towards adaptive phenotypes.

I examined potential cis-regulatory variants of *D. melanogaster* aldehyde dehydrogenase enzyme gene (*DmAldh*). *DmAldh* is an indispensable part of the genetic system that controls resistance to ethanol – an ecological adaptation in *D. melanogaster* (Fry and Saweikis, 2006). Increased resistance to ethanol is adaptive in temperate region *D. melanogaster* fly populations across all major continents (David and Bocquet, 1975; Chakir et al., 1993; Parkash et al., 1999). In contrast to their tropical African ancestors, which are easily killed by 6-8% ethanol, flies from temperate regions can survive even in presence of 12-15% ethanol. The biochemical basis of such high ethanol tolerance comprises of an efficient, two step ethanol detoxification process: the enzyme alcohol dehydrogenase (DmADH) first converts ethanol to more toxic intermediate acetaldehyde (Geer et al., 1985; Anderson and Barnett, 1991), which is then converted to less toxic acetate either by the enzyme aldehyde dehydrogenase (DmALDH) or by DmADH (Heinstra et al., 1989; Anderson and Barnett, 1991; Fry and Saweikis, 2006). Previously, multiple polymorphisms in coding and non-coding part of *DmAldh* gene had been
implicated in increasing the rate of ethanol detoxification, either through increased DmADH concentration or due to improved catalytic efficiency of DmADH (David et al., 1986; Laurie et al., 1991; Stam and Laurie, 1996). The high expression variants and catalytically superior coding variant appear to contribute jointly in determining the high DmADH activity in ethanol tolerant temperate populations, where the variants are considered adaptive (Laurie et al., 1991; Berry and Kreitman, 1993; Laurie and Stam, 1994). Nonetheless, much evidence suggests that genetic variants at other loci contribute to latitudinal variation in ethanol resistance in natural populations of *D. melanogaster*.

Previously, while searching for sequence variants in other candidate genes underlying ethanol resistance, an amino acid polymorphism in *DmAldh* associated with higher turnover rate for acetaldehyde was discovered in natural populations of *D. melanogaster* (Fry et al., 2008). The derived allele, which substitutes an ancestral leucine (Leu) residue at position 479 of the enzyme by a phenylalanine (Phe), is nearly absent in ethanol sensitive tropical fly populations. But frequency of the Phe allele reaches up to 20% in temperate populations (Fry et al., 2008), likely because Phe allele is more advantageous relative to the ancestral Leu allele in detoxification of acetaldehyde derived from ethanol (Chapter 2). Nonetheless, frequency of the Phe allele in temperate populations remains low, suggesting that its contribution to variation in ethanol resistance is likely to be small.

Interestingly, activity of DmALDH has been found to vary considerably among the *DmAldhLeu/DmAldhLeu* homozygotes (Fry et al., 2008), indicating that other nonreplacement polymorphisms may contribute to determining DmALDH activity. If
increased concentration of DmALDH confers increased resistance to ethanol, nonreplacement polymorphisms associated with increased DmALDH concentration should be under positive selection in ethanol resistant fly populations.

Increase in enzyme concentration can be achieved via cis- or trans-regulatory changes (Latchman, 2010). I investigated polymorphisms which have potential cis-regulatory contributions in determining the DmAldh transcript level and hence DmALDH enzyme activity. By association mapping using genomewide SNPs, I have found at least one expression SNP (eSNP) located within a DmAldh intron. The high expression allele is uncommon in ethanol sensitive ancestral African populations, but found at higher frequencies in temperate European and North American populations. Interestingly, the haplotype bearing the derived allele has undergone a selective sweep in a European population, indicating beneficial role of the variant in ethanol resistant temperate populations. Additionally, a deletion polymorphism within 3’ UTR detected previously (Fry & Stahlhut, unpublished data) also contributes to variation in DmAldh transcript level.

**Results**

**Variation in transcript level and enzyme activity**

To examine if increased DmAldh expression evolved in temperate fly populations, I compared DmAldh expression level between male flies from European (Vienna and Netherlands) and ancestral African populations (Cameroon and Tanzania) by real time PCR. Transcript levels of DmAldh of European flies on average is 2 fold higher (t= 26.31,
Figure 3-1. a) Relative expression level of DmAldh (DmAldh/βtub56D) in two African and two European populations. b) Relative DmAldh (normalized with respect to lowest transcript level) transcript level in 40 DGRP inbred lines. Relative acetaldehyde turnover rate in DmAldh$^{Leu}$/DmAldh$^{Leu}$ DGRP homozygotes. Turnover rates for each line was normalized by the lowest acetaldehyde turnover rate. d) Regression of acetaldehyde turnover rate on DmAldh transcript level in 40 DGRP lines.
df =2, p =0.0007) compared to the flies expression level also varies among the North American flies – as indicated by microarray based comparison of DmAldh transcript level in 40 inbred lines from Raleigh (North Carolina, U.S.A.) (Fig 3-1b). Regression of acetaldehyde turnover rate on transcript level of DmAldh among DmAldh^{Leu}/DmAldh^{Leu} homozygotes in these lines shows that acetaldehyde dehydrogenase activity of DmALDH is positively correlated with transcript level (r=0.64, n=32, df= 1, p=1.2×10^{-4}, t test) (Fig 3-1d). Hence, if any sequence variant increased DmAldh transcript level, the variant would also increase acetaldehyde turnover rate.

**Association of SNPs with high expression in DGRP lines**

To locate the candidate cis-regulatory polymorphisms for DmAldh high expression, a genome-wide association mapping was performed using 1450197 SNPs present in the genomes of the 40 DGRP lines (Ayroles et al., 2009; Mackay et al., 2012). The SNP (SNP4809) showing strongest association (p = 1.65X10^{-9}) is located within the 3\textsuperscript{rd} intron of DmAldh (Fig 3-2a, 3-2b) (Table 3-1). The SNP4809 is the only SNP among the 1450197 SNPs that shows statistically significant association after a Bonferonni correction for multiple testing. Among the top seven hits (p<10^{-6}) from the genomewide scan for SNPs associated with high DmAldh expression, five other DmAldh SNPs were also present (Table 3-1). Three among these –SNP5298, SNP5325, and SNP5940 – are synonymous coding changes, and two others – SNP6695 and SNP7377 – are located at the 3’ intergenic region (Fig 3-2b) (Table 3-1). Another SNP (SNP6131) located within the 3’ UTR of DmAldh is
Figure 3-2. a) Association probabilities of the 1450197 SNPs used in association mapping. The DmAldh SNP (SNP4809; see Table 3-1 and text for more details) showing strongest association with increased DmAldh expression level has been marked with arrow. b) Relative positions of the top hits from association mapping on DmAldh gene. Positions 1 and 8179 correspond to the 5’ and 3’ ends of DmAldh genomic region.
Table 3-1. The genomic locations and frequencies of the potential DmAldh eSNPs in DPGP and DGRP populations.

<table>
<thead>
<tr>
<th>Name/relative location</th>
<th>Genomic position</th>
<th>African freq. (n=110)</th>
<th>European freq. (n=9)</th>
<th>DGRP freq. (n=162)</th>
<th>Ancestral/derived</th>
<th>Association with high expression (p)</th>
<th>Effect size</th>
<th>Expression Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>4809</td>
<td>9390197</td>
<td>0.12</td>
<td>1.0</td>
<td>0.3</td>
<td>G/T</td>
<td>1.65x10^{-09}</td>
<td>-0.37</td>
<td>1.68</td>
</tr>
<tr>
<td>5298</td>
<td>9390686</td>
<td>0.11</td>
<td>0.3</td>
<td>G/A</td>
<td>8.7 x10^{-07}</td>
<td>-0.34</td>
<td>1.59</td>
<td></td>
</tr>
<tr>
<td>5325</td>
<td>9390713</td>
<td>0.12</td>
<td>1.0</td>
<td>0.3</td>
<td>C/T</td>
<td>8.7 x10^{-07}</td>
<td>-0.34</td>
<td>1.59</td>
</tr>
<tr>
<td>5940</td>
<td>9391328</td>
<td>0.2</td>
<td>0.89</td>
<td>0.29</td>
<td>A/G</td>
<td>8.7x10^{-07}</td>
<td>-0.34</td>
<td>1.59</td>
</tr>
<tr>
<td>6131</td>
<td>9391519</td>
<td>0.23</td>
<td>0.89</td>
<td>0.41</td>
<td>A/T</td>
<td>1.02x10^{-06}</td>
<td>-0.33</td>
<td>1.58</td>
</tr>
<tr>
<td>6220</td>
<td>9391608</td>
<td>NC</td>
<td>0.89</td>
<td>0.43</td>
<td>C/T</td>
<td>3.75 x10^{-05}</td>
<td>-0.31</td>
<td>1.58</td>
</tr>
<tr>
<td>6695</td>
<td>9392083</td>
<td>0.15</td>
<td>0.89</td>
<td>0.33</td>
<td>G/C</td>
<td>7.5 x10^{-07}</td>
<td>-0.33</td>
<td>1.57</td>
</tr>
<tr>
<td>6795</td>
<td>9392183</td>
<td>0.11</td>
<td>0.89</td>
<td>0.33</td>
<td>G/T</td>
<td>1.52 x10^{-05}</td>
<td>-0.33</td>
<td>1.53</td>
</tr>
<tr>
<td>7377</td>
<td>9392765</td>
<td>0.11</td>
<td>0.89</td>
<td>0.29</td>
<td>C/T</td>
<td>1.21x10^{-06}</td>
<td>-0.33</td>
<td>1.58</td>
</tr>
</tbody>
</table>
also present among the top ten SNPs (p < 10^{-5}) showing strongest association. Nonetheless, after accounting for the variation in \textit{DmAldh} due to SNP4809, none of the other SNPs showed in Table 3-1 showed significant association (two-way factorial ANOVA, p>0.05) with \textit{DmAldh} expression level.

**Distribution of SNP4809T in different populations**

If SNP4809 is advantageous in ethanol resistant temperate populations, they should be present at high frequencies in the temperate European populations but rare in the ethanol sensitive African populations. The allele showing the strongest association with high \textit{DmAldh} expression in DGRP lines –SNP4809T – is present at high frequencies in two French populations but rare in most ancestral fly populations in Africa (Table 3-1; Fig 3-3b). Ethanol resistance in North America is positively correlated with latitude: flies collected from the higher latitudes show higher resistance to ethanol than flies collected from lower latitudes. Consistent with the beneficial role of SNP4809T in ethanol resistance, frequency of the allele is higher in three North American populations (North Carolina, Pennsylvania, and Maine) than a North American population (Florida) located at a lower latitude relative to the other three (Fig 3-3a).

**Selective sweep in French population**

High frequency of 4809T allele in the French populations indicate that spread of the high expression allele in French populations might have been facilitated by positive
Figure 3-3. a) Distribution of derived SNP4809T (black) and SNP4809G (grey) in four different populations located at different latitudes (FL - 25°32', NC - 35°8', PA - 39°53', ME - 44°) in North America. b) Frequency of SNP4809T and SNP4809G in two French populations and sub-Saharan ancestral African populations.
Figure 3-4. Distribution of composite likelihood ratio (CLR) for selective sweep across the DmAldh genomic region calculated using 100 grids. Significance level is determined from the distribution of likelihood of sweep in simulated neutral samples. 4809 is located very close to the inferred peak and the CLR of the grid containing 4809 is significant at 9% level.
selection. I tested the hypothesis that SNP4809T has increased to high frequency in French populations under positive selection in the past. I used the composite likelihood ratio (CLR) method which compares the likelihood of selective sweep to the likelihood of neutral evolution as the cause of the observed site frequency spectrum for a sample of sequences (Kim and Stephan, 2002; Nielsen et al., 2005). Sweep is inferred if the maximum likelihood (ML) value for selective sweep is higher than ML value for sweep in 95% of the neutral samples. I conducted the test using French sequences from the DPGP sample. Maximum composite likelihood within the 8kb window of DmAldh genomic sequence is statistically significant for the grids near 5300 bp (p=.02) (Fig 3-4). Interestingly, SNP4809 was located close to the peak, having higher composite likelihood ratio than 91% of the simulated neutral samples.

**Candidate deletion polymorphisms**

Several overlapping deletions ($\Delta^{UTR}$) located within DmAldh 3’ UTR, which were first discovered in North American populations (Fry & Stahlhut, unpublished data), are present in the 40 DGRP lines. The DGRP sequences are assembled from short reads, so inferring deletion boundary from these sequences can be ambiguous. Therefore, to determine the frequency of the deletions in these lines, DmAldh 3’ UTR from 40 DGRP lines were re-sequenced by Sanger method. The $\Delta^{UTR}$ length polymorphism in DGRP sample was similar to the $\Delta^{UTR}$ length polymorphism found previously in a larger set of DmAldh sampled from different worldwide populations (Fry & Stahlhut, unpublished data) (Fig 3-5a). A test of association with high expression based on a simple linear model
suggested that $\Delta^{\text{UTR}}$ was associated with high $DmAldh$ expression (one way ANOVA, $F=33.19, \text{df}=1, p=1.21 \times 10^{-6}$). The effects of different deletion lengths on expression level were not significantly different ($F=0.67, \text{df}=2, p=0.53$), suggesting that deletion of a cis-regulatory element common to all deletions was responsible for effects of different $\Delta^{\text{UTR}}$ on $DmAldh$ transcript level. For this analysis the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ deletion in Figure 5a were combined into one group due to their similar structure. The association between $\Delta^{\text{UTR}}$ and high expression was still significant ($\text{df}=1, p=0.029$, two way factorial ANOVA) after accounting for the variation in expression due to SNP4809 ($\text{df}=1, p=8.7 \times 10^{-11}$, two way ANOVA). Interestingly, the positive effects of $\Delta^{\text{UTR}}$ and SNP4809 on $DmAldh$ expression interacted in a negative (diminishing returns) manner ($p<0.01$, two way factorial ANOVA).

**Functional significance of $\Delta^{\text{UTR}}$**

Because 3’ UTR is a common target of microRNAs, I investigated if the deletion removed any microRNA (miRNA) target site. Two programs, Targetscan and Miranda, independently predicted that $D. melanogaster$ miRNA miR980 is likely to bind within $DmAldh$ 3’ UTR. The predicted binding site is at the 5’ boundary of a 29bp deletion ($\Delta^{\text{UTR}29}$) (Fig 5a) that is present in 10 of the total 12 sequences that possess the 3’ UTR deletions among the 40 DGRP lines. The deletions ($\Delta^{\text{UTR}65}$) in the remaining 2 sequences were located right next to miR980 binding site. So those deletions can potentially influence binding of miR980.
Figure 3-4. a) Schematic diagrams of different types of deletions present in 3’ UTR of DmAldh. The predicted binding sites for miR980 and miR992 are also shown. The bottommost 3’ UTR lacking any deletion represents the ancestral state. b) Acetaldehyde turnover rates in flies with overexpressed miR-992 and miR992. Acetaldehyde turnover rate goes down (p=0.019) upon overexpression of miR980.
Removal of microRNA binding site from 3’UTR upregulates expression of a gene (He and Hannon, 2004), so deletion of miR980 binding site from DmAldh 3’ UTR should increase DmAldh transcript level and DmALDH enzyme activity. To investigate the role of miR980 in controlling DmALDH activity, I over-expressed miR980 from a heat shock inducible expression construct of miR980. Consistent with the predicted role of miR80 in DmAldh transcription regulation, overexpression of miR-980 reduced DmALDH by approximately 20% (t = 4.74, df=2, p= 0.019) (Fig 5b). As a control for enzyme activity changes resulting from heat shock and microRNA over-expression, I over-expressed another microRNA – miR992. MiR992 was predicted by Targetscan, but not Miranda, to bind within DmAldh 3’ UTR. Nonetheless, no reduction in DmALDH activity (t test, df=2, p= 0.5) was observed when miR992 was over-expressed (Fig 3-5b). Although these results are preliminary, they suggest that the deletions might contribute to natural variation in DmALDH activity level by disrupting the miR980 binding site.

Discussion

DmAldh can potentially contribute to increased resistance to ethanol through faster detoxification of acetaldehyde. This can be achieved either by replacement mutations which improve catalytic efficiency of the enzyme for acetaldehyde, or by cis-regulatory mutations that increase in vivo concentration of the enzyme. A DmAldh replacement polymorphism was described previously that increases turnover rate of DmALDH for acetaldehyde (Fry et al., 2008) (Chapter 2). As shown here, turnover rate of DmALDH for acetaldehyde may also be improved by increasing the transcript level of DmAldh. Because
increased DmALDH activity is potentially advantageous for ethanol metabolism, increased *DmAldh* transcript level should be beneficial in temperate *D. melanogaster* populations, where enhanced resistance to ethanol is adaptive. Consistent with this prediction, I have found that *DmAldh* is expressed at approximately 2 fold higher level in the European populations relative to the considerably more ethanol sensitive African populations. Using genomewide SNP data and *DmAldh* transcript level in 40 North American inbred lines, I have identified an intronic SNP that is associated with high *DmAldh* expression.

The high expression allele is rare in ethanol sensitive African populations but present at high frequencies in French populations. A test for selective sweep based on composite likelihood ratio suggests that frequency of the high expression haplotype in the French population increased rapidly under selection in recent past. Similar but less striking frequency differentiation of SNP4908 between ethanol resistant and relatively ethanol sensitive populations is also observed among North American *D. melanogaster* populations. Frequency of 4908T is lowest in a Florida where fly populations show less resistance to ethanol than the other three North American populations. Frequency of the SNP4908-T allele also appears to be different between a Northern (Queensland, q= 0.28) and a Southern (Tasmania, q=0.6) Australian populations. Fly populations from Southern Australia are more resistant to ethanol than their counterparts from Northern Australia. Therefore, relatively higher frequency of SNP4809T in Southern Australia is also consistent with beneficial role of SNP4809T in ethanol resistant populations. However, these estimates of allele frequencies are based on low read depth (see Materials and Methods) and therefore may be less reliable as estimates for population allele frequencies.
Nonetheless, these observations indicate that increased *DmAldh* transcript level is under positive selection in ethanol resistant populations, likely due to its advantageous role in ethanol metabolism.

Apart from SNPs, I have also tested if a previously discovered 3’ UTR deletion polymorphism (Fry & Stahlhut, unpublished data) is associated with variation in *DmAldh* expression level. The deletions not only appear to contribute to increase in DmAldh transcript level, but its effect interacts with that of SNP4809. The deletion removes a predicted target site for the miRNA miR980. MiRNA binding typically downregulates transcript level. So deletion of a miRNA binding site would be consistent with the prediction that the deletion increases *DmAldh* transcript level. Accordingly, I have found that overexpression of miR980 reduces DmALDH activity. However, two sequences among 40 DGRP lines had a 65 bp deletion that is located immediately upstream to the miR980 predicted binding site (Fig 3-5a). Base pair modification in the vicinity of miRNA binding site is known to alter transcription regulation by miRNA (Haas et al., 2012). It is therefore possible that this 65bp deletion influences miR980 binding. Furthermore, although the deletions appear to be present in French populations, their actual frequency in those populations remains unknown. Therefore, it is currently unclear if they are adaptive in ethanol resistant *D. melanogaster* populations.

The results described here, along with the previously described amino acid polymorphism, provide a rare glimpse of alternative pathways of phenotypic evolution accomplished via the same gene. The derived replacement variant of *DmAldh* – Phe – is beneficial in ethanolophilic temperate populations, yet its frequency is low in these
populations. Inability of the Phe allele to increase in frequency in temperate populations is attributed to functional trade off: Phe enzyme detoxifies large, harmful aldehydes generated as byproducts of normal respiration more slowly than the ancestral leu enzyme (Chapter 2). The resulting fitness disadvantage of phe in absence of ethanol is also reflected in its near absence in ethanol sensitive tropical African, North American, and Australian populations (Fry et al., 2008). While phenotypic effect of DmAldh high expression is similar to the effect of DmAldhPhe on acetaldehyde detoxification rate and possibly on ethanol resistance, their overall fitness effects do not appear to be similar. The Phe allele improves turnover rate of the enzyme only for acetaldehyde, but compromises turnover rates of the enzyme for larger aldehydes. In contrast, increase in enzyme concentration increases turnover rate of the enzyme for all of its substrates, without incurring any fitness defects through enzyme function. Evidence in favor of this hypothesis comes from the fact that high expression haplotype is much more abundant in ethanol resistant temperate populations than the Phe allele.

Expression changes can modify phenotypic and fitness effects of amino acid variants (Olson-Manning et al., 2012). It seems plausible that high expression partially offsets the fitness cost associated with the Phe allele, by improving its turnover rate for large aldehydes. This hypothesis is supported by the observations that Phe allele and 4908T are in linkage disequilibrium in the DGRP lines. Whether this genetic association reflects functional compensation remains to be further investigated. Nonetheless, increasing expression of a gene itself can also have fitness cost, particularly when the gene is expressed constitutively at a high level (Lang et al., 2009; Cherry, 2010). As DmAldh is
expressed at a high level ubiquitously (Graveley et al., 2010), such a cost may explain why 4809T is rare in ancestral African populations or other ethanol sensitive populations.
Materials and Methods

Gene expression analysis

Real time PCR

Total RNA of 2-4 days old male and female flies were extracted by RNAeasy mini kit (Qiagen) and cDNA was synthesized by iScript cDNA synthesis kit (Biorad). Abundance of Aldh transcript was measured using the internal standard βTub56D as the reference. Real time PCR reactions were performed in Applied Biosystems 7300 real time PCR systems using TaqMan probes (Dm01809880_g1 for Aldh and Dm02362299_u1 for βTub56D) following manufacturer’s protocol (Life Technologies). Relative expression of Aldh (ΔC_T) obtained from the African flies was compared to that of the European flies.

Microarray

The microarray dataset used is same as described in (Ayroles et al., 2009). I used the R package affy for extracting the expression data for DmAldh from all 40 lines. To control for probe-target mismatches caused by sequence polymorphisms, I removed probes overlapping known polymorphic sites before calculating expression values from probe intensities. The background correction was performed by gcma method to control for sequence specific background noise in probe intensities.

Sequence data

The coordinates for DmAldh genomic region (9385389-9393567, release 5.53) was obtained from Flybase (Marygold et al., 2013). The first base in this sequence is denoted as position 1 in Figure 2b and the last base is denoted as position 8139. This numbering
convention is used throughout the text. The DGRP sequences were downloaded from. The sequences of African and one of the French populations are from Drosophila Populations Genomics Project. The short read data for Florida, Pennsylvania, and Maine populations were downloaded from the source as mentioned in (Fabian et al., 2012). The short reads for the other French population were downloaded from http://bergmanlab.smith.man.ac.uk/?p=1685 (Haddrill and Bergman 2012). Reads for Tasmanian and Queensland populations were downloaded from the source as mentioned in (Kolaczkowski et al., 2011). Due to the low read depth covering this SNP position (5 for Tasmania and 7 for Queensland), allele frequencies for SNP4809T obtained from these Australian populations are not presented in the Results section. They are briefly mentioned in the Discussion section for comparison with other populations. The short reads were mapped to DmAldh genomic region using Bowtie2 (Langmead and Salzberg, 2012). The alignments were sorted and viewed using SAMtools (Li et al., 2009).

Sanger sequencing of DmAldh 3’ UTR from 40 DGRP lines was performed using the forward primer 5’ CCGGCGATAATCTACATCTC 3’ and reverse primer 5’ GCACACTCTGAACTGAATGT C 3’ designed by Julie Stahlhut.

**Association mapping**

For finding SNPs associated with high DmAldh expression, the association mapping tool available on DGRP server was used. The microarray expression data from males of 40 inbred lines [as mentioned in (Ayroles et al., 2009)] were used to detect the SNPs showing significant association. To verify that top DmAldh SNPs found by male expression data do not have male-specific effects, association mapping was also performed
using both male and female expression data. The top SNPs returned by the analysis were same as found from analysis based on the male expression data alone, with no significant interaction between sex and genotype. Association between deletion and high expression tested using a simple linear model implemented in Mathematica 9.0 (Wolfram Research, Inc.). To test for interaction between 4809T and $\Delta^{UTR}$, two way factorial ANOVA was performed.

**Test for selective sweep**

Test for selective sweep was performed using the Sweepfinder program (Nielsen et al., 2005). Significance of the maximum composite likelihood ratio was computed from the distribution of CLR statistic obtained from neutral samples simulated by $ms$ program (Hudson, 2002). For generating neutral samples, population mutation and recombination parameters estimated from the French sequences by Dnasp5.0 (Librado and Rozas, 2009) were used.

**Statistical analysis**

All statistical tests were performed in Mathematica 9.0 (Wolfram Research, Inc., Champaign, IL). For t-tests, the method (TTest) was used. Regression of acetaldehyde turnover rate on $DmAldh$ transcript level was performed using LinearModelFit. All ANOVA were performed using the method ANOVA.
microRNA binding site prediction

The list of *D. melanogaster* miRNA were downloaded from Flybase (release 5.53). To predict miRNA binding site using Targetscan (Ruby et al., 2007), sequence alignment of 3’ UTR of *Aldh* from *D. melanogaster, D. simulans,* and *D. sechelia* were used. For miRanda (Enright et al., 2003), a score cutoff of 120 and energy cutoff of -12 Kcal/M were used.

Overexpression of microRNA

Fly stocks carrying Gal4 inducible expression constructs of miR980 (stock no. 41191) and miR992 (stock no. 41130) were obtained from Bloomington Drosophila Stock Center (Bloomington, IN). The fly stock carrying heat shock inducible GAL4 expression construct (stock no. 2077) was also obtained from Bloomington stock center. Flies carrying GAL4 construct and microRNA construct were crossed and 0-2 day old F1 males from each miRNA cross were collected and placed in two separate vials. One of the vials was placed at 37° C for heat shock. The other vial was kept at room temperature. 4 hours after the heat shock flies from both vials were frozen and used for enzyme activity assays.

Enzyme activity

Flies killed by freezing were macerated in grinding buffer (0.25M sucrose, 5mM EDTA, 15 mM Triton X-100, 5 mM DTT). Protease inhibitor (Roche Applied Science) was added to the grinding buffer to prevent proteolytic degradation of the enzymes. The unbroken body parts and other insoluble debris were removed by centrifuging the extracts
at 16,000xg for 20 minutes. The supernatant obtained following centrifugation was used in enzyme activity assays. Protein concentration of the supernatant was measured by Lowry method following manufacturer’s (Sigma-Aldrich, St. Louis, MO) protocol. Enzyme assays were carried out at room temperature in cuvettes of 1 cm path length in 1 ml volume of Na₂HPO₄/NaH₂PO₄ buffer containing 1mM DTT (pH = 8.5). Pyrazole (0.02M) was added to the assay buffer to inhibit aldehyde dehydrogenase activity of the enzyme alcohol dehydrogenase (ADH). Acetaldehyde was purchased from Sigma-Aldrich (St. Louis, MO) and added to the assay mixture at a final concentration of 1mM ([s]>>K_M) (Rothacker and Ilg, 2008). Turnover rates of ALDH were estimated by change in NAD+ concentration as measured by change in absorbance at 340 nm (ε₃₄₀ = 6220 M⁻¹cm⁻¹) using Ultrospec spectrophotometer (GE Healthcare Bio-Sciences, Pittsburgh, PA).
References


Chapter 4

Parallel functional changes in independent testes-specific
duplicates of Aldehyde Dehydrogenase in *Drosophila*
Abstract

A large proportion of duplicates in animal genomes, originating from ubiquitously expressed genes, acquire testis biased expression. Identifying the underlying cause of this observation requires determining whether the duplicates have altered molecular functions relative to the parental genes. Typically, statistical methods are used to test for positive selection, signature of which in protein sequence of duplicates implies functional divergence. When assumptions are violated, however, such tests can give rise to false inference of positive selection. More convincing evidence for naturally selected functional changes would be the occurrence of parallel, functionally relevant amino acid substitutions in independent duplicates of the same gene. We investigated two testis-specific duplicates of the broadly expressed enzyme gene *Aldehyde dehydrogenase (Aldh)* that arose in different *Drosophila* lineages. The duplicates show a typical pattern of accelerated amino-acid substitutions relative to their broadly expressed paralogs, with statistical evidence for positive selection in one case. Importantly, in both duplicates, a site known *a priori* to influence substrate specificity, and otherwise conserved throughout the genus *Drosophila*, underwent a replacement from the ancestral threonine to a bulkier, non-polar amino acid. The adjacent site also went parallel changes, with alanine being replaced in both cases by larger, polar amino acids. Protein structure modeling indicates that both pairs of substitutions are likely to reduce the size of the enzyme’s substrate entry channel, shifting substrate specificity toward smaller aldehydes. Expression data suggest that the evolution of the testis-specific duplicates was accompanied by reduction in expression of the
ancestral Aldh in males, supporting the hypothesis that the duplicates may have helped resolve intralocus sexual conflict over Aldh function.
Introduction

A large number of duplicates in animal genomes originating from ubiquitously expressed parental genes have been observed to acquire testis specific expression pattern (Vinckenbosch et al., 2006; Bai et al., 2007; Gallach et al., 2010; Baker et al., 2012). Identification of the underlying cause of such a pattern depends on whether and how testis-limited duplicates have diverged functionally from their parental counterparts (Gallach and Betran, 2011). Although statistical methods are typically employed to detect amino acid sites under positive selection—a proxy for functional divergence— their inferences often lead to misidentification of functional changes in the protein (Wong et al., 2004; Nozawa et al., 2009; Gharib and Robinson-Rechavi, 2013). More convincing evidence in this context can come from parallel amino acid substitutions in independent duplicates of the same gene, because parallel amino acid changes in proteins often represent parallel occurrence of adaptive functional changes (Christin et al., 2010; Zhen et al., 2012). Although parallel amino acid substitutions in independent testis-specific duplicates of the same gene have been noted previously, their functional significance has not been investigated (Vemuganti et al., 2007). Therefore, to gain insight into the process that underlies evolution of testis-specialized duplicates, we investigated effect of parallel amino acid changes on protein structure and function in independent duplicates of Aldehyde dehydrogenase (Aldh) enzyme gene in Drosophila.

The Drosophila aldehyde dehydrogenase encoded by DmAldh is a conserved homolog of the vertebrate mitochondrial ALDH (ALDH2). In D. melanogaster, it is capable of catalyzing oxidation of a wide range of aldehydes (Rothacker and Ilg, 2008).
Broad substrate specificity of DmALDH is probably necessary to protect from various toxic aldehydes generated by lipid peroxidation within mitochondria (Esterbauer et al., 1991; Chakraborty and Fry, 2011), as well as acetaldehyde derived from dietary ethanol (Fry and Saweikis, 2006). All Drosophila homologs of DmALDH possess mitochondrial leader peptides and share highly conserved sequence which suggest that the enzyme performs a conserved function within mitochondria. Because most Drosophila species are not resistant to ethanol, protection from aldehydes produced by lipid peroxidation is perhaps the conserved function of Drosophila ALDH (Chapter 2). These toxic aldehydes are typically large, so any new adaptation that requires ALDH enzyme with improved detoxification rate for small aldehydes requires alteration of the conserved specificity of the enzyme for large aldehydes. Such alteration of substrate specificity in ALDH, however, is likely to compromise ancestral substrate specificity of the enzyme (Vasiliou et al., 2000).

Substrate specificity of Drosophila ALDH and its conserved homologs in other eukaryotes is controlled by three main residues: Thr144, Ala323, and Leu479 (numbered with reference to DmALDH prior to cleavage of the mitochondrial leader) (Sobreira et al., 2011). These residues are located in the substrate entry channel (SEC), an intramolecular tunnel within the enzyme that guides the substrate to the active site. Altering side chains of these residues alters substrate specificity of ALDH predictably: a side chain larger than threonine at position 144 or leucine at position 479 shifts the substrate specificity of the enzyme towards smaller aldehydes (Moore et al., 1998; Sobreira et al., 2011; Chapter 2). All of these three residues are conserved in Drosophila, suggesting that the residues are preserved by strong purifying selection. Thus, evolution of any adaptive trait that requires
an ALDH with different substrate specificity will be opposed by the negative fitness effect of the substitution. This prediction is supported by a naturally occurring Leu479-Phe479 polymorphism in *D. melanogaster*. The derived form of the enzyme, DmALDH^{Phe479}, has increased turnover rate of for acetaldehyde relative to the ancestral form DmALDH^{Leu479}, but has reduced turnover rates for larger aldehydes compared to the ancestral counterpart.

However, fixation of an Aldh allele containing a substitution at one of the three residues and resulting altered substrate specificity, can be achieved if the allele resides in a duplicated copy of the Aldh gene. The allele can then evolve independently of the ancestral, conserved copy of the enzyme and thus bypass the deleterious fitness consequences. In two subgenera of *Drosophila*, independent duplications of Aldh have created new gene copies which have diverged from the parental genes in terms of both amino acid sequence and expression pattern. We investigated if the altered expression pattern and amino acid sequence of the duplicates is due to optimization of Aldh by natural selection for a testis specific function in both lineages. In particular, we looked for parallel changes affecting the substrate specificity controlling residues of the two duplicates which, if present, might have been responsible for parallel functional divergence in the Aldh duplicates.

**Results**

Throughout the rest of the text, we refer to the diverged duplicates of Aldh as Aldh-*dup* and conserved duplicates as Aldh. Aldh-*dup* in the Obscura group was created by tandem duplication which preserved the exon-intron structure of the autosomal ancestral gene in the duplicate (Fig 4-1C). The duplicate is absent in *D. melanogaster*, so we infer
Figure 4-1. A) Phylogenetic relationships between Aldh and Aldh-dup genes. Branch lengths represent maximum likelihood estimate of nonsynonymous substitution rate per codon. The position of duplication events (solid squares) in the Obscura and Drosophila subgenus are identified by reconciling known Drosophila phylogenetic tree with Aldh gene tree. Following duplication, rate of amino acid substitution in one copy of the duplicates had increased in both subgenera, as indicated by longer branch lengths for the Aldh-dup branches relative to the corresponding Aldh branches. B) Residues present at one of the locations controlling substrate entry (144), and 145th position in the ALDH enzymes encoded by the corresponding Aldh genes. C) Gene structure of Aldh and Aldh-dups. Solid rectangles and the lines represent exons and introns, respectively. Due to their origin by retroduplication, duplicates within Drosophila subgenus lack introns.
that the duplication event occurred at least 4-5 million years ago – before D. lowei had diverged from D. miranda (Fig 4-1A) (Gao et al., 2007). We found the occurrence of the second duplication event in the subgenus Drosophila, which consists of Hawaiian Drosophila and species belonging to the Virilis-Repleta group (Fig 4-1A). This duplicate, unlike its counterpart in Obscura group, originated by retroduplication and thus lost the introns and genomic location of the parental Aldh (Fig 4-1C). Presence of the duplicate in D. virilis, D. mojavensis, and D. grimshawi suggests that the duplication occurred approximately 30 million years ago, before the split between the ancestors of Hawaiian and Virilis-Repleta groups (Morales-Hojas and Vieira, 2012).

**Expression of Duplicate and Ancestral Genes**

*Obscura group*-- To find out if the duplicates in this group were associated with a tissue specific role of the enzyme, we analyzed the expression level of Aldh and Aldh-dup in head, testis, and ovary. Comparison of the expression patterns of Aldh and Aldh-dup in D. miranda and D. pseudoobscura ovary and testis from publicly available RNAseq data indicates that the Aldh-dups have a highly testis-biased expression pattern, whereas expression of Aldhs are female-biased (Table 4-1). Sex-biased expression of Aldh-dup appears to be present in head and thorax of D. pseudoobscura as well, although at a much lesser extent (Table 4-1). In contrast, female biased expression of DpseAldh appears to be solely due to the higher expression of the gene in ovaries compared to testes (Table 4-1).

*Drosophila subgenus* -- We tested if expression of Aldh-dups in Drosophila
Table 4-1. Relative expression level of Aldh and Aldh-dup in different species (* = p< 0.01).

<table>
<thead>
<tr>
<th>Group</th>
<th>Species id</th>
<th>Gene</th>
<th>Tissue</th>
<th>Fold difference (log(_2)(♂♂/♀♀))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obscura</td>
<td>Dmir</td>
<td>Aldh</td>
<td>Testis/Ovary</td>
<td>-3.1667*</td>
</tr>
<tr>
<td></td>
<td>Dmir</td>
<td>Aldh-dup</td>
<td>Testis/Ovary</td>
<td>11.5447*</td>
</tr>
<tr>
<td></td>
<td>Dpse</td>
<td>Aldh</td>
<td>Testis/Ovary</td>
<td>-2.0492*</td>
</tr>
<tr>
<td></td>
<td>Dpse</td>
<td>Aldh</td>
<td>Carcass without gonad</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Dpse</td>
<td>Aldh-dup</td>
<td>Testis/Ovary</td>
<td>8.4064*</td>
</tr>
<tr>
<td></td>
<td>Dpse</td>
<td>Aldh-dup</td>
<td>Carcass without gonad</td>
<td>1.78*</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Dvir</td>
<td>Aldh</td>
<td>Whole body</td>
<td>-0.5992*</td>
</tr>
<tr>
<td></td>
<td>Dvir</td>
<td>Aldh-dup</td>
<td>Whole body</td>
<td>6.1495*</td>
</tr>
<tr>
<td></td>
<td>Dmoj</td>
<td>Aldh</td>
<td>Whole body</td>
<td>-2.3468*</td>
</tr>
<tr>
<td></td>
<td>Dmoj</td>
<td>Aldh-dup</td>
<td>Whole body</td>
<td>4.5264*</td>
</tr>
<tr>
<td></td>
<td>Dmel</td>
<td>Aldh</td>
<td>Testis/Ovary</td>
<td>2.1800*</td>
</tr>
<tr>
<td></td>
<td>Dmel</td>
<td>Aldh</td>
<td>Head and thorax</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Dmel</td>
<td>Aldh</td>
<td>Whole body</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td>Dwil</td>
<td>Aldh</td>
<td>Abdomen</td>
<td>0.2</td>
</tr>
</tbody>
</table>
subgenus showed tissue specificity similar to Aldh-dups of Obscura. We compared the
eexpression pattern of Aldh and Aldh-dup between males and females of D. mojavensis and
D. virilis. Even though gene structure and genomic location of the Aldh-dups of Drosophila
subgenus are different from those of Aldh-dups of Obscura group, we found that similar to
Obscura Aldh-dup and Aldh, expression of Aldh-dup and Aldh in Drosophila subgenus are
male- and female- biased, respectively (Table 4-1). Analysis of an independent RNAseq
dataset suggested that male biased expression of DvirAldh-dup was due to higher
expression of the gene in testis (Ahmed et al., unpublished data).

The expression pattern of Aldhs in Obscura and Drosophila subgenus differed from
expression patterns of their homologs in D. willistoni and D. melanogaster, which lacked
Aldh-dup. In these two species, Aldh was expressed in nearly unbiased or testis-biased
patterns (Table 4-1). The male-specific expression pattern of Aldh in D. melanogaster
appeared to be due solely to the expression differences of Aldh between testes and ovaries,
because expression levels of the gene in head and thorax were indifferent between the
sexes.

**Test for positive selection**

*Obscura group* – Comparison of the amino acid sequences of the duplicates to that
of the ALDH of D. melanogaster, and to their broadly-expressed paralogs, suggested that
the Obscura group ALDH-dup had diverged considerably from the ancestral sequence of
ALDH (Fig. 4-1). To test if this divergence was driven by positive or relaxed selection, we
analyzed ratio of non-synonymous and synonymous substitutions (ω = dn/ds) in the
Table 4-2. \( dN/dS \) for different site classes (Zhang, et al. 2005) of ALDH and ALDH-dup in *Obscura* group.

<table>
<thead>
<tr>
<th>Site class</th>
<th>0</th>
<th>1</th>
<th>2a</th>
<th>2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>.82</td>
<td>.06</td>
<td>.11</td>
<td>.01</td>
</tr>
<tr>
<td>Background ( \omega )</td>
<td>.04</td>
<td>1.00</td>
<td>.04</td>
<td>1.00</td>
</tr>
<tr>
<td>Foreground ( \omega )</td>
<td>0.04</td>
<td>1.00</td>
<td>6.36</td>
<td>6.36</td>
</tr>
</tbody>
</table>
duplicates. The $dn/ds$ ratio between for the duplicate and ancestral $Aldh$ is 0.9, a value typically considered to be consistent with neutral evolution.

However, most residues of $Drosophila$ ALDH are conserved, so a more realistic test for selection ($\omega>1$) should be adjusted for the ALDH specific low background substitution rates. To accomplish that, we tested for positive selection in ALDH-dup following the branch-site model: a likelihood ratio based test for site-specific increase in $\omega$ under positive selection, along a specified phylogenetic branch, termed the “foreground” branch (Zhang et al., 2005). The model estimates $\omega$ for four classes of sites (Table 4-2). Site class 0 denotes codons that are conserved throughout the tree. Site class 1 includes codons that are evolving neutrally ($\omega=1$) throughout the tree. Classes 2a and 2b comprise codons that are conserved or neutral on the background branches, but are under positive selection on the foreground branches. We specified the duplicate lineage as the foreground branch and rest of the lineages as background. The test showed that 12% of the neutral ($\omega=1$) or conserved ($0<\omega<1$) sites in ALDH accumulated amino acid substitutions in ALDH-dup under historical positive selection ($\omega>1$, $df=1$, $p=10^{-5}$, log likelihood ratio test) (Table 4-2). The observation that ALDH-dup of $D. lowei$, $D. miranda$, $D. pseudoobscura$, and $D. persimilis$ are closely related to each other (96-99% amino acids sequence identity) suggested that evolution of amino acid sequence may have slowed down following the initial burst of adaptive substitutions (Fig 4-1A).

$Drosophila$ subgenus – In the $Drosophila$ subgenus, the $Aldh$-dup accumulated amino acid substitutions much faster compared to its conserved counterpart after the duplication event (Fig 4-1A). Although the hypothesis that such an accelerated rate of
protein evolution occurred under positive selection could potentially be tested by the “branch site model” mentioned in the previous section, we refrained from using the method due to the age of ALDH-dups in this subgenus (Fig 4-1A).

**Functional analysis of amino acid substitutions**

*Obscura*—To find out if any member of the substrate specificity controlling triad (Thr144, Ala323, and Leu479) has been altered in ALDH-dup, we aligned the sequences and structures of the two enzymes. Both duplicates retained the ancestral Ala323 and Leu479. In contrast, both duplicates experienced substitutions of Thr144 with non-polar residues with larger R groups, Methionine in the Obscura group, and Leucine in the *Drosophila* subgenus (Fig. 4-1C). In the model structure of DpseALDH-dup, the bulkier sidechain of Met144 protruded into the channel mouth (Fig. 4-2C), which seem to have constricted the channel mouth of the enzyme relative to the DpseALDH (Fig 4-2B). Smaller channel mouth would sterically hinder entry of large aldehydes into the active site of the enzyme. To find out if Thr144Met substitution was responsible for this constriction, we mutated Thr144 by Met144 in the ancestral enzyme *in silico*. The substitution, however, did not reduce the diameter of the channel mouth sufficiently in the model structures (Table 4-3). Upon closer inspection, the side chain of the Met144 seemed to face away from the channel mouth (Fig 4-2F). We predicted that other residues in DpseALDH-dup might contribute to the conformation of Met144 side chain observed in the duplicate (Fig 4-2C). We investigated the role of three other residues – Ala145, Ala478, and Ala480 which are present near the SEC in ALDH (Fig 4-2A) and have been substituted by Thr145, Ile478,
Figure 4-2. A) Front view of the entrance/mouth of the substrate entry tunnel of DpseALDH. The residues Thr144 (blue), Ala145 (red), Ala478 (purple), and Ala480 (turquoise) are highlighted to show their positions relative to the channel mouth. B) Side view of substrate entry channel (SEC) of the DpseALDH. C) Side view of SEC of DpseALDH-dup showing the constricted mouth; the Met144 can be seen hanging from the top. D) Side view of SEC of DmojALDH. Note the wider mouth compared to E) SEC of DmojALDH-dup (side view). The Leu144 can be seen hanging from the top. F) Front view of SEC of mutant DpseALDH showing the conformation of Met144 in presence (green) and absence (turquoise) of Thr145 (FV). In presence of Thr145, Met144 assumes the “stretched” conformation required to constrict the channel mouth. The arrows designate width/diameter of the SEC mouth (See Table 4-3 for lengths of the arrows). The solvent
side is on right for the side view and on the viewer’s side for the front view. The black surfaces represent the walls of the SEC.
and Pro480 in the DpseALDH-dup, respectively. We mutated the residues in DpseALDH individually first, and then in combination with the Thr144Met substitution, to find out if any of them affected the channel mouth diameter. None of the substitutions had any effect on the channel mouth in absence of Met144. In contrast, the difference in channel mouth diameter between DpseALDH and DpseALDH-dup was recovered when Ala145 in the ancestral enzyme was replaced by Ser145 in combination with the Thr144Met substitution (Table 4-3). The Ala480Pro mutation in presence of Met144 similarly constricted the diameter of the channel mouth in DpseALDH. On the other hand, Ala476Ile substitution in presence of Met144 completely closed the entrance to SEC. Nonetheless, combining all these four substitutions in DpseALDH constricted the SEC mouth to the level of the DpseALDH-dup. It seemed that Thr144Met was the key substitution needed to constrict the SEC mouth, while additional residues, especially Thr145 and probably Pro480, helped stabilize the “protruding” conformation of the Met144 sidechain (Fig 4-2F). Consistent with the predicted key role of Thr144Met substitution in transformation of the ancestral channel mouth into that of the duplicate, we found that substitution of Met144 in DpseALDH-dup by Thr144 widened the SEC mouth of the enzyme to the level of its ancestral counterpart.

*Drosophila* – In contrast with the duplicates in the *Obscura* group, ALDH-dups in *Drosophila* subgenus shared very few residues among themselves which differed from residues at the corresponding positions in the ALDHs. However, in addition to the ancestral Thr144 (or possibly Val144, the residue in D. grimshawi ALDH; Fig. 4-1B) being replaced by Leucine in the testis-specific duplicates, Ala145, substitution of which contributed to
Table 4-3. Diameter of the SEC mouth in DpseALDH, DpseALDH-dup, DmojALDH, and their mutants created in silico.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Mutation</th>
<th>Diameter (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DpseALDH</td>
<td>wild type</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>M145</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>T146</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>M145+T146</td>
<td>5.8</td>
</tr>
<tr>
<td>DpseALDH-dup</td>
<td>wild type</td>
<td>5.9</td>
</tr>
<tr>
<td>DmojALDH</td>
<td>wild type</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>L145</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td>S146</td>
<td>8.05</td>
</tr>
<tr>
<td></td>
<td>L145+S146</td>
<td>5.8</td>
</tr>
<tr>
<td>DmojALDH-dup</td>
<td>wild type</td>
<td>6.5</td>
</tr>
</tbody>
</table>
wide to narrow SEC mouth transition in *Obscura* duplicates, had also been altered in duplicates of *Drosophila* sub-genus, all of which carry Ser145.

The similarity between the substituent residues in the duplicates of the two subgenera was remarkable: methionine and leucine both have sidechains larger than threonine, while serine and threonine both possess hydroxyl group in their sidechains. Accordingly, we predicted that SEC mouth of ALDH-dups would be constricted compared to the SEC mouth of ALDH. We found that indeed is the case (Fig 4-2D & 4-2E). We further predicted that a combination of Thr144Leu and Ala145Ser substitutions would constrict the mouth of the SEC in ancestral ALDH. Consistent with this prediction, when Thr144 of DmojALDH was replaced by Leu144, SEC mouth of the enzyme was transformed into the SEC mouth of DmojALDH-dup (Table 4-3). This, however, seemingly contrasted with the second part of our hypothesis because presence of Ala145Ser substitution was not necessary in DmojALDH to attain the SEC mouth of DmojALDH-dup. To further examine if Ala145Ser substitution had any effect on Thr144Leu substitution, we compared the stability of DmojALDH\textsuperscript{Leu144} and DmojALDH\textsuperscript{Leu144Ser145} enzymes. The structure of the later was found to be more stable ($\Delta G = -678.28$ Kcal) compared to the former ($\Delta G = -677.25$ Kcal). This suggests that while the Thr144Leu substitution introduced the physical change in the SEC mouth, the Ala145Ser substitution energetically reinforced the constricted conformation of the SEC mouth.

We next examined if larger diameter of SEC mouth is an ancestral and conserved feature of drosophila ALDH. We reconstructed the ancestral channel mouth diameter using the maximum likelihood method based on channel mouth diameter of ALDH across
Figure 4-3. Normalized diameter of the channel mouth in ALDH and duplicated ALDH in different Drosophila species and the maximum likelihood estimates of their ancestral states predicted by the program ape (log likelihood =19.52). Diameter of the solid circles are proportional to the diameters of the channel mouth. The narrow channel mouth of the duplicates evolved twice independently from the ancestral and conserved wider channel mouth of Drosophila ALDH.
species which possessed the duplicates as well as species which did not have the duplicates. The maximum likelihood estimate of the ancestral state suggests that large SEC mouth is the ancestral and conserved feature of ALDH, from which the narrow SEC mouth has evolved independently in duplicates of *Obscura* and *Drosophila* subgenus (Fig 4-3).

Inferences from Enzyme Activity

If substrate specificity of DpseALDH-dup for small aldehydes was improved relative to DpseALDH, then ratio of turnover rates for a small aldehyde and a large aldehyde would be higher in testis than in ovary of *D. pseudoobscura*. To test this, we compared ratio of turnover rates of abdominal enzyme extracts from male and female *D. pseudoobscura* for acetaldehyde – a small aldehyde, and hexanal – a larger aldehyde which is also a known by-product of oxidative damage within mitochondria (Frankel et al., 1989) (Fig 4-4). Higher ratio of turnover rates for acetaldehyde to hexanal in males compared to females suggested that male abdomen had an extra aldehyde dehydrogenase – possibly DpseALDH-dup – specialized for faster oxidation of acetaldehyde. Such difference in acetaldehyde turnover rates between two sexes of *D. pseudoobscura*, however, was not observed in *D. melanogaster* (Fig 4-4), which lacks ALDH-dup.

**Discussion**

A large number of duplicates, originating from broadly expressed genes, acquire testis biased expression pattern and often undergo accelerated protein evolution. Although a general explanation of such a pattern is lacking, it has been hypothesized that duplicates acquire testis biased expression pattern to mitigate intralocus sexual conflict (IASC)
Figure 4-4. Acetaldehyde dehydrogenase activity of extracts from abdomens of *D. pseudoobscura* and *D. melanogaster* normalized by turnover rate of the extracts for hexanal, a large aldehyde.
(Gallach and Betran, 2011). IASC results from the conflict between the selection pressure that attempts to preserve the ubiquitous function of the ancestral gene and the selection pressure that tries to optimize the gene for a testis specific function. The duplicate mitigates the conflict because it becomes optimized for testis specific function, relaxing the selection pressure that attempted to optimize the ancestral gene for testis-specific function. The alternative hypotheses argue that testis-specific duplicates evolve to escape from meiotic sex chromosome inactivation (MSCI) or simply to compensate for dosage effects (Meisel et al., 2009; Tracy et al., 2010; Parsch and Ellegren, 2013).

The fact that both Aldh and Aldh-dup are autosomal rules out the possibility that duplication of Aldh facilitated MSCI or dosage compensation. To examine whether ALDH-dups resolved IASC, we investigated three main parts of the IASC hypothesis: 1) ALDH-dups underwent functional changes relevant to optimization for testis specific function; 2) ALDH cannot be optimized for the testis-specific function ALDH-dups are selected for without incurring any fitness consequences; 3) following origin of testis specific Aldh-dups, Aldh was freed from the selection pressure to evolve testis specific function.

First, we investigated if ALDH-dups have undergone functional changes relative to ALDH and if their function is testis specific. Occurrence of parallel substitutions in Aldh-dups of the two subgenera and their overlap with at least one residue of known functional importance indicate that Aldh-dups may have diverged with respect to the function of their parental genes. In particular, one of the parallel substitutions replaced Thr144, the homolog of which is known to control substrate specificity in vertebrate as well as invertebrate ALDH (Sobreira et al., 2011). The residue is replaced by Met144 and Leu144 in the
*Obscura* and *Drosophila* duplicates of ALDH, respectively. The prediction that substitutions of Thr144 reduced the width of the channel mouth, was confirmed by mutating Thr144 by Met144 and Leu144 in ancestral ALDH *in silico*. Because expression of both *Aldh-dups* is highly testis-biased, the change in substrate specificity is presumably relevant to a testis specific function of the enzyme.

Second, we asked if substitutions necessary for the testis specific functional change in ALDH-dups perturb the conserved function of ALDH. If it did then we can infer that ALDH could not have been optimized for the new function without compromising its existing function. The reduction in the diameter of ALDH channel mouth is known to shift substrate specificity of the enzyme from large to small aldehydes (Moore et al., 1998; Sobreira et al., 2011). Therefore, occurrence of Met144 and Ser145 or Leu144 and Thr145 in the ancestral enzyme would have compromised the conserved substrate specificity of ALDH. Altering ancestral substrate specificity of DmALDH, a close homolog of *Obscura* and *Drosophila* ALDHs, is known to impair protection by the enzyme from respiratory oxidative stress, and reduce organismal fitness (Chapter 2). Therefore, testis specific substrate specificity of ALDH-dups could not have evolved in ALDH without incurring fitness defects.

Third, we investigated if evolution of *Aldh-dups* relaxed the selection that acted on the ancestral copy to optimize it for testis specific function. In *D. melanogaster*, which lacks *Aldh-dup*, expression of *Aldh* is testis biased, suggesting that selection favors high expression in the testis of *Aldh* in *D. melanogaster* (cf. Connallon and Clark, 2011). It is therefore possible that before the origin of *Aldh-dup*, *Aldh* in *Obscura* and *Drosophila* was
expressed in testis biased pattern due to the selection favoring testis specific function of the gene. Following the origin of testis specific Aldh-dup, expression of Aldh was no longer under selection for the testis specific function. Females biased expression of Aldh-dups is consistent with this prediction.

Presently, we neither know what specific biological function within testis the duplicates were selected for, nor we can rule out the possibility that additional residues might have contributed to this function. Comparison of ALDH enzyme activity between testes and ovaries of D. pseudoobscura with ALDH activity between males and females of D. melanogaster suggest that the duplicate could help to increase the rate of acetaldehyde detoxification in testes. In mammals acetaldehyde is detoxified primarily by ALDH2 – the homolog of Drosophila ALDH – and is known to compromise male fertility via mechanisms that remain unclear (Muthusami and Chinnaswamy, 2005). Whether acetaldehyde, and ALDH, affects male fitness of drosophila in nature depends on the exposure of these species to acetaldehyde, which can either be derived from exogenous sources such as dietary ethanol (Fry et al., 2004) or produced in vivo as in mammals (Langevin et al., 2011). Owing to the limited knowledge of the natural ecology of most the Drosophila species mentioned here, it is unclear how much dietary ethanol and acetaldehyde they encounter in nature.

In summary, we have shown that amino acid residues with conflicting structural and functional effects may underlie evolution of testis specific duplicates. By providing an explanation that is independent of lineages and genomic location of the ancestral and duplicated gene copies, our results complements previous studies which predicted,
primarily based on gene expression pattern (Gallach et al., 2010; Baker et al., 2012), that intralocus sexual conflict could be responsible for evolution of testis-biased expression in many gene duplicates. It is known that testis is a home of the fastest evolving proteins in *Drosophila* and other organisms (Swanson and Vacquier, 2002; Nielsen et al., 2005; Haerty et al., 2007). It is therefore plausible that structural constraints in conserved proteins of testis, like ALDH (Dorus et al., 2006), and the demand to increase the rate of adaptation for testis specific functions frequently creates adaptive conflicts between testis specific and conserved functions of the proteins. Such conflicts are likely to predispose the conserved genes to generate duplicates that are maintained to mitigate IASC. Consequently, duplicates originating from the conserved genes in such cases should be less likely to get lost than those which are originating from less conserved genes. Previous observations seem to support this hypothesis (Davis and Petrov, 2004; Gallach et al., 2010).
Materials and methods

Protein and DNA sequences

DNA and protein sequences of ancestral ALDH and its duplicates in D. pseudoobscura, D. persimilis, D. virilis, D. mojavensis, and D. grimshawi were obtained from Flybase (Marygold et al., 2013) (Table 4-4). DNA sequences of the ancestral and duplicated Aldh of D. lowei were obtained from Pseudobase (McGaugh et al., 2012). Sequences of D. miranda Aldh and Aldh-dup were obtained by aligning the sequences of the two genes available from Pseudobase. Protein sequence of ALDH of D. melanogaster and D. willistoni was obtained from Flybase. Orthologs of Aldh and Aldh-dup were identified by reciprocal BLAST. Homologous amino acid residues were identified by aligning amino acid sequences of all fly ALDH mentioned here with amino acid sequence of human ALDH2 in CLUSTALW.

Test for selection

Nucleotide sequences of the coding region of Aldh and Aldh-dup were aligned guided by amino acid sequence alignment. The gene tree was created based on this alignment by phylogeny.fr (Dereeper et al., 2008) and reconciled with the known species phylogeny (Powell, 1997) by the program Notung (Durand et al., 2005). Lineage specific dn/ds was calculated using the “free ratios model” implemented in CODEML program of PAML 4.7 package. Positive selection in Obscura duplicates was tested using the “branch sites model” included in the same package (Zhang et al., 2005). In the gene tree provided to CODEML,
Table 4-4. Aldh genes and the source of their corresponding RNAseq data.

<table>
<thead>
<tr>
<th>Flybase ID</th>
<th>Gene name</th>
<th>RNAseq (NCBI accession id)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA25309</td>
<td>DpseAldh-dup</td>
<td>SRX102928, SRX102924</td>
</tr>
<tr>
<td>GA17661</td>
<td>DpseAldh</td>
<td></td>
</tr>
<tr>
<td>GL26144</td>
<td>DperAldh-dup</td>
<td></td>
</tr>
<tr>
<td>GL26145</td>
<td>DperAldh</td>
<td></td>
</tr>
<tr>
<td>GJ16294</td>
<td>DvirAldh-dup</td>
<td></td>
</tr>
<tr>
<td>GJ17859</td>
<td>DvirAldh</td>
<td>SRX247003, SRX247002, SRX054489, SRX054488</td>
</tr>
<tr>
<td>GI18106</td>
<td>DmojAldh-dup</td>
<td>SRX054487, SRX054486, SRX054485, SRX054484</td>
</tr>
<tr>
<td>GI17663</td>
<td>DmojAldh</td>
<td></td>
</tr>
<tr>
<td>GH13227</td>
<td>DgriAldh-dup</td>
<td></td>
</tr>
<tr>
<td>GH13324</td>
<td>DgriAldh</td>
<td></td>
</tr>
<tr>
<td>GK24201</td>
<td>DwilAldh</td>
<td>SRX095362, SRX095361</td>
</tr>
<tr>
<td>CG3725</td>
<td>DmelAldh</td>
<td></td>
</tr>
<tr>
<td>Pseudobase</td>
<td>DmirAldh-dup</td>
<td>SRX106031, SRX106026</td>
</tr>
<tr>
<td>Pseudobase</td>
<td>DmirAldh</td>
<td></td>
</tr>
<tr>
<td>Pseudobase</td>
<td>DlowAldh-dup</td>
<td></td>
</tr>
<tr>
<td>Pseudobase</td>
<td>DlowAldh</td>
<td></td>
</tr>
</tbody>
</table>
we specified the branch leading to the duplicates as the foreground branch, which facilitates detection of branch specific accelerated rate of protein evolution under positive selection. Then we evaluated the null hypothesis of no selection ($\omega=1$) against the alternative hypothesis of positive selection ($\omega > 1$) by the log likelihood ratio test ($\chi^2$, df=1).

Expression analysis

RNAseq raw reads were downloaded from NCBI short read archive (Table 4-4). The reads were mapped to the genomic sequence of Aldh and Aldh-dup by Tophat (Trapnell et al., 2012). Exon-intron boundaries and UTRs were identified by Cufflinks. Expression difference was calculated using the Cuffdiff program. The expression data for DvirAldh and DvirAldh-dup were confirmed from RNAseq dataset obtained from an independent experiment (Ahmed et al., unpublished data). Expression data for D. melanogaster was obtained from modENCODE (Graveley et al., 2010).

Protein structure modeling and ancestral channel mouth reconstruction

Protein structure models were constructed by Populus server (Offman et al., 2008) based on the structures of human ALDH2 (PDB id 1O04) and sheep liver ALDH1 (PDB id 1BXS). The top structure model for each enzyme returned by the server was used for further analysis. The substrate entry channel was inspected and the surrounding residues were identified using the software Fred Receptor. The key residue positions were identified
by aligning the models with the structure of human ALDH2 in Pymol (Schrödinger). The distance between the residues lining the diametrically opposite sides of the neck of the SEC were measured in Pymol. To control for model-specific random structural variations, the aforementioned measurements were normalized by the diameter (distance between the conserved residues Val141 and Leu479) of the SEC region preceding the mouth. To reconstruct the ancestral channel mouth diameter, we added Aldh of D. willistoni to the gene tree used for PAML analysis and reconstructed the channel mouth diameter of the ALDH enzyme of the ancestor by the maximum likelihood method implemented in the R package ape (Paradis et al., 2004).

**Enzyme Assay**

To compare ALDH activity of *D. pseudoobscura* testis and ovary, testis and ovary of the 2-4 days old flies (genome strain MV2-25) were dissected and macerated in grinding buffer (0.25M sucrose, 5mM EDTA, 15 mM Triton X-100, 5 mM DTT). Protease inhibitor (Roche Applied Science) was added to the grinding buffer to prevent proteolytic degradation of the enzymes. Next, the unbroken body parts were removed by centrifuging the extracts at 16,000xg for 20 minutes. The supernatant obtained following centrifugation was used in enzyme activity assays. For *D. melanogaster* (genome strain), total protein from males and females were extracted following the same procedure as was used for *D. pseudoobscura*. Protein concentration of the supernatant was measured by Qubit fluorometer following manufacturer’s (Invitrogen, Grand Island, NY) protocol. Enzyme
assays were carried out at 25°C in cuvettes of 1 cm path length in 1 ml volume of Na₂HPO₄/NaH₂PO₄ buffer containing 1mM DTT (pH = 8.5). Pyrazole (0.02M) was added to the assay buffer to inhibit aldehyde dehydrogenase activity of the enzyme alcohol dehydrogenase (ADH). Acetaldehyde and hexanal were purchased from Sigma-Aldrich (St. Louis, MO) and added to the assay mixture at a final concentration of 1mM ([s]>>KM) (Rothacker and Ilg, 2008). Turnover rates of ALDH were estimated by change in NAD+ concentration as measured by change in absorbance at 340 nm (ɛ₃₄₀ = 6220 M⁻¹cm⁻¹) using Ultrospec spectrophotometer (GE Healthcare Life Sciences, Pittsburgh, PA).
References


Meisel, R.P., Han, M.V., Hahn, M.W., 2009. A complex suite of forces drives gene traffic from drosophila x chromosomes. Genome Biology and Evolution 1, 176-188.


Schrödinger, L., The pymol molecular graphics system, 1.5.0.4 ed.


Tracy, C., Rio, J., Motiwale, M., Christensen, S.M., Betran, E., 2010. Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in drosophila. Genetics 184, 1067-U1296.


Chapter 5

Conclusion
**Adaptive constraint: Leu-Phe polymorphism**

Although natural selection optimizes individual traits, mutations often affect more than one phenotype. When effects of a mutation on two traits have opposing effects on fitness, the resulting tradeoff can constrain spread of the mutation. While such tradeoffs are predicted to play a major part in maintenance of heritable phenotypic variation, we know little about the molecular basis of such tradeoffs – which are also referred to as adaptive constraints.

Fry et al. (2008) described an amino acid polymorphism in *D. melanogaster* Aldehyde dehydrogenase gene (*DmAldh*) that replaces a conserved leucine (Leu) residue at position 479 by a phenylalanine (Phe) residue. Based on several lines of evidence, they hypothesized that Leu-Phe polymorphism was perhaps maintained because it had opposing effects on fitness: one through its advantage in ethanol metabolism and another through its disadvantage in a second, then unknown, function of *DmAldh* (Fry et al., 2008). The Leu-Phe polymorphism hence provided an excellent opportunity to investigate the molecular and phenotypic basis of fitness tradeoff.

My results based on purified recombinant Leu and Phe enzymes confirm enzyme kinetics results of Fry et al. (2008) based on crude fly extracts: the Phe variant shows higher turnover rate relative to the ancestral Leu variant for acetaldehyde, a harmful intermediate formed during metabolism of dietary ethanol. I also found that, supporting the prediction of Fry et al. (2008), the Phe allele provides better protection from harmful effects of
acetaldehyde and ethanol relative to the Leu allele. Phe is therefore beneficial in *D. melanogaster* temperate populations, where increased resistance to ethanol is adaptive.

However, frequencies of the Phe allele in these temperate region population are low. Based on this observation, it was hypothesized (Fry et al., 2008) that Phe does not spread in these populations because it is disadvantageous for the ancestral, conserved function of DmALDH: detoxification of large, reactive aldehydes generated by lipid peroxidation during aerobic respiration within mitochondria (Chakraborty and Fry, 2011). This accorded with the enzyme kinetics results of Fry et al. (2008) as they had found that crude enzyme extracts of DmALDH\textsuperscript{Phe} oxidized large aldehydes more slowly than DmALDH\textsuperscript{Leu} crude extracts (Fry et al., 2008).

My *in silico* investigation of structures of Leu and Phe variants revealed the likely molecular basis of this tradeoff between substrate specificity: the substitution of leucine by bulkier phenylalanine at 479\textsuperscript{th} position of the enzyme constricts the channel that guides aldehydes to the catalytic site of the enzyme, thus allowing small aldehydes like acetaldehyde to pass easily to the site of catalysis, but hindering access of larger aldehydes to the catalytic site. In accord with this, I found that, recombinant, purified Phe enzyme detoxifies large aldehydes, including those produced by lipid peroxidation, at a much slower rate than the Leu enzyme. I predicted that Leu provided better protection from these aldehydes than Phe, giving Leu allele fitness advantage relative to Phe allele in absence of ethanol. Indeed, I showed that Leu provides better protection from hyperoxic conditions than Phe, likely because it detoxifies reactive aldehydes derived from lipid peroxidation
faster than Phe, and endows flies with higher overall fitness than Phe under normoxic conditions.

Although opposing effects of Phe on fitness through ethanol resistance and oxidative stress resistance is the likely reason of why it does not spread in temperate populations of D. melanogaster, the population genetic mechanism underlying its maintenance as a polymorphism is unknown. It is possible that varying concentration of ethanol across natural habitats of flies leads to spatially varying selection pressure on Phe and Leu alleles, with the result that the polymorphism is maintained by balancing selection resulting from marginal overdominance (see Introduction). Further experiments are required to test this hypothesis.

**Rescuer of adaptive constraint: Cis-regulatory variants of DmAldh**

Gene expression polymorphism is considered as one of the mechanisms by which fitness disadvantages associated with an enzyme variant can be overcome (Olson-Manning et al., 2012). Because the Phe variant increases resistance to ethanol due to its increased turnover rate for acetaldehyde, cis-regulatory variants that increase DmAldh transcript level and hence DmALDH activity should also increase resistance to ethanol. Such cis-regulatory variants should also be adaptive in ethanol resistant temperate populations. Furthermore, if such variants exist they might be able to offset fitness disadvantage of Phe.

A SNP (eSNP) within the 3rd intron of DmAldh is associated with high DmAldh transcript level. Frequency of the derived allele (eSNP-T), which increases DmAldh
expression, has increased under positive selection in an ethanol resistant European population, suggesting that the allele is adaptive likely due to its contribution in ethanol resistance. Furthermore, the variant is rare in ethanol sensitive ancestral African populations, but its frequency is consistently higher in relatively ethanol resistant North American populations. All these observations reinforce the adaptive contribution of the eSNP through increased ethanol resistance. Apart from the eSNP, a set of deletions within 3’ UTR (Fry & Stahlhut, unpublished data) was also found to be associated with high expression. The deletions overlap a predicted miRNA binding site and supporting this prediction, DmALDH activity is reduced when the miRNA (miR980) is overexpressed.

I also found that the Phe allele is in linkage disequilibrium with the derived eSNP allele. This suggests that Phe allele might be more beneficial in high expression background, likely because high expression decreases fitness disadvantage associated with the Phe allele. Whether increased expression actually offsets fitness disadvantage of the Phe allele in absence of ethanol requires further experimental work.

The microRNA binding site deletions are likely to be adaptive in ethanol resistant temperate populations because they also increase DmAldh expression level. However, further population genetic evidence is required to infer the beneficial role of the deletions in ethanol resistant D. melanogaster populations. Furthermore, a more comprehensive analysis of relative phenotypic contributions of amino acid polymorphism, eSNP, and deletion polymorphisms would shed more light on the co-evolutionary history of these variants. This would provide a rare perspective on how origin and spread of typical phenotypically important mutations is influenced by other such mutations.
Rescuer of adaptive constraint: duplicates of Aldh

One situation that would allow the Phe allele to spread in ethanol resistant populations, despite its fitness disadvantage, is one in which Leu and Phe allele both are present as separate genes – each compensating for the fitness disadvantage of the other. Although such duplicates are absent in D. melanogaster, similar duplicates are present in Drosophila subgenus and Obscura species group. The duplicates are similar to Leu, Phe hypothetical duplicates in the sense that each duplicate has a residue at 144th position that is predicted to affect substrate specificity of the enzyme in a similar fashion as Leu/Phe at position 479 does. The conserved residue threonine at 144th residue, similar to Leu479, is needed for ancestral function of the enzyme, whereas the bulkier Leu144 or Met144, similar to Phe479, is predicted to improve kinetic rate of the enzyme for smaller aldehydes. While Thr144 is present in the conserved copy, Leu144 or Met144 are present in the duplicates which appear to have diverged to specialize on a testis specific function. Because Leu144 or Met144 alters substrate specificity of the ancestral enzyme, it was inferred that Leu144 or Met144 could not have evolved in ALDH if Aldh was not duplicated. Thus duplication of Aldh facilitated resolution of the constraint that prevented optimization of ALDH protein for the testis specific function.

Understanding how heritable phenotypic variation is preserved in natural populations and how organisms adapt to new environments using variants from this pool has been a long-sought goal for evolutionary biologists. Although we have sufficient theoretical models to explain these phenomena, experimental data to examine these models in light of phenotypic evolution in natural populations is still modest (Barton and
Keightley, 2002). The new age of large scale sequencing endeavors is revealing unexpected patterns about the contribution of mutations to phenotypic evolution (Clark et al., 2007; Coop et al., 2009; Burke et al., 2010). Nonetheless, to gain a complete understanding of the process of phenotypic evolution, discovery of such patterns and predictions of theoretical models need to be complemented by functional data on genetic variants. The Leu-Phe amino acid polymorphism, *DmAldh* cis-regulatory polymorphisms, and *Aldh* duplicates, by demonstrating how natural selection exploits the same genetic resource (or gene) in different ways to maintain and erode variation, takes us one step closer to that goal.
References


