Empirical approaches to detecting the action of natural selection in *Drosophila*

John Baines

München 2004

Empirical approaches to detecting the action of natural selection in *Drosophila*

John Baines

Dissertation
der Fakultät für Biologie
der Ludwig-Maximilians-Universität
München

vorgelegt von
John Baines
aus Maryland, USA

München, den 23. Dezember 2003

Erstgutachter: Prof. Dr. Wolfgang Stephan
Zweitgutachter: Prof. Dr. John Parsch
Tag der mündlichen Prüfung: 23.04.2004

CURRICULUM VITAE

John Baines was born on February 26, 1976 in Silver Spring, Maryland, USA. He attended the University of Maryland from 1994 to 1998 and received a Bachelor of Science degree in Biochemistry with honors. He began graduate study in the Department of Biology at the University of Rochester from 1998 to 2000 and received a Master of Science degree under the direction of Professor Wolfgang Stephan. He began his PhD research at the University of Munich in 2000 under the direction of Professor Wolfgang Stephan.

LIST OF PUBLICATIONS

BAINES, J.F., A. DAS, S. MOUSSET and W. STEPHAN, (in press) The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. Genetics.

BAINES, J.F., J. PARSCH and W. STEPHAN, 2004 Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the *Drosophila Adh* gene. Genetics. **166:** 237-242.

BAINES, J.F., Y. CHEN, A. DAS and W. STEPHAN, 2002 DNA sequence variation at a duplicated gene: excess of replacement polymorphism and extensive haplotype structure in the *Drosophila melanogaster bicoid* region. Mol. Biol. Evol. **19:** 989-998.

CHEN, Y., D.B. CARLINI, **J.F. BAINES**, J. PARSCH, J.M. BRAVERMAN, S. TANDA and W. STEPHAN, 1999 RNA secondary structure and compensatory evolution. Genes Genet. Syst. **74:** 271-286.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Dr. Wolfgang Stephan, for introducing me to the field of population genetics and molecular evolution, for giving me the opportunity to conduct my PhD research in his lab, and his support and guidance through the projects he suggested. I am especially thankful to Dr. John Parsch for showing me the ropes of *Drosophila* transformation genetics and always graciously sharing his excellent insight and expertise.

I owe a tremendous debt of gratitude to all the fellow graduate students and post-docs whom with I have shared my career in science. My lab mates, Dr. Yuseob Kim, Dr. Ying Chen and Dr. David Carlini made the lab a wonderful place to work and learn in. Dr. Ying Chen provided exceptional advice on all matters scientific, both on and off the lab bench. My fellow classmates Jon Bollback, Andrea Betancourt and Kelly Dyer made even the Rochester winters fun. I owe the entire Munich group the same amount of thanks for their support and the positive working environment they provided. In particular I would like to thank Stephan Hutter for his assistance in the laboratory and Dr. David De Lorenzo for his cheerfulness and willingness to solve any problem that is Macintosh. I would like to thank Dr. Daven Presgraves for being the most excellent resource to bounce scientific ideas for the few short months we overlapped in the lab. I must also thank Dr. Aparup Das, whose enthusiasm and expertise in the study of *D. ananassae* made the level of work on this species possible.

I thank my parents for encouraging me to learn at a young age and providing their love, care and support every step of the way.

And finally, I thank my wife, Judith, for her love, support and inspiration.

ABSTRACT

This dissertation examines two aspects of how natural selection shapes the amount and pattern of genetic variation within and between species: (1) the role of positively selected alleles in shaping the variation within and between subpopulations of a subdivided species and (2) the influence of epistatic selection operating on RNA secondary structures. First, the role of natural selection in shaping the pattern of variation within and between populations of the subdivided species *Drosophila ananassae* is investigated. To delimit the spread of positively selected alleles and characterize the role of natural selection in genetic differentiation, sequence data was collected from a locus in a region of low recombination for 13 populations, spanning a majority of the species range of *D. ananassae*. The migration behavior of this selected locus is compared to that of 10 independent neutrally evolving loci and tested against alternative models of natural selection. Second, nucleotide variation at the *D. melanogaster bicoid* locus is examined. The presence of a large, conserved secondary structure in the 3' untranslated region enables the relationship between RNA secondary structure and patterns of standing variation in natural populations to be explored. Variation within this structure is analyzed with respect to models of compensatory evolution and recent improvements of these models. Evidence suggests that *bicoid* may be the result of a relatively recent gene duplication in the Dipteran lineage, thus, variation in the *bicoid* coding region is also analyzed with respect to the evolutionary processes that may be ongoing if this gene is still undergoing diversification and/or refining of its function. Finally, long-range compensatory interactions between the two ends of *Drosophila* alcohol dehydrogenase (*Adh*) mRNA are investigated by experimental manipulation. Site-directed mutations were introduced in the *D. melanogaster Adh* gene in an effort to explain why previous

mutational analysis failed to fit Kimura's classical model of compensatory evolution. The results of the mutational analysis indicate that a classical result was not observed due to the pleiotropic effect of changing a nucleotide involved in both long-range base pairing *and* the negative regulation of gene expression.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Adh	alcohol dehydrogenase
ANOVA	analysis of variance
bcd	bicoid
bp	base pair(s)
DNA	deoxyribonucleic acid
fw	furrowed
kb	kilobase(s), kilobase pairs
mRNA	messenger ribonucleic acid
PCR	polymerase chain reaction
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SE	standard error
SSCP	single-strand conformation polymorphism
tRNA	transfer ribonucleic acid
UTR	untranslated region
v	vermilion
w	white
y	yellow

FOREWORD

The research presented in this thesis was conducted by myself, except for the following contributions: The C code written for Monte Carlo sampling of allele frequencies in the analysis of clinal variation in Chapter 1 was kindly provided by Dr. Sylvain Mousset. Dr. Aparup Das kindly provided use of his data from 10 neutral loci from the same *D. ananassae* populations that were sequenced for the *furrowed* locus. Dr. Ying Chen helped with the collection of sequence data and Dr. Aparup Das assisted with the inversion analysis in Chapter 2. In Chapter 3, the data from *Wa-f* control and *E2mut* lines, from the making the mutant construct to generating transformed lines, was generated by Dr. John Parsch. Dr. John Parsch provided the *Δ3* lines in the second portion of Chapter 3.

INTRODUCTION

Understanding the evolutionary forces that shape the amount and pattern of genetic variation is a fundamental goal of population genetics and crucial to the understanding of the evolutionary process. A theory of profound influence in the field of molecular population genetics and the first to explain a large body of population genetics data is Kimura's neutral theory of molecular evolution (KIMURA 1968; 1983). This theory holds that while most mutations are strongly deleterious and are eliminated from populations, those that are observed are selectively equivalent in fitness and are governed by random genetic drift rather than Darwinian selection. Thus, under the neutral theory, the level of genetic variation within a population is determined by the population's effective size and the neutral mutation rate. The amount of divergence between species is determined by the neutral mutation rate and the time since the splitting of the species, and a positive correlation between variation within species and divergence between species is expected for different gene regions. Because of the simplicity and mathematical tractability of the strictly neutral model, a number of specific and straightforward predictions can be made, making it a useful and powerful null hypothesis for disentangling the forces that shape genetic variation within and between species. In this dissertation, several different aspects of how natural selection influences genetic variation are investigated using *Drosophila* as a model system, each of which relying or drawing upon predictions or ideas stemming from the neutral theory.

Non-neutral evolution of DNA sequences: With the arrival of some of the first substantial data of variation at the DNA sequence level in the 1980's, it became clear that some observations were inconsistent with the predictions of the neutral model. Notably, studies of genetic variation in *Drosophila* found that levels of genetic

variation are markedly reduced in regions of low recombination in comparison to those with normal to high rates of recombination (AGUADÉ *et al.* 1989; STEPHAN and LANGLEY 1989). However, levels of divergence between closely related species were not affected by recombination (BEGUN and AQUADRO 1991,1992). A lack of correlation between levels of variation and divergence is inconsistent with the predictions of a constant-rate neutral model (KIMURA 1983). Thus, models involving natural selection were invoked.

Two alternative models proposed to explain the reduction of variability in regions of low recombination are the hitchhiking (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; STEPHAN *et al.* 1992) and background selection (CHARLESWORTH *et al.* 1993; HUDSON and KAPLAN 1995; CHARLESWORTH 1996) models. The genetic hitchhiking, or "selective sweep" model considers the effect of a strongly selected beneficial mutation on linked neutral variation. As a beneficial allele increases in frequency, only those neutral variants linked to this allele "come along for the ride", thus displacing other non-linked neutral variants. In contrast, the background selection model considers the effects of frequent, strongly deleterious mutations on linked neutral variants. Because neutral variants linked to deleterious mutations experience indirect purifying selection, these variants may also be removed from the population. If selection is sufficiently common (*i.e.* recurrent hittchiking), both models predict an overall correlation between levels of variation and the recombination rate (WIEHE and STEPHAN 1993; HUDSON and KAPLAN 1995). In other words, a higher rate of recombination provides more opportunity for neutral variants to recombine away from a beneficial allele, or escape the fate of deleterious mutations (removal from the population), respectively. Distinguishing between the relative

contributions of these models to the level and pattern of variation within species has been a major focus of research in the past decade.

Compensatory evolution and RNA secondary structure: In 1985, Kimura proposed a model of compensatory neutral evolution, in which his classic neutral model is extended to include pairs of mutations at two sites (KIMURA 1985). Under this model, compensatory mutations are defined as a pair of mutations at different loci that are individually deleterious, but neutral in appropriate combinations. For example, STEPHAN (1996) extended this model and applied it to nucleotide sites involved in Watson-Crick (WC) base pairing. In this case, individual mutations occurring at sites involved in base pairing may be deleterious if an important double-stranded (stem) region of a secondary structural element is disrupted, whereas the fitness can be restored if a second, compensatory mutation occurs at the appropriate complementary position. Examples of sequences that may evolve in such a manner are the secondary structures in rRNAs (NOLLER and WOESE 1981), tRNAs (SPRINZL *et al.* 1987), catalytic RNAs (PACE *et al.* 1989) and mRNAs (STEPHAN and KIRBY 1993; KIRBY *et al.* 1995; PARSCH *et al.* 2000).

Two main approaches are available for predicting RNA secondary structure: the thermodynamic and phylogenetic approaches. The method of thermodynamic predictions is based on free energy minimization (ZUKER 1989), but is generally only reliable for short sequences (WALTER *et al.* 1994). Currently, the phylogenetic approach is most reliable for predicting large RNA secondary structures (JAMES *et al.* 1989). This method relies on the comparison of interspecific sequence data, for which the underlying process of compensatory evolution described above serves as a basis. For example, an RNA secondary structure may be inferred by DNA sequence comparison using the Woese-Noller criterion (FOX and WOESE 1975; NOLLER and

WOESE 1981), which considers a putative helix as 'proven' if two or more covariations, caused by independently occurring base substitutions, are detected in sequence comparisons. A more rigorous method proposed by MUSE (1995) relies on a likelihood-ratio test to identify putative pairing regions displaying constraints for WC interactions. This method, however, has only been applied to putative structures that were previously identified. PARSCH *et al.* (2000) combined this method with a program (HAN and KIM 1993) that searches for secondary structures, allowing structures to be predicted from a sequence alignment.

Scope of this dissertation: In chapter one, the role of natural selection in shaping the pattern of variation within and between populations is investigated. *Drosophila ananassae* is a cosmopolitan species in the *melanogaster* group showing significant population structure (STEPHAN and LANGLEY 1989; Stephan *et al.* 1998). Previous studies on a limited geographic scale found compelling evidence for the action of natural selection at loci in regions of low recombination in this species (STEPHAN *et al.* 1998; CHEN *et al.* 2000). A statistical test designed to distinguish between the background selection and selective sweep models rejected the former, though the mode and geographic distribution of the sweep remained unclear. To further delimit the spread of positively selected alleles in this subdivided species and characterize the role of natural selection in genetic differentiation, sequence data was collected from a locus in a region of low recombination for 13 populations, spanning a majority of the species range of *D. ananassae*. The migration behavior of this selected locus is compared to that of 10 independent neutrally evolving loci (DAS *et al.*, submitted) and tested against alternative models of natural selection. Evidence of two independent sweeps restricted to specific regions of the species range is found, and the frequencies of the respective allele classes significantly correlate with latitude. These results

provide the first example of a cline in a region of low recombination and suggest a significant role for natural selection in genetic differentiation.

In chapter 2, nucleotide variation at the *D. melanogaster bicoid* (*bcd*) locus is examined. The motivation behind this study is two-fold. First, the presence of a large, conserved secondary structure in the 3' untranslated region (UTR) makes the *bcd* gene a good candidate for studying compensatory evolution and the relationship between RNA secondary structure and patterns of standing variation in natural populations. Variation within this structure is analyzed with respect to models of compensatory evolution (KIMURA 1985; STEPHAN 1996) and recent improvements of these models (INNAN and STEPHAN 2001). Second, several studies in *Drosophila* and closely related insects suggest that *bcd* may be the result of a relatively recent gene duplication in the Dipteran lineage. Thus, variation in the *bcd* coding region is also analyzed with respect to the evolutionary processes that may be ongoing if this gene is still undergoing diversification and/or refining of its function. For the purpose of understanding how these evolutionary forces are shaping variation within a single population, a population of *D. melanogaster* from Zimbabwe was chosen because it represents an ancestral population presumably closer to mutation-drift equilibrium, enabling the selective forces determining DNA sequence variation to be more easily elucidated (DAVID and CAPY 1988; BEGUN and AQUADRO 1995b).

In chapter 3, long-range compensatory interactions between the two ends of *Drosophila* alcohol dehydrogenase (*Adh*) mRNA are explored by experimental manipulation. Phylogenetic comparison of a wide range of *Drosophila Adh* sequences predicted a pairing region between a region in close proximity to the start codon and a conserved region of the 3' untranslated region (PARSCH *et al.* 1997). This suggests that selection maintaining WC base pairing between these regions of the transcript has

shaped the pattern of variation seen between these species. However, previous mutational analysis of this interaction failed to meet the conditions of KIMURA's (1985) classical model of compensatory evolution (PARSCH *et al.* 1997). In order to further investigate and verify long-range pairing in *Drosophila Adh* with respect to models of compensatory evolution, site-directed mutations were introduced in the *D. melanogaster Adh* gene. Alternative hypotheses for why previous analysis of long-range compensatory interactions failed to fit the classical model are explored. The results of the mutational analysis indicate that a classical result was not observed due to the pleiotropic effect of changing a nucleotide involved in both long-range base pairing *and* the negative regulation of gene expression.

CHAPTER 1

The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*

INTRODUCTION

Recent large-scale studies of genetic variation are beginning to confirm that species range expansion and the colonization of previously uninhabited territories is accompanied by genetic adaptation to changes in environmental conditions, the signature of which may be detected at the molecular level (HARR *et al*. 2002; GLINKA *et al*. 2003). In the case of *Drosophila melanogaster*, such an expansion is believed to have originated in Africa ~10,000-15,000 years ago (DAVID and CAPY 1988; LACHAISE *et al*. 1988). *D. ananassae*, another cosmopolitan species in the *melanogaster* group, is thought to have its origin in Southeast (SE) Asia (TOBARI 1993). A recent multilocus study of worldwide populations of *D. ananassae* substantiates this claim, defining the ancestral range of this species to be a region of SE Asia that existed as a single landmass (Sundaland) during the late Pleistocene (~18,000 years ago), while other populations including those in more temperate regions appear to be more recent colonizations (DAS *et al*., submitted). Thus, a similar scenario is emerging for this species, with the invasion of new climatic zones providing *a priori* expectation that local populations have adapted to their new environments. However, in contrast to *D. melanogaster*, *D. ananassae* is a species displaying significant population structure, enabling the footprints of natural selection at the DNA level to be analyzed in a subdivided population.

Previous studies of four *D. ananassae* populations (Nepal, Myanmar, India, Sri Lanka) found compelling evidence for the action of natural selection at loci in regions of low recombination (STEPHAN *et al.* 1998; CHEN *et al.* 2000). At both the *vermilion* (*v*) and *furrowed* (*fw*) loci, a pattern of homogenization of allele frequencies *within*, but differentiation *between* geographic regions [*i.e.* North (Nepal, Myanmar) *vs.* South (India, Sri Lanka)] was found. In both studies, this homogenization of allele frequencies in the northern populations rejected a model of background selection against deleterious mutations (CHARLESWORTH *et al.* 1993), instead favoring a model of the spreading of a beneficial allele (the selective sweep model) (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; STEPHAN *et al.* 1992). At the *fw* locus, the background selection model was rejected for the southern populations as well (CHEN *et al.* 2000), raising several important questions about the mode of selective sweeps in this subdivided species. Namely, is this pattern best explained by a single sweep (SLATKIN and WIEHE 1998), or have two independent sweeps occurred? Furthermore, the geographic distribution of the sweep(s) is unknown, as is whether it is associated with local adaptation to novel environments. Given that *D. ananassae* is highly structured and occupies a wide range of climatic zones, answers to these questions would also shed light on the role of natural selection in genetic differentiation.

For these reasons, we have expanded the study of nucleotide variation at the *fw* locus to include 13 populations, spanning a majority of the species range of *D. ananassae*. In contrast to previous studies, polymorphism data was collected by PCR and direct sequencing rather than single-strand conformation polymorphism (SSCP) and stratified sequencing. The migration behavior of this selected locus is compared to that of 10 independent neutrally evolving loci (DAS *et al.*, submitted), which alleviates the potential stochasticity of single-locus estimates of the migration rate.

The pattern of differentiation between pairs of populations is tested against alternative models of selection by the F_{ST} test of background selection (STEPHAN *et al*. 1998; CHEN *et al*. 2000). To further understand the nature of the selective forces shaping variation at *fw*, the distribution of *fw* haplotypes is analyzed with respect to population latitude.

MATERIALS AND METHODS

Population samples: A total of 126 isofemale lines were sampled from 13 locations in India, SE Asia, Australia and Japan. The location, abbreviation, number of sampled lines and date of collection is listed for each population in Table 1.1.

DNA extraction, PCR amplification and direct sequencing of individual *fw* alleles: To obtain sequence data from individual X chromosomes, genomic DNA was extracted from individual male flies using the PUREGENE™ DNA isolation kit (Gentra Systems, Minneapolis, MN). Oligonucleotides for amplification and direct sequencing were designed based on previously published *D. ananassae fw* sequence of the R1 (AF185289) and R9 and R42 (combined; AF185290) *Eco*RI restriction fragments described by CHEN *et al*. (2000). The R1 fragment covers part of the 5' untranslated region (UTR) and exons 1-9; R9/R42 covers exon 12, the 3' UTR and 3' flanking region. A 5.1-kb region (1.2-kb of R1 and 3.9-kb of R9 and R42) corresponding to the 5.7-kb *fw* fragment of CHEN *et al*. (2000) (minus 600 bp of 5' sequence) was amplified in three separate PCR reactions (FIGURE 1.1). Due to the presence of stretches of repetitive sequence, the R11 fragment was not sequenced (CHEN *et al*. 2000). Products were purified with QIA-quick columns (Qiagen), and both strands were subsequently sequenced using primers spaced ~400-500 bp apart.

Sequencing was performed on a Megabace 1000 automated DNA sequencer (Amersham Biosciences, Buckinghamshire, UK).

Sequence analysis: Sequences were edited with SeqMan and aligned with MegAlign (DNAStar Inc., Madison, WI). The DnaSP program version 3.51 (ROZAS and ROZAS 1999) was used for most intraspecific analyses. Nucleotide diversity, θ, was estimated according to WATTERSON (1975) and π according to NEI (1987).

Pairwise HKA tests: The HKA test was performed for all pairwise comparisons between loci [11 loci (*fw* + 10 neutral loci) → 55 comparisons], for each of the 13 sampled populations. For each population, the probability of observing at least *i* significant tests at the *fw* locus given that *n* paired tests were performed and *k* were significant between the *l* loci was calculated by the following equation (S. Mousset; personal communication):

$$p = \sum_{j=i}^{\min(l-1,k)} \frac{\binom{k}{j}\binom{n-k}{l-1-j}}{\binom{n}{l-1}} \qquad (1)$$

***F*$_{ST}$ test of the background selection model:** The original development of this test is described in STEPHAN *et al*. (1998) and was modified by CHEN *et al*. (2000). In summary, this test takes into account the effect of background selection and recombination on the effective population size of the locus of interest, enabling the effect of background selection on neutral variation in a subdivided population to be approximated by simulating the neutral coalescent under a model of population structure. In these simulations, the finite island model (CROW 1986, chap. 3.4) is used. The per-locus nucleotide diversity θ_S, the migration rate M_S and the recombination

rate R_S at the locus of interest are specified along with the number of subpopulations, *k*.

The migration rate at the locus putatively under selection, M_S, is estimated from the data:

$$M_S = M_O \frac{\overline{\theta}_S}{\overline{\theta}_O} f_{SO} \quad (2)$$

where M_O is the migration rate at neutrally evolving reference loci. M_O is estimated for each pair of populations as in CHEN *et al*. (2000), but is now obtained by taking the average over 10 neutrally evolving loci (introns). θ_S and θ_O are the arithmetic means of the per-site nucleotide diversities in the two subpopulations at the locus putatively under selection and the average of 10 neutral loci, respectively. The factor f_{SO} takes differences in the neutral mutation rate between loci into account (CHEN *et al*. 2000).

Analysis of clinal variation: To assess the association of allele frequency with population sample latitude, a linear regression analysis was performed. If selection affecting the observed distribution of *fw* haplotypes is attributable to an environmental gradient covarying with latitude, allele frequencies at *fw* may be expected to display a latitudinal cline. This analysis was performed on both a haplotype and a site-by-site basis following the design of BERRY and KREITMAN (1993). To distinguish between the effects of selection and population history, clinal variation at *fw* was compared to that observed at 10 neutrally evolving loci.

To assess the statistical significance of clinal variation, haplotype and SNP frequencies were first arcsine-transformed and then regressed on population latitude (measured as distance from the equator). The significance of the observed squared

correlation coefficient, r^2, was then estimated by generating 10,000 randomized data sets by binomial sampling under the expected frequency (the overall mean in the entire sample) of a SNP or haplotype. This generates 10,000 new frequencies for each subpopulation, for which 10,000 r^2 values are then computed to determine the significance of the observed r^2.

In addition, we performed an analysis to investigate the extent to which clinal variation at one site can be explained by the amount of linkage disequilibrium to another site as described by BERRY and KREITMAN (1993). In this approach, each site in turn is considered as the "governing" site, for which the clinal variation of every other "affected" site within a given locus may be explained by linkage to this site. For example, consider site **X** as the governing site and an affected site **Y**. For the entire pooled sample, the nucleotide **T** at site **Y** is present in 50% of the chromosomes in which the nucleotide **A** is present at site **X**, and 25% of the chromosomes which lack **A** at site **X**. If **A** is present at site **X** in 8 out of 12 chromosomes in a given subpopulation, the expected frequency of **T** at site **Y** in this subpopulation is (0.5 x 8) + (0.25 x 4) = 5/12. The expected frequency is computed in this manner for each individual subpopulation, from which 10,000 simulated frequencies are generated for each subpopulation. The significance is then determined by performing regressions on each of the 10,000 simulated sets of frequencies as described above. Thus, if the r^2 falls within the 95% confidence interval of the simulated r^2 values, the clinal variation of **T** at site **Y** may be explained by linkage with **A** at site **X**.

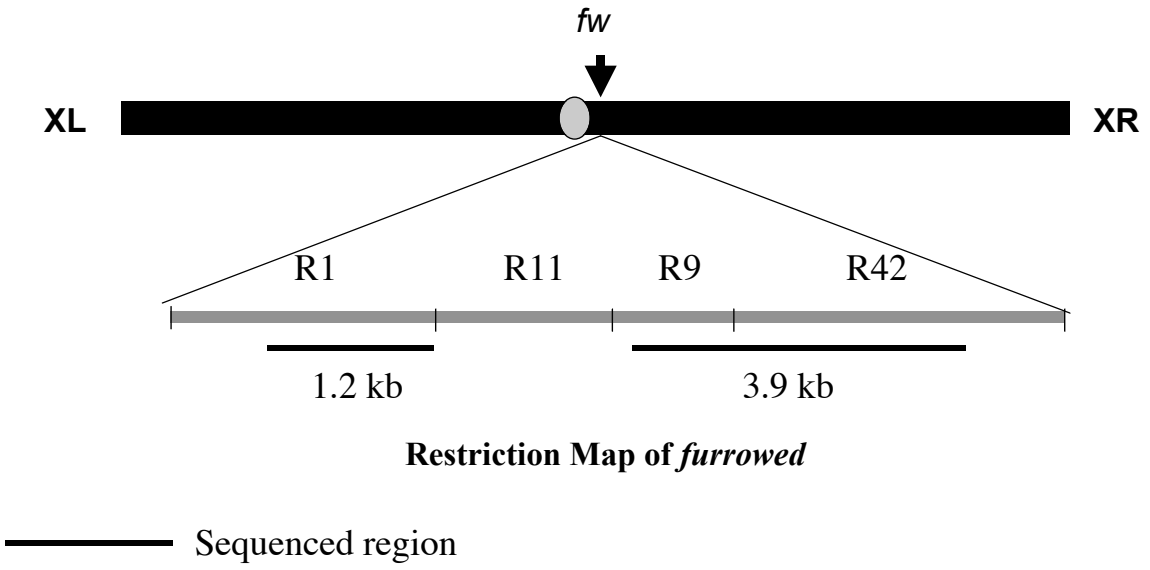

FIGURE 1.1 Restriction map of *furrowed* and location of the region sequenced in this study. R1-R42 are *Eco*RI restriction fragments described by CHEN *et al*. 2000. The R1 fragment covers part of the 5 ‘ UTR and exons 1-9; R9/R42 covers exon 12, the 3’ UTR and 3’ flanking region. A 5.1-kb region (1.2-kb of R1 and 3.9-kb of R9 and R42) corresponding to the 5.7-kb *fw* fragment of CHEN *et al*. 2000 (minus 600 bp of 5’ sequence) was amplified in three separate PCR reactions and subjected to direct sequencing.

TABLE 1.1

Population samples of *D. ananassae* used in this study

	Sampling Location	Country	Symbol	No. of Isofemale Lines	Collection Date
1.	Chennai	India	CH	9	2000
2.	Puri	India	PUR	8	2000
3.	Bhubaneswar	India	BBS	9	2000
4.	Kathmandu	Nepal	KATH	10	2000
5.	Mandalay	Myanmar	MAN	10	1994
6.	Chiang Mai	Thailand	CNX	10	2002
7.	Bangkok	Thailand	BKK	8	2002
8.	Kota Kinabalu, Borneo	Malaysia	KK	8	2002
9.	Bogor, Java	Indonesia	BOG	16	2001
10.	Darwin and Kakadu	Australia	DAR	9	1995
11.	Cebu	Philippines	CEB	9	2002
12.	Manila	Philippines	MNL	10	2002
13.	Kumejima, Okinawa	Japan	KMJ	10	2000

RESULTS

DNA polymorphism at *fw*: A region totaling 5.1 kb including most of the 3' half of the *fw* transcriptional unit and a large portion of the 3' flanking region was subjected to PCR and direct sequencing (FIGURE 1.1). On average, 10 lines per population were sequenced for 13 populations, giving a total of 126 sequenced lines (Table 1.1). A total of 54 nucleotide and 11 length polymorphisms were detected in this sample. Representative polymorphism data is shown in FIGURE 1.2. Of the three nucleotide polymorphisms in the coding region, only one changes the amino acid sequence (Glu to Gln at position 1113 of the R1 fragment), and this occurs only once in the sample (line 95 from BOG). The estimates of average nucleotide diversity, π and θ, at silent sites are low for each population (Table 1.2), on average more than 10-fold lower than estimates at 10 neutral loci in regions of normal to high recombination ($\pi_{fw} = 0.00066$; $\pi_{neutral} = 0.0086$) (DAS *et al.*, submitted). Notably, populations from the northernmost range of the sampled locations (Nepal, Myanmar and Japan) show the lowest levels of diversity, the most extreme being Nepal, which is monomorphic at *fw*.

Polymorphism and divergence: The average silent divergence between *D. ananassae* and its sibling species *D. pallidosa* at *fw* was 0.0055, while the average value of the 10 neutral loci was 0.0148 (DAS *et al.*, submitted). Under a constant-rate, neutral model of molecular evolution, levels of polymorphism and divergence should be correlated. To test this hypothesis, the method of Hudson, Kreitman and Aguadé (the HKA test) (HUDSON *et al.* 1987) was performed for all pairwise comparisons between loci [11 loci (*fw* + 10 neutral loci) → 55 comparisons], for each of the 13 sampled populations. For each population, the probability of observing at least *i* significant tests at the *fw* locus given that *n* paired tests were performed and *k* were significant between the *l* loci was calculated using equation (1) (see MATERIALS and

TABLE 1.2

Summary of polymorphism at *fw*

Population	Diversity, π	Diversity, θ	Tajima's *D*	Divergence
CH	0.00140	0.00132	0.32	0.00563
PUR	0.00068	0.00065	0.26	0.00556
BBS	0.00049	0.00070	-1.36	0.00574
KATH	0	0	-	0.00591
MAN	0.00023	0.00022	0.10	0.00585
CNX	0.00110	0.00089	1.04	0.00563
BKK	0.00132	0.00114	0.80	0.00541
KK	0.00077	0.00098	-1.07	0.00507
BOG	0.00034	0.00082	-2.28**	0.00502
DAR	0.00077	0.00093	-0.80	0.00530
CEB	0.00053	0.00077	-1.49	0.00511
MNL	0.00077	0.00112	-1.44	0.00585
KMJ	0.00013	0.00022	-1.56	0.00592

Nucleotide diversity π was estimated accordting to Nei (1987), and θ according to WATTERSON (1975). The value of *D* was obtained by TAJIMA's (1989) method. $^{**}P < 0.01$.

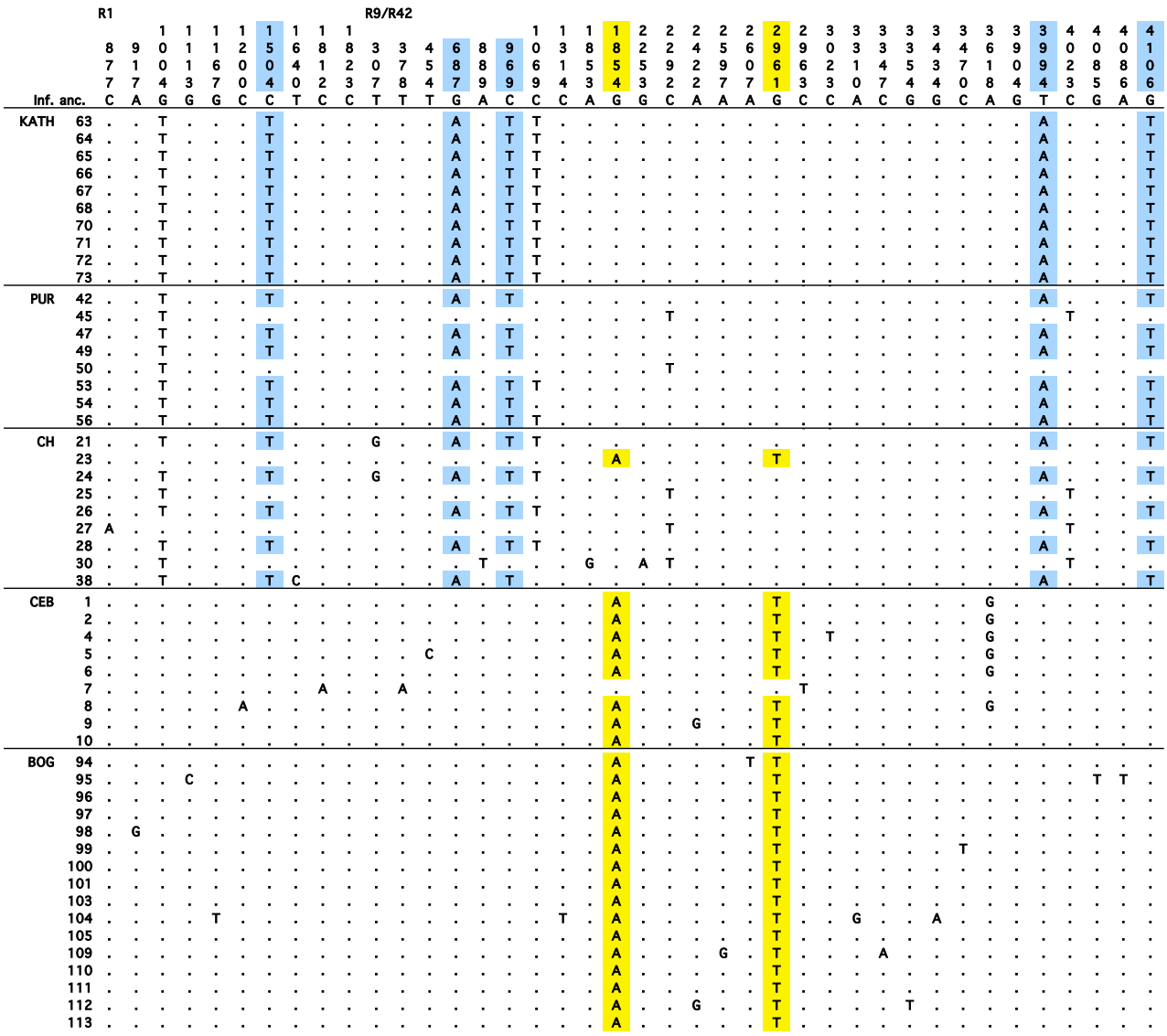

FIGURE 1.2 Representative polymorphism at *fw*. Length polymorphisms are not shown. The standard sequence is based on the inferred ancestral sequence as determined by *D. pallidosa*. All nucleotides shown as letters represent the derived state of the polymorphism. Polymorphisms distinguishing the Northern haplotype class (sites 1504 of R1 and 687, 969, 3994 and 4106 of R9/R42) are highlighted in blue. Polymorphisms distinguishing the Southern haplotype class (sites 1854 and 2961 of R9/R42) are highlighted in yellow.

METHODS). The number of comparisons deviating from the neutral expectation was significantly higher than expected for all northernmost populations (PUR, BBS, KATH, MAN, KMJ), as well as several populations in the south (KK, DAR, CEB). Thus, a constant-rate, neutral model of molecular evolution is rejected for these populations. These results are summarized in Table 1.3.

TABLE 1.3

Results of pairwise HKA tests between *fw* and 10 neutral loci

Population	Significant comparisons with *fw*	Total significant comparisons	*P*
CH	0	4	1
PUR	5	5	**7.2E-05**
BBS	4	4	**0.00062**
KATH	10	10	**3.4E-11**
MAN	9	16	**1.6E-05**
CNX	1	1	0.18182
BKK	2	3	0.08176
KK	2	2	**0.0303**
BOG	0	0	1
DAR	3	3	**0.00457**
CEB	4	4	**0.00062**
MNL	1	1	0.18182
KMJ	7	9	**1.9E-05**

The HKA test was performed for all pairwise comparisons between loci (11 loci (*fw* + 10 neutral loci) → 55 comparisons), for each of the 13 sampled populations. For each population, the probability of observing at least *i* significant tests at the *fw* locus given that *n* paired tests were performed and *k* were significant between the *l* loci was calculated using equation (1) (see MATERIALS and METHODS).

Haplotype structure: Of the 37 haplotypes observed in our data set, two major haplotype classes are apparent and are distinguishable by unique, high-frequency derived polymorphisms in complete linkage disequilibrium with one another. The

"Northern" haplotype class, which is in high frequency or fixed within the northern range of the sampled locations (overall frequency = 49.2%), is distinguished from all other haplotypes by a "T" at position 1504 of the R1 fragment and "A", "T", "A" and "T" at positions 687, 969, 3994 and 4106 of the R9/R42 fragment, respectively (FIGURE 1.2). Likewise, the "Southern" haplotype class is in high frequency or fixed within the south (overall frequency = 43.7%), and is distinguished from all other haplotypes by "A" and "T" at positions 1854 and 2961 of the R9/R42 fragment, respectively (FIGURE 1.2). The remaining haplotypes comprise 7.1% of the sample and do not contain any of these diagnostic derived polymorphisms. These are collectively more variable than the Northern or Southern haplotype classes and are likely representative of ancestral polymorphism at *fw* ($\pi_{other} = 0.00078$). The haplotype classes in high frequency, in particular the Northern class, harbor less variation ($\pi_{Northern} = 0.00024$; $\pi_{Southern} = 0.00045$). The geographic distribution of Northern, Southern and other haplotypes is shown in FIGURE 1.3.

Analysis of clinal variation: The relationship of allele frequency with population latitude is plotted for each haplotype class in FIGURE 1.4. A significant correlation (r^2) between transformed haplotype frequency and population latitude was found for both the Northern ($r^2 = 0.841$; $P < 0.0001$) and Southern ($r^2 = 0.669$; $P < 0.001$) haplotype classes.

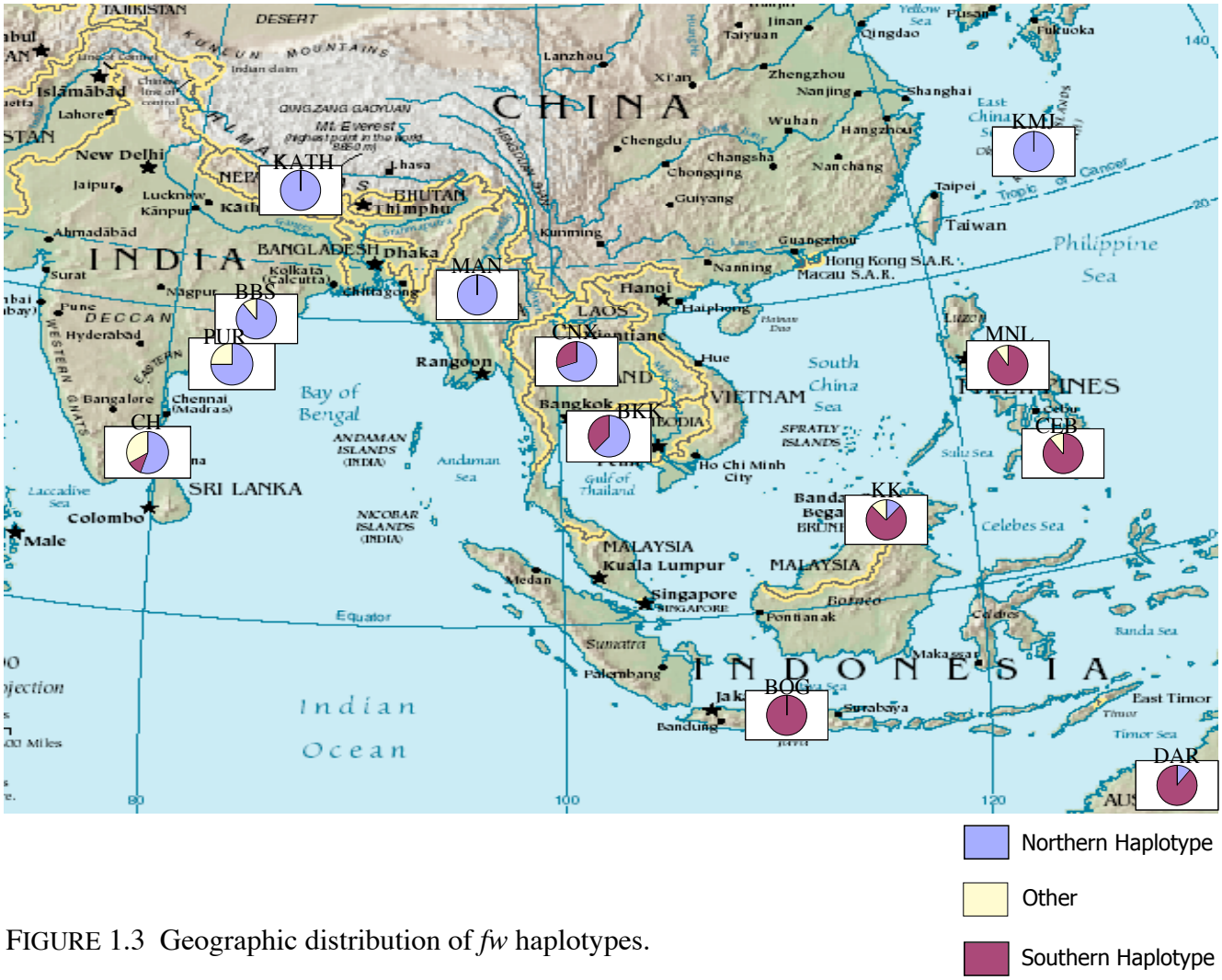

FIGURE 1.3 Geographic distribution of *fw* haplotypes.

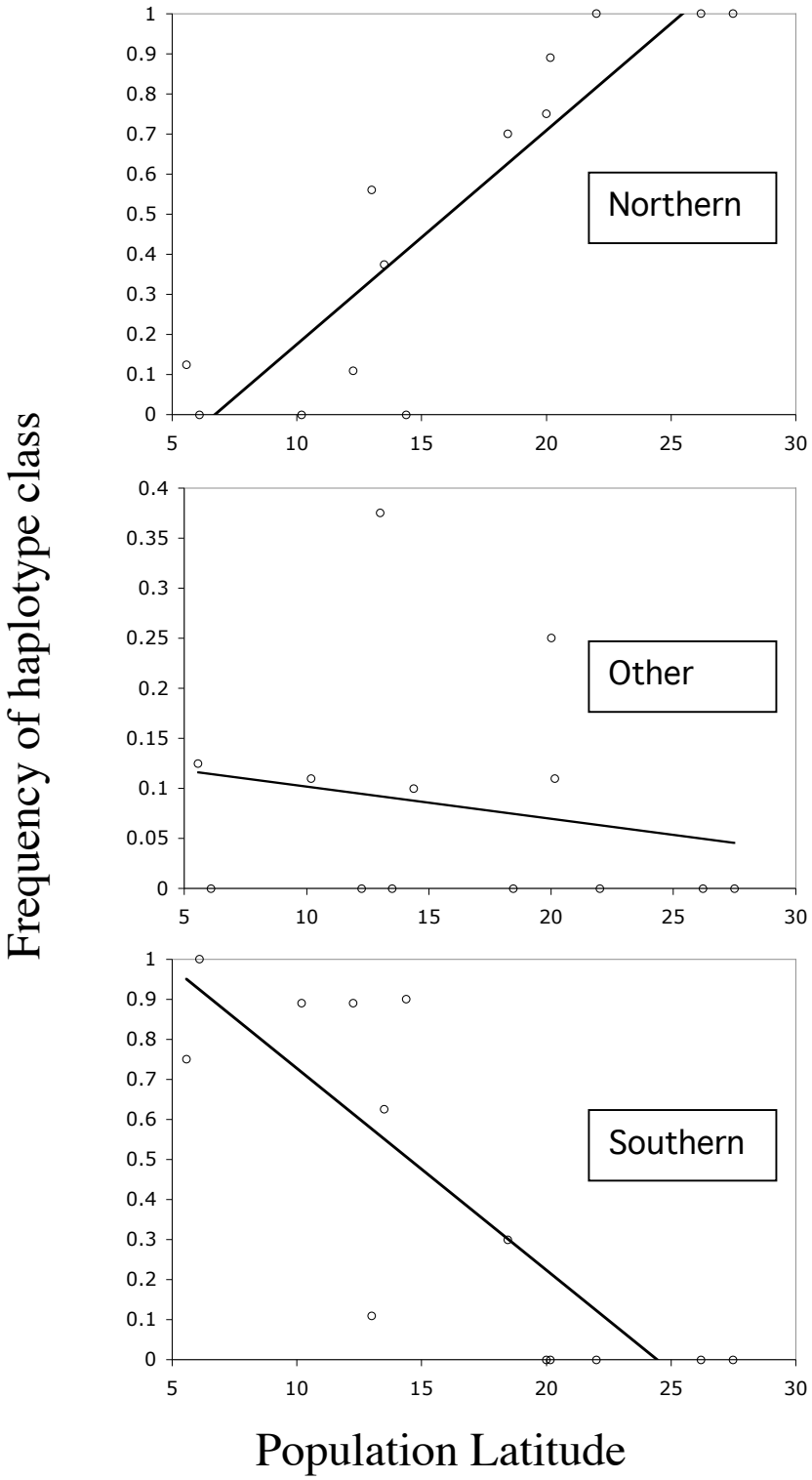

FIGURE 1.4 Relationship of nontransformed haplotype frequency and population latitude (measured as distance from the equator). Regressions (r^2) and slopes (m) are based on transformed frequencies: Northern $r^2 = 0.841$***, $m = 10.224$; Southern $r^2 = 0.669$**, $m = -10.115$. **$P < 0.001$ ***$P < 0.0001$.

To further investigate the observed clinal variation at *fw* and distinguish between the effects of natural selection *versus* population structure and/or history, a linear regression analysis was performed on a site-by-site basis for both *fw* and 10 neutrally evolving loci, following the design of BERRY and KREITMAN (1993). If selection acting on a site(s) linked to *fw* is responsible for the observed cline, the expectation is to observe clines only at *fw* or other sites linked to the target(s) of selection. In contrast, if population history is responsible, clines may be observed at loci across the entire genome. Thus, we compared the clinal variation of polymorphic sites at *fw* with that found at 10 unlinked, neutrally evolving loci. Of the 25 polymorphic sites (singletons were eliminated) subjected to regression analysis at *fw*, nine were significantly correlated with latitude with an average correlation of $r^2 = 0.751$. In comparison, of a total of 326 polymorphic sites tested at the neutral loci, 19 were significantly correlated with latitude with an average correlation of $r^2 = 0.324$. The results of the regression of polymorphic site frequency with population latitude for these 11 loci are summarized in Table 1.4.

In addition, we analyzed the relationship between allele frequency and latitude in various subsets of the sampled populations. If selection is responding to environmental factors covarying with latitude, the same linear relationship of allele frequency with latitude should be observable in multiple latitudinal transects (*i.e.* parallel clines), as seen with the F/S polymorphism of *D. melanogaster Adh* (OAKESHOTT *et al.* 1982). We divided the population samples into subsets labeled India (KATH, BBS, PUR and CHN), SE Asia (BUR, CNX, BKK, BOG) and "Easternmost" (KMJ, MNL, CEB, KK, BOG, DAR). The five sites diagnostic of the Northern haplotype class (R1: 1504 and R9R/42: 687, 969, 3994 and 4106) remain significant in all three subsets. Likewise, the two sites diagnostic of the Southern

haplotypes (R9/R42: 1854 and 2961) remain significant in the two subsets where they exist in appreciable frequency (the Southern haplotype occurs only once in India). In contrast, of the 37 polymorphic sites at the neutral loci displaying significant clinal variation in at least one set of populations (entire sample or one of the three subsets), 30 are significant in only one set, while the remaining seven are significant in only two sets (Table 1.4). Thus, while polymorphic sites associated with the Northern and Southern haplotype classes display significant clinal variation for the entire data set as well as independent subsets, the clinal variation observed at the neutral loci is inconsistent across the data set and more likely caused by chance on a more local scale.

Though the overall scheme and rationale of our analysis of clinal variation at *fw* follows that of BERRY and KREITMAN (1993), our data set differs in an important way. Previous studies applying this design (BERRY and KREITMAN 1993; VERRELLI and EANES 2001) have focused on distinguishing and identifying the target(s) of clinal selection (*i.e.* sites such as amino acid polymorphisms displaying significant clinal variation that could not be explained by linkage to other sites were identified as putative targets). Although linkage disequilibrium should technically be calculated only for individual populations, extensive non-independence between polymorphic sites exists across the entire surveyed region. In particular, the derived polymorphisms characterizing the Northern and Southern haplotypes are in complete linkage disequilibrium. This is not surprising, given that *fw* resides in a region of very low recombination (STEPHAN and MITCHELL 1992), the size of the region of which linked neutral variation is affected by selection may be very large. Given that 53 of 54 segregating mutations in this data set are silent and the single nonsynonymous mutation occurs only once in the sample, the target(s) of selection is unlikely to reside

within the region sequenced in this survey. However, the analysis of clinal variation with respect to linkage disequilibrium to other sites applied to this data set is informative nonetheless, as it reaffirms that sites distinguishing the Northern and Southern haplotypes are responding to clinal selection in a non-independent manner (FIGURE 1.5), most likely due to the target of this clinal selection being located outside of the sequenced region.

TABLE 1.4

Summary of clinal variation of polymorphic sites at *fw* and 10 neutral loci

				r^2			
Locus	Site	Frequency	Slope	All pop's	India	SE Asia	Easternmost
fw	834	0.07	9.79	n.s.	-	-	0.795*
(25)	1004	0.53	8.84	0.782***	n.s.	0.886*	0.806*
	1504	0.49	10.46	0.857***	0.929*	0.886*	0.806*
	687	0.49	10.46	0.857***	0.929*	0.886*	0.806*
	969	0.49	10.46	0.857***	0.929*	0.886*	0.806*
	1069	0.19	8.9	0.305*	n.s.	n.s.	-
	1854	0.44	-10.57	0.693***	-	0.987**	0.731*
	2292	0.06	-6.05	n.s.	0.589*	-	-
	2961	0.44	-10.57	0.693***	-	0.987**	0.731*
	3994	0.49	10.46	0.857***	0.929*	0.886*	0.806*
	4023	0.05	-11.8	n.s.	0.948*	-	-
	4106	0.49	10.46	0.857***	0.929*	0.886*	0.806*
1	51	0.14	6.81	n.s.	n.s.	n.s.	0.689*
(35)	59	0.19	5.74	n.s.	n.s.	n.s.	0.736*
	348	0.14	-20.91	0.335*	n.s.	n.s.	n.s.
2	94	0.05	21.56	n.s.	-	-	0.873**
(63)	157	0.12	-23.67	0.283*	n.s.	n.s.	n.s.
	159	0.10	-48.77	0.424*	-	n.s.	n.s.
	310	0.78	5.46	n.s.	-	0.933**	n.s.
	321	0.88	10.27	0.290*	-	0.857*	n.s.
	338	0.02	-94.96	0.270*	-	0.852*	n.s.
	340	0.05	-5.48	n.s.	-	0.852*	n.s.
	352	0.02	77.84	0.308*	n.s.	-	-
3	12	0.31	5.35	n.s.	0.928*	n.s.	0.757*
(60)	24	0.01	114.99	0.280*	-	-	-
	26	0.01	104.11	0.251*	-	-	-
	32	0.15	3.96	n.s.	0.994*	n.s.	n.s.

	45	0.05	10.80	n.s.	-	0.840*	n.s.
	95	0.01	114.99	0.280*	-	-	-
	219	0.01	114.99	0.280*	-	-	-
4 (34)	425	0.07	23.81	n.s.	-	n.s.	0.681*
5	6	0.18	-18.56	n.s.	0.936*	n.s.	n.s.
(27)	58	0.05	-57.94	0.343*	-	n.s.	n.s.
6	17	0.10	22.68	n.s.	n.s.	n.s.	0.664*
(10)	119	0.18	28.16	0.398*	n.s.	n.s.	0.664*
7	31	0.35	11.33	n.s.	n.s.	n.s.	0.725*
(19)	145	0.10	8.36	n.s.	n.s.	n.s.	0.736*
	207	0.01	-126.33	0.370**	-	-	-
8 (15)	89	0.03	-50.13	0.305*	-	n.s.	-
9	9	0.03	35.08	n.s.	-	-	0.795*
(29)	205	0.25	-7.13	n.s.	n.s.	n.s.	n.s.
	242	0.91	18.28	0.454**	-	n.s.	0.942*
	257	0.33	-4.82	n.s.	0.931*	n.s.	n.s.
10	28	0.01	-129.66	0.356*	-	0.852*	-
(34)	150	0.25	12.62	0.315*	0.943*	n.s.	n.s.
	152	0.86	4.74	n.s.	0.935*	n.s.	n.s.
	327	0.27	13.13	0.303*	n.s.	n.s.	n.s.
	395	0.16	10.41	n.s.	n.s.	n.s.	0.890*
	453	0.37	-10.48	0.313*	n.s.	n.s.	n.s.

The numbers of polymorphic sites analyzed for clinal variation (singletons were eliminated) at each locus are indicated in parentheses in column 1. Only sites displaying significant clinal variation in one or more the subsets (see below) are shown. The frequency of individual sites is calculated for the entire pooled sample, based on the derived state of the polymorphism as determined by the outgroup *D. pallidosa*. The slopes are computed from transformed data based on the entire pooled sample. Regressions (r^2) of transformed allele frequencies on latitude were performed for all the populations combined, as well as the following subsets: India (KATH, BBS, PUR, CH), SE Asia (MAN, CNX, BKK, BOG) and Easternmost (KMJ, MNL, CEB, KK, BOG, DAR). Polymorphic sites monomorphic or occurring only once in individual subsets are indicated by dashes. "n.s." indicates no significant clinal variation. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

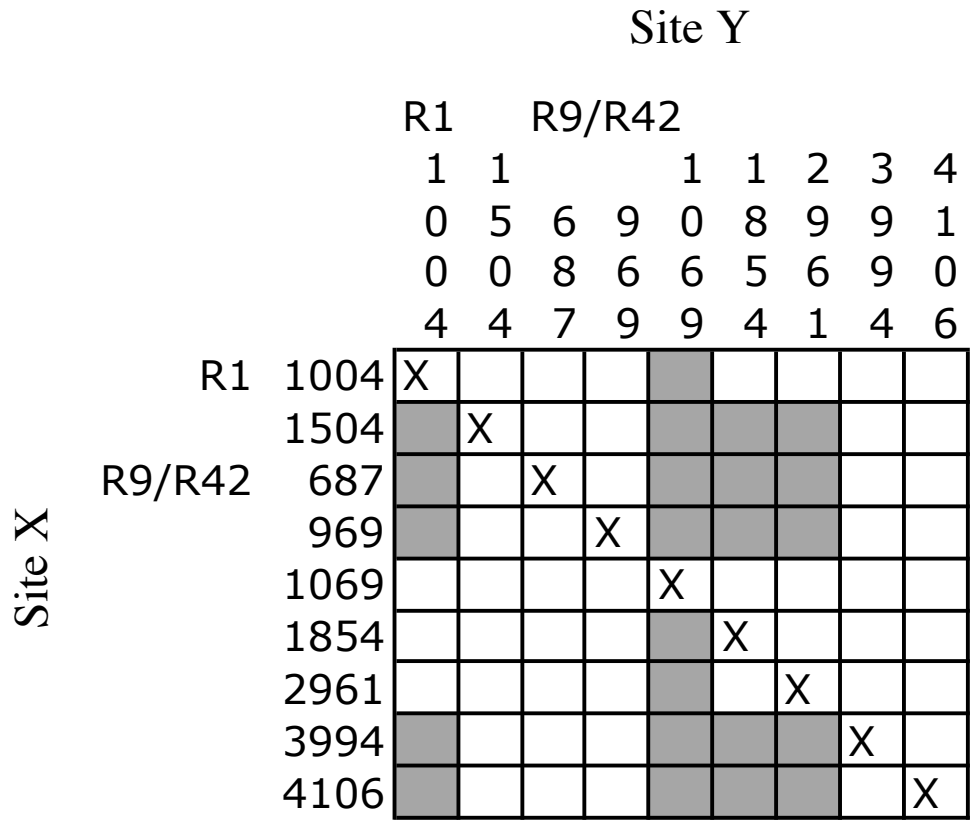

FIGURE 1.5 Summary of clinal variation at *fw* for all populations. Only sites with significant clinal variation are shown (see Table 1.4). Shaded boxes refer to significant clinal variation at site Y that cannot be explained by linkage to site X.

Test of the background selection model: The above results of the HKA test indicate that for several populations, the level of polymorphism at *fw* is too low to be explained by a constant rate, neutral model. Two alternative models proposed to explain the reduction of variability in regions of low recombination are the hitchhiking (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; STEPHAN *et al.* 1992) and background selection (CHARLESWORTH *et al.* 1993; HUDSON and KAPLAN 1995; CHARLESWORTH 1996) models. The hitchhiking model describes the effect of rare, strongly selected beneficial mutations on linked neutral polymorphism, while the background selection model considers the effects of frequent, strongly deleterious mutation on linked neutral variants. In the following, we applied the method of STEPHAN *et al.* (1998), which utilizes the unique prediction of background selection operating in a subdivided population to distinguish between these two alternative models. Because the effective size of local demes is reduced in regions of low recombination relative to that in regions of normal to high recombination, a fewer number of effective migrants is expected to increase F_{ST} (CHARLESWORTH *et al.* 1997).

To test the null hypothesis that background selection is responsible for the observed pattern of differentiation between pairs of populations throughout the *D. ananassae* species range, we generated a probablity density of F_{ST} values under the finite island model for k demes and a migration rate M_S, mutation parameter θ_S and per locus recombination rate R_S at the locus putatively under selection (*fw*). A range of values was chosen for the unknown parameters k and R_S, while M_S and θ_S were estimated from the data (see MATERIALS and METHODS). The probability of obtaining a value of F_{ST} less than or equal to the observed F_{ST} under background selection is given for representative pairwise comparisons between populations in Table 1.5. For

several comparisons between pairs of populations in the north and in the south, F_{ST} values are too low to be explained by the background selection model for various values of k and R_S, whereas almost all remaining values within these regions approached significance. Although less conservative, higher values of k are likely more realistic for *D. ananassae* (DAS *et al.*, submitted) and produced lower P values. In addition, in contrast to the previous study of *fw*, evidence of intragenic recombination was found by the four-gamete rule (HUDSON and KAPLAN 1985) in this data set, indicating that a non-zero level of recombination may be appropriate, which also produces lower P values. Thus, the low level of differentiation between populations in the north and south may be indicative of the spread of positively selected alleles in these regions.

TABLE 1.5

Probability of obtaining the observed or lower values of F_{ST} under the background selection model

Population 1	Population 2	Region of Comparion	k=100		k=500	
			R=0	R=0.1	R=0	R=0.1
KATH	MAN	N-N	0.068	0.057	0.070	**0.032**
KATH	BBS	N-N	**0.039**	**0.027**	**0.039**	**0.012**
KATH	KMJ	N-N	0.080	0.077	0.080	0.078
MAN	BBS	N-N	**0.025**	**0.029**	**0.025**	**0.005**
MAN	KMJ	N-N	0.051	**0.035**	**0.050**	**0.019**
KATH	BOG	N-S	0.522	0.536	0.525	0.556
MAN	DAR	N-S	0.456	0.458	0.461	0.437
BBS	DAR	N-S	0.516	0.498	0.509	0.473
BBS	BOG	N-S	0.747	0.766	0.737	0.770
KMJ	KK	N-S	0.342	0.322	0.332	0.299
DAR	BOG	S-S	0.107	0.074	0.107	**0.030**
DAR	CEB	S-S	**0.048**	**0.034**	**0.045**	**0.017**
DAR	KK	S-S	0.065	**0.046**	0.065	**0.019**
BOG	CEB	S-S	0.056	**0.038**	0.057	**0.013**
BOG	KK	S-S	0.064	**0.039**	0.069	**0.012**

DISCUSSION

Overview: In this study, we have re-examined the pattern of nucleotide variation at *fw* on a much larger scale, using PCR and direct sequencing as opposed to SSCP and stratified sequencing. Though in most cases new population samples were used (only the Myanmar sample was also used by CHEN *et al.* (2000)), the overall level of polymorphism at *fw* was found to agree between these two methods. In addition, the added advantage of a detailed knowledge of population history from 10 neutrally evolving loci was available (DAS *et al.*, submitted). The major goals of this study were to elucidate the pattern and distribution of selective sweeps at this locus and help

establish the role of natural selection in differentiation between populations. In the following, we discuss several lines of evidence for natural selection playing a significant role, in particular with respect to recent range expansions and potential adaptation to new environments.

Selection *vs*. demography: In addition to providing a control for nonadaptive processes in the analysis of clinal variation, detailed analysis of population structure based on10 neutral loci has revealed other interesting aspects of the population history of *D. ananassae* that shed light on the pattern of variation observed at *fw* (DAS *et al*., submitted). First, the method of VOGL *et al*. (2003) applied to these loci has enabled these populations to be characterized as either central or peripheral by the inference of the migration-drift parameter, θ_P. In short, this is the probability that two sequences randomly drawn from a population coalesce before migration. High values of θ_P are indicative of populations being highly differentiated due to drift (and thus peripheral), while low values indicate the population is closer to the central, ancestral species distribution (VOGL *et al*. 2003). The populations from five SE Asian localities [BKK, KL (Kuala Lumpur, not included in the *fw* survey), BOG, KK and MNL] display high variability and low estimates of θ_P, and are inferred to be central populations likely representative of an ancestral population of *D. ananassae*. The other populations showed lower variability and higher estimates of θ_P, indicating that these populations are more peripheral. Due to the consistent ~10-fold lower variation at *fw* in comparison to the neutral loci, estimates of θ_P are systematically higher at *fw*. However, the relative difference in these estimates between populations differs at *fw* and the 10 neutral loci in several cases (FIGURE 1.6). In particular, the CH population has one of the highest estimates of θ_P at the neutral loci, in contrast to the lowest at *fw*. Thus, though CH appears to be one of the most peripheral of all the populations based

on the neutral loci, a large number of haplotypes are segregating at *fw* relative to the other populations. A likely explanation for this discrepancy is the intermediate latitude of CH and allele frequencies at *fw* being largely governed by selection responding to an environmental gradient correlated with latitude.

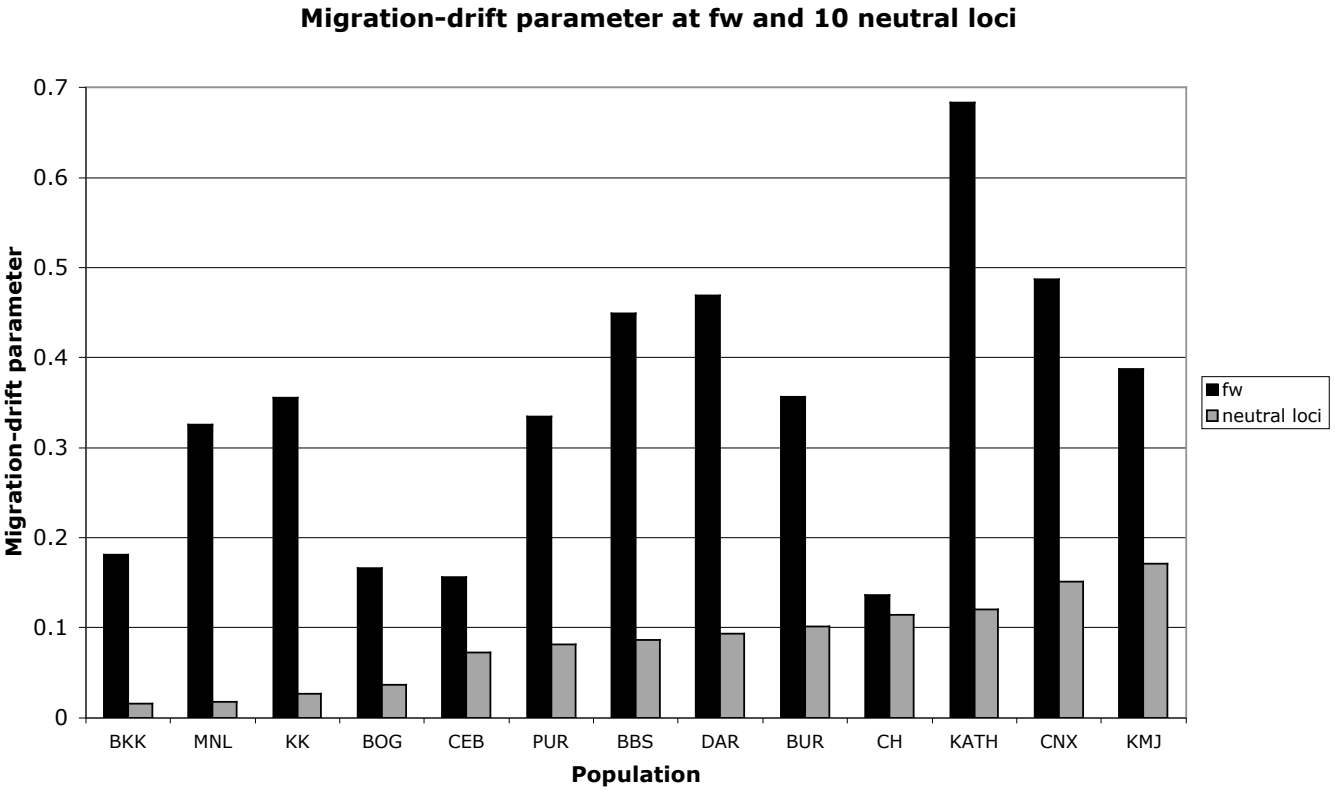

FIGURE 1.6 Comparison of the migration-drift parameter, θ_P (VOGL *et al*. 2003) at *fw* and 10 neutral loci.

Second, analysis of the ancestry of these populations is suggestive of selection influencing the distribution of haplotypes at *fw*. Based on both the model-based clustering algorithm of the program Structure (PRITCHARD *et al*. 2000) and a Neighbor-Joining population tree (based on F_{ST}), the 10 neutral loci reveal a close relationship between the Indian populations (BBS, PUR, CH), KATH and MAN and

the sample from Australia (DAR). Similarly, the samples from Java (BOG) and Japan (KMJ) are closely related, suggesting a common ancestral origin for these pairs of populations. In contrast, these pairs of populations are highly differentiated at *fw*, being fixed or nearly fixed for the Northern and Southern haplotypes in these respective regions. Thus, the composition of the ancestral populations from which current peripheral populations are sampled does not appear to have solely determined the current pattern observed at *fw*.

Selective sweeps in a subdivided population: Previous analysis of polymorphism at *fw* in four populations (Nepal, Myanmar, India, Sri Lanka) considered several possible scenarios of a selective sweep in a subdivided population (CHEN *et al*. 2000). One possibility is that the pattern of homegenization of allele frequencies *within*, but differentiation *between* geographic regions [North (Nepal, Myanmar); South (India Sri Lanka)], was caused by independent selective sweeps in each region (the two-sweep model). Alternatively, if more than one haplotype became associated with the selected allele *via* recombination, differential migration of these two haplotypes could result in a similar pattern (the single-sweep model) (SLATKIN and WIEHE 1998). A third scenario not mutually exclusive of the above two models is that of local adaptation, where a selective sweep may be restricted to certain regions of a species range.

The significantly expanded sampling of this current survey greatly facilitates distinguishing between alternative models. Similar to the study of CHEN *et al*. (2000), a pattern of homogenization *within*, but differentiation *between* geographic regions is observed. However, two important differences are the scale on which this is observed and the cline of allele frequencies between these two regions. The Northern haplotype is fixed or in high frequency in all populations of higher latitude, and a cline of

decreasing frequency is found throughout the entire sample. A similar pattern is observed with the Southern haplotype, though the pattern of clinal variation is not as strong: the Northern haplotype also decreases in frequency in the absence of high frequencies of the Southern haplotype (*i.e.* in India), thus, the cline of Southern haplotype frequency in the opposite direction may be a secondary effect (see section on clinal variation and below). The model of SLATKIN and WIEHE (1998) predicts that differential migration of two different haplotypes linked to the same selected allele will lead to the fixation of only one of these haplotypes in any given population. In addition, should this single-sweep model be invoked, the selective advantage of the beneficial allele should also be necessarily unconditional. Thus, under this model, given that populations in the north and south are fixed or nearly fixed for their respective haplotypes, populations located in intermediate locations (*i.e.* CH, CNX, BKK) should also be fixed for one haplotype or the other. In contrast to this prediction, the Northern haplotype coexists with other haplotypes, the degree to which being determined by latitude. For this reason, the single-sweep model is unlikely to explain the data. Thus, it is most plausible that two independent sweeps have occurred in the Northern and Southern regions.

Given the strong evidence for clinal variation of the Northern haplotype, it seems that minimally this sweep is a candidate for a locally favored substitution. We hypothesize that the regional high frequency of the Southern haplotype is most likely due to the spread of an unconditionally favorable allele [*i.e.* some populations showing evidence of this sweep are part of the ancestral range of *D. ananassae* (DAS *et al.*, submitted)], though this has not spread throughout the species range because a second, independent sweep associated with a locally favored allele has occurred in the North.

Target(s) of selection: Traits such as cold tolerance are known to vary with latitude in several species, including *D. ananassae* (GILBERT and HUEY 2001), and it was recently shown that high altitude Himalayan strains of this species have evolved a temperature-dependency to the rhythmicity of eclosion (KHARE *et al.* 2002). Although the pattern of differentiation at *fw* suggests that positively selected mutations have occurred at linked sites, the size of the fragment displaying reduced variation may be quite large due to the low recombination of the region containing *fw*. Though numerous chromosomal rearrangements have occurred since *D. ananassae* and *D melanogaster* last shared a common ancestor, gene order on a more local scale is more likely to be preserved. In *D. melanogaster*, *fw* lies in a region of normal to high recombination that is relatively gene-rich (~10 genes in a 100 kb window around *fw*). Thus, it is reasonable to expect that many potential targets of selection are linked to *fw*. The availability of the genome sequence of *D. ananassae* in the near future will greatly facilitate the identification of mutation(s) involved in this sweep, as well as providing the necessary background for studying adaptation at the genome level in another species. The parallels between the recent evolutionary history of the two cosmopolitan species *D. melanogaster* and *D. ananassae* (*e.g.* the invasion of temperate regions from an ancestral tropical environment) provide an exciting opportunity for comparative studies of adaptation at the genome level.

CHAPTER 2

DNA sequence variation at a gene of relatively recent origin: Excess of replacement polymorphism and extensive haplotype structure in the *Drosophila melanogaster bicoid* gene

INTRODUCTION

The *bicoid* (*bcd*) gene of *Drosophila* has played an important role in understanding the system of developmental genes that regulate pattern formation in the early fly embryo. *bcd* mRNA is maternally transcribed and localized to the anterior pole of the embryo (BERLETH et al. 1988). The *cis*-acting sequences necessary for localization fall within a large (~700 nucleotides), phylogenetically conserved and well-characterized secondary structural element in the 3' untranslated region (UTR) (MACDONALD and STRUHL 1988; MACDONALD 1990). Translation of this localized transcript gives rise to a gradient of Bicoid protein that controls development of the head and thorax (DRIEVER and NÜSSLEIN-VOLHARD 1988; FROHNHÖFER, LEHMANN and NÜSSLEIN-VOLHARD 1986). This is achieved by two important functions of Bicoid. First, it transcriptionally activates zygotic segmentation genes in a threshold-dependent fashion. Second, it functions as a translational repressor of uniformly distributed *caudal* mRNA by binding to its 3' end, thus resulting in a gradient of Caudal protein in the opposite direction (DUBNAU and STRUHL 1996).

In addition to being crucial in understanding the determination of anterior-posterior polarity in the fruit fly, two aspects of *bcd* make it an interesting candidate for evolutionary analysis. First, recent studies in *Drosophila* and closely related

insects suggest that *bcd* may be unique in insect developmental systems in terms of function and evolutionary history. Numerous laboratories have consistently failed in attempts to isolate *bcd* from insects other than schizophoran flies, despite the usual ease in cloning other developmental homologs even from distantly related species (STAUBER, JÄCKLE and SCHMIDT-OTT 1999). One hypothesis suggests that *bcd* may simply be a rapidly evolving homeobox gene, and failure to clone it is due to technical difficulties (PATEL 2000; SCHRÖDER and SANDER 1993). However, when *bcd* was cloned from a basal cyclorrhaphan fly (*Megaselia abdita*) it was found to be most closely related to the *Megaselia zerknüllt* (*zen*) gene, suggesting that *bcd* may be the result of a recent gene duplication and diversification, leading to a novel regulatory protein (STAUBER, JÄCKLE and SCHMIDT-OTT 1999). Recently, SCHAEFFER *et al*. (2000) found functional redundancy between Bicoid and the terminal system's role in thorax development, supporting the recent evolution of an anterior morphogenetic center comprised of both Bicoid and the terminal system. Second, the presence of a large, conserved secondary structure in the 3' UTR makes the *bcd* gene a good candidate for studying compensatory evolution and the relationship between RNA secondary structure and patterns of standing variation in natural populations (CHEN *et al*. 1999).

Despite these intriguing observations, a population-level analysis has until now not been performed on *bcd*. In this study, DNA sequence variation was examined for a 4-kb region of the *bcd* gene, including a portion of the 5' UTR, the entire coding region and the 3' UTR, for 25 *D. melanogaster* isofemale lines from Lake Kariba, Zimbabwe. This population was chosen because it has been previously shown to harbor more than twice the amount of genetic variation and lower levels of linkage disequilibrium than that of other non-African populations *of D. melanogaster* (BEGUN

and AQUADRO 1993, 1995a, 1995b). Thus, Zimbabwe may represent an ancestral population closer to mutation-drift equilibrium, enabling the selective forces determining DNA sequence variation to be more easily elucidated (DAVID and CAPY 1988; BEGUN and AQUADRO 1995b). The goals of this study are (1) to test whether there is evidence for natural selection attributable to the relatively recent origin of *bcd* in the Dipteran lineage and (2) to test whether the pattern of variation in the 3' UTR is consistent with the presence of a large conserved mRNA secondary structure.

MATERIALS AND METHODS

***Drosophila* strains:** A total of 25 *D. melanogaster* isofemale lines collected from Lake Kariba, Zimbabwe (kindly provided by C. Aquadro) were used in this survey. To isolate individual *bcd* alleles, a male from each of the isofemale lines was crossed to virgin females from a *Df(3R)MAP117 pp / TM3, Sb1* line (obtained from the Bloomington Fly Stock Center), which contains a deleted *bcd* region. Single male offspring with wild type bristles was again crossed to the *Df(3R)MAP117 pp / TM3, Sb1* virgin females, and offspring with wild type bristles were used for genomic DNA extraction. For the interspecific comparison, one *D. simulans* stock collected in Davis, California was used (kindly provided by H. A. Orr). All flies were raised and crossed at room temperature, and on standard media.

DNA extraction, PCR amplification and direct sequencing of the *bcd* alleles: Genomic DNA was extracted from homozygous whole flies with the DNeasy tissue kit (Qiagen). Oligonucleotides for amplification and direct sequencing were designed based on a previously published *D. melanogaster bcd* sequence (GenBank accession number X07870). These primers were used in PCR reactions to amplify a 4-kb region of *bcd*, comprising 450 bp of 5' flanking region, the entire coding region, and 1 kb of

3' flanking region. PCR products were purified with QIAquick columns (Qiagen), and both strands were subsequently sequenced using primers spaced ~400 bp apart. Sequencing was performed on an ABI377 automated sequencer with the Dye Terminator chemistry (Perkin-Elmer). The homologous region of *D. simulans* was amplified using the PCR primers designed for *D. melanogaster*, and new *D. simulans* primers were designed based on the available *D. simulans* sequence and used if the *D. melanogaster* primers failed in the sequencing reactions due to mismatches. The *D. melanogaster* sequences are deposited in GenBank as a population set with accession numbers AF466621-45, and the accession number of the *D. simulans* sequence is AF465792. The coordinates according to the reference sequence (GenBank accession X07870) are used throughout this paper.

Sequence analysis: Sequences were assembled and aligned with the SeqEd program (Perkin-Elmer), and all variable sites were checked manually and verified in both strands. The *bcd* gene of *D. melanogaster* is alternatively spliced at intron 2 (positions 2216 - 2270, 55 bp; or positions 2216 - 2255, 40 bp). Since smaller introns (< 51 bp) are usually spliced less efficiently (MOUNT *et al.* 1992), the assignment of coding and non-coding regions are according to the major transcript; *i.e.*, positions 2256 - 2270 are regarded as non-coding region. The homologous region of *D. simulans* was aligned to the *D. melanogaster bcd* sequence, and gaps in the alignment were not used in the sequence analysis. After the sequence alignment, the coding and non-coding regions of *D. simulans bcd* were assigned according to the *D. melanogaster* sequence. The DnaSP program version 3.50 (ROZAS and ROZAS 1999) was used for most intraspecific and interspecific analyses. Nucleotide diversity, θ, was estimated according to WATTERSON (1975), and π according to NEI (1987). Nucleotide

divergence, κ, between *D. melanogaster* and both *D. simulans* and *D. pseudoobscura* was estimated according to NEI (1987).

The following neutrality tests were performed using the program DnaSP (ROZAS and ROZAS 1999): The HKA test (HUDSON, KREITMAN and AGUADÉ 1987), TAJIMA's (1989) test and the MCDONALD and KREITMAN (1991) test. The probabilities for the MCDONALD and KREITMAN (1991) test were obtained by both the two-tailed Fisher's exact test and the G-test. To detect heterogeneity in the ratio of polymorphism to divergence in the region surveyed, the program DNA slider (MCDONALD 1996, 1998) was used.

Coalescent simulations for obtaining the probabilities of the number of haplotypes and haplotype diversity (DEPAULIS and VEUILLE 1998) were performed using the program DnaSP (ROZAS and ROZAS 1999), and the haplotype test of HUDSON *et al.* (1994) was performed using a program written by J. Braverman (kindly provided by J. Parsch). The test of HUDSON *et al.* (1994) determines the probability of observing a subset of alleles of size *i* with *j* or fewer segregating sites given an overall sample of *n* alleles with *S* segregating sites. The recombination parameter, *R*, was estimated by three methods, including two based on polymorphism data (HUDSON 1987; HEY and WAKELEY 1997) and one determined from experimental laboratory crosses (COMERON *et al.* 1999). The program of COMERON *et al.* (1999) (kindly provided by J. Comeron) estimates recombination rates in *D. melanogaster* based on cytological map position using polynomial curves (KLIMAN and HEY 1993) as a function of the amount of DNA in each chromosomal division *vs.* the change in cytological map position (SORSA 1988). For the coalescent simulations, 10,000 replicates were performed for each test.

Inversion analysis: Larvae were grown in yeast-rich corn meal-sugar media at 18°C without larval crowding. Late third instar larvae were dissected in a drop of insect Ringer's solution. Both salivary glands were placed under a cover slip with a drop of lacto-aceto-orcein stain. After five minutes the glands were squashed and polytene chromosomes were analyzed under a phase contrast microscope at 100X. Inversion break-points were determined according to the photographic map of LEFEVRE (1976).

RESULTS

A total of 25 *D. melanogaster bcd* alleles derived from isofemale lines from Lake Kariba, Zimbabwe and one *D. simulans* allele were sequenced for a 4019-bp region (coordinates 971 – 4990, according to GenBank accession X07870; position 3599 of the reference sequence is deleted in our 25 lines), spanning the 3' portion of the 5' UTR and nearly the entire transcript length. A total of 40 nucleotide and five length polymorphisms were detected in the 13 *D. melanogaster* haplotypes found in this sample (FIGURE 2.1). Of the 40 nucleotide polymorphisms, seven occurred in the coding region and 33 in non-coding regions (5' flanking region, intron 1, intron 3, 3' UTR and 3' flanking region). Interestingly, out of the seven polymorphic sites in the coding region, one was synonymous compared to six nonsynonymous polymorphisms. A large length polymorphism (a 7-bp insertion (AGGGAAG) followed by a perfect 138-bp duplication of positions 3427 to 3564 at position 3565) in intron 3 was found in four alleles. A second complex change in intron 1 is only found in one allele and involves multiple indels: ATATGAATTGTGGGGCAAC in the reference sequence at positions 1828 to 1846 are replaced by TACGTTATTGTTATAATTGTTAA in line 377. Interspecific comparison with *D. simulans* revealed 32 synonymous and six nonsynonymous differences in the coding

region, with divergence at silent and non-coding sites falling within the range of several previously surveyed genes (EANES *et al*. 1993).

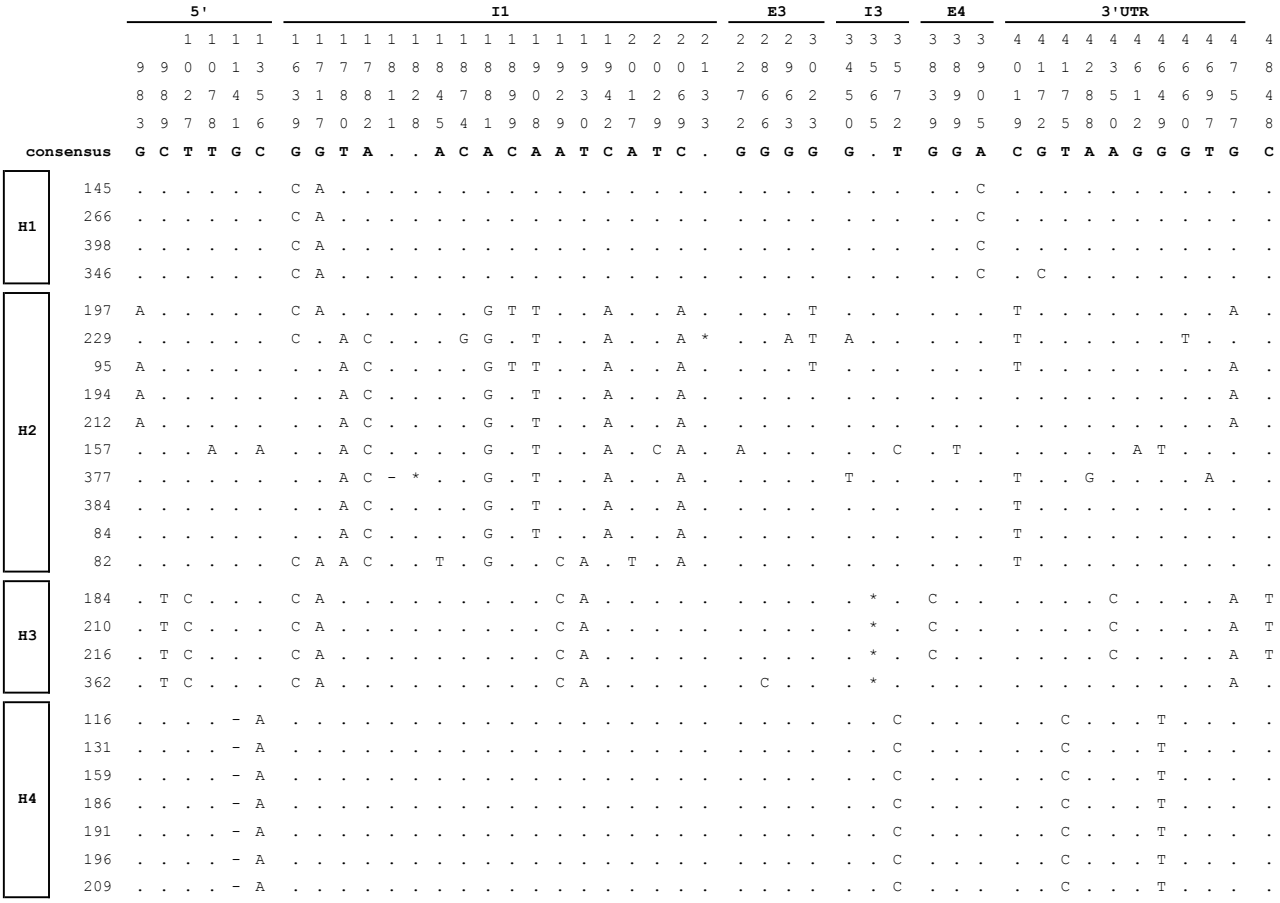

FIGURE 2.1 DNA polymorphisms of the *bcd* gene region found in 25 lines of *D. melanogaster*. The line number of each strain is indicated on the far left side, and the haplotype classes they belong to are shown as H1 to H4. The nucleotides of the consensus sequence are shown along the top. The coordinates above the sequence represent the positions of each segregating site according to the reference sequence (GenBank accession X07870). The five length polymorphisms are noted as follows: a: deletion of G at position 1141; b: deletion of GTA at positions 1811-1813; c: complex changes (multiple indels) at positions 1828-1846; d: T_4 in place of T_3 at positions 2133-2135; e: a 7-nt insertion (AGGGAAG) followed by a perfect 138-nt duplication from 3427 to 3564 at position 3565.

Silent polymorphism and divergence: Estimates of π (NEI 1987) and θ (WATTERSON 1975) are based on the number of equivalent silent sites according to ROZAS and ROZAS (1999) (Table 2.1). The nucleotide diversity for non-coding regions (θ = 0.0035, π = 0.0038) is lower than average values for *D. melanogaster* (~0.011, MORIYAMA and POWELL 1996), probably due to the lower than average recombination rate of the *bcd* region. In particular, the level of polymorphism at synonymous sites in the coding region (θ = 0.0008, π = 0.0002) was substantially lower than the average values of θ = 0.014 and π = 0.013 (MORIYAMA and POWELL 1996). On the other hand, the estimates of nucleotide diversity are two- to threefold higher than the values found in regions of restricted recombination rates at the tip and base of the X chromosome in a Zimbabwe population (BEGUN and AQUADRO 1995a). Divergence estimates for both introns (κ = 0.102) and the coding region (κ = 0.094) are typical for *D. melanogaster* and *D. simulans* (MORIYAMA and POWELL 1996) while the 3' UTR, which contains an important secondary structural element involved in mRNA localization (MACDONALD 1990), was more conserved (κ = 0.040).

TABLE 2.1

Polymorphism and divergence in the *bcd* gene

	5' UTR	Intron 1	Intron 3	3' UTR	Coding region	Total
Silent sites	444	467	364	795	341.85	2649.85
Segregating sites	5 (1)[a]	13 (4)	3 (2)	10 (5)	1 (1)	33 (13)
Diversity θ	0.0030	0.0074	0.0022	0.0033	0.0008	0.0033
Diversity p	0.0031	0.0096	0.0017	0.0029	0.0002	0.0034
Divergence k	0.1341	0.1014	0.1028	0.0396	0.0937	0.0815
p / k	0.0231	0.0949	0.0164	0.0742	0.0025	0.0422

Note: [a] Singletons are given in parentheses.

Two tests of neutrality were performed on the silent polymorphism of this data set. First, to test whether the frequency spectrum of silent polymorphism significantly deviates from the neutral expectation, TAJIMA's (1989) test was applied. The value of the *D* statistic (*D* = -0.073) did not significantly deviate from zero, thus the null hypothesis that the silent polymorphisms are selectively neutral cannot be rejected.

A second test of neutrality, the HKA test (HUDSON *et al*. 1987), examines the prediction of the neutral mutation hypothesis that levels of intraspecific polymorphism are positively correlated with levels of interspecific divergence. We tested six possible pair-wise comparisons using 5' UTR, intron 1, intron 3 and the 3' UTR. Though none of these tests were significant, comparisons between the 5' and 3' UTR (*P*= 0.065) and between intron 3 and the 3' UTR (*P* = 0.053) approach significance (Table 2.2). Comparisons between both the 5' UTR and intron 3 with intron 1 also give small *P* values of around 0.1. These may suggest lower levels of polymorphism in 5' UTR and intron 3 when compared to the two other reference non-coding loci, and heterogeneity in the ratio of polymorphism to divergence along the *bcd* DNA sequence. However, the HKA test was applied somewhat *post hoc*, so the results should be interpreted with caution.

Replacement polymorphism and divergence: A surprising observation was that six out of seven polymorphic sites in the coding region are replacement polymorphisms. Under the hypothesis of neutral protein evolution, the ratio of replacement to synonymous fixed differences between species should be the same as the ratio of replacement to synonymous polymorphisms within species (MCDONALD and KREITMAN 1991). The comparison between *D. melanogaster* and *D. simulans* revealed 32 synonymous and six replacement fixed differences. However, within our surveyed population only one out of seven polymorphic sites was a synonymous

polymorphism, the rest being nonsynonymous polymorphisms (Table 2.3). Using the statistical test developed by MCDONALD and KREITMAN (1991), we found a highly significant difference between these two ratios ($P = 0.0007$ by Fisher's exact test). This is not expected under the standard neutral model and may be interpreted as either an excess of replacement or a deficiency of synonymous polymorphism within the *D. melanogaster* population. To distinguish between these two possibilities, we used the method of TAVARÉ (1984) and HUDSON (1990). We showed that our observation of one synonymous polymorphism is not inconsistent with a value of $\theta = 0.0033$ for the coding region ($P = 0.115$). This suggests that if our estimate of $\theta = 0.0033$ obtained for the entire *bcd* region is representative for the coding region as well, the significant result of the MCDONALD-KREITMAN test is mainly due to an excess of replacement polymorphism.

TABLE 2.2

Pair-wise HKA tests between various regions of *bcd*

Regions	***S***	***D***	X^2	***P* value**
5' UTR	5	59.52		
Intron 1	13	47.36	2.672	0.102
5' UTR	5	59.52		
Intron 3	2	37.40	0.209	0.648
5' UTR	5	59.52		
3' UTR	10	31.48	3.399	**0.065**
Intron 1	13	47.36		
Intron 3	2	37.40	3.098	**0.078**
Intron 1	13	47.36		
3' UTR	10	31.48	0.049	0.824
Intron 3	2	37.40		
3' UTR	10	31.48	3.747	**0.053**

Note: *S*: observed number of silent polymorphisms in *D. melanogaster*;
D: observed number of fixed differences between *D. melanogaster* and *D. simulans*;
X^2: test statistic;
Marginally significant *P* values are highlighted.

TABLE 2.3

Result of the McDonald-Kreitman test

	Fixed sites	Polymorphic sites
Synonymous sites	32	1
Replacement sites	6	6

Note: $G_W = 12.09$; $P = 0.0005$;
G_W indicates G value with Williams' correction.

To further investigate the forces determining the population dynamics of these polymorphisms, we analyzed their frequency spectrum. All six replacement polymorphisms are at low frequency (≤ 4/25), three of which are singletons. Though TAJIMA's (1989) test applied specifically to these six replacement polymorphisms did not significantly deviate from zero (D = -1.184), its negative value is indicative of an excess of low-frequency variants, consistent with the hypothesis that they are slightly deleterious.

Another noteworthy observation is the distribution of both replacement fixed differences and polymorphisms along the Bicoid protein (FIGURE 2.2). All six replacement polymorphisms and five of six replacement fixed differences cluster within one of two regions of the protein. The first region (amino acids 249-332) contains a cluster of three polymorphisms and four fixed differences and overlaps *opa*-like repeats and portions of the protein with no known function (*i.e.* linker or hinge regions) (SEEGER and KAUFMAN 1990). The second region (amino acids 433-455) contains a cluster of three polymorphisms and one fixed difference and overlaps a region which some suggest contains an RNA recognition motif, though this has never been experimentally proven (SEEGER and KAUFMAN 1990). In an interspecific

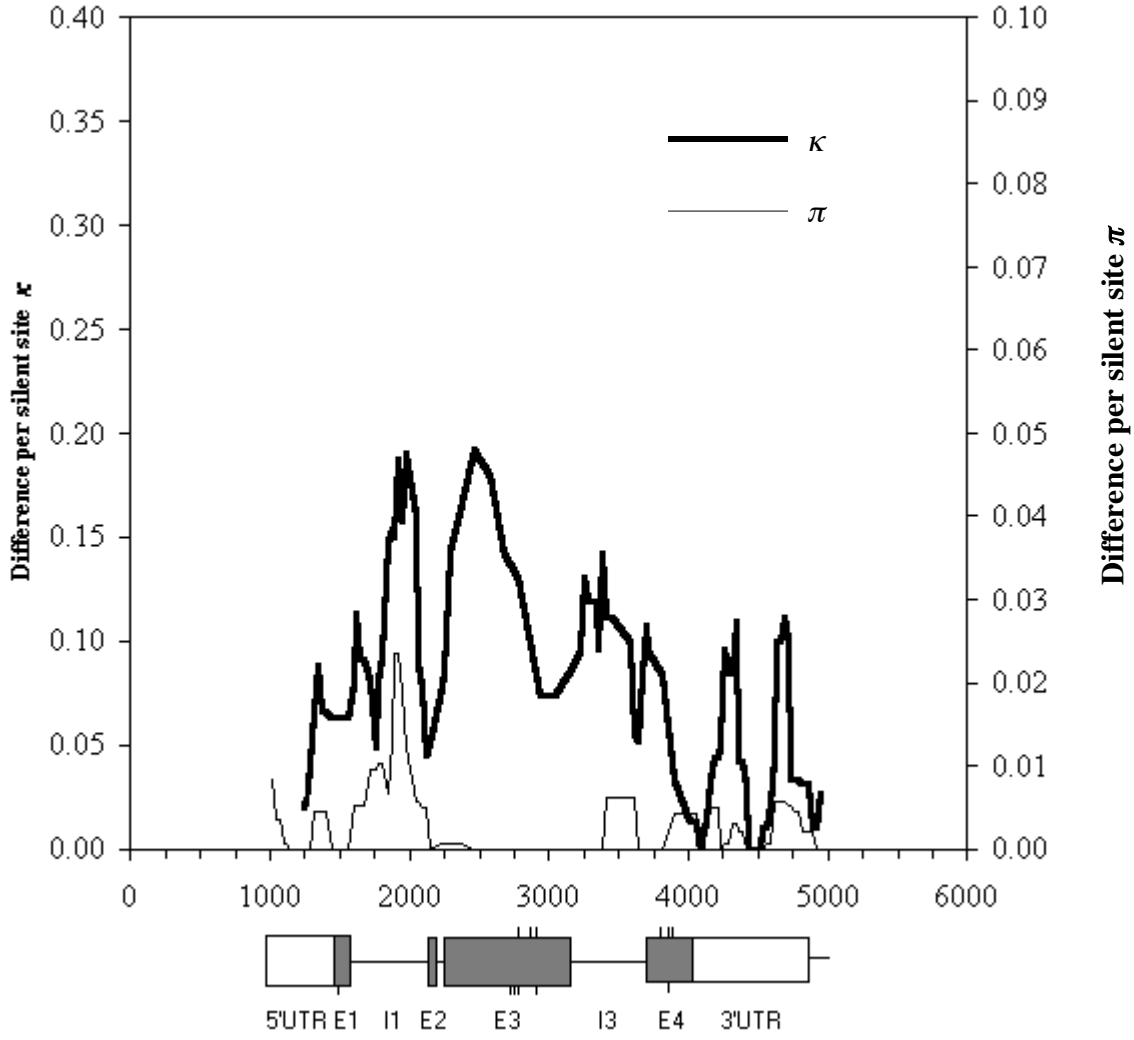

FIGURE 2.2 Sliding window plot of silent nucleotide diversity (p) within *D. melanogaster* and silent nucleotide divergence (k) between *D. melanogaster* and *D. simulans*. Note that different scales for *p* and k are used. The size of the sliding window is 100 silent sites, and step size is 25 sites. The structure of the *bcd* gene is represented below the plot. Exons are shown as bars, and translated regions are shaded. The vertical lines above the gene structure indicate the positions of amino acid replacement polymorphisms within *D. melanogaster*; the vertical lines below the gene structure indicate the positions of fixed amino acid differences between species. The nucleotide numbering is according to the reference sequence (GenBank accession X07870).

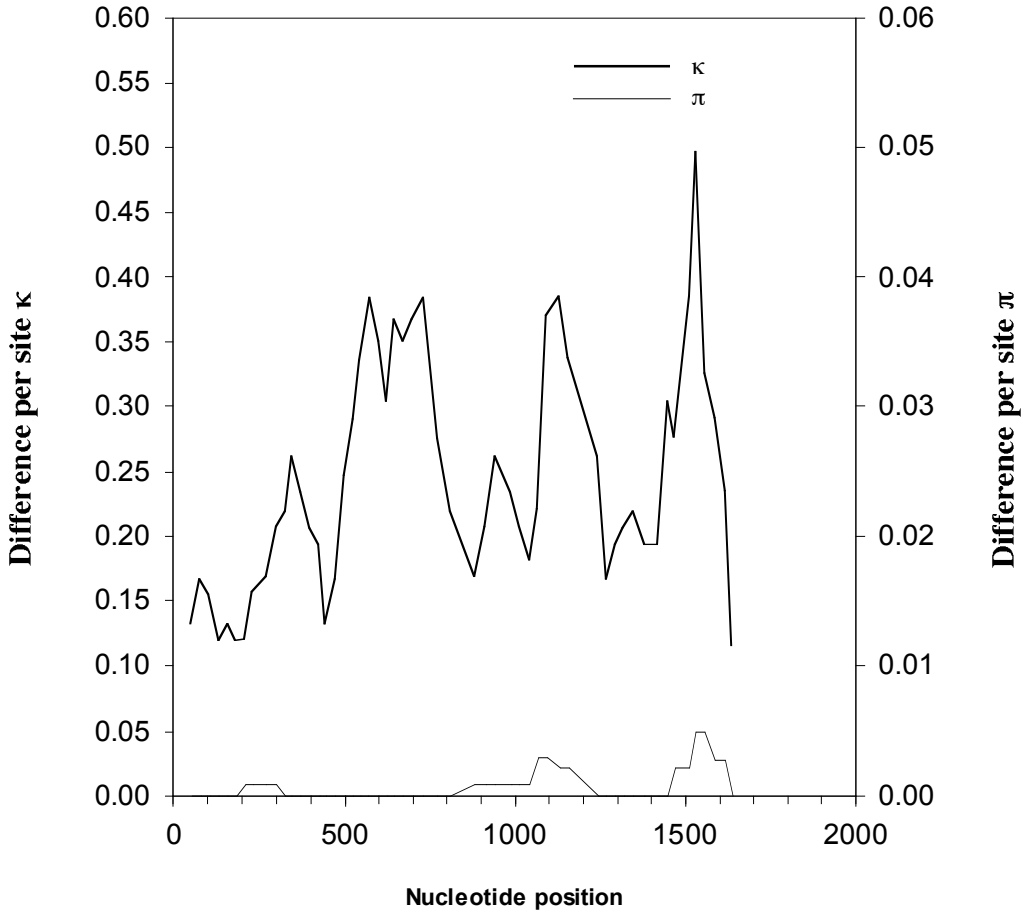

FIGURE 2.3 Sliding window plot of the divergence of replacement sites between the *D. melanogaster* and *D. pseudoobscura* coding regions. Note that only the coding portion is shown because of difficulties with the alignment of the non-coding regions. The coordinates at the bottom refer to nucleotide positions in the coding region (not to the reference sequence, GenBank accession X07870). The size of the sliding window is 100 nt, and step size is 25 nt.

comparison of the *D. melanogaster* sequence with *D. pseudoobscura*, a sliding window analysis of divergence at replacement sites revealed a peak of divergence in this location (FIGURE 2.3), and the ratio of replacement to silent substitutions, κ_a/κ_s, was nearly 4 times greater ($\kappa_a/\kappa_s = 0.554$) than the value for the entire coding region ($\kappa_a/\kappa_s = 0.139$).

Haplotype structure: On observation of the 13 haplotypes in our data set, the pattern of linkage disequilibrium easily lends itself to a classification of haplotypes, which we designate H1 through H4 (FIGURE 2.1). Classes H1, H3 and H4 are monomorphic or show few polymorphisms within class, but are distinct from one another at at least six sites. The remaining haplotypes comprise the H2 class. Linkage disequilibrium is extensive and includes nearly the entire 4019-bp region. Out of a total of 300 pair-wise comparisons made between 25 polymorphic sites (only informative sites are considered), 102 are significant by Fisher's exact test, 24 of which are significant after a Bonferroni correction (FIGURE 2.4), while 119 are significant by the χ^2 test, with 40 significant after a Bonferroni correction. Some of the most highly significant ($P < 0.001$ after Bonferroni correction) comparisons are between sites 3 kb apart. Though the average correlation over all pair-wise comparisons, ZnS (KELLY 1997), did not significantly deviate from the neutral expectation, linkage disequilibrium does not decay with distance, as expected under neutrality.

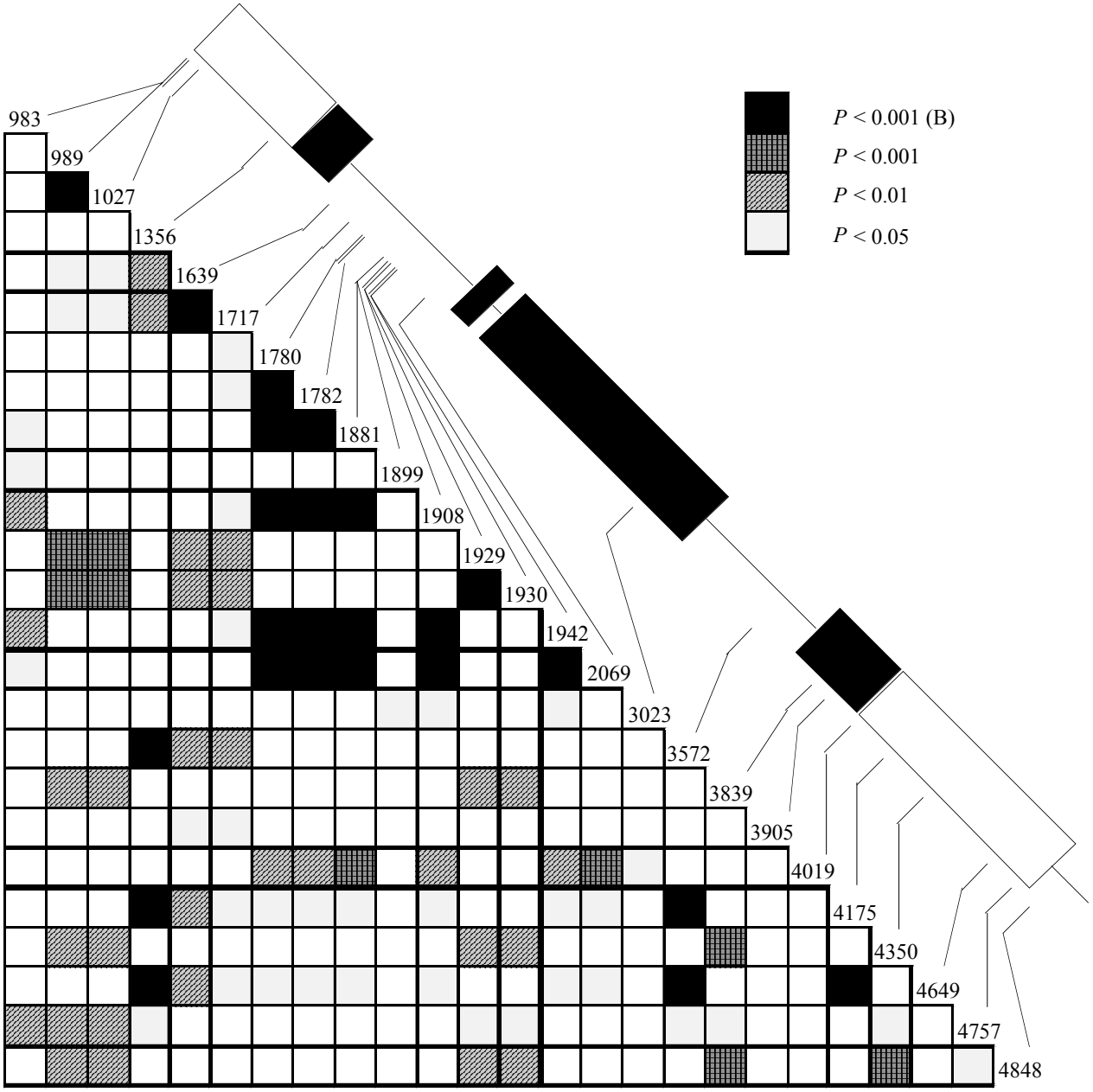

FIGURE 2.4 Linkage disequilibria between polymorphic sites in the *bcd* gene from 25 *D. melanogaster* lines. Only the informative sites were used in Fisher's exact test. Squared, lined and dotted boxes indicate the significance levels at 0.001, 0.01 and 0.05, respectively. Black boxes indicate significant linkage disequilibria after Bonferroni correction. The comparisons indicated by the white boxes are not significant.

We applied several statistical tests to determine whether the apparent structuring of haplotypes deviated from the neutral expectation. The recombination parameter, *R*, was estimated by three methods (see MATERIALS and METHODS). The method of HUDSON (1987) yielded an estimate of $R = 25$, the method of HEY and WAKELEY (1997) 43, and that of COMERON *et al*. (1999) 39. Using the more conservative estimate of $R = 25$, we first performed coalescent simulations using the computer program DnaSP (ROZAS and ROZAS 1999). We found a significant reduction in both the number of haplotypes ($P = 0.032$) and haplotype diversity ($P = 0.021$) (DEPAULIS and VEUILLE 1998) when compared to the results of the coalescent simulations. Significantly low values indicate a structuring of polymorphic sites into a small number of haplotypes (DEPAULIS and VEUILLE 1998). Second, the haplotype test of HUDSON *et al*. (1994) indicates a significantly lower number of segregating sites within haplotype 13 (H4 class) than expected under a neutral equilibrium model ($P = 0.039$).

Given the pattern of linkage disequilibrium and significant structuring of haplotypes, we further investigated our data in the context of two possible explanations: associations with polymorphic inversions and/or RNA secondary structure. First, since *D. melanogaster* is known to be polymorphic for well over 300 inversions (DAS and SINGH 1991; LEMEUNIER and AULARD 1992), four of which are cosmopolitan and reach appreciable frequencies, we investigated whether certain haplotype classes (*i.e*. H1 through H4) may be associated with chromosomal inversion types. Though the cosmopolitan inversion *In(3R)P* (breakpoints 89C to 96A) was found segregating at approximately 14%, there is no association between this inversion and haplotype classes, and no inversions associated with the cytological interval containing *bcd* (84A5) were detected.

Second, we analyzed linkage disequilibrium with respect to the large, conserved secondary structural element in the 3' UTR (MACDONALD 1990). This structure plays an essential role in the localization of *bcd* mRNA and has been well characterized by both mutational (FERRANDON *et al.* 1997; MACDONALD and KERR 1998) and phylogenetic analysis (PARSCH *et al.* 2000). However, there seems to be no obvious association between the observed haplotype structure and the predicted mRNA secondary structure in the 3' UTR (further discussed below).

DISCUSSION

Main observations: We surveyed nucleotide variation of the *bcd* transcriptional unit (from the 3' portion of the 5' UTR to the end of the 3' UTR) of a population of *D. melanogaster* from Zimbabwe. Our survey produced several salient features that will each be discussed in turn. The first and most striking observation is the significant excess of replacement polymorphism found by the MCDONALD-KREITMAN test. Second, extensive linkage disequilibria across the region and a structuring of polymorphic sites into haplotype classes are observed. Third and less significant, evidence for heterogeneity in the ratio of polymorphism to divergence (π / κ) along the gene region is detected. In the following, we will discuss the evolutionary forces that might underlie these patterns of variation.

Evidence for relaxed purifying selection: A significant excess of replacement polymorphisms is usually interpreted as a relaxation of selective constraints. The effect of purifying selection may be weaker on some amino acid replacement mutations than others, and slightly deleterious polymorphisms may persist at low frequencies within a population for a period of time due to genetic drift, but are unlikely to either rise in frequency or become fixed (KIMURA 1983; OHTA 1992).

Under this scenario slightly deleterious mutations contribute more to intraspecific polymorphism than to interspecific fixed differences (KIMURA 1983; OHTA 1992). Several other studies have reported similar cases of an excess of intraspecific replacement polymorphism in mitochondrial DNA (BALLARD and KREITMAN 1994; NACHMAN *et al*. 1994, 1996; RAND and KANN 1996; WISE *et al*. 1998) and in one case a nuclear gene, *Pgm* (VERRELLI and EANES 2000, 2001), and interpretation has ranged from that of slightly deleterious mutations to positive selection.

TABLE 2.4

Polymorphic replacement changes in the *bcd* gene

Position	Nucleotide polymorphism	Amino acid replacement	Amino acid property change
2866	CA**G** → CA**C**	Gln → His	neutral to basic
2963	**G**AG → **A**AG	Glu → Lys	acidic to basic
3023	**G**CA → **U**CA	Ala → Ser	hydrophobic to hydrophilic
3839	**G**CC → **C**CC	Ala → Pro	no change
3899	**G**UG → **U**UG	Val → Leu	no change
3905	**A**UG → **C**UG	Met → Leu	no change

Note: The polymorphic nucleotide sites are highlighted;
the coordinates in the left column represent the positions of each replacement change according to the reference sequence (GenBank accession X07870).

At least some of our observed amino acid polymorphisms at the *bcd* locus appear to be slightly deleterious. All of the polymorphisms are at relatively low frequency (4-16%), three of which are singletons. Three of these polymorphisms also cause drastic changes in amino acid property (Table 2.4). It is usually hard to imagine that replacement polymorphisms, especially those that drastically change encoded amino acid properties, are only slightly deleterious. Therefore, we investigated the

protein regions where the replacement polymorphisms are found. The three found in exon 3 are located within an *opa*-like repeat region and regions with no known function (possibly linker or hinge regions in the polypeptide chain) (SEEGER and KAUFMAN 1990). So despite the changes in amino acid property, their effect could be minimal due to the functional insignificance of their location. The other three replacement polymorphisms are located in exon 4, within a region containing a putative, but ill-characterized RNA recognition motif (SEEGER and KAUFMAN 1990). Sliding window analysis of divergence between *D. melanogaster* and *D. pseudoobscura* shows that this region overlaps with a peak in both divergence and the replacement to silent substitution ratio ($\kappa_a / \kappa_s = 0.554$) (FIGURE 2.3). κ_a / κ_s ratios are usually kept low by purifying selection ($\kappa_a / \kappa_s = 0.139$ for the entire coding region). The rise in κ_a / κ_s ratio may suggest that this region is under less constraint than other parts of the molecule, and the polymorphisms (sites 3839, 3899 and 3905) found here are due to a relaxation of purifying selection.

It is possible that selection on parts of *bcd* is relaxed because this gene was apparently created by a relatively recent gene duplication event. *bcd* homologs have been identified only in *Drosophila* and some of its close relatives (higher Dipterans). It was originally thought that "rapid" evolution of the homeoprotein led to a subsequent deterioration of homology. However, *caudal* genes, as well as some other homeobox genes of the *Drosophila* segmentation gene cascade have been found to be well conserved in evolution (PATEL 2000). This poses the question of how *bcd* has evolved and adopted its function in anterior patterning. Recent studies suggest that *bcd* may have originated from a duplication of *zen*, which is downstream of *bcd* in the *hox* gene cluster (STAUBER *et al*. 1999; DEARDEN and AKAM 1999; PATEL 2000).

If *bcd* arose through tandem duplication from *zen*, then it would initially have a function very similar to that of *zen*. This would allow deleterious mutations to rise in frequency without being eliminated by purifying selection. At the same time, positively selected mutations may also arise, go to fixation and thus give *bcd* new functions. But the functions of *bcd* and *zen* may still be similar enough that adaptive sweeps and relaxed selection in localized regions of the gene occur simultaneously. In the following, we discuss the evidence that, in addition to the relaxed purifying selection, positive selection has also occurred at *bcd*. This evidence is for the most part based on the other two observations, namely the extensive haplotype structure and the apparent heterogeneity in the polymorphism-to-divergence ratio along the gene.

Evidence for positive selection: First, we observed extensive linkage disequilibria and distinct haplotype structure in our sample of *bcd* alleles, and positive natural selection is likely involved in maintaining this structure. We found a significantly smaller number of haplotypes than the neutral expectation ($P = 0.032$) by the DEPAULIS and VEUILLE (1998) test of haplotypes. Coalescent simulations with the most conservative value of the recombination parameter, R, gave the range for the number of haplotypes as [13, 22], and we observed 13 haplotypes in our sample. It appears that polymorphic sites structure into a few haplotypes. Since the Zimbabwe population typically exhibits less linkage disequilibria than non-African populations and is thought to be a panmictic population close to mutation-drift equilibrium (BEGUN and AQUADRO 1993), demographic and bottleneck effects should be minimal. Thus, it seems likely that balancing selection or partial selective sweeps of haplotypes are contributing to this pattern. In the latter model, variants are positively selected for,

but fail to go to fixation because of "traffic" with haplotypes where selection is acting on other sites (KIRBY and STEPHAN 1996).

In particular, the H4 haplotype appears to be the target of positive selection operating at or near the *bcd* locus. This haplotype has a frequency of 28% and has no within-class variation. By applying a statistical test for high-frequency haplotypes (HUDSON *et al*. 1994), our results show that there is too little variation within the H4 class, given the level of variation in the rest of the sample, and cannot be explained by a neutral equilibrium model of mutation and drift. This pattern suggests that this haplotype has arisen recently, and is being pulled to high frequency due to directional selection at or near the region we surveyed.

A second phenomenon possibly attributable to positive selection is the heterogeneity in the ratio of polymorphism to divergence along the *bcd* region. The π / κ ratio is high in the intron 1 region (0.0949), while low in the coding region (0.0025) (Table 2.1), which is suggestive of balancing selection on sites within intron 1 and/or a selective sweep that occurred in the region of exon 2 to exon 4. Comparisons between gene regions involving intron 1 by the HKA test approach significance (Table 2.2). In addition, the run's test (MCDONALD 1996, 1998), a measure of heterogeneity, produces a marginally significant result ($P \approx 0.05$). However, the results of these tests alone do not put forth convincing evidence for such underlying selective mechanisms. It is difficult to distinguish this pattern of heterogeneity from merely neutral fluctuation of polymorphism along the region (KIM and STEPHAN 2002).

Haplotype structure and mRNA secondary structure in the *bcd* 3' UTR: The *cis*-acting sequences necessary for the localization of *bcd* mRNA fall within a large, phylogenetically conserved and well characterized secondary structural element in the

3' UTR (MACDONALD and STRUHL 1988; MACDONALD 1990; SEEGER and KAUFMAN 1990; FERRANDON *et al*. 1997; MACDONALD and KERR 1998; PARSCH *et al*. 2000). We investigated whether the observed linkage disequilibrium and haplotype structure is related to the maintenance of the mRNA secondary structure in the 3' UTR.

Of the 40 nucleotide polymorphisms observed within the entire *bcd* gene region, nine fall within the region of the localization signal in the 3' UTR, and only two (A4350C and G4612A) are located within the pairing regions supported by both phylogenetic study (PARSCH *et al*. 2000) and mutational analysis (FERRANDON *et al*. 1997; MACDONALD and KERR 1998). Both of these substitutions cause mismatches in the original Watson-Crick pair of the mRNA secondary structure. No covariations (compensatory mutations) with the pairing regions were observed in our sample of the Zimbabwe population. This is in qualitative agreement with the theoretical results developed under a two-locus, two-allele, reversible mutation compensatory model (INNAN and STEPHAN 2001). According to the predictions of this model, the populations spend most of the time in the first stage of waiting for a successful double mutant to appear in the population. The second stage of fixing the successful double mutant in the population is much shorter than the first stage (INNAN and STEPHAN 2001). Their simulations showed that almost no linkage disequilibrium due to compensatory interactions is expected during the first stage, so it is unlikely to observe much linkage disequilibrium or covariations caused by epistatic selection on mRNA secondary structures. Thus, the strong haplotype pattern and linkage disequilibria of our dataset cannot be explained by epistatic selection on the *bcd* 3' UTR.

CHAPTER 3

Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the *Drosophila Adh* gene

INTRODUCTION

KIMURA's (1985) classical model of compensatory evolution and its application to nucleotide sites involved in Watson-Crick (WC) base pairing within RNA secondary structures (STEPHAN 1996; INNAN and STEPHAN 2001) typically assume a symmetrical interaction between alleles at two interacting loci. Under this scenario, with alleles *A* and *a* present at the first locus and *B* and *b* at the second, the genotypes *Ab* and *aB* are both considered deleterious, while the wild-type *AB* and double mutant *ab* are selectively neutral. In terms of RNA secondary structure, genotypes *ab* and *AB* would represent WC paired nucleotides, while *Ab* and *aB* would represent mismatches. Because of its simplicity, this model is generally used to predict the outcome of experimental tests of secondary structure (*e.g.*, HAAS *et al.* 1991; CHEN and STEPHAN 2003). However, it is becoming increasingly clear that in many cases (*e.g.*, when helices do not meet the condition such that single base changes destabilize, but do not destroy the overall structure) the classical model may be too simple (CHEN *et al.* 1999). For example, the model does not adequately explain the data of SCHAEFFER and MILLER (1993), where the compensatory process between two divergent haplotypes in the introns of the *D. pseudoobscura* alcohol dehydrogenase gene (*Adh*) likely involved significant rearrangements (insertions and deletions of bases) (INNAN and STEPHAN 2001).

Experimental evidence also suggests that the compensatory process is more complex than previously modeled. PARSCH *et al.* (1997) extended the phylogenetic analysis of STEPHAN and KIRBY (1993) and predicted a long-range, tertiary contact between a region just downstream of the start codon and a conserved region of the 3' untranslated region (UTR) in *Drosophila Adh* (FIGURE 3.1). Site-directed mutagenesis was used to test a long-range WC base pairing between positions 819 and 1756. A synonymous mutation in exon 2 (C to T at position 819, designated *mutC819T*) resulted in a significant reduction in ADH activity, while a second, compensatory mutation in the 3' UTR (G to A at position 1756, designated *mutG1756A*) restored activity to that of wild-type *Adh* levels. However, *mutG1756A* alone did not significantly differ from that of wild-type levels, thus not fitting classical models of compensatory evolution where *both* intermediate states should be deleterious (KIMURA 1985; STEPHAN 1996; INNAN and STEPHAN 2001).

It is well known that communication between the 5' and 3' ends of mRNA plays an important role in the initiation of translation in eukaryotes (GALLIE 1991; SACHS *et al.* 1997). These interactions have been well documented at the protein-protein level (TARUN and SACHS 1996; HENTZE 1997; WELLS *et al.* 1998), and it has been demonstrated that a viral transcript which lacks a 5' cap and poly(A) tail achieves 5'-3' communication and initiation of translation in a similar manner (WANG *et al.* 1997; GUO *et al.* 2000). In this case, the 5'-3' interaction occurs via direct RNA-RNA base pairing between the 5' and 3' UTRs (GUO *et al.* 2001). Such long-range RNA-RNA interactions have been predicted for a large number of eukaryotic transcripts (KONINGS *et al.* 1987; STEPHAN and KIRBY 1993; PARSCH *et al.* 1998), but the results of PARSCH *et al.* (1997) offer the first experimental evidence of such interactions. Thus, it is important to further investigate and verify long-range pairing

in *Drosophila Adh* with respect to models of compensatory evolution and its potential functional role.

In the present study, we experimentally investigate two alternative explanations for the results of PARSCH *et al.* (1997), namely, the respective local structure/function of the nucleotides neighboring positions 819 and 1756. In the first experiment, the importance of the local secondary structure in exon 2 is investigated. Free energy minimization analysis (ZUKER 2003) indicates that the phylogenetically predicted RNA secondary structure of exon 2 may be displaced by a more thermodynamically stable alternate structure when position 819 is changed from C to T (PARSCH *et al.* 1997). One particular pairing region with phylogenetic support that is not present in this alternate structure is that comprising the central stem (positions 793-797/833-837). We directly test the importance of this structure by introducing two synonymous mutations that disrupt the central stem (FIGURE 3.1). Thus, if the effect seen by C819T is due to the disruption of the native exon 2 structure, these mutations should produce a similar effect.

In a second set of experiments, a conserved region of the *Adh* 3' UTR encompassing position 1756 is investigated. Previous studies identified a highly conserved 8-base regulatory sequence at positions 1762-1769 (PARSCH *et al.* 1999, 2000). Deletion of this sequence resulted in a two-fold increase in ADH activity due to an underlying two-fold increase in mRNA levels, suggesting a functional role in the negative regulation of mRNA (PARSCH *et al.* 1999, 2000). It is hypothesized that the conserved portion of the 3' UTR upstream of this 8-base sequence may play a dual role in both long-range pairing and the negative regulation of *Adh* expression, so that any reduction in ADH activity due to a disruption of long-range pairing may be masked by a decrease in the negative regulation of mRNA (CHEN *et al.* 1999). To

investigate whether the proximity of position 1756 to the 8-bp sequence confounds the ability to measure its involvement in long-range pairing with exon 2, a series of compensatory mutations is made in a background of a deletion of positions 1762-1769. The results indicate that the local structure predicted in exon 2 has no significant effect on *Adh* expression, whereas the conserved region of the 3' UTR likely plays a role in both long-range base pairing and the negative regulation of *Adh* mRNA levels.

MATERIALS AND METHODS

Site-directed mutagenesis and plasmid construction: All constructs were derived from an 8.6-kb *Sac*I-*Cla*I fragment of the *D. melanogaster Adh Wa-f* allele (KREITMAN 1983). A pUC18 plasmid containing the 8.6-kb fragment was subjected to mutagenesis using the Quick-change mutagenesis kit (Stratagene, La Jolla, CA). Two point mutations (T to C at positions 834 and 837) were introduced simultaneously into this wild-type background to create the construct designated as *E2mut*. In order to test the long-range pairing between sites 819 and 1756, individual mutations (C to T at position 819 and G to A at position 1756) were introduced into a *Wa-F* allele with bases 1762-1769 deleted (*Δ3*; PARSCH *et al.* 1999). These constructs were designated *Δ3mut2* and *Δ3mut1*, respectively. A final construct (*Δ3mut3*) contained both of the above mutations together in the *Δ3* background. Desired mutations were verified by sequencing before proceeding further.

***P*-element mediated germline transformation:** For each mutant construct, the respective mutant *Sac*I-*Cla*I fragment was inserted into the polycloning region of the YES transformation vector (PATTON *et al.* 1992). This is a *P*-element vector containing the *D. melanogaster yellow* (*y*) gene as a selectable marker and *suppressor*

of Hairy-wing binding sites flanking the target DNA, which serve as an insulator of chromosomal position effects (PATTON *et al.* 1992). Constructs were introduced into an ADH-null background by microinjection of *y w*; Adh^{fn6}; Δ2-3, *Sb*/TM6 embryos (RUBIN and SPRADLING 1982; SPRADLING and RUBIN 1982). For each mutant construct, five to six independent transformed lines were generated by microinjection. To increase the number of lines, insertions on the *X* chromosome were mobilized to new genomic locations by crosses utilizing the Δ2-3 *P*-element as a source of transposase (ROBERTSON *et al.* 1988; PARSCH *et al.* 1997). Insertions on the third chromosome containing the source of transposase (the Δ2-3, *Sb* third chromosome) are not suitable for maintaining as transformed stocks and were thus also mobilized to new genomic locations. All lines were crossed to a *y w*; Adh^{fn6} stock following transformation or mobilization to remove the source of transposase. Lines containing single insertions were determined by Southern blotting (PARSCH *et al.* 1997). Due to possible dosage compensation effects, only lines containing autosomal insertions were used for further analysis (LAURIE-AHLBERG and STAM 1987; PARSCH *et al.* 1997).

ADH assays: ADH enzymatic activity was measured by the method of MARONI (1978), using 2-propanol as the substrate. Assays were performed on five 6- to 8-day old males heterozygous for the respective *Adh* insertion following the procedure of PARSCH *et al.* (1997). The total protein of the extracts was determined by the method of LOWRY *et al.* (1951), and units of activity are represented as micromoles of NAD reduced per minute per milligram of total protein. Differences in activity between *Adh* genotypes were tested by analysis of variance (ANOVA), using a model that accounts for position-effect variation (LAURIE-AHLBERG and STAM 1987).

RESULTS

Analysis of the local secondary structure of exon 2: In our first experiment, we investigate the mechanism by which *mutC819T* causes a significant reduction in ADH activity. Although this mutation changes a preferred alanine codon to an unpreferred codon, it is unlikely that a single synonymous codon replacement could lead to such a large difference in gene expression (CARLINI and STEPHAN 2003). Thus, given that involvement in secondary structure is a more likely explanation, the next question is whether the effect of *mutC819T* is due to the disruption of the local structure of exon 2, or involvement in long-range pairing with the 3' UTR.

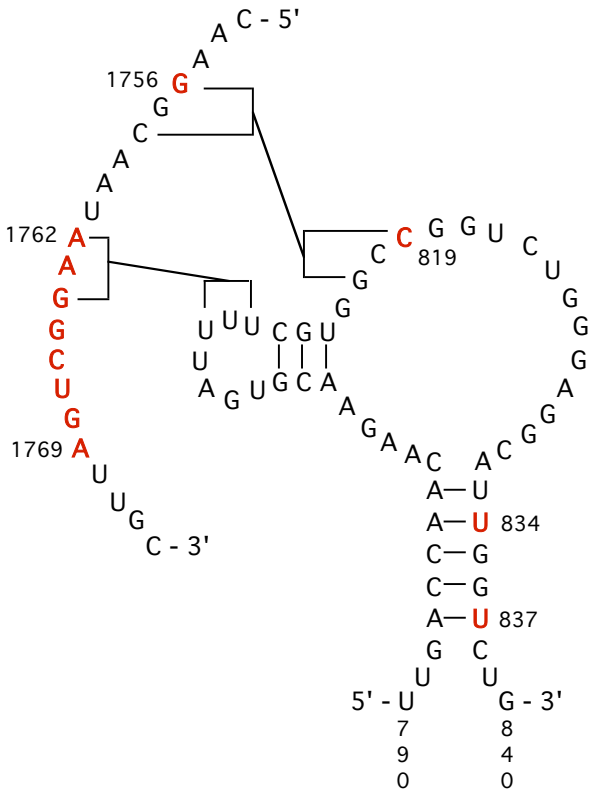

FIGURE 3.1 Phylogenetically predicted secondary structure of wild-type exon 2 and its putative tertiary contacts with the 3' UTR (modified from PARSCH *et al.* (1997)). All positions used in mutational analysis are labeled and shown in red. Positions 819, 1756 and 1762-1769 were investigated in previous studies (PARSCH *et al.* 1997, 1999, 2000). Brackets connected by lines indicate predicted pairing regions.

To test the functional significance of the local structure of exon 2, site-directed mutations were made at degenerate codon positions 834 and 837, thus disrupting the central, phylogenetically predicted pairing region 793-797/833-837 (FIGURE 3.1). This genotype, designated *E2mut* (T834C-T837C) (FIGURE 3.2), enables an indirect test of the mechanism by which *mutC819T* reduces ADH activity. Our logic is as follows. Given that *mutC819T* creates the potential for an alternative structure in exon 2 and results in a significant reduction in ADH activity (PARSCH *et al.* 1997), loss of the hairpin structure depicted in FIGURE 3.1 may be the underlying reason for this reduction. Under this scenario, targeted disruption of this structure (see sites 834 and 837 in FIGURE 3.1) should result in a similar or more extreme phenotype (*i.e.*, a *decrease* in *Adh* expression).

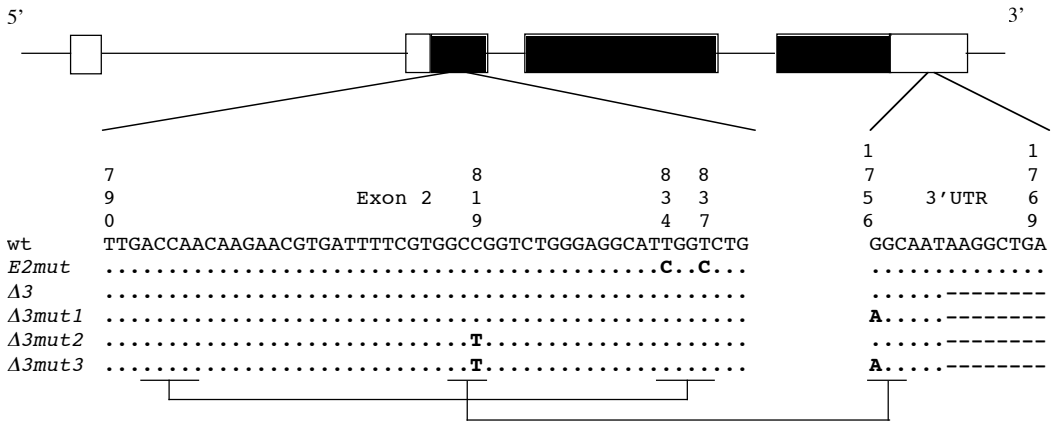

FIGURE 3.2 Location of mutations along the *Adh* transcript. Point mutations are shown by nucleotides differing from the wild-type sequence. Deletion of bases 1762-1769 is represented by dashes. Brackets indicate phylogenetically predicted pairing regions (STEPHAN and KIRBY 1993; PARSCH *et al.* 1997).

Lines containing the *E2mut* allele were generated by *P*-element-mediated germline transformation and compared to lines transformed with a wild-type *Wa-f* control. The results indicate that *E2mut* lines do not have reduced ADH activity relative to wild-type (FIGURE 3.3). In fact, the *E2mut* lines have slightly *higher* activity (the mean ADH activity ± SE of six *E2mut vs*. seven wild-type lines was 118.6 ± 6.5 units *vs*. 108.9 ± 4.6 units), though this difference is not significant (F = 1.99, P = 0.16). Thus, the more severe disruption of the local structure of exon 2 caused by *E2mut* in comparison to *mutC819T* does not appear to significantly affect *Adh* expression, leaving long-range pairing as the more likely explanation.

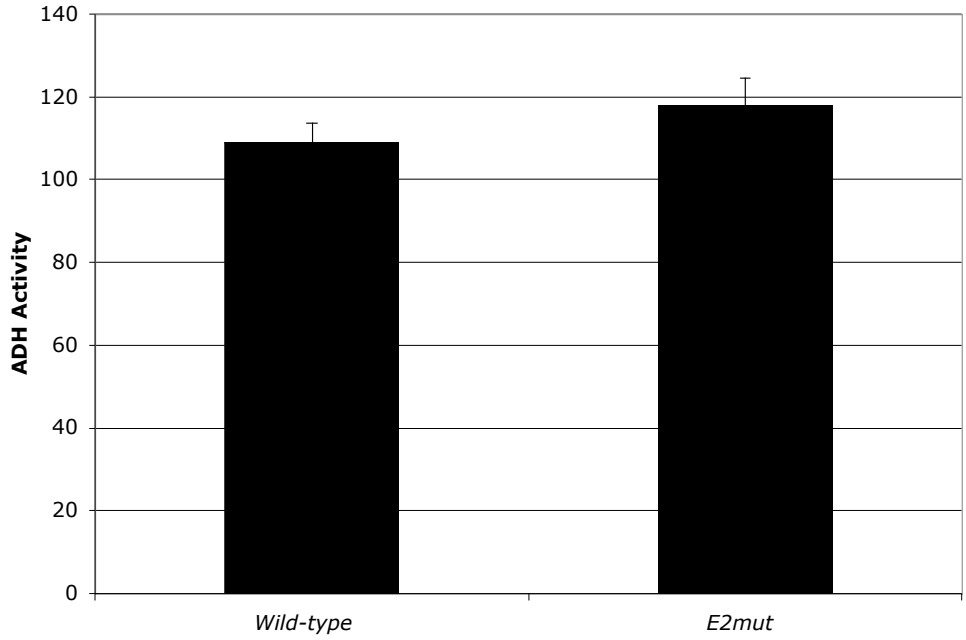

FIGURE 3.3 Average ADH activity of wild-type and *E2mut* lines. Values represent the mean of seven and six transformed lines, respectively. ADH activity is given in units of micromoles of NAD reduced per minute per milligram of total protein (multiplied by 100). Error bars represent ±1 SE. Differences between genotypes are not statistically significant.

Analysis of long-range pairing in a deletion background: To test the hypothesis that a classical compensatory effect was not observed between the predicted paired nucleotides 819/1756 due to the proximity of a highly conserved 8-base regulatory element in the 3' UTR (FIGURE 3.1), a series of compensatory mutations was made in a background of a deletion of this sequence. As a control, a deletion of bases 1762-1769 was used (PARSCH *et al.* 1999), and is designated *Δ3*. The compensatory-mutant alleles, *Δ3mut1*, *Δ3mut2* and *Δ3mut3* contain the mutations G1756A, C819T and C819T+G1756A, respectively (FIGURE 3.2). Thus, alleles *Δ3* and *Δ3mut3* allow for WC base pairing, whereas *Δ3mut1* and *Δ3mut2* are mismatches. The mean ADH activity ± SE of lines transformed with *Δ3* (15 lines), *Δ3mut1* (16 lines), *Δ3mut2* (17 lines) and *Δ3mut3* (12 lines) alleles was 228.3 ± 6.1, 210.8 ± 5.5, 209.3 ± 8.9 and 230.3 ± 7.1, respectively (FIGURE 3.4). Note that these activity values are approximately two-fold higher than those in the first experiment, due to the absence of the 8-base negative regulatory element in the *Δ3* background. Consistent with our hypothesis, the alleles with mutations causing mismatches (*Δ3mut1* and *Δ3mut2*) had significantly lower ADH activity in comparison to the *Δ3* control, whereas the compensatory double mutant *Δ3mut3* did not significantly differ from the control. Comparisons between the compensatory double mutant *Δ3mut3* and the mismatch alleles *Δ3mut1* and *Δ3mut2* are also significant, though only approach significance after Bonferroni correction (Table 3.1). Thus, it appears that the nucleotides within the conserved region of the 3' UTR (1756-1769) are involved in both WC base pairing and the negative regulation of *Adh* mRNA, and a complete compensatory interaction may be seen only in a background in which this latter function is removed.

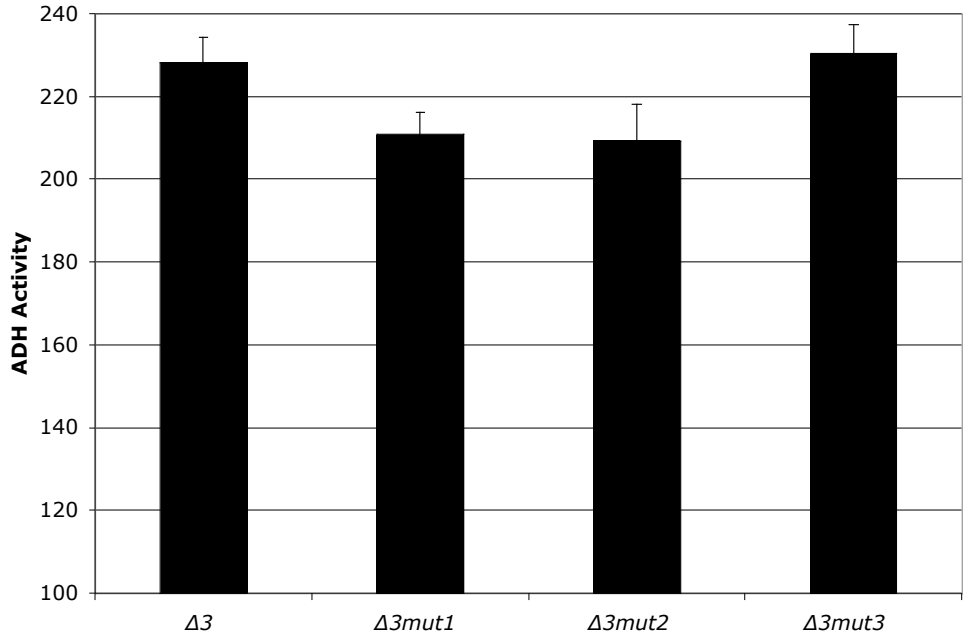

FIGURE 3.4 Average ADH activity of *Δ3*, *Δ3mut1*, *Δ3mut2* and *Δ3mut3* lines. Values represent the mean of 15, 16, 17 and 12 transformed lines, respectively. Units of activity are given as in FIGURE 3.3. Error bars represent ±1 SE. Tests of significance among genotypes are found in Table 3.1.

DISCUSSION

In this study, we examined long-range compensatory interactions between the two ends of *Adh* mRNA in more detail. Previous mutational analysis by PARSCH *et al.* (1997) suggested a long-range interaction between positions 819 and 1756, though the results did not completely conform to the classical model of compensatory evolution. We have now extended the analysis of PARSCH *et al.* (1997) with respect to this scenario and asked the question, do these results not fit the model (1) because the first deleterious intermediate (*mutC819T*) causes a reduction in ADH activity via a mechanism other than long-range base pairing, or (2) because the deleterious status of the second intermediate (*mutG1756A*) is more complex due to pleiotropy?

TABLE 3.1

Results of statistical analysis of ADH activity between mutant genotypes

Genotype 1	Genotype 2	Pairing	ΔADH Activity (Gen1 – Gen2)	*F*	*P*
Δ3	*Δ3mut1*	WC *vs*. mismatch	17.5	38.46	0.004*
Δ3	*Δ3mut2*	WC *vs*. mismatch	19.0	65.74	0.006*
Δ3mut3	*Δ3mut1*	WC *vs*. mismatch	19.5	40.61	0.011
Δ3mut3	*Δ3mut2*	WC *vs*. mismatch	21.0	69.63	0.015
Δ3	*Δ3mut3*	WC *vs*. WC	-2.0	4.74	0.790
Δ3mut1	*Δ3mut2*	mismatch *vs*. mismatch	1.5	12.29	0.623

For description of the genotypes, see FIGURE 3.2. The pairing column indicates whether WC base pairing or a mismatch is present between sites 819/1756 in the comparisons between genotypes.
* Significant after Bonferroni correction, for which $P = 0.0083$ corresponds to the 5% significance level.

To address the first possibility, we investigated an alternative explanation for the 15% reduction in ADH activity seen in *mutC819T* lines (PARSCH *et al.* 1997). More specifically, should disruption of the local structure of exon 2 be the reason for this change, disruption of this structure by two targeted mutations should produce a similar or more extreme phenotype. However, though the difference is not significant, *E2mut* lines have on average *higher* levels of ADH activity than lines transformed with the wild-type *Wa-f*. Given that T834C and T837C each change an unpreferred codon to a preferred codon (for isoleucine and glycine, respectively), another possible explanation is that any decrease in *Adh* expression caused by a disruption of the structure of exon 2 may be masked by an increase in expression due to the use of

preferred codons (CHEN *et al.* 1999). However, this would require the change of two synonymous codons to result in a >15% increase in expression. The recent results of CARLINI and STEPHAN (2003) suggest that such a difference may require on average seven to eight codon changes, making this scenario very unlikely. The lack of a measurable effect by disrupting the local structure of exon 2 is also in qualitative agreement with the analysis of CARLINI *et al.* (2001), who show a lower potential for secondary structure in the highly expressed *Adh* relative to the lowly expressed *Adhr* gene.

In contrast, closer inspection of a conserved region in the 3' UTR has yielded results consistent with our hypothesis concerning position 1756. Namely, *mutG1756A*'s role as a putative deleterious intermediate may be understood only in the context of the sum of the functional roles in which it and its neighboring nucleotides are involved. Detailed information regarding an 8-base regulatory element in the 3' UTR has proven particularly important. Previous studies have shown this sequence to be completely conserved across all *Drosophila* species examined (spanning the subgenera *Sophophora* and *Drosophila*, as well as the genus *Scaptodrosophila*), and deleting the first four bases, last four bases, or the entire sequence results in the same phenotype (a two-fold increase in ADH activity due to an underlying two-fold increase in mRNA) (PARSCH *et al.* 1997, 1999). Despite the apparent positive selection for increased ADH activity in the wild (*i.e.*, the S -> F amino acid replacement) (OAKESHOTT *et al.* 1982; BERRY and KREITMAN 1993; MERCOT *et al.* 1994), there appears to be strong purifying selection against changes in this sequence. Indeed, PARSCH *et al.* (2000) have demonstrated that transformed lines lacking this 8-base sequence (*Δ3*) have a significantly delayed development time, likely due to the presence of excessive amounts of *Adh* mRNA.

Given our knowledge of this 8-base sequence and our results from *Δ3mut1*, *Δ3mut2* and *Δ3mut3* lines, we propose that *mutG1756A* produces a partial *Δ3* phenotype. Under this scenario, the conserved sequences upstream of positions 1762-1769 also play a role in the negative regulation of mRNA, though to a lesser degree. One possibility is that positions 1762-1769 are *essential* to the binding of some *trans*-acting regulatory factor, whereas the conserved sequences upstream only *facilitate* this binding and hence produce only a partial phenotype. As demonstrated by the deletion analysis of PARSCH *et al.* (1999), disruption of any part of bases 1762-1769 does produce a full phenotype, enabling position 1756's role in long-range pairing to be investigated in the absence of its role in the regulation of mRNA (*i.e.,* there will be no pleiotropic effect of changing this nucleotide to determine its role in long-range pairing). Indeed, in a background of deleting positions 1762-1769, we show that positions 819/1756 do fit a classical model of compensatory evolution; *i.e.*, *both* intermediate states show a reduction in activity (KIMURA 1985; STEPHAN 1996; INNAN and STEPHAN 2001).

Additional support for a dual functional role of bases 1756-1761 comes from two sources. First, phylogenetic comparisons indicate that these six nucleotides are conserved within the Sophophora subgenus, including the distantly related *D. pseudoobscura* and *D. ambigua* (PARSCH *et al.* 1997). Such strong conservation in a non-coding region suggests functional constraint. Second, the experiments of PARSCH *et al.* (1997) show that the single mutation G1756A results in higher ADH activity than both wild-type and the compensatory double mutant, C819T-G1756A. Although this difference is not significant in both of the above comparions, the qualitative pattern of ADH activity in these mutant constructs is in agreement with the above hypothesis.

CONCLUSION

In chapter 1, DNA sequence variation at the *fw* locus of *D. ananasssae* was analyzed for 13 populations, representing almost the entire range of this species. In comparison to 10 neutrally evolving loci, the pattern at *fw* appears strikingly different. Two major haplotype classes distinguished by unique, high-frequency derived polymorphisms were fixed or nearly fixed in opposite regions of the species range. A significant reduction in levels of polymorphism and a greater than expected homogeneity of allele frequencies in these regions given estimates of gene flow is consistent with the action of natural selection in these populations. In particular, the haplotype class found in populations in the North displays a cline of decreasing frequency with increasing proximity to the equator, suggesting an association with mutation(s) that are locally favored. This pattern is also inconsistent with the single-sweep model proposed by SLATKIN and WIEHE (1998), instead favoring two independent sweeps, one in the North and one in the South. These results provide evidence for natural selection playing a significant role in genetic differentiation, and are particularly interesting with respect to recent range expansions and potential adaptation to new environments. The future availability of the *D. ananassae* genome will provide the exciting opportunity to map genes involved in adaptation.

In chapter 2, DNA sequence variation was examined for a 4-kb region of the *bicoid* gene of 25 *D. melanogaster* isofemale lines from Zimbabwe and one allele from *D. simulans*. Statistical tests revealed a significant excess of replacement polymorphisms in the *D. melanogaster* lineage that are clustered in two putative linker regions of the Bicoid protein. These are likely due to relaxed purifying selection on functionally unimportant regions of *bcd*. In addition, we found extensive linkage disequilibria across the region and a significantly smaller number of

haplotypes than the neutral expectation, which may suggest that positive selection has also played a role in the recent history of this newly duplicated gene. In addition, the recent results of GLINKA *et al.* (2003) suggest the Zimbabwe population of *D. melanogaster* has undergone a population size expansion. Because linkage disequilibrium levels are expected to be low in an expanding population (PRITCHARD and PRZEWORSKI 2001), tests of haplotype structure are more conservative, making the case for the involvement of selection stronger. Similarly, the apparent heterogeneity in the ratio of polymorphism to divergence (π / κ) along the gene region may be attributed to spatially varying negative as well as positive selective forces. On the other hand, we did not find any relationship between the strong haplotype pattern of our data set and epistatic selection maintaining the mRNA secondary structure in the *bcd* 3' UTR. However, the pattern of variation is consistent with the predictions of the model of INNAN and STEPHAN (2001), which predicts that linkage disequilibrium associated with the transition of one stable WC base pair to another one should be observed only rarely.

In chapter 3, long-range compensatory interactions between the two ends of *Drosophila* alcohol dehydrogenase (*Adh*) mRNA were investigated experimentally. Alternative hypotheses for why previous analysis of long-range compensatory interactions failed to fit KIMURA's (1985) classical compensatory model were tested. The results of the mutational analysis indicate that a classical result was not observed due to the pleiotropic effect of changing a nucleotide involved in both long-range base pairing *and* the negative regulation of gene expression. This reaffirms that long-range base pairing in *Drosophila Adh* mRNA does play a functional role, and future experiments may help to further define this. These results also show that more complex scenarios than those typically considered by KIMURA's (1985) classical

model of compensatory evolution should be considered in the study of compensatory interactions.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607-615.

BALLARD, J. W., and M. KREITMAN, 1994 Unraveling selection in the mitochondrial genome of *Drosophila*. Genetics **138:** 757-772.

BEGUN, D. J., and C. F. AQURADO, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete region*. Genetics **129:** 1147-1158.

———, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature **356:** 519-520.

———, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548-550.

———, 1995*a* Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. Mol. Biol. Evol. **12:** 382-390.

———, 1995*b* Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140:** 1019-1032.

BERLETH, T., M. BURRI, G. THOMA, D. BOPP, S. RICHSTEIN *et al*., 1988 The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. EMBO J. **7:** 1749-1756.

BERRY, A., and M. KREITMAN, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134:** 869-893.

CARLINI, D. B., Y. CHEN and W. STEPHAN, 2001 The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the Drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. Genetics **159:** 623-633.

CARLINI, D. B., and W. STEPHAN, 2003 *In vivo* introduction of unpreferred synonymous codons into the Drosophila *Adh* gene results in reduced levels of ADH protein. Genetics **163:** 239-243.

CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. **68:** 131-149.

CHARLESWORTH, B., M. NORDBORG and D. CHARLESWORTH, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155-174.

CHEN, Y., D. B. CARLINI, J. F. BAINES, J. PARSCH, J. M. BRAVERMAN *et al.*, 1999 RNA secondary structure and compensatory evolution. Genes Genet. Syst. **74:** 271-286.

CHEN, Y., B. J. MARSH, and W. Stephan, 2000 Joint effects of natural selection and recombination on gene flow between *Drosophila ananassae* populations. Genetics **155:** 1185-1194.

CHEN, Y., and W. STEPHAN, 2003 Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. Proc. Natl. Acad. Sci. USA **100:** 11499-11504.

COMERON, J. M., M. KREITMAN and M. AGUADÉ, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151:** 239-249.

CROW, J. F., 1986 *Basic Concepts in Population, Quantitative, and Evolutionary Genetics*. Freeman, New York.

DAS, A., S. MOHANTY and W. STEPHAN, 2003 Out of Sundaland: Population structure and demographic history of *Drosophila ananassae*. Submitted.

DAS, A., and B. N. SINGH, 1991 Chromosomal polymorphism in Indian natural populations of *Drosophila melanogaster*. Korean J. Genet. **13**: 97-112.

DAVID, J. R., and P. CAPY, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106-111.

DEARDEN, P., and M. AKAM, 1999 Developmental evolution: Axial patterning in insects. Curr. Biol. **9:** R591-594.

DEPAULIS, F., and M. VEUILLE, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788-1790.

DRIEVER, W., and C. NÜSSLEIN-VOLHARD, 1988 The *bicoid* protein determines position in the *Drosophila* embryo in a concentration-dependent manner. Cell **54:** 95-104.

DUBNAU, J., and G. STRUHL, 1996 RNA recognition and translational regulation by a homeodomain protein. Nature **379:** 694-699.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475-7479.

FERRANDON, D., I. KOCH, E. WESTHOF and C. NÜSSLEIN-VOLHARD, 1997 RNA-RNA interaction is required for the formation of specific *bicoid* mRNA 3' UTR-STAUFEN ribonucleoprotein particles. EMBO J. **16:** 1751-1758.

FOX, G. E., and C. R. WOESE, 1975 5S rRNA secondary structure. Nature **256:** 505-507.

FROHNHÖFER, H. G., R. LEHMANN and C. NÜSSLEIN-VOLHARD, 1986 Manipulating the anteroposterior pattern of the *Drosophila* embryo. J. Embryol. Exp. Morphol. **97** (Suppl.)**:** 169-179.

GALLIE, D. R., 1991 The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. Genes Dev. **5:** 2108-2116.

GILBERT, P. and R. B. HUEY, 2001 Chill-coma temperature in Drosophila: effects of developmental temperature, latitude, and phylogeny. Physiol Biochem Zool. **74:** 429-434.

GILLESPIE, J. H., 1991 *The causes of molecular evolution*. Oxford University Press, New York.

GILLESPIE, J. H., 1994 Substitution processes in molecular evolution. III. Deleterious alleles. Genetics **138:** 943-952.

GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN and D. DE LORENZO, 2003 Demography and natural selection have shaped genetic variation in Drosophila melanogaster. Genetics **165:** 1269-1278.

GUO, L., E. ALLEN and W. A. MILLER, 2000 Structure and function of a cap-independent translation element that functions in either the 3' or the 5' untranslated region. RNA **6:** 1808-1820.

GUO, L., E. M. ALLEN and W. A. MILLER, 2001 Base-pairing between untranslated regions facilitates translation of uncapped, nonpolyadenylated viral RNA. Mol. Cell **7:** 1103-1109.

HAN, K. and H. J. KIM, 1993 Prediction of common folding structures of homologous RNAs. Nucleic Acids Res. **21:** 1251-1257.

HARR, B., M. KAUER and C. SCHLÖTTERER, 2002 Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949-12954.

HAAS, E. S., D. P. MORSE, J. W. BROWN, F. J. SCHMIDT and N. R. PACE, 1991 Long-range structure in ribonuclease P RNA. Science **254:** 853-856.

HENTZE, M. W., 1997 eIF4G: a multipurpose ribosome adapter? Science **275:** 500-1.

HEY, J., and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833-846.

HUDSON, R. R. and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147-164.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245-250.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Survey Evol. Biol. **7:** 1-44.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329-1340.

HUDSON, R. R. and N. L. KAPLAN, 1995 Deleterious background selection with recombination. Genetics **141:** 1605-1617.

INNAN, H., and W. STEPHAN, 2001 Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics **159:** 389-399.

JAMES, B. D., OLSEN, G. J. and N. R. PACE, 1989 Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol. **180:** 227-239.

KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989 The 'hitchhiking effect' revisited. Genetics **123:** 887-899.

KELLY, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146:** 1197-1206.

KHARE, P. V., R. J. BARNABAS, M. KANOJIYA, A. D. KULKARNI and D. S. JOSHI, 2002 Temperature dependent eclosion rhythmicity in the high altitude Himalayan strains of *Drosophila ananassae*. Chronobiol Int. **19:** 1041-1052.

KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking on a recombining chromosome. Genetics **160:** 765-777.

KIMURA, M., 1956 A model of a genetic system which leads to closer linkage by natural selection. Evolution **10:** 278-287.

———, 1968 Evolutionary rate at the molecular level. Nature **217:** 624-626.

———, 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, Massachusetts.

———, 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7-19.

KIRBY, D. A., MUSE, S. V. and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047-9051.

KIRBY, D. A. and W. STEPHAN, 1996 Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. Genetics **144:** 635-645.

KLIMAN, R. M., and J. HEY, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239-1258.

KONINGS, D. A., L. P. VAN DUIJN, H. O. VOORMA and P. HOGEWEG, 1987 Minimal energy foldings of eukaryotic mRNAs form a separate leader domain. J. Theor. Biol. **127:** 63-78.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412-417.

LACHAISE, D., M. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS et al., 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup, pp. 159–225 in Evolutionary Biology, edited by M. K. HECHT, B. WALLACE and G. T. PRANCE. Plenum, New York.

LAURIE-AHLBERG, C. C., and L. F. STAM, 1987 Use of P-element-mediated transformation to identify the molecular basis of naturally occurring variants affecting *Adh* expression in *Drosophila melanogaster*. Genetics **115:** 129-140.

LEFEVRE, F., 1976 A photographic representation and interpretation of polytene chromosomes of *Drosophila melanogaster* salivary glands. Pp. 31-66 *in* M. ASHBURNER and E. NOVITSKI, eds. The genetics and biology of Drosophila. Academic Press, New York.

LEMEUNIER, F., and S. AULARD, 1992 Inversion polymorphism in *Drosophila melanogaster*. Pp. 339-405 in C. B. Krimbas and J. R. Powell, eds. *Drosophila* inversion polymorphism. CRC Press, Boca Raton, Fla.

LEWIN, B., 1997 *Genes VI*. Oxford University Press, New York.

LEWONTIN, R. C., 1974 *The genetic basis of evolutionary change*. Columbia University Press, New York.

LOWRY, O. H., N. J. ROSEBROUGH, A. L. FARR and R. J. RANDALL, 1951 Protein measurements with the Folin phenol reagent. J. Biol. Chem. **193:** 265-275.

MACDONALD, P. M., 1990 *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development **110:** 161-171.

MACDONALD, P. M., and K. KERR, 1998 Mutational analysis of an RNA recognition element that mediates localization of *bicoid* mRNA. Mol. Cell Biol. **18:** 3788-3795.

MACDONALD, P. M., and G. STRUHL, 1988 *cis*-acting sequences responsible for anterior localization of *bicoid* mRNA in *Drosophila* embryos. Nature **336:** 595-598.

MAYNARD SMITH, J. and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.

MARONI, G. 1978 Genetic control of alcohol dehydrogenase levels in *Drosophila*. Biochem. Genet. **16:** 509-523.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652-654.

MCDONALD, J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **13:** 253-260.

———, 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377-384.

MERCOT, H., D. DEFAYE, P. CAPY, E. PLA and J. R. DAVID, 1994 Alcohol tolerance, ADH activity, and ecological niche of *Drosophila* species. Evolution **48:** 746-757.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13:** 261-277.

MOUNT, S. M., C. BURKS, G. HERTZ, G. D. STORMO, O. WHITE *et al.*, 1992 Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. **20:** 4255-4262.

MUSE, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139:** 1429-1439.

NACHMAN, M. W., S. N. BOYER and C. F. AQUADRO, 1994 Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. Proc. Natl. Acad. Sci. USA **91:** 6364-6368.

NACHMAN, M. W., W. M. BROWN, M. STONEKING and C. F. AQUADRO, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142:** 953-963.

NAKAMURA, Y., T. GOJOBORI and T. IKEMURA, 2000 Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. **28:** 292.

NEI, M., 1987 *Molecular evolutionary genetics*. Columbia University Press, New York.

NOLLER, H. F., and C. R. Woese, 1981 Secondary structure of 16S ribosomal RNA. Science **212:** 403-411.

OAKESHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON *et al.*, 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. Evolution **36:** 86-96.

OHTA, T., 1982 Linkage disequilibrium due to random drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:** 1940-1944.

———, 1992 Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. Genetics **130:** 917-923.

PACE, N. R., D. K. SMITH, G. J. OLSEN, and B. D. JAMES, 1989 Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA - a review. Gene **82:** 65-75.

PARSCH, J., S. TANDA and W. STEPHAN, 1997 Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **94:** 928-933.

PARSCH, J., W. STEPHAN and S. TANDA, 1998 Long-range base pairing in *Drosophila* and human mRNA sequences. Mol. Biol. Evol. **15:** 820-826.

———, 1999 A highly conserved sequence in the 3'-untranslated region of the *Drosophila Adh* gene plays a functional role in *Adh* expression. Genetics **151:** 667-674.

PARSCH, J., J. M. BRAVERMAN and W. STEPHAN, 2000 Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics **154:** 909-921.

PARSCH, J., J. A. RUSSELL, I. BEERMAN, D. L. HARTL and W. STEPHAN, 2000 Deletion of a conserved regulatory element in the Drosophila *Adh* gene leads to increased alcohol dehydrogenase activity but also delays development. Genetics **156:** 219-227.

PATEL, N. H., 2000 It's a bug's life. Proc. Natl. Acad. Sci. USA **97:** 4442-4444.

PATTON, J. S., X. V. GOMES and P. K. GEYER, 1992 Position-independent germline transformation in Drosophila using a cuticle pigmentation gene as a selectable marker. Nucleic Acids Res. **20:** 5859-5860.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: Models and data. Am. J. Jum. Genet. **68:** 1-14.

RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. **13:** 735-748.

ROBERTSON, H. M., C. R. PRESTON, R. W. PHILLIS, D. M. JOHNSON-SCHLITZ, W. K. BENZ *et al.*, 1988 A stable genomic source of *P* element transposase in *Drosophila melanogaster*. Genetics **118:** 461-470.

ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

RUBIN, G. M., and A. C. SPRADLING, 1982 Genetic transformation of Drosophila with transposable element vectors. Science **218:** 348-353.

SACHS, A. B., P. SARNOW and M. W. HENTZE, 1997 Starting at the beginning, middle, and end: translation initiation in eukaryotes. Cell **89:** 831-838.

SCHAEFFER, S. W. and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541-552.

SCHAEFFER, V., D. KILLIAN, C. DESPLAN and E. A. WIMMER, 2000 High *bicoid* levels render the terminal system dispensable for *Drosophila* head development. Development **127:** 3993-3999.

SCHRÖDER, R., and K. SANDER, 1993 A comparison of transplantable *bicoid* activity and partial *bicoid* homeobox sequences in several *Drosophila* and blowfly species (Calliphoridae). Roux's Arch. Dev. Biol. **203:** 34-43.

SEEGER, M. A., and T. C. KAUFMAN, 1990 Molecular analysis of the *bicoid* gene from *Drosophila pseudoobscura*: identification of conserved domains within coding and noncoding regions of the *bicoid* mRNA. EMBO J. **9:** 2977-2987.

SLATKIN, M. and T. WIEHE, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155-160.

SORSA, V., 1988 *Chromosome maps of Drosophila*. CRC, Boca Raton, Florida.

SPRADLING, A. C., and G. M. RUBIN, 1982 Transposition of cloned P-elements into Drosophila germ line chromosomes. Science **218:** 341-347.

SPRINZL, M., T. HARTMANN, F. MEISSNER, J. MOLL, and T. VORDERWÜLBECKE, 1987 Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. **15**(Suppl.)**:** r53-r188.

STAUBER, M., H. JÄCKLE and U. SCHMIDT-OTT, 1999 The anterior determinant *bicoid* of *Drosophila* is a derived *Hox* class 3 gene. Proc. Natl. Acad. Sci. USA **96:** 3786-3789.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* population. I. Contrasts between the *vermilion* and *forked* loci. Genetics **121:** 89-99.

STEPHAN, W. and S. J. MITCHELL, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics **132:** 1039-1045.

STEPHAN, W., T. H. E. WIEHE and M. W. LENZ, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237-254.

STEPHAN, W. and D. A. KIRBY, 1993 RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. Genetics **135:** 97-103.

STEPHAN, W., 1996 The rate of compensatory evolution. Genetics **144:** 419-426.

STEPHAN, W., L. XING, D. A. KIRBY, and J. M. BRAVERMAN, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95:** 5649-5654.

TACHIDA, H., 1994 Decay of linkage disequilibrium in a finite island model. Genet. Res. **64:** 137-144.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TARUN, S. Z. and A. B. SACHS, 1996 Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. EMBO J. **15:** 7168-77.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119-164.

TOBARI, Y. N., 1993 Drosophila ananassae - *Genetical and Biological Aspects*. Japan Scientific Societies Press, Tokyo, and Karger, Basel.

VERRELLI, B. C., and W. F. EANES, 2000 Extensive amino acid polymorphism at the *pgm* locus is consistent with adaptive protein evolution in *Drosophila melanogaster*. Genetics **156:** 1737-1752.

———, 2001 Clinal variation for amino acid polymorphisms at the *Pgm* locus in *Drosophila melanogaster*. Genetics **157:** 1649-1663.

VOGL, C., A. DAS, M. BEAUMONT, S. MOHANTY and W. STEPHAN, 2003 Population subdivision and molecular sequence variation: theory and analysis of Drosophila ananassae data. Genetics **165:** 1385-1395.

WALTER, A. E., D. H. TURNER, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker, 1994 Coaxial stacking of helixes enhances binding of

oligoribonucleotides and improves predictions of RNA folding. Proc. Natl. Acad. Sci. USA **91:** 9218-9222.

WANG, S. P., K. S. BROWNING and W. A. MILLER, 1997 A viral sequence in the 3' untranslated region mimics a 5' cap in facilitating translation of uncapped mRNA. EMBO J. **16:** 4107-4116.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

WELLS, S. E., P. E. HILLNER, R. D. VALE and A. B. SACHS, 1998 Circularization of mRNA by eukaryotic translation initiation factors. Mol. Cell **2:** 135-140.

WIEHE, T. and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842-854.

WISE, C. A., M. SRAML and S. EASTEAL, 1998 Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics **148:** 409-421.

ZUKER, M., 1989 On finding all suboptimal foldings of an RNA molecule. Science **244:** 48-52.

ZUKER, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:** 1-10.