The Pennsylvania State University

The Graduate School

Eberly College of Science

MOLECULAR EVOLUTIONARY GENETICS OF INVERSION BREAKPOINT REGIONS IN *DROSOPHILA PSEUDOOBSCURA*

A Dissertation in

Biology

by

Andre G. Wallace

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010

The dissertation of Andre G. Wallace was reviewed and approved* by the following:

Hong Ma
Distinguished Professor of Biology
Chair of Committee

Stephen W. Schaeffer
Associate Professor of Biology
Dissertation Adviser

Mary Poss
Professor in Biology and in Veterinary and Biomedical Sciences

Avery August
Distinguished Professor of Immunology

Douglas Cavener
Professor and Head of Biology

*Signatures are on file in the Graduate School.

ABSTRACT

Chromosomal inversions are present on the chromosomes of many organisms. The effects of inversions on most of these organisms are frequently and extensively studied. *Drosophila pseudoobscura* is known to have more than 30 different inversions on its third chromosome and is one of the most frequently studied inversion systems. We employed this inversion system to study the evolutionary history of the gene arrangements that resulted from some of these inversions. The standard and most widely accepted phylogeny of the *D. pseudoobscura* gene arrangements was generated in the early 1900's and the phylogeny was established using cytological analyses. The fairly recent sequencing of the *D. pseudoobscura* genome enabled us to develop a large molecular dataset to investigate the phylogeny of the inversion system. The dataset contained regions near inversion breakpoints because these regions are potentially informative because recombination is suppressed at inversion breakpoints. Several non-breakpoint regions were also included in the data set. There were three key questions addressed in this study. 1) Is the origin of the inversion polymorphism monophyletic? 2) Is the molecular phylogeny consistent with the cytological phylogeny? 3) Which arrangement is the ancestral in the *D. pseudoobscura* population? We were able to provide support for a monophyletic origin of this inversion polymorphism and our results also provided support for the cytological phylogeny. Though further experimentation is necessary to identify the exact ancestor, we were able to determine that the ancestral arrangement is either the Santa Cruz or the Hypothetical arrangement.

A population genetics study was also performed using the same dataset. In this study, we investigated how nucleotide diversity within and between *D. pseudoobscura* inversions was altered by differences in the levels of genetic exchange across the chromosome. Consistent with

theoretical predictions, regions within the proximal segment had lower levels of recombination than regions in the distal segment of the chromosome. We also identified one inversion (PP) that shared more genetic information with all other arrangements than any other arrangement studied. Finally, our results indicated that positive selection is acting on this species and that this species has recently experienced a population expansion.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 2

Chapter 3

ACKNOWLEDGEMENTS

I would like to thank Dr. Stephen Schaeffer for opening up his lab to me so that I was able to conduct my doctoral thesis research. Dr. Schaeffer has maintained a positive working environment which made it easier to accomplish my daily goals. I would also like to thank him for always keeping his door open and being readily available to provide advice.

I would also like to express my sincere gratitude to all members of my committee (Dr. Avery August, Dr. Hong Ma, and Dr. Mary Poss) who served faithfully from the very first meeting. The guidance provided by my doctoral thesis committee helped to make this a great scientific experience for me. In addition to serving on my thesis committee, I would like to thank Dr. August for serving as my fellowship advisor and for providing valuable advice when needed.

I would like to express my appreciation to Dr. Charles Fisher who showed great interest in my graduate education. When I needed a new lab to conduct my doctoral studies, Dr. Fisher was loyal in helping me and even provided a link between myself and Dr. Schaeffer. If it was not for Dr. Fisher's help, the transition to a new lab would not be as smooth.

Special thanks to the Penn. State Department of Biology for assisting with funding for my graduate assistantship. I would also like to thank the Alfred P. Sloan foundation for additional funding that helped throughout my studies.

I would like to thank my family and friends for their prayers and support that helped to keep me grounded throughout this process. Specifically, I would like to thank my mother and father, Mr. and Mrs. George Wallace, for sacrificing their time and happiness, at times, to ensure that I stayed focused throughout my educational journey.

Finally, I would like to thank Dr. Randen Patterson and members of the Patterson-lab.

CHAPTER 1

LITERATURE REVIEW

Background on Inversions

The presence of inversions on the chromosomes of organisms has been documented for decades. Inversions have been identified in mammalian systems as well as non-mammalian systems. In fact, inversions identified within the human genome have been linked to various diseases (ANTONACCI *et al.* 2009; CHEN *et al.* 2010; ESTER *et al.* 2006). Inversions were discovered in *Drosophila* in the early 1900's when Sturtevant proposed that crossover rates were being affected in heterokaryotypes of *Drosophila melanogaster* (STURTEVANT 1917). It was not until 1926 when he noticed that gene order could change between related species that he finally understood that chromosomal inversions were responsible for this reduction in crossover rates (STURTEVANT 1926). Inversions occur due to a double break on the chromosome that results in the broken segment being rotated an angle of 180 degrees, followed by rejoining of the fragment, which ultimately results in the order of genes being reversed. There are two types of chromosomal inversions, pericentric and paracentric. Pericentric inversions involve the centromere and a break occurs on each arm of the chromosome, while paracentric inversions do not involve the centromere and both breaks occur on the same arm of the chromosome. The *Drosophila* genus is known to be extremely polymorphic for paracentric inversions (Sperlich and Pfreim 1986). For example, *Drosophila melanogaster* has been shown to have greater than 350 inversions (Lemeunier and Aulard 1992). Inversions were first visualized using a cytogenetic technique made popular by Painter in the 1930's. He utilized larval salivary squashes to view polytene chromosomes and study banding and looping patterns. Polytene chromosomes occur when chromosomal DNA replicates, but does not separate via karyokinesis (KING 1985). Polytene chromosomes are limited to a small number of taxa. Homologous chromosomes in polytenes from the *Drosophila* larval salivary gland pair as they would during meiosis so that if

an individual is heterozygous for different gene arrangements, one will see characteristic loops. This approach utilized by Painter was effective in identifying looping patterns that could be linked to a rearranged chromosome (PAINTER 1933).

The presence of inversions in all chromosomes of *D. pseudoobscura* has been observed but the third chromosome is significantly more polymorphic for paracentric inversions (DOBZHANSKY 1944). More than 30 different gene arrangements have been identified on the 3rd chromosome of *Drosophila pseudoobscura*. These gene arrangements are suggested to have formed as a result of overlapping paracentric inversions. Dobzhansky and Sturtevant (1938) described some of these arrangements and provided views of the 3rd chromosome configurations. The arrangements described originated from various localities and were categorized based on banding and looping patterns of the chromosomes (DOBZHANSKY and STURTEVANT 1938).

Inversions and phylogeny.

The *D. pseudoobscura* inversion-generated gene arrangements represented the first use of genetic data set to produce an evolutionary phylogeny (DOBZHANSKY and STURTEVANT 1938). This data set included gene arrangements which are located on the third chromosome of *D. pseudoobscura*. The banding pattern of an arbitrarily chosen chromosome, the Standard arrangement, was chosen to develop the standard cytogenetic map with numbered sections and lettered subsections and breakpoints of new arrangements were identified relative to the Standard arrangement, and the location where the first new chromosome type was collected was used to name the arrangement. The main chromosomal arrangements used in our studies are named Standard (ST), Arrowhead (AR), Pikes Peak (PP), Santa Cruz (SC), Chiricahua (CH), and Tree Line (TL). **Figure 1** depicts a revised version of this phylogeny and includes the Hypothetical

(HY) arrangement which is so named because it has never been collected in nature. The gene arrangements were sorted phylogenetically by examining polytene chromosomes of *D. pseudoobscura* with the assumption that each arrangement was linked to another by unique inversion events. The only exception is the link between the Standard and Santa Cruz arrangements that require two inversion events unless one allows for the Hypothetical arrangement. The trouble with the cytological phylogeny is that one cannot determine the ancestral arrangement by comparing the maps among related species because the banding and puffing patterns change so rapidly that one cannot know the exact homologies of the maps.

The use of inversions to infer evolutionary relationships is not limited to the *Drosophila* organism. Chromosomal rearrangements generated by inversions have been used to help understand the evolutionary forces affecting mammalian species. Inversion-induced rearrangements were used to reconstruct the genomic architecture of ancestral mammals (BOURQUE *et al.* 2004). The genomes of human and mouse were shown to contain a large number of small synteny blocks that are difficult to identify except through the linked inversion breakpoints. These breakpoints frequently cluster and disproves a model of mammalian rearrangement evolution that suggests that there is a single rearrangement site linking any two conserved sites shared by humans and mouse (PEVZNER and TESLER 2003). Inversion breakpoints were also shown to influence mammalian evolution because the distribution of breakpoints along the chromosome provided information about the order of chromosomal rearrangements. More notably polytene chromosome analysis was used to infer phylogenetic relationships in *Anopheles gambiae*, a type of mosquito. Phylogenetic relationships were inferred from fixed inversions in the 2R arm of the chromosome (COLUZZI *et al.* 2002). The use of

inversions to study evolutionary relationships continues to expand and span across a wide variety of organisms.

Origin of Inversions

Since being proposed by Dobzhansky and Sturtevant (1938), numerous studies have supported the monophyletic origin of the *D. pseudoobscura* gene arrangements. The cytological phylogeny generated by Dobzhansky and Sturtevant (**Figure** 1) was constructed with the assumption that each arrangement is connected to each other by single inversion events. This view of a unique origin for the inversion polymorphism is characterized by many as the "traditional view." The unique origin of inversions in *D. pseudoobscura* was verified by Aquadro *et al*. (1991) with sequence data from the amylase (*Amy*) gene region. The *Amy* gene, which is found in different gene arrangements, was used to determine the age of the different arrangements, providing a linear correlation with the phylogenetic order of the gene arrangements (AQUADRO *et al.* 1991). The results of the study not only supported a unique origin of the inversion polymorphism but also validated the proposed cytological phylogeny. Though Aquadro *et al*. (1991) provided support for the unique origin of inversions, the strength of the study may have been affected by the small sample size of the endonuclease sites across one genetic locus in a small number of strains. Our lab developed a large data set of 11 inversion breakpoint regions and seven non-breakpoint regions which was used to study the phylogeny of *D. pseudoobscura* arrangements. These regions were sequenced in up to 100 flies from six different gene arrangements. Our results validate the unique origin of *D. pseudoobscura* inversions and provided strong support for the proposed cytological phylogeny (Chapter 2). We were also able to use a molecular clock approach to date the different gene arrangements.

For inversions to have a unique origin, their breakpoints must be used a single time. Because gene arrangements in *D. pseudoobscura* are suggested to occur from multiple overlapping paracentric inversions, the inversion chromosomal breaks would have to occur sequentially and not all at the same time. Chromosomal breaks are very rare and so the chances of two such events occurring simultaneously are very low. However, if these breaks should occur the new rearrangement must be able to succeed genetically in the population and this is very unlikely to happen. If a situation such as this occurs where a rearrangement resulting from multiple simultaneous breaks becomes established in the population or species, the monophyletic origin of the inversion polymorphism would be challenged.

The possibilities of a polyphyletic origin in *Drosophila* inversions have been frequently argued. Some studies have suggested that there are "hot" points located on chromosomes in *Drosophila*. These "hot" points are regions of multiple breaks and are suggested to occur only in some species of *Drosophila*. The *Drosophila* obscura is one of the species that was shown to contain such "hot" points (BREHM and KRIMBAS 1991). Transposable elements are widely distributed throughout the *Drosophila* genus and their discovery also challenged the theory of a monophyletic origin for inversions. Before transposable elements were identified in *Drosophila* (GREEN 1980), they were initially described as chromosomal breaking agents in corn by Barbara McClintock. A group of transposable elements known as P elements, were shown to initiate single and multiple breaks in *D. melanogaster* (ENGELS and PRESTON 1984; GREEN 1980). These multiple breaks could support a polyphyletic origin of inversions if they rejoin simultaneously on the same chromosome. For these polyphyletic events to occur, the multiple breaks must occur within a few hundred nucleotides of each other rather than too close or too far (ENGELS and PRESTON 1984). It is not clear whether these P element-induced multiple breaks occur in all

species of *Drosophila* but so far none has been discovered in *D. pseudoobscura*. Another argument introduced to propose a polyphyletic origin is the theory of re-inversions (EMMENS 1937; GRUNEBERG 1936). This phenomenon would likely be true if the same inversions were found to be produced repeatedly. These re-inversions have been generated in laboratory experiments but have never been identified in natural populations. Therefore, their existence has been doubted (KAUFMANN 1942). Recent comparisons of complete *Drosophila* genomes have suggested that intervals between some genes could be used as breakpoints multiple times, but the exact location of the chromosomal breaks could differ (BHUTKAR *et al.* 2008; VON GROTTHUSS *et al.* 2010).

Selection in *Drosophila pseudoobscura*

Drosophila pseudoobscura inversions have formed the basis for a variety of studies about the adaptive significance of these arrangements. For example, *D. pseudoobscura* have been consistently used since the early 1900's to look for evidence of selection. There are several studies that discuss the theory of selection acting on *D. pseudoobscura* arrangements. Dobzhansky showed that *D. pseudoobscura* gene arrangements responded to seasonal changes with altering frequencies (DOBZHANSKY 1943). The genetic composition of the population was measured for different seasons over a number of years and it was shown that the frequency of different fly strains fluctuated based on the seasons. In the Mount San Jacinto (California) locality, the Standard (ST) arrangement was shown to occur at its highest frequency in the winter and early spring and lowest in the summer time. On the other hand, the Arrowhead (AR) and Chiricahua (CH) arrangements behaved in a manner opposite to the ST arrangement. It was later shown that a force of selection which is adaptive may be responsible for the maintenance of

these inversion polymorphisms (Wright and Dobzhansky 1946). Dobzhansky later supported the hypothesis that selection influences *D. pseudoobscura* gene arrangements in his famous "population cage" experiments that showed supremacy of fitness values in inversion heterozygotes over homozygotes (Dobzhansky 1948; 1950). Although no exact genes have been identified as targets of selection, inversions have the ability to 1) change the order of genes on a chromosome, 2) eliminate genes, and 3) possibly introduce new genes to a chromosome. These factors could ultimately result in a chromosomal rearrangement that may or may not be favored by selective pressures. Selection in inversions is not limited to *D. pseudoobscura* but this system is one of the most frequently studied systems and so it is used here to discuss the topic.

Ancestry of *Drosophila pseudoobscura* Arrangements

It was first suggested by Sturtevant and Dobzhansky (1936) that the *D. pseudoobscura* inversion polymorphism was introduced to the population about one million years ago. This hypothesis did gain strong support until molecular approaches were utilized (AQUADRO *et al.* 1991). The findings of Aquadro *et al*. (1991) were able to support an ancestral age of 2.1 million years ago for the *D. pseudoobscura* third chromosome inversion polymorphism. This age for the inversion polymorphism is now widely accepted but within the last several decades, the ancestral arrangement of these *D. pseudoobscura* inversions has been challenged. Based on banding patterns, geographic distribution, and the position of the arrangements in the phylogeny, the TL was labeled as the ancestral arrangement (WALLACE 1966). The proposal of TL being the oldest arrangement triggered an experiment in which a series of crosses were put together to test pairing of the TL arrangement with the second and third oldest collected *D. pseudoobscura* arrangements (ST, SC) and also the *D. miranda* strain (MORROW 1970). The TL was shown to

have significantly better pairing with the *D. miranda* than any other arrangement which means that the TL is likely the oldest arrangement. The longer a gene arrangement exists in a population the more likely it is that it will extend geographically. It is also more likely that it will initiate the formation of new arrangements. The standard and widely accepted phylogeny of *D. pseudoobscura* gene arrangements show more single step inversions stemming from the TL arrangement than any other (OLVERA *et al.* 1979). Based on the proposed age of the TL arrangement and the number of single step inversions arising from it, a supportive argument was given in favor of the TL arrangement being ancestral (OLVERA *et al.* 1979). These early observations were all made using the standard phylogeny of the gene arrangements generated from cytological data. Aquadro *et al*. (1991) provided an analysis on the ancestry of the arrangements using sequence data from the amylase gene which is distributed across the different arrangements. The TL arrangement appeared to be the most divergent and it is reasoned that the older an arrangement, the more divergent it will be. This study should be viewed as informative but not taken as a final analysis since the study employed a small sample size and one gene region across the arrangements. Updated studies to the Aquadro *et al*. (1991) study have determined that the ancestral arrangement is either the TL or the SC arrangement (POPADIĆ and ANDERSON 1994). This conclusion was made after analyzing sequencing and restriction mapping data. It is also important to point out that Popadic and Anderson (1994) ruled out the ST and HY arrangements as the possible ancestral arrangements.

The effects of paracentric inversions on recombination

Paracentric inversions are frequently studied due to their ability to suppress recombination in heterozygotes. As suppressors of recombination, inversions can alter the

position of genes on the chromosome which could result in genetic hitchhiking. A finite population model was developed by Kaplan *et al*. (1989) to explain the possible effects of low rates of recombination on genetic hitchhiking. The model predicts that regions of low recombination rates harbor strongly selected mutations that can substantially reduce the levels of polymorphisms in a population which are expected under a neutral model (KAPLAN *et al.* 1989). It was also proposed that as suppressors of recombination, inversions can also reduce nucleotide variation at linked loci due to deleterious background selection (HUDSON and KAPLAN 1995; KIM and STEPHAN 2000). Therefore, variation can be reduced in regions of low recombination either through genetic hitchhiking associated with favorable mutations, but also through deleterious background selection. Using the genetic hitchhiking models, it was predicted that nucleotide variation can be positively linked to levels of recombination (BEGUN and AQUADRO 1992).

Recombination is an important process in all organisms since it has a significant impact on genetic variation because of its ability to homogenize diversity between chromosomes. Lower levels of recombination can lead to genetic diversification. In fact, recombination is suggested to be responsible for approximately 25% of genetic variation in *Drosophila* (MORIYAMA and POWELL 1996). The ability of recombination to affect levels of nucleotide variability is not restricted to *Drosophila*; recombination has been shown to affect variation in multiple organisms. For example, though the exact values of recombination rates are not known, recombination was shown to affect heterozygosity levels in humans and the house mouse (NACHMAN 1997; NACHMAN *et al.* 1998). In *Drosophila*, inversions are shown to reduce rates of recombination through the production of inviable gametes during crossing over. Evidence was provided when increased mortality was observed in inversion heterozygotes when compared to homozygote pairing in *Drosophila* (STURTEVANT and BEADLE 1936). Since then, numerous

studies have provided a link between inversions and rates of crossing over in the *Drosophila* genus. Another reason linked to the production of inviable gametes during meiosis is the partial chiasmata inhibition caused by inversion loops.

The production of inviable gametes from crossing over as a result of inversions is not always the end result in *Drosophila*. Instead, double crossing over can occur to influence the production of viable gametes during meiosis. It is also important to note that the double crossing over in inversion heterozygotes will likely eliminate the partial chiasmata inhibition which occurs as a result of inversion loops. For double crossovers to occur, an inversion must be large enough to allow for both exchanges to occur within the boundaries of the inverted region. In addition to crossing over, gene conversion is another process that allows genetic exchange among chromosomes in *Drosophila*. The α-amylase (*Amy*) gene found on the 3rd chromosome of *D. pseudoobscura* is present in various *D. pseudoobscura* gene arrangements and provides evidence of gene conversion in this species (POPADIĆ and ANDERSON 1995). Multiple copies of the *Amy* gene were sequenced in three different gene arrangements and gene conversion was detected among the duplicate genes. Results showed that gene conversion led to higher similarity among duplicated *Amy* genes than was expected given the age of the duplication. These results are consistent with the findings of gene conversion in the *Amy* locus in *D. melanogaster* (HICKEY *et al.* 1991). It is uncommon for gene conversion to influence recombination more than crossing over but there are documented instances of this pattern. Gene conversion and not crossing over was shown to cause a greater rate of genetic exchange in the rosy locus found within the central region of an inversion on the 3rd chromosome of *D. melanogaster* (CHOVNICK 1973; HILLIKER *et al.* 1994). An explanation was provided using the Counting Model, suggesting a one order of magnitude greater rate of recombination caused by gene conversion rather than crossing over

(NAVARRO *et al.* 1997). To determine the impact of gene conversion as a force in recombination, it is important to consider the size of the inversion. The effects of gene conversion on nucleotide variability is suggested to be more pronounced in smaller inversions than in larger inversions because the probability of two crossovers occurring in a small inversion is less than that for a large inversion (BETRAN *et al.* 1997; NAVARRO *et al.* 1997). The exact mechanisms governing the actions of gene conversion events on recombination rates remain unclear.

Drosophila inversion polymorphisms have been established as a potent model system to study genetic variation due to their effects on recombination. These effects have been extensively studied in *D. melanogaster* where a large portion of the variation in nucleotide diversity is attributed to the rate of recombination (AQUADRO *et al.* 1994). Three gene regions across an inversion (*In(3L)*) in *D. melanogaster* were employed to study nucleotide variation (HASSON and EANES 1996). Regions close to the inversion breakpoints showed low levels of genetic exchange compared to the regions in the central part of the inverted segment. Low genetic variation was demonstrated by the presence of fixed differences between arrangements and the absence of shared nucleotide polymorphisms. This display of low nucleotide variability at the inversion breakpoints were explained by lack of recombination between the gene arrangements in concert with genetic hitchhiking. This inversion was shown to emerge less than half a million years ago and therefore may not have been in existence long enough to share its mutations with the rest of the population. This explanation is detailed in the "origin of inversions" section.

Inversion breakpoints have also been a centerpiece for understanding patterns of nucleotide variability in *D. pseudoobscura*. Navarro *et al*. (2000) employed the coalescent approach to explore the effects of chromosomal inversions on nucleotide variability in *D. pseudoobscura* gene arrangements. Nucleotide variability was simulated for new and old

inversion polymorphisms under two separate conditions. First it was assumed that the frequency of the inversion remained constant across generations which would be the result of a balanced polymorphism. The second assumption was that the polymorphism has been recently established and is therefore not at equilibrium. The results of the simulations showed that the breakpoints of new inversions contained lower recombination rates than within the inverted segment. They also showed that once an inversion has reached equilibrium (old), the recombination rates at the breakpoints are greater than the recombination rates within the inverted segment. Overall, these results suggest that new inversions have the power to significantly reduce some variability in the population which is only regained once the inversion polymorphism is well on its way to achieving equilibrium. The sequencing of the *D. pseudoobscura* genomic sequence has provided a means to study patterns of nucleotide variability at precise locations relative to the breakpoints of the inverted chromosomal regions.

Linkage Disequilibrium in Inversions

Inversions may also be biologically informative because the reduced recombination rates are also expected to lead to significant linkage disequilibrium (LD). LD is the nonrandom associations among nucleotide sites. Levels of LD are likely to be highest where recombination is lowest and vice versa. One expects LD between adjacent sites because the probability of crossovers between these tightly linked sites is small, while more distant sites have a high probability of exchange occurring. This leads to the general prediction of a decline of LD with distance (SCHAEFFER and MILLER 1993). LD has been studied in a variety of *Drosophila* species but the most detailed studies have been carried out in the *D. melanogaster*, *D. subobscura*, *D. pavani*, *D. robusta* and *D. pseudoobscura* systems (KRIMBAS and POWELL 1992; SCHAEFFER *et*

al. 2003). Studies in these model systems have provided a major observation that regions not associated with an inversion rarely show evidence of significant levels of LD. The regions that tend to show significant LD are the regions within the inverted segment or close to the inverted segment. The involvement of inversions in determining nonrandom nucleotide associations was further supported when significant LD was shown in the *Amy* region found on the 3rd chromosome of *D. pseudoobscura* arrangements (AQUADRO *et al.* 1991). LD between different loci is proposed to be reduced as the genetic distances increase. However, this pattern was not consistently observed when eight gene regions on the third chromosome of *D. pseudoobscura* were sequenced and tested for significant inter-locus and intra-locus LD (SCHAEFFER *et al.* 2003). Another inconsistency observed in this study is the lack of significant LD between adjacent regions and high levels of LD between regions separated by larger distances. This finding refutes the hypothesis that regions within an inversion show significant LD. This observation may be due to the age of the region since LD decreases with age, but it could also be due to the inability of the data set to consistently detect significant LD. Being able to efficiently study patterns of LD is also important because significant levels of LD would likely suggest that selection is acting on these regions.

Thesis Overview

There are two experimental chapters in this thesis that address some of the key concepts in the literature that are yet to be clarified. The *Drosophila pseudoobscura* inversion polymorphism is one of the most frequently studied inversion systems and we have employed this system to perform our analyses. The majority of the conclusions about this inversion system have been drawn utilizing the standard phylogeny generated with cytological data. The

availability of the *D. pseudoobscura* genomic sequence has provided an updated medium through which we can explore this inversion polymorphism. We have generated what we consider a strong data set of *D. pseudoobscura* breakpoint regions and gene regions sequenced across a number of arrangements in 80-100 strains to study the biology of this inversion polymorphism.

The first experimental project (Chapter 2) addressed the phylogenetic relationships of the *D. pseudoobscura* third chromosome inversions. In this section, we investigated the evolutionary history of the different arrangements and estimated the ages of the different inversion events. We also investigated the origin of the arrangements and determined the validity of the previously generated cytological phylogeny. Finally, we addressed the question of ancestry and provide intriguing arguments on the future of this area of research.

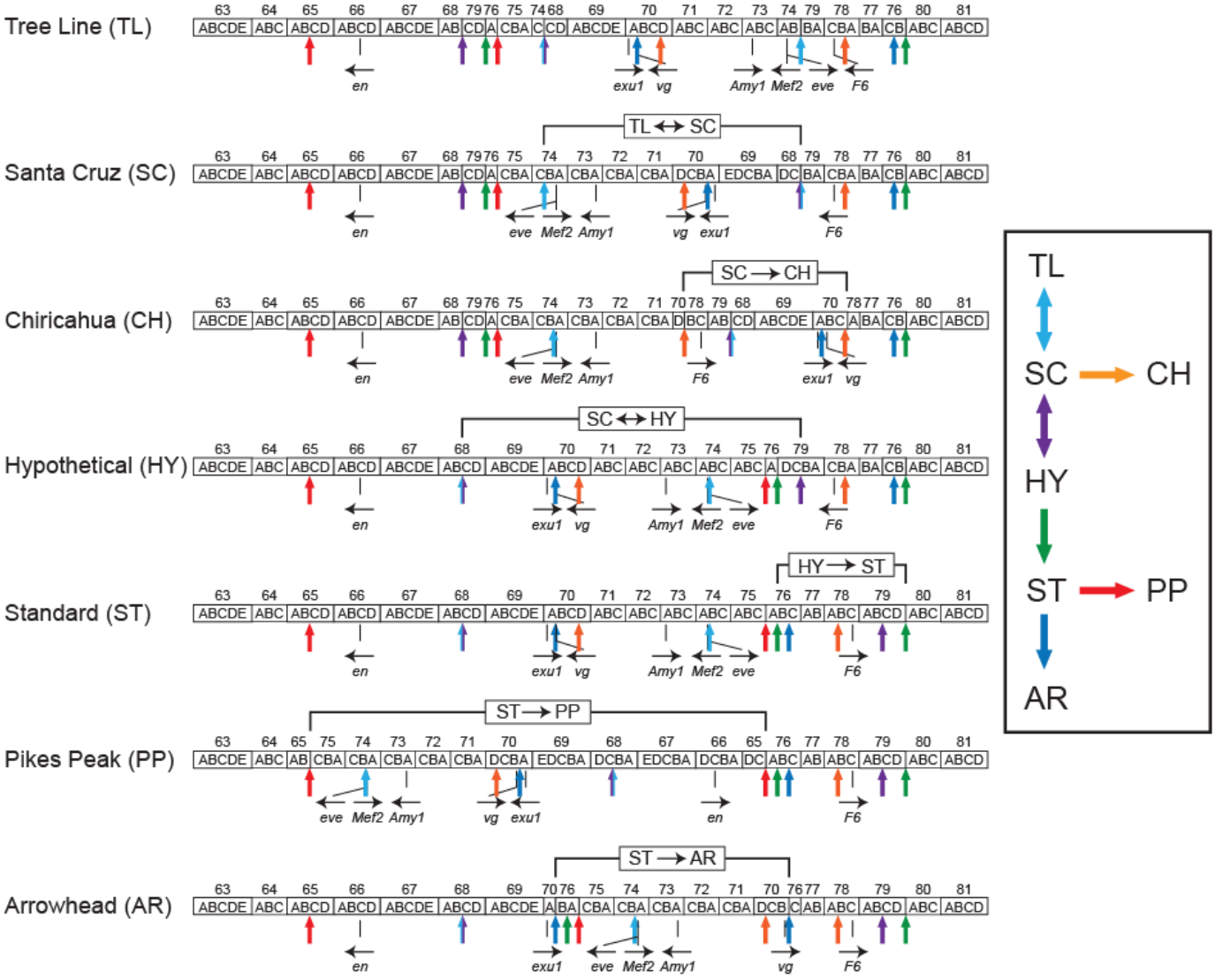
The second experimental study (Chapter 3) examined the population genetics of the *D. pseudoobscura* gene arrangements. Navarro *et al.* (1997, 2000) have developed population genetic theories about how inversions affect genetic exchange among inverted chromosomes and how this will affect levels of nucleotide diversity. Essentially, when an inversion is formed it has low levels of diversity because a new inversion is likely to occur on a single chromosome leading to a genetic bottleneck within this chromosomal type. This new inverted chromosome will recover variation through the accumulation of new nucleotide mutations over time and through allowed genetic exchanges among arrangements. Based on the predictions of Navarro et al (2000), we generated sequence data from regions close to the inversion breakpoints to determine patterns of nucleotide diversity in *D. pseudoobscura* arrangements. We studied patterns of nucleotide diversity in terms of the position of markers on the chromosome and in terms of the age of the gene arrangements. In addition, we investigated levels of LD within loci

and between loci. Breakpoint regions were highlighted in this study because they are proposed to be informative regarding rates of recombination. Recombination rates were shown to be directly proportional to rates of nucleotide diversity and significant LD is associated with low rates of recombination.

In chapter 4, we provide a brief summary of the conclusions drawn from our analyses and provide experimental insights for future experiments.

Figure 1. The order of breakpoints on the chromosomes of *Drosophila pseudoobscura* gene arrangements and proposed cytological phylogeny. The left side of the figure shows the different gene arrangements and the order of the breakpoints in each arrangement depicted by the colored arrows. The breakpoints appear in pairs and are color-coded according to the inversion phylogeny depicted on the right side of the figure. For example, red represents the PP arrangement. The inversion phylogeny on the right side of the figure is an abbreviated version of the proposed cytological phylogeny only showing the arrangements we study.

Figure 1.


Tree Line (TL)
Santa Cruz (SC)
TL ↔ SC
Chiricahua (CH)
SC → CH
Hypothetical (HY)
SC ↔ HY
Standard (ST)
HY → ST
Pikes Peak (PP)
ST → PP
Arrowhead (AR)
ST → AR
en
exu1
vg
Amy1
Mef2
eve
F6
TL
SC
CH
HY
ST
PP
AR

References

ANTONACCI, F., J. M. KIDD, T. MARQUES-BONET, M. VENTURA, P. SISWARA *et al.*, 2009 Characterization of six human disease-associated inversion polymorphisms. Human Molecular Genetics **18:** 2555-2566.

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*. Non-neutral evolution. Theories and molecular data.**:** 46-56.

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proceedings of the National Academy of Sciences of the United States of America **88:** 305-309.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

BETRAN, E., J. ROZAS, A. NAVARRO and A. BARBADILLA, 1997 The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics **146:** 89-99.

BOURQUE, G., P. A. PEVZNER and G. TESLER, 2004 Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons From Human, Mouse, and Rat Genomes. Genome Research **14:** 507-516.

BREHM, A., and C. B. KRIMBAS, 1991 Inversion polymorphism in *Drosophila obscura*. Journal of Heredity **82:** 110-117.

CHEN, J.-M., D. N. COOPER, C. FÉREC, H. KEHRER-SAWATZKI and G. P. PATRINOS, 2010 Genomic rearrangements in inherited disease and cancer. Seminars in Cancer Biology **20:** 222-233.

CHOVNICK, A., 1973 Gene conversion and transfer of genetic information within the inverted region of inversion heterozygotes. Genetics **75:** 123-131.

COLUZZI, M., A. SABATINI, A. DELLA TORRE, M. A. DI DECO and V. PETRARCA, 2002 A Polytene Chromosome Analysis of the *Anopheles gambiae* Species Complex. Science **298:** 1415-1418.

DOBZHANSKY, T., 1943 Genetics of natural populations. Genetics **28:** 162-186.

DOBZHANSKY, T., 1948 Genetics of natural populations. Genetics **33:** 588-602.

DOBZHANSKY, T., 1950 The genetics of natural populations. Genetics **35:** 288-302.

DOBZHANSKY, T., and A. H. STURTEVANT, 1938 Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics **23:** 28-64.

EMMENS, C. W., 1937 Salivary gland cytology of roughest[3] inversion and reinversion, and roughest[2]. Journal of Genetics **34:** 191-202.

ENGELS, W. R., and C. R. PRESTON, 1984 Formation of chromosome rearrangements by P factors in *Drosophila*. Genetics **107:** 657-678.

ESTER, A., V. FRANCESCA, E. JOSEP and B. JOAN, 2006 Genetic reproductive risk in inversion carriers. Fertility and sterility **85:** 661-666.

GREEN, M. M., 1980 Transposable elements in *Drosophila* and other Diptera. Annual Review of Genetics 14:109-120.

GRUNEBERG, H., 1936 A case of complete reversion of a chromosomal rearrangement in *Drosophila melanogaster*. Nature **138:** 508.

HASSON, E., and W. F. EANES, 1996 Contrasting histories of three gene regions associated with In(3L)Payne of *Drosophila melanogaster*. Genetics **144:** 1565-1575.

HICKEY, D. A., L. BALLY-CUIF, S. ABUKASHAWA, V. PAYANT and B. F. BENKEL, 1991 Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **88:** 1611-1615.

HILLIKER, A. J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.*, 1994 Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics **137:** 1019-1026.

HUDSON, R. R., and N. L. KAPLAN, 1995 Deleterious Background Selection With Recombination. Genetics **141:** 1605-1617.

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

KAUFMANN, B. P., 1942 Reversion from roughest to wild type in *Drosophila melanogaster*. Genetics **27:** 537-549.

KIM, Y., and W. STEPHAN, 2000 Joint Effects of Genetic Hitchhiking and Background Selection on Neutral Variation. Genetics **155:** 1415-1427.

KRIMBAS, C. B., and J. R. POWELL, 1992 *Drosophila* Inversion Polymorphism. Boca Raton, FL: CRC Press. p. 1-52.

LEMEUNIER, F. A. A., S., 1992 Inversion polymorphism in *Drosophila melanogaster*.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Molecular Biology and Evolution **13:** 261-277.

MORROW, D., 1970 A Cytological Analysis of the Position of Tree Line in the Phylogeny of Gene Arrangements of the Third Chromosome of *Drosophila pseudoobscura* pp. Cornell University, Ithaca, NY.

NACHMAN, M. W., 1997 Patterns of DNA Variability at X-Linked Loci in *Mus domesticus*. Genetics **147:** 1303-1316.

NACHMAN, M. W., V. L. BAUER, S. L. CROWELL and C. F. AQUADRO, 1998 DNA Variability and Recombination Rates at X-Linked Loci in Humans. Genetics **150:** 1133-1141.

NAVARRO, A., E. BETRAN, A. BARBADILLA and A. RUIZ, 1997 Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics **146:** 695-709.

OLVERA, O., J. R. POWELL, M. E. DE LA ROSA, V. M. SALCEDA, M. I. GASO *et al.*, 1979 Population genetics of Mexican *Drosophila*. Evolution **33:** 381-395.

PAINTER, T. S., 1933 A new method for the study of chromosome rearrangements and the plotting of chromosome maps. Science **78:** 585-586.

PEVZNER, P., and G. TESLER, 2003 Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America **100:** 7672-7677.

POPADIĆ, A., and W. W. ANDERSON, 1994 The history of a genetic system. Proceedings of the National Academy of Sciences of the United States of America **91:** 6819-6823.

POPADIĆ, A., and W. W. ANDERSON, 1995 Evidence for gene conversion in the amylase multigene family of *Drosophila pseudoobscura*. Molecular Biology and Evolution **12:** 564-572.

SCHAEFFER, S. W., M. P. GOETTING-MINESKY, M. KOVACEVIC, J. R. PEOPLES, J. L. GRAYBILL *et al.*, 2003 Evolutionary genomics of inversions in *Drosophila pseudoobscura*: Evidence for epistasis. Proceedings of the National Academy of Sciences of the United States of America **100:** 8319-8324.

STURTEVANT, A. H., 1917 Genetic factors affecting the strength of linkage in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **3:** 555-558.

STURTEVANT, A. H., 1926 A crossover reducer in *Drosophila melanogaster* due to inversion of a section of the third chromsome. Biologisches Zentralblatt **46:** 697-702.

STURTEVANT, A. H., and G. W. BEADLE, 1936 The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. Genetics **21:** 554-604.

WALLACE, B., 1966 Chromosomes, giant molecules, and evolution. New York, NY: W.W. Norton & Co.

WRIGHT, S., and T. DOBZHANSKY, 1946 Genetics of Natural Populations. XII. Experimental reproduction of some of the Changes Caused by Natural Selection in Certain Populations of *Drosophila pseudoobscura*. Genetics **31:** 125-156.

CHAPTER 2

EVOLUTIONARY HISTORY OF THE THIRD CHROMOSOME GENE ARRANGEMENTS OF *DROSOPHILA PSEUDOOBSCURA* INFERRED FROM INVERSION BREAKPOINTS

Research Article

Keywords: Chromosomal inversion, breakpoints, gene conversion, *Drosophila pseudoobscura*

Running Head: Phylogeny of *Drosophila* Inversions

Submitted to Molecular Biology and Evolution for Publication: MBE-10-0923

Andre G. Wallace, Donald Detweiler and Stephen W. Schaeffer

Department of Biology

The Pennsylvania State University

Abstract

The third chromosome of *Drosophila pseudoobscura* is polymorphic for numerous gene arrangements that form classical clines in North America. The polytene salivary chromosomes isolated from natural populations revealed changes in gene order that allowed the different gene arrangements to be linked together by paracentric inversions representing one of the first cases where genetic data was used to construct a phylogeny. Although the inversion phylogeny can be used to determine the relationships among the gene arrangements, the cytogenetic data is unable to infer the ancestral arrangement or the age of the different chromosome types. These are both important properties if one is to infer the evolutionary forces responsible for the spread and maintenance of the chromosomes. Here we employ the nucleotide sequences of 18 regions distributed across the third chromosome in 80 to 100 *D. pseudoobscura* strains to test whether five gene arrangements are of unique or multiple origin, what the ancestral arrangement was, and what are the ages of the different arrangements. Each strain carried one of six commonly found gene arrangements and the sequences were used to infer their evolutionary relationships. Breakpoint regions in the center of the chromosome supported monophyly of the gene arrangements while regions at the ends of the chromosome gave phylogenies that provided less support for monophyly of the chromosomes either because the individual markers did not have enough phylogenetically informative sites or genetic exchange scrambled information among the gene arrangements. A data set where the genetic markers were concatenated strongly supported a unique origin of the different gene arrangements. The inversion polymorphism of *D. pseudoobscura* is estimated to be about a million years old. We have also shown that the generated phylogeny is consistent with the cytological phylogeny of this species. In addition, the data presented here support Hypothetical as the ancestral arrangement. One of the youngest

arrangements, Arrowhead, has one of the highest population frequencies suggesting that selection has been responsible for its rapid increase.

Introduction

A number of broad hypotheses have been suggested for why chromosomal inversions emerge in natural populations (Kirkpatrick and Barton 2006). The first class of models for the spread of inversions is indirect effects due to recombination suppression (Sturtevant and Beadle 1936). Inversions are favored in these models either because inversions capture a chromosome relatively free of deleterious mutations (Nei et al. 1967; Ohta and Kojima 1968), capture a suite of beneficial genes (Charlesworth and Charlesworth 1973; Dobzhansky 1950; Wallace 1968) or capture a set of locally adapted genes (Kirkpatrick and Barton 2006). A second class of models suggests that underdominant inversions spread through a population by random genetic drift (Lande 1984). The third class of models posit that inversions spread in populations due to direct effects such as position effects (Sperlich and Pfreim 1986b) or meiotic drive (Novitski 1951; 1967). Finally, the last class of models suggests that inversion mutations spread because they capture a gene that is selectively sweeping through the population (Schaeffer and Aguade 2000).

Drosophila is an excellent group to test models of chromosomal evolution. Species within the genus *Drosophila* have significant levels of inversion polymorphisms within populations and many species groups have inversion differences that distinguish among the species (Sperlich and Pfreim 1986b). *D. melanogaster* has over 350 gene arrangements that have been detected in its populations with most chromosome arms having two major arrangements with cosmopolitan distribution (Lemeunier and Aulard 1992). Only nine different arrangements have been detected in *D. melanogaster*'s sibling species, *D. simulans*, although both species have a cosmopolitan distribution. *D. pseudoobscura* has over 30 different gene arrangements that are segregating on the third chromosome and two on its X chromosome (Dobzhansky and Sturtevant 1938; Sturtevant and Dobzhansky 1936a). D. subobscura has extensive inversion polymorphisms

on all of its major chromosomal arms (Krimbas and Powell 1992). Why one species can have inversions on all major chromosomal arms and its close relative does not is a curious puzzle about how new inversions originate. To better understand how inversions evolve in natural populations, it is important to understand the evolutionary history of the arrangements.

The inversions of the third chromosome of *D. pseudoobscura* were the first genetic data set used to generate an evolutionary phylogeny (Dobzhansky and Sturtevant 1938) (**Figure 1**). One chromosome was chosen as the arbitrary Standard arrangement and new arrangements were named based on the locality where the new chromosome type was collected. The major chromosomes were named Standard (ST), Arrowhead (AR), Pikes Peak (PP), Santa Cruz (SC), Chiricahua (CH), and Tree Line (TL). One gene arrangement in the central part of the phylogenetic network has never been collected in nature and was named the Hypothetical (HY) arrangement. By examining polytene chromosomes of *D. pseudoobscura*, it was possible to link the different gene arrangements together in an unrooted network. A central assumption of Dobzhansky and Sturtevant's analysis was that each inversion had a unique or monophyletic origin. Aquadro *et al.* (1991) used restriction fragment length polymorphisms in the Amylase gene region across different inversions to test the validity of the monophyly hypothesis, to estimate the age of the inversion polymorphism, and to verify the cytological phylogeny. They provided support for the cytogenetic phylogeny and confirmed the monophyletic origin hypothesis. The inversion polymorphism was estimated to be about 2 million years old. A minor weakness of the Aquadro *et al.* (1991) study was that it sampled a small set of restriction endonuclease site polymorphisms from a single genetic region of the chromosome in a small number of strains. Though there is strong evidence supporting a monophyletic origin of the inversions, there are arguments proposing that gene arrangements could be polyphyletic.

Transposable elements are suggested as a causative agent for chromosomal breakage resulting in polyphyletic origins of gene arrangements. In *Drosophila melanogaster*, P elements have been shown to initiate multiple breaks that were coincident (Engels and Preston 1984; Green 1980; Kidwell et al. 1977). It was also suggested that these breaks occur within a few hundred nucleotides of each other rather than at the exact same locations. Repetitive sequences have been observed at a pair of inversion breakpoints in *D. pseudoobscura* (Richards et al. 2005), but these repeats are not related to any known transposable element families. The presence of repetitive sequences raises the possibility that the *D. pseudoobscura* arrangements could have a polyphyletic origin. Phylogenetic analysis will determine whether the breakpoint repeats can serve as the substrate for multiple rearrangements.

The inversion phylogeny developed from the bands and puffs of the polytene chromosomes can not be used to infer the ancestral gene arrangement or its age. Powell (1992) described the current evidence for the ancestral arrangement that we summarize here. TL, SC, HY, and ST have all been proposed as the possible common ancestor of all other chromosomes. Four reasons are given for TL being the ancestral arrangement. First, the TL arrangement has a wide geographic distribution. The TL chromosome has been collected from the northern most populations in British Columbia, Canada, to the southern most populations in Mexico (Anderson et al. 1991; Olvera et al. 1979; Wallace 1966). Second, the TL arrangement has given rise to more new gene arrangements than any of the other arrangements (Olvera et al. 1979). The assumption is that the longer an arrangement exists the more time it has to mutate to form new arrangements. Third, the TL arrangement pairs best with the third chromosome homolog of the outgroup species *D. miranda* in comparison to either the ST or SC arrangements (Morrow 1970; Wallace 1966). Finally, Popadic and Anderson (1994) suggested that either the TL or the SC

arrangement was ancestral based on a phylogenetic analysis of the Amylase gene, a gene located in the middle of most gene arrangements. Their analysis concluded that more evidence supports SC as the ancestral arrangement (Popadic and Anderson 1994).

Dobzhansky and Sturtevant favored the HY arrangement as the ancestral arrangement (Dobzhansky 1944; Sturtevant and Dobzhansky 1936b). The central location of the HY arrangement in the cytogenetic phylogeny and the perceived similarity of the HY banding patterns with the *D. miranda* homolog led them to suggest that HY was the ancestral arrangement. Unfortunately, the HY arrangement has never been detected in samples from nature preventing experimental verification of the pairing in *D. pseudoobscura* and *D. miranda* hybrids. Bartolome and Charlesworth (2006) suggested that either the HY or ST arrangements were the ancestral arrangement based on comparative mapping of 25 *in situ* hybridization markers in *D. miranda* and *D. pseudoobscura* chromosomes as well as a phylogenetic analysis of nucleotide sequences of eight gene sequences obtained from Schaeffer *et al.* (2003).

Dobzhansky and Sturtevant (1938) suggested that the origin of the *D. pseudoobscura* inversion polymorphism occurred about a million years ago. Other early studies argued that the proposed age of the inversion polymorphism was underestimated by more than 6 million years (Stebbins 1945). Molecular data have provided strong evidence that the inversion polymorphism is about 2 million years old (Aquadro et al. 1991).

The phylogenetic analyses of nucleotide sequences performed to date (Aquadro et al. 1991; Bartolome and Charlesworth 2006b; Popadić and Anderson 1994) may not reflect the true relationships among the *D. pseudoobscura* gene arrangements. Most of the genes examined in these studies are located within the central region of the inverted segments where genetic exchange may occur more frequently (Navarro et al. 1997; Schaeffer et al. 2003). Genetic

exchange among arrangements may mask the true relationships between the arrangements. Regions near inversion breakpoints may be a better indicator of true phylogenetic relationships among the chromosomes because genetic flux among arrangements is likely to be lowest near the breakpoints (Navarro et al. 1997).

We have sequenced inversion breakpoints from the *D. pseudoobscura* third chromosome in flies that carry different gene arrangements: 1) to test whether the gene arrangements are of monophyletic or polyphyletic origin; 2) to determine the common ancestral arrangement; (3) to estimate the age of each inversion; and 4) to determine if the molecular phylogeny of the gene arrangements agrees with the cytogenetic phylogeny. Genes within inversion breakpoints that were previously sequenced (Schaeffer et al. 2003) were also included in our analysis. Here, we present a molecular phylogenetic analysis of 18 genetic markers that show that gene arrangements of *D. pseudoobscura* are monophyletic and that the gene tree agrees with the cytogenetic phylogeny except for the Pikes Peak arrangement where frequent gene conversion events have altered relationships. We present a linkage chain analysis (Bhutkar et al. 2008) of gene adjacency information for the *D. pseudoobscura* third chromosome that indicates that the HY arrangement is ancestral, which agrees with the gene tree data. In addition, we used a molecular clock analysis of the phylogeny to estimate the ages of the different arrangements. These data suggest that the gene arrangement polymorphism of *D. pseudoobscura* is at least one million years old.

Materials and Methods

Fly strain and DNA extraction. In 1998, isofemale lines of *Drosophila pseudoobscura* were collected by S.W. Schaeffer and W.W. Anderson (University of Georgia) from four separate locations in the southwestern United States (Schaeffer et al. 2003). The localities include: Davis Mountain, TX; James Reserve, CA; Mount Saint Helena, CA; and Kaibab National Forest, AZ. The third chromosome was made homozygous from the isofemale strains using balancer crosses (Dobzhansky and Queal 1938). Genomic DNA was extracted from 101 isochromosomal lines of *D. pseudoobscura* with a single fly DNA extraction protocol (Gloor and Engels 1992). In the data set, 78 strains had sequences for all loci and were used to generate the concatenated data set. Between 91 and 101 strains were sequenced at individual breakpoint regions and were used to construct single-marker phylogenies.

PCR Primer Design. This study examined the molecular evolution of regions near six pairs of inversion breakpoints in *D. pseudoobscura*, Arrowhead (AR), Standard (ST), Pikes Peak (PP), Santa Cruz (SC), Chiricahua (CH), and Tree Line (TL). We named the breakpoint regions based on the assumption that the Hypothetical arrangement was the ancestral gene order (Bartolome and Charlesworth 2006b; Dobzhansky and Epling 1944), an assumption that is tested and verified in this study. The name consists of a proximal or distal designation (p or d) for each end of the inversion as well as the ancestral and derived gene arrangement. For example, pSTAR would be the proximal breakpoint for the inversion that converted Standard into Arrowhead. We developed a single region for the pHYSC and dSCTL breakpoints because the cytogenetic map indicates that the HYSC and SCTL inversions had one breakpoint in common (**Figure 1**). The pHYSC broke between 68B and 68C, while dSCTL split between 68C and 79B where 68C was the cytological region in common to the two breakpoints. This reduced the number of breakpoint

regions to 11. The vestigial locus was used as the marker for the dSTAR inversion breakpoint because this locus was mapped within 20 kb of the distal break leaving a remainder of 10 regions where new pairs of PCR primers needed to be developed (Richards et al. 2005; Schaeffer et al. 2003). PrimerSelect (DNASTAR, Madison, WI) was used as a guide to locate pairs of oligonucleotides for PCR amplification of the 10 breakpoint regions (**Table 1** and **Figure 1**). These primers were used to amplify 350-900 base pair fragments from 10 regions near the distal and proximal breakpoints of six *D. pseudoobscura* inversions. It should be clear that the regions amplified do not straddle the breakpoint, but are near the mapped locations of the chromosomal breaks. The approximate locations of the six pairs of breakpoints were inferred from the genome sequence of the third chromosome of *D. pseudoobscura* and the correlation of the nucleotide map with the polytene map (Schaeffer et al. 2008). We wanted to sequence regions that were likely to be less constrained so we used the comparison of the *D. persimilis* and *D. pseudoobscura* genome sequences from the third chromosome as a guide (Noor et al. 2007). In addition to the 11 breakpoint regions, oligonucleotide primers used to amplify the seven loci examined by Schaeffer *et al* (2003) were used to sequence the fly strains used in this study. These seven regions are located further away from the inversion breakpoints.

PCR Amplification and Nucleotide Sequencing. Each PCR reaction was a total of 50µl in volume and contained 10x Polymerase buffer with $MgCl_2$, both forward and reverse primers, dNTP mix, Taq DNA polymerase and the isolated target DNA. 5ul of the finished PCR product was run on 1% agarose gel to verify the expected product length. After verification of the fragment size, the remainder of the PCR product was treated with ExoSAP-IT (USB, Affymetrix Corporation) to remove contaminants such as unused dNTPs and primers remaining in the PCR product. The amplified fragments were then sequenced at the Penn State University Nucleic Acid

Facility (University Park, PA) using an ABI Hitachi 3730XL DNA Analyzer. Samples were sequenced in both forward and reverse directions. The forward and reverse sequences were assembled and conflicts between the two reads were resolved using the SEQMANII program (DNASTAR, Madison, WI). The sequences used for analysis in this study were submitted to GenBank (Accession numbers XXXXXX-XXXXXX).

Nucleotide Sequence Alignment and Phylogenetic Analysis. Nucleotide sequences for each breakpoint and non-breakpoint region were aligned using the MEGALIGN program (DNASTAR, Madison, WI). Alignments were performed manually and inspected visually to ensure that indels were scored consistently among the different sequences. Complete data sets containing aligned sequences were imported into MEGA 4.0 (Kumar et al. 2008) where phylogenetic analysis was performed. Phylogenetic trees were constructed by the Neighbor Joining method and confidence values for clades were generated by 1000 bootstrap replicates (MEGA 4). These values are depicted on the branches of the phylogenies as percentages. Indels were completely removed from the analysis using the complete deletion option within MEGA.

Linkage Chain Analysis. We used linkage chain analysis as an additional tool to infer the ancestral arrangement for the third chromosome. Linkage chain analysis as described in Bhutkar *et al.* (2008) was used to identify regions of the *D. pseudoobscura* third chromosome that had lineage specific breaks that correlated with the cytological position of the rearrangement breakpoints. The Arrowhead arrangement was sequenced by Richards *et al.* (2005) in the *D. pseudoobscura* genome project. The Arrowhead arrangement is a derived arrangement within the *D. pseudoobscura* third chromosome cytological phylogeny. Therefore, inferred synteny breaks between the *D. pseudoobscura* Arrowhead chromosome and other *Drosophila* species reflect breakpoints from the common ancestral arrangement and the derived Arrowhead arrangement. If

either the Tree Line or Santa Cruz arrangements are the common ancestral gene order, then the TL or SC breakpoint locations will be reflected in the Arrowhead arrangement and can be recovered from synteny breaks inferred from comparisons of AR with outgroup species. Synteny breaks near the inferred locations of the TL or SC cytological breaks (SC: 68B-68C and 79B-79C; TL 68B-79B and 74A-74B) on the nucleotide sequence were tested as candidates for the ancestral gene order (**Figure 1**). The ancestral arrangement will be the last arrangement whose synteny breaks are joined together in linkage chain analysis.

Results

Origin of the *D. pseudoobscura* Gene Arrangements. We tested the hypothesis that the gene arrangements of *D. pseudoobscura* were generated by unique versus multiple events by constructing phylogenies of each breakpoint separately and of a variety of concatenated data sets. The Amylase restriction map phylogeny of the *D. pseudoobscura* arrangements show that the origin of each gene arrangement resulted from a unique event (Aquadro et al. 1991; Dobzhansky and Sturtevant 1938). Neighbor joining trees were constructed with individual markers to examine phylogenetic relationships (**Figure S1**). The majority of the clades in these trees were not well supported in the bootstrap replicates. To better observe phylogenetic relationships, we concatenated all 11 breakpoint regions along with the eight gene regions previously sequenced by Schaeffer *et al.* (2003). The concatenated files were used to construct phylogenetic trees based on different partitions of the data set. The analyses included: 1) All loci (**Figure 2**), 2) Proximal, distal and central loci separately (**Figure 3**), 3) Breakpoint regions only (**Figure 4A**), 4) Non-breakpoint regions only (**Figure 4B**). The proximal, distal and central markers were mapped based on their position on the AR chromosome (**Figure 1**), although this designation is a bit arbitrary because gene region change their context through the inversion phylogeny. The phylogeny generated for all the concatenated data (**Figure 2**) showed monophyly of all arrangements except for PP where one PP strain (Davis Mountains strain DM1068) clustered with the AR clade. It is also important to note that the CH strains appear paraphyletic due to the presence of the SC strain within the CH clade. The CH and SC arrangements have been shown to be extremely close in age and previous phylogenetic studies have provided support for their close genetic relationship (Aquadro et al. 1991). Markers in the proximal segment of the AR arrangement (**Figure 3A**) failed to show any significantly supported clades within the

phylogenetic tree. The phylogenetic tree constructed using markers in the central region of the AR chromosome (**Figure 3B**) showed complete monophyly of all arrangements, except for the PP arrangement where the PP strain DM1068 clustered within the AR clade. In the phylogenetic tree generated with distal markers (**Figure 3C**), none of the arrangements formed complete monophyletic groups.

We tested whether breakpoint regions have a stronger phylogenetic signal than non-breakpoint loci. Regions near breakpoints are expected to have lower levels of genetic flux compared to regions away from the breakpoints (Navarro et al. 1997) and would be expected to reflect the true relationships of the arrangements. The phylogenies generated with the breakpoint regions would be monophyletic except for six strains, one CH strain (Kaibab strain KB888) and five PP strains. One PP strain (Davis Mountains strain DM1068) clusters within the ST clade and four PP strains (James Reserve strain JR83; and Davis Mountains strains DM1038, DM1041, and DM1053) cluster with the TL clade **(Figure 4A**). The phylogenetic tree for the non-breakpoint loci (**Figure 4B**) also shows monophyly of the arrangements.

Concordance of the Molecular Phylogenies with the Cytogenetic Phylogeny. The cytogenetic phylogeny for *Drosophila pseudoobscura* was first generated by Dobzhansky and Sturtevant (1938). They inferred that each of the gene arrangements can be linked by single inversion events based on the number of loops formed in heterozygotes from gene arrangement pairs. A short revised version of this phylogeny is shown in **Figure 1**. We examined the phylogenies constructed with the individual markers to determine if the phylogeny is consistent with the proposed cytogenetic phylogeny.

The individual marker phylogenies that showed complete monophyly of the arrangements are not consistent with the proposed cytological phylogeny (**Figure S1**). These molecular

phylogenies show PP being more closely related to TL and CH rather than to ST, PP's ancestor in the cytological tree. The phylogeny generated using the dSCTL and dSCCH markers show the PP arrangement clustering with the TL arrangement, which is not consistent with the proposed cytological phylogeny. For the dSTPP marker, PP is most closely related to the CH arrangement which is also inconsistent with the cytological phylogeny. All other trees showed weak monophyly of gene arrangements and provide no basis to evaluate the concordance with the cytogenetic phylogeny. All the other markers suggest that these *D. pseudoobscura* arrangements have a polyphyletic origin which is not consistent with the cytological phylogeny.

To ensure that our findings are not biased due to a possible lack of resolving power of the individual markers, we analyzed the concatenated data files to see if the phylogenies agree with the proposed cytological phylogenies. Among the trees that showed monophyly of the arrangements, two appeared to be consistent with the cytological phylogeny. The phylogeny constructed with the non-breakpoint data set (**Figure 4B**) and the complete concatenated data set (**Figure 2**) both agree with the cytogenetic phylogeny. One tree that did not strictly agree with the cytological phylogeny was the one generated with concatenated breakpoint regions (**Figure 4A**). In this tree, one and four PP strains clustered with the AR and the TL arrangements, respectively. Other than these exceptions, the breakpoint tree was concordant with the cytogenetic tree.

Linkage Chain Analysis of *D. pseudoobscura* breakpoints. The analysis of gene order data in the regions for the Standard to Arrowhead inversion and the Hypothetical to Standard inversion have interspecific breakpoints in the genome sequence of the Arrowhead chromosome that are correlated with the position of the cytological breakpoints for these mutations (see Richards *et al.* 2005 for the mapping of the Standard to Arrowhead inversion breakpoints). In

both cases, we found pairs of breaks that mapped to the locations of cytological breaks (**Figure S2**). Inter-specific breakpoints in the Arrowhead chromosome near the expected cytological locations of either the Santa Cruz or the Tree Line breakpoints failed to be linked in a single *D. pseudoobscura* chain of breakpoints (S. W. Schaeffer, unpublished data). These data suggest that the HY arrangement and not SC or TL is the common ancestral arrangement of the *D. pseudoobscura* gene arrangements.

Age of the *D. pseudoobscura* Inversion System. We used the inferred divergence time between *D. miranda* and *D. pseudoobscura* of two million years (Aquadro et al. 1991; Babcock and Anderson 1996) to estimate the ages of the different gene arrangements. We used a molecular clock in MEGA (Takezaki et al. 1995) and the phylogenetic tree with the concatenated data set of markers (**Figure 5**). The Hypothetical arrangement is estimated to diverge from *D. miranda* about 1.12 million years ago (mya). The time of origin of the five major arrangements is: SC =1.06 mya; ST = 0.95 mya; PP = 0.95 mya; AR = 0.57 mya; TL 0.91 mya; and CH = 0.58 mya.

Discussion

Phylogenies and the Origin of the *D. pseudoobscura* Gene Arrangements. A fundamental assumption made by Dobzhansky and Sturtevant (1938) when they constructed the phylogeny of the *D. pseudoobscura* gene arrangements was that each chromosome was of unique origin. Aquadro *et al.* (1991) used restriction map variation around the Amylase gene in a small number of strains to show that *D. pseudoobscura* arrangements result from single and not multiple events (Aquadro et al. 1991; Dobzhansky and Sturtevant 1938). Here, we used a study of nucleotide sequence markers distributed across the third chromosome to test whether the gene arrangements are of unique origin. Theoretical models have shown that recombination is reduced near inversion breakpoints (Navarro et al. 2000; Navarro et al. 1997). Recombination acts as a homogenizing force that can mask the true relationships among different alleles. Thus, nucleotide sequences near breakpoints were chosen as candidate regions to uncover the relationships among the *D. pseudoobscura* arrangements.

We employed breakpoint regions as well as previously studied gene regions (Schaeffer et al. 2003) to more rigorously test the monophyly hypothesis with an examination of nucleotide sequence phylogenies of five *D. pseudoobscura* gene arrangements (TL, AR, ST, PP, CH). We included gene sequences from one Santa Cruz strain to help polarize the events in the tree. Single markers generated phylogenies that failed to show monophyly of breakpoint regions in the proximal and distal segments of the chromosome (**Figure S1**). Breakpoint regions in the central region of the chromosome indicated that the gene arrangements were likely monophyletic. This pattern was verified when the phylogenies were generated using concatenated sequences of the proximal, distal and central breakpoint regions (**Figure 3**). The central regions are likely to be monophyletic because suppression of recombination in this system of overlapping inversions is a

strong isolating force. As a result, mutations that accumulate on one chromosomal background are not likely to spread to other gene arrangements. Phylogenies generated with the individual regions did not support the unique origin of inversions because the number of phylogenetically informative nucleotide sites is low.

When we concatenated all genetic markers together, the phylogeny strongly supports a unique origin of the *D. pseudoobscura* gene arrangements (**Figure 2**). The single exception is the Pikes Peak arrangement where one PP strain clusters with the Arrowhead clade. Of all the arrangements, strains with the CH arrangement tended to cluster together more frequently in the majority of individual trees. This pattern is likely observed because the CH arrangement is one of the older arrangements and has accumulated a significant number of nucleotide differences from the other arrangements. There was one strain of CH (KB888) that was more distantly related to other Chiricahua strains. The most likely explanation for why PP DM1068 and KB888 had odd placements in the phylogenies is gene conversion (Schaeffer and Anderson 2005). Gene conversion occurs at double strand breaks introduced during meiosis and causes 200 to 300 bp tracts of sequence to be exchanged between arrangements (Betran et al. 1997; Schaeffer and Anderson 2005).

The PP arrangement shows the most evidence for gene conversion. The PP inversion covers a significantly larger portion of the chromosome than any other inversion so the opportunity for gene conversion is higher in PP than any other arrangement. The data indicate that PP has had frequent exchanges with other members of the SC phylad as well as with the AR arrangement. The size of PP and its co-occurrence with the AR and SC phylad arrangements provide the opportunity for genetic exchange with other arrangements leading to the development of odd relationships. For example, the gene conversion detection approach used by

Betran *et al.* (1997) found evidence that the PP DM1068 strain received tracts of sequence from an Arrowhead strain. The AR/PP heterozygotes could reach frequencies as high as 30% in Texas based on the arrangement frequencies and the assumption of Hardy-Weinberg equilibrium. This would provide sufficient opportunity for gene conversion to take place between these arrangements.

Concordance of the Molecular Genetic and Cytogenetic Phylogenies. Here the constructed phylogenies of six *D. pseudoobscura* arrangements using molecular markers across the polymorphic third chromosome were used to determine if the molecular phylogenies were in concordance with the cytogenetic phylogenies. A schematic of the proposed cytological phylogeny is provided in **Figure 1**. The trees constructed from each of the 18 regions of the third chromosome failed to show strong concordance with the cytogenetic phylogeny. This is not surprising given that these regions had modest numbers of phylogenetically informative SNPs.

The tree constructed from the concatenated data of all 18 regions agreed with the cytogenetic phylogeny (**Figure 2**). We can infer the common ancestral arrangement from this tree to infer the expected gene arrangements at the nodes coupled with the cytogenetic information (**Figure 5**). These data indicate that the Hypothetical arrangement is the ancestral arrangement because the common ancestral node occurs at the split between the Standard and Santa Cruz phylads, the expected position of the Hypothetical arrangement. This is supported by the linkage chain analysis where interspecific breakpoints between the Arrowhead genome sequence and outgroup species could map the breakpoints for the Standard to Arrowhead and the Hypothetical to Standard inversion events but could not find evidence for Santa Cruz to Hypothetical or Tree Line to Santa Cruz inversions breakpoints (**Figure S2**). These data are consistent with the proposal of Bartolome and Charlesworth (2006) who provided supported for

either ST or HY as the ancestral arrangement. Our data indicate that HY and not ST is the common ancestral gene arrangement.

Age of *Drosophila pseudoobscura* arrangements. The phylogenetic tree with all concatenated data was selected to estimate the age of the arrangements because it is the most resolved phylogeny. According to the molecular clock, the HY arrangement is about 1.1 million years old, the ST arrangement is less than a million years old, and the TL arrangement about a million years old (**Figure 5**). One of the interesting aspects that emerges from the phylogeny is that the age of the inversions is not correlated with the observed population frequencies (Dobzhansky and Epling 1944). If we were to use population frequency to estimate the age of the different gene arrangements, we might choose either Arrowhead or Tree Line. Tree Line is found at high frequencies in Mexico (Olvera et al. 1979) and the levels of variation found within the Tree Line arrangement are consistent with its old age. This is also consistent with the large number of different arrangements that trace their ancestry to Tree Line. Arrowhead, on the other hand, is observed at high frequencies in the southwestern United States reaching frequencies as high as 95% in Arizona. The Arrowhead arrangement is estimated to be 570,000 years old or is one of the younger arrangements. These data support a model where AR has rapidly increased in frequency through the action of natural selection. Recent numerical analysis have suggested that this is due to local adaptation which would indicate that indirect effects associated with recombination suppression are driving the spread of new inversions (Schaeffer et al. 2008). Thus, the data presented here suggest that frequency is not a good proxy for inversion age.

The estimates of the ages of the *D. pseudoobscura* gene arrangements will now allow us to use this system to empirically test models of breakpoint evolution (Navarro et al. 2000; Navarro et al. 1997). These models make predictions about the levels of variation predicted at

inversion breakpoints given the age of the inversions. Young inversions are expected to have lower levels of variation at breakpoints than older inversions. We can now examine these predictions based on our time estimates.

Breakpoints as Phylogenetic Markers. We chose to use breakpoint markers because we expected them to show little evidence for recombination. Our data showed that some recombination in the form of gene conversion does occur. This genetic exchange leads to clustering of strains independent of the gene arrangement that they carry. Overall, breakpoint regions provided a strong phylogenetic signal to determine the relationships among the *D. pseudoobscura* arrangements, but some genetic flux occurs close to inversion breakpoints.

Table 1. Primer Sequences Used for PCR Amplification

Primer Name	Primer Sequence	Coordinate Interval	Length (bp)	Cytological Position
pSTPP_f	GAT ACC ACT CGG CAA GCA GAA G	2,292,730…2,293,129	354	64C
pSTPP_r	CGC CTC AGT TAA TTA GCC CAC AAA			
pHYSC_f	TGG TGT TGA GTA TCT GCC GTG GTT	6,432,192…6,432,612	376	68C
pHYSC_r	CTG CTG CCG CTG CTC CTA TCA			
pSTAR_f	CCT GAT ACC CAC GGA GTC TTC	8,900,335…8,900,844	468	76B
pSTAR_r	TCG CTA CAG GGA TCA GGT TTT			
pHYST_f	CTT ATT CCC GCC TCT TGT GTA GC	9,140,888..9,141,693	806	76B
pHYST_r	GAC GGC CCT CAG ACG ATA GTT G			
dSTPP_f	ATC GGT ACA ACA GCC AGG GAC AAC	9,832,232…9,832,852	573	75B
dSTPP_r	ACT TCG TGG GAT CGC TGG CAT AAT			
dSCTL_f	ATG GCG ATG GAG TCC TCT GTC TAT	10,830,454...10,830,904	404	74B
dSCTL_r	ACT GGC GCC ATG TCT CTG TCT CG			
pSCCH_f	AAC CGG CAT ACA CCC TCA TTC	14,259,585…14,260,006	377	70C
pSCCH_r	GTT GCG CAT TAT TTA TTC CCT GTA			
dSCCH_f	TCC GGA GAT CGC AAA ACT GTC G	15426271…15426670	379	77B
dSCCH_r	TAT GCG CTG CTT CTG ATG CTT GAT			
dHYSC_f	GAG CCC GGG CCA GGT CCA T	17,444,472…17,444,876	362	79C
dHYSC_r	TAT CGT GCG TTG TGC GTA ATC AGC			
dHYST_f	ACA AGA TCC GGG GTA TTA	17,705,213..17,706,152	898	79D/80A

dHYST_r	CTG TTC CGG GTA GAT GTA TTC GTA

The abbreviated name represents the location (p, proximal or d, distal) of the breakpoint on the chromosome, and the last four letters represent the two arrangements involved in the inversion. The first two letters are the ancestral arrangement and the last two letters are the derived arrangement. For example, the STPP notation is for the breakpoints that converted the ancestral Standard arrangement into the derived Pikes Peak arrangement. Primer names are the abbreviated region names with the addition of "f" and "r" for forward or reverse. The breakpoint regions sequenced included less than 50% coding nucleotide sequence for the pHYST marker and non-coding sequences for all other markers. The coordinates are the location of the genetic marker in the genome strain (Richards et al. 2005), which carries the Arrowhead arrangement.

Figures

Figure 1. Proposed inversion phylogeny of *Drosophila pseudoobscura* inversion arrangements is shown on the lower left portion of the figure. The top portion of the figure is a representation of how breakpoints are organized on the chromosome of various *D. pseudoobscura* gene arrangements. The colored arrows represent the breakpoint regions and the figure legend matches the colors with the inversions breakpoints. The arrows are in pairs with the left one representing the proximal arrangement and the right one the distal arrangement. For example, red represents the PP arrangement. The AR chromosome is shown in more details because it is the genomic strain and it formed the basis for our analysis.

Figure 1

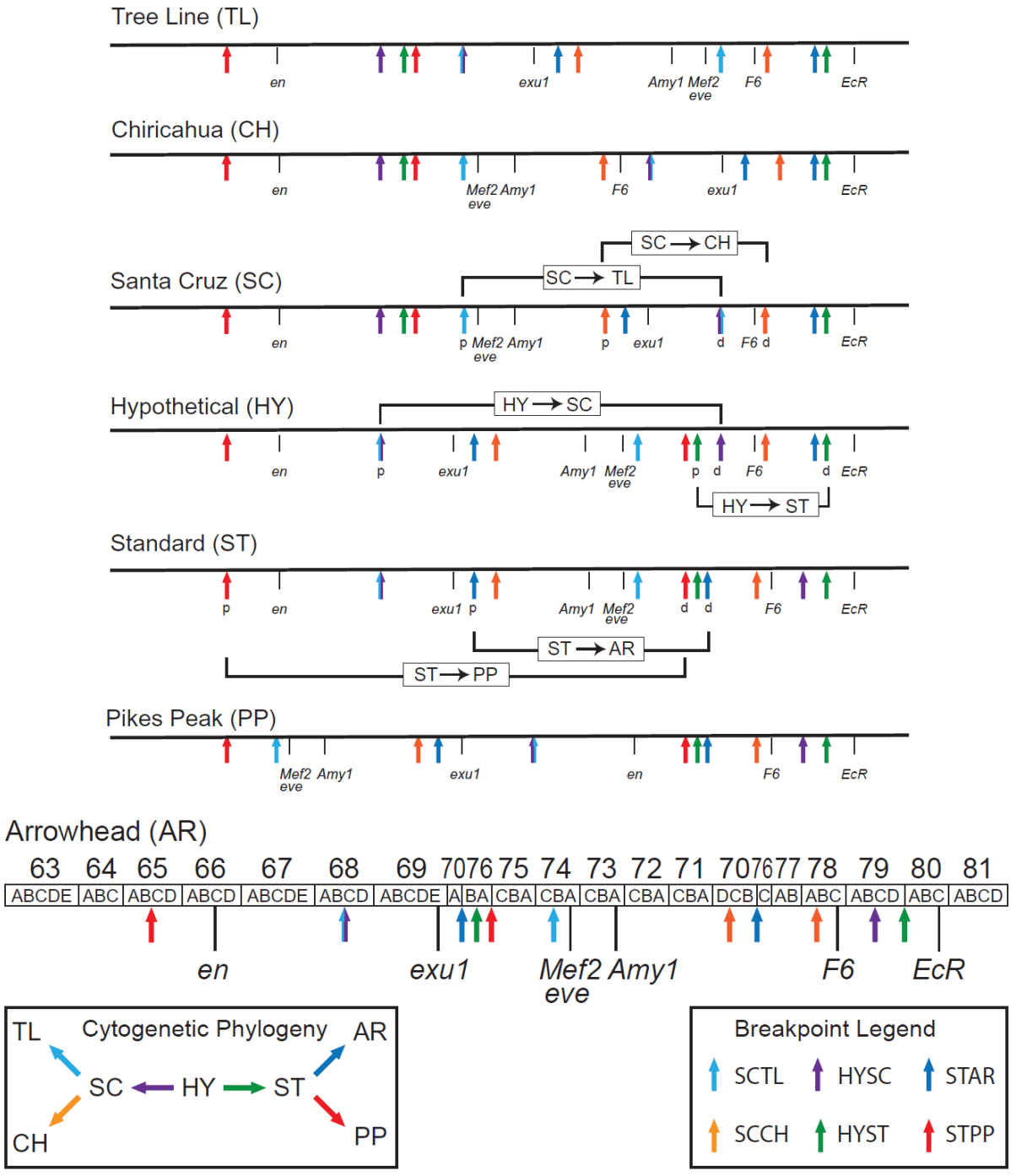

Tree Line (TL)
en
exu1
Amy1 Mef2 eve
F6
EcR
Chiricahua (CH)
en
Mef2 Amy1 eve
F6
exu1
EcR
SC → CH
SC → TL
Santa Cruz (SC)
en
p Mef2 Amy1 eve
p
exu1
d
F6 d
EcR
HY → SC
Hypothetical (HY)
en
p
exu1
Amy1 Mef2 eve
p d
F6
d EcR
HY → ST
Standard (ST)
p
en
exu1 p
Amy1 Mef2 eve
d d
F6
EcR
ST → AR
ST → PP
Pikes Peak (PP)
Mef2 Amy1 eve
exu1
en
F6
EcR
Arrowhead (AR)
63 64 65 66 67 68 69 7076 75 74 73 72 71 707677 78 79 80 81
ABCDE ABC ABCD ABCD ABCDE ABCD ABCDE A BA CBA CBA CBA CBA CBA DCB C AB ABC ABCD ABC ABCD
en
exu1
Mef2 Amy1 eve
F6
EcR
TL
Cytogenetic Phylogeny
AR
SC
HY
ST
CH
PP
Breakpoint Legend
SCTL
HYSC
STAR
SCCH
HYST
STPP

Figure 2. Phylogenetic relationships of *D. pseudoobscura* gene arrangements generated from the concatenation of 79 taxa across all 18 markers studied. Evolutionary relationship was inferred using the Neighbor-Joining method (Saitou and Nei 1987) and evolutionary distances were calculated using the Maximum Composite Likelihood method (Tamura et al. 2004). The phylogenetic trees were created in MEGA4 (Tamura et al. 2007) and bootstrapped 1000 times to test for statistical significance. The percentages from the bootstrap replicates are depicted on the branches. Only bootstrap values of 70% or greater are shown. The *D. miranda* sequence is rooted on the midpoint as an outgroup. The numbers in parentheses represent the number of taxa clustering on a particular branch.

Figure 2

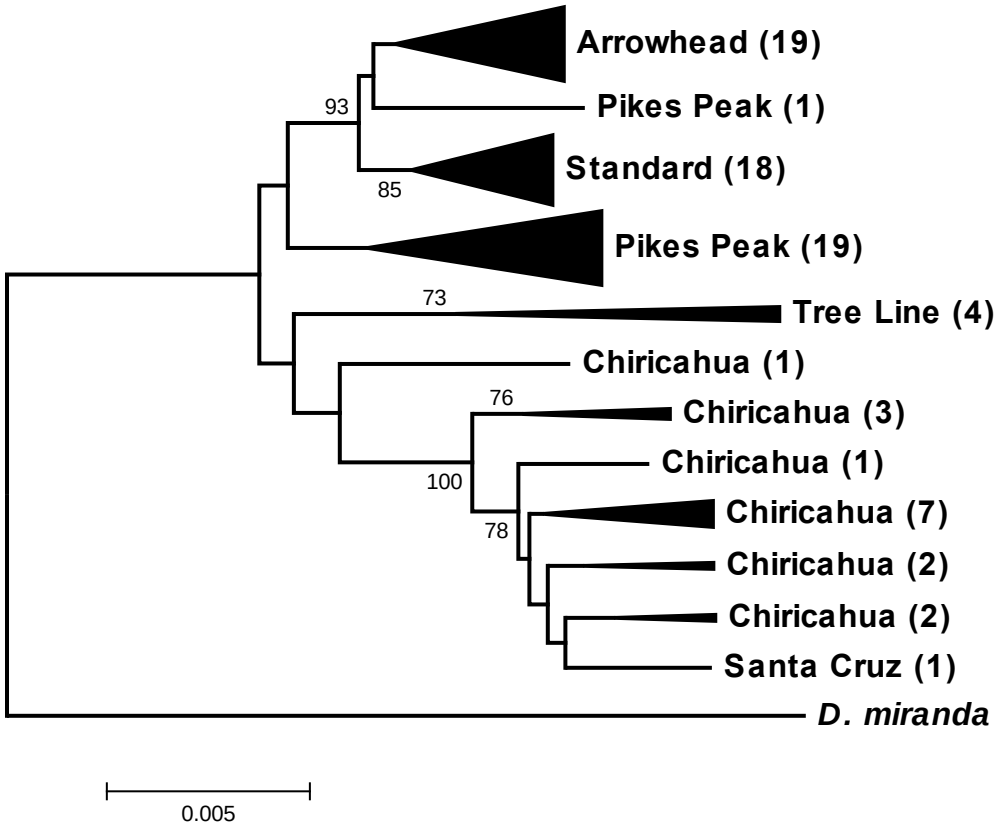
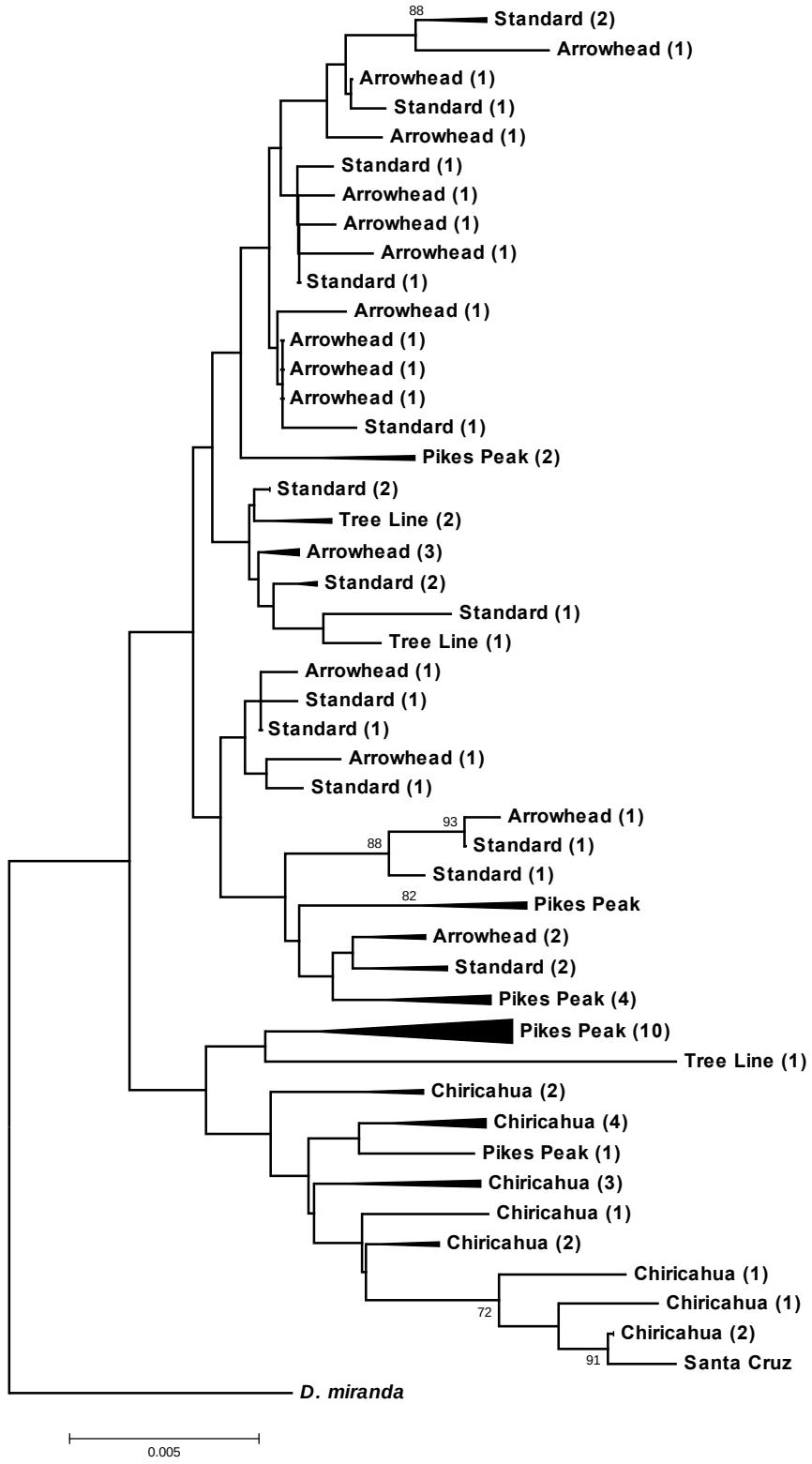

Arrowhead (19)
93
Pikes Peak (1)
Standard (18)
85
Pikes Peak (19)
73
Tree Line (4)
Chiricahua (1)
76
Chiricahua (3)
Chiricahua (1)
100
Chiricahua (7)
78
Chiricahua (2)
Chiricahua (2)
Santa Cruz (1)
D. miranda
0.005

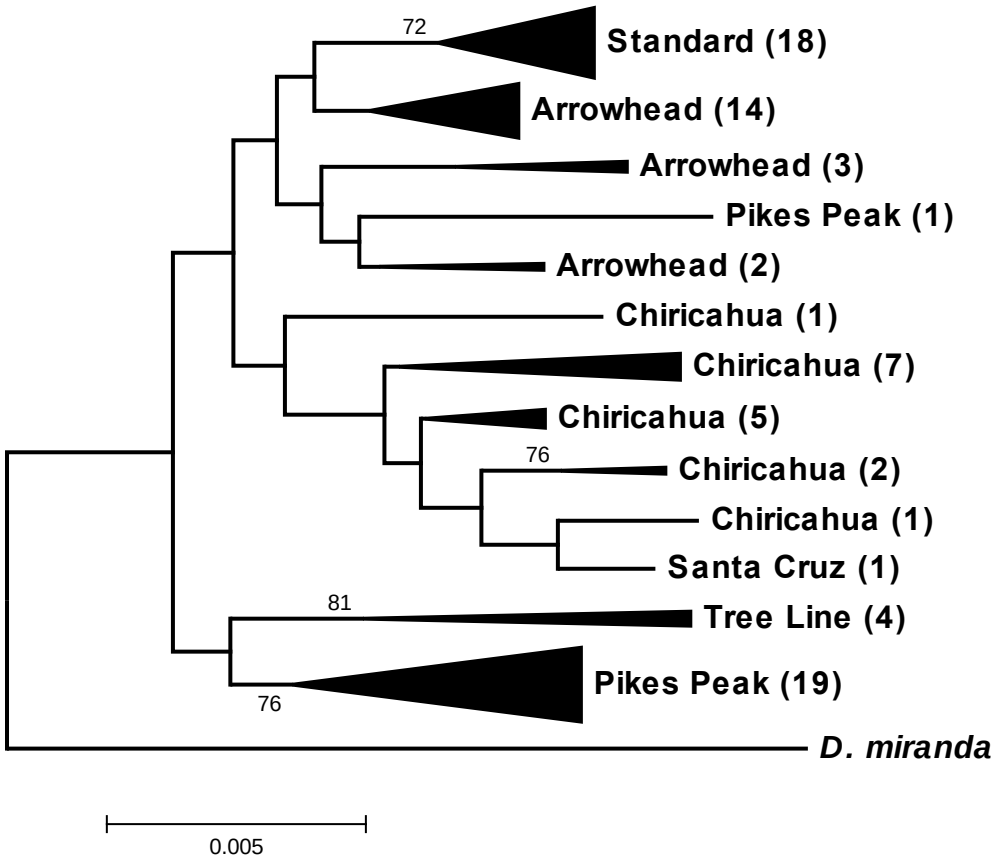
Figure 3. Phylogeny based on the position of the markers on the chromosome. Based on their position on the AR chromosome, markers were separated into: (A) proxima1 (pSTPP, pHYSC, *en*), (B) central (*exu1*. pSTAR, pHYST, dSTPP, *eve*, *mef2*, dSCTL, *Amy1*) and (C) distal (pSCCH, dSTAR, dSCCH, dHYSC, dHYST, *F6*, *EcR*) sub-groups. The three groups were formed from the concatenated dataset used in **Figure 2** and the trees were generated as specified in Figure 2.

Figure 3

A


88
Standard (2)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Pikes Peak (2)
Standard (2)
Tree Line (2)
Arrowhead (3)
Standard (2)
Standard (1)
Tree Line (1)
Arrowhead (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
93
Arrowhead (1)
88
Standard (1)
Standard (1)
82
Pikes Peak
Arrowhead (2)
Standard (2)
Pikes Peak (4)
Pikes Peak (10)
Tree Line (1)
Chiricahua (2)
Chiricahua (4)
Pikes Peak (1)
Chiricahua (3)
Chiricahua (1)
Chiricahua (2)
Chiricahua (1)
Chiricahua (1)
72
Chiricahua (2)
91
Santa Cruz
D. miranda
0.005

B


72
Standard (18)
Arrowhead (14)
Arrowhead (3)
Pikes Peak (1)
Arrowhead (2)
Chiricahua (1)
Chiricahua (7)
Chiricahua (5)
76
Chiricahua (2)
Chiricahua (1)
Santa Cruz (1)
81
Tree Line (4)
Pikes Peak (19)
76
D. miranda
0.005

C

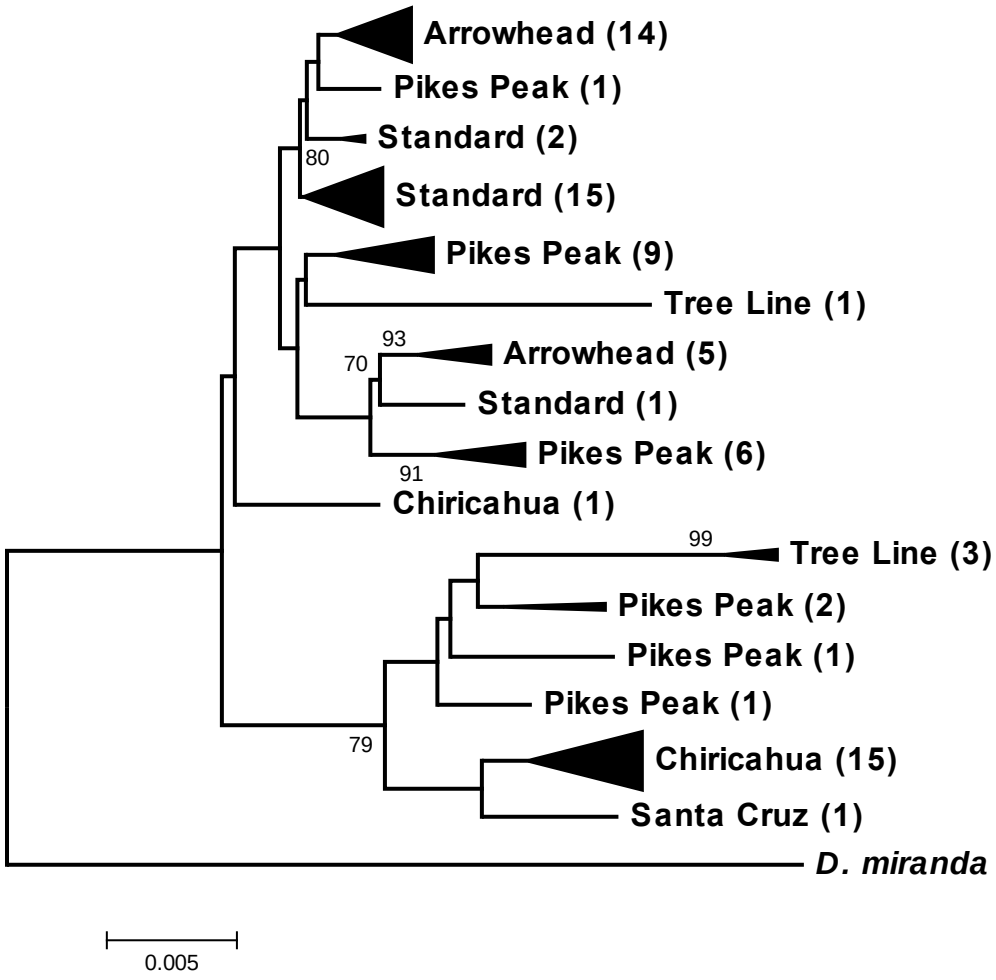
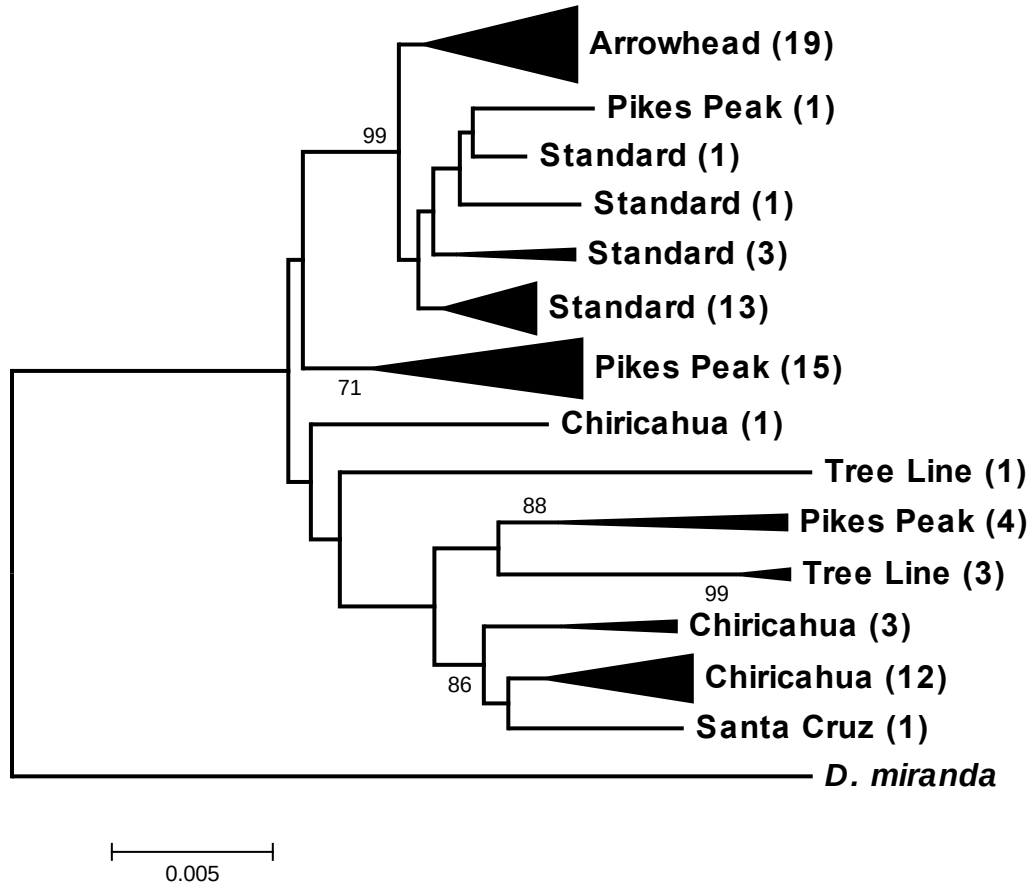

Arrowhead (14)
Pikes Peak (1)
Standard (2)
80
Standard (15)
Pikes Peak (9)
Tree Line (1)
93
70
Arrowhead (5)
Standard (1)
Pikes Peak (6)
91
Chiricahua (1)
99
Tree Line (3)
Pikes Peak (2)
Pikes Peak (1)
Pikes Peak (1)
79
Chiricahua (15)
Santa Cruz (1)
D. miranda
0.005

Figure 4. Phylogenetic relationships of inversion arrangements generated by concatenated data sets separated into (A) Breakpoint regions only, and (B) Non-breakpoint regions. Trees were generated according to the specifications described in **Figure 2**.

Figure 4

A


Arrowhead (19)
Pikes Peak (1)
99
Standard (1)
Standard (1)
Standard (3)
Standard (13)
Pikes Peak (15)
71
Chiricahua (1)
Tree Line (1)
88
Pikes Peak (4)
Tree Line (3)
99
Chiricahua (3)
86
Chiricahua (12)
Santa Cruz (1)
D. miranda
0.005

B

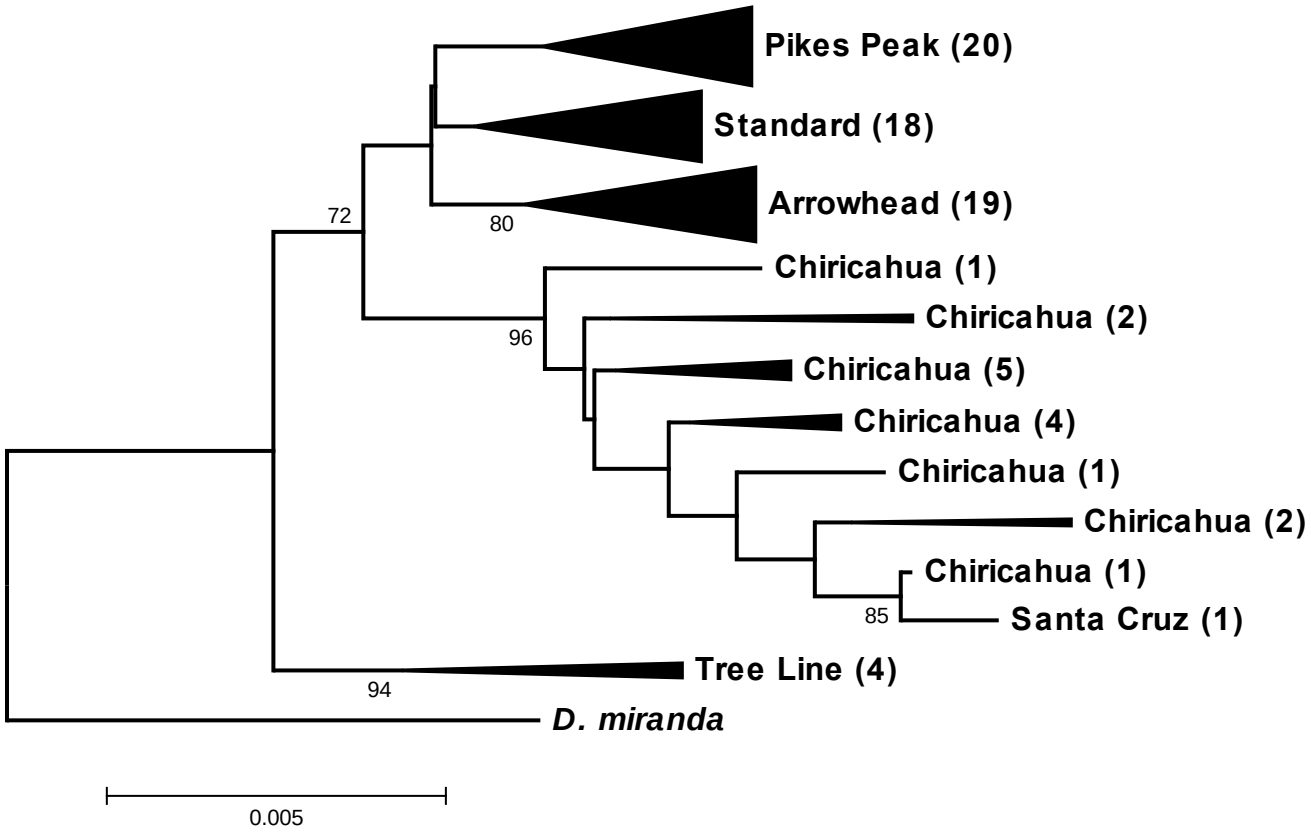
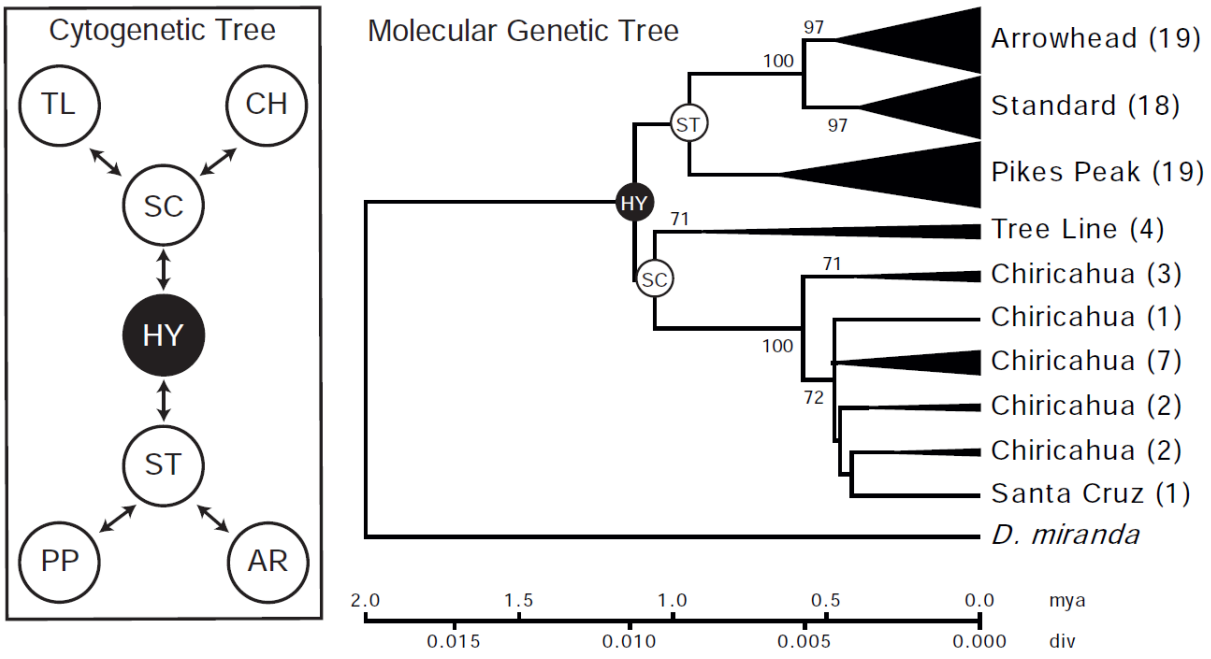

Pikes Peak (20)
Standard (18)
Arrowhead (19)
72
80
Chiricahua (1)
Chiricahua (2)
96
Chiricahua (5)
Chiricahua (4)
Chiricahua (1)
Chiricahua (2)
Chiricahua (1)
85
Santa Cruz (1)
Tree Line (4)
94
D. miranda
0.005

Figure 5. Phylogenetic tree showing the age of the *D. pseudoobscura* gene arrangements obtained using a molecular clock approach in MEGA4. This is a linearized tree generated from the complete concatenated data set. The three central nodes (HY, SC and ST) are highlighted on the tree. To the left of the phylogenetic tree is a schematic of the proposed phylogeny of the gene arrangements.

Figure 5


Cytogenetic Tree
TL
CH
SC
HY
ST
PP
AR
Molecular Genetic Tree
97
100
97
ST
HY
71
SC
71
100
72
Arrowhead (19)
Standard (18)
Pikes Peak (19)
Tree Line (4)
Chiricahua (3)
Chiricahua (1)
Chiricahua (7)
Chiricahua (2)
Chiricahua (2)
Santa Cruz (1)
D. miranda
2.0
1.5
1.0
0.5
0.0
mya
0.015
0.010
0.005
0.000
div

Acknowledgement

We would like to thank Dr. Mohamed Noor (Duke University, Durham, NC) for providing data from the MA 1959 TL strain and the *D. miranda* SPE138 strain. Funding for this study was provided by the Pennsylvania State University Department of Biology (University Park, PA).

References

Anderson WW, Arnold J, Baldwin DG, Beckenbach AT, Brown CJ, Williams GO, Bryant S, Coyne J, Harshman LG, Heed WB et al. . 1991. Four decades of inversion polymorphism in *Drosophila pseudoobscura*. Proceedings of the National Academy of Sciences of the United States of America 88(22):10367-10371.

Aquadro CF, Weaver AL, Schaeffer SW, Anderson WW. 1991. Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proceedings of the National Academy of Sciences of the United States of America 88:305-309.

Babcock CS, Anderson WW. 1996. Molecular evolution of the Sex-Ratio inversion complex in *Drosophila pseudoobscura*: analysis of the Esterase-5 gene region. Molecular Biology and Evolution 13(2):297-308.

Bartolome C, Charlesworth B. 2006. Rates and Patterns of Chromosomal Evolution in *Drosophila pseudoobscura* and *D. miranda*. Genetics 173(2):779-791.

Betran E, Rozas J, Navarro A, Barbadilla A. 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146(1):89-99.

Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. 2008. Chromosomal Rearrangement Inferred From Comparisons of 12 *Drosophila* Genomes. Genetics 179(3):1657-1680.

Charlesworth B, Charlesworth D. 1973. A study of linkage disequilibrium in populations of *Drosophila melanogaster*. Genetics 73(2):351-359.

Dobzhansky T. 1944. Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*. Carnegie Institute of Washington Publication 554:47-144.

Dobzhansky T. 1950. The genetics of natural populations. Genetics 35:288-302.

Dobzhansky T, Epling C. 1944. Taxonomy, geographic distribution, and ecology of *Drosophila pseudoobscura* and its relatives. Carnegie Institute of Washington Publication 554:1-46.

Dobzhansky T, Queal ML. 1938. Genetics of natural populations. II. genic variation in populations of *Drosophila pseudoobscura* inhabiting isolated mountain ranges Genetics 23(5):463-484.

Dobzhansky T, Sturtevant AH. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics 23:28-64.

Engels WR, Preston CR. 1984. Formation of chromosome rearrangements by P factors in *Drosophila*. Genetics 107:657-678.

Gloor G, Engels W. 1992. Single-fly DNA preps for PCR. *Drosophila* Information Service 71:148-149.

Green MM. 1980. Transposable elements in *Drosophila* and other Diptera. Annual Review of Genetics 14:109-120.

Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila melanogaster:* A syndrome of aberrant traits including mutation, sterility, and male recombination. Genetics 86(4):813-833.

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. Genetics 173(1):419-434.

Krimbas CB, Powell JR. 1992. *Drosophila* Inversion Polymorphism. Boca Raton, FL: CRC Press. p. 1-52.

Kumar S, Nei M, Dudley J, Tamura K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9(4):299-306.

Lande R. 1984. The expected fixation rate of chromosomal inversions. Evolution 38(4):743-752.

Lemeunier F, Aulard S. 1992. Inversion polymorphism in *Drosophila melanogaster*.339-405.

Morrow D. 1970. A Cytological Analysis of the Position of Tree Line in the Phylogeny of Gene Arrangements of the Third Chromosome of *Drosophila pseudoobscura* [MS. Thesis]. [Ithaca, NY]: Cornell University.

Navarro A, Barbadilla A, Ruiz A. 2000. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. Genetics 155(2):685-98.

Navarro A, Betran E, Barbadilla A, Ruiz A. 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146(2):695-709.

Nei M, Kojima KI, Schaffer HE. 1967. Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57:741-750.

Noor MAF, Garfield DA, Schaeffer SW, Machado CA. 2007. Divergence Between the *Drosophila pseudoobscura* and *D. persimilis* Genome Sequences in Relation to Chromosomal Inversions. Genetics 177(3):1417-1428.

Novitski E. 1951. Non-random disjunction in *Drosophila*. Genetics 36(3):267-280.

Novitski E. 1967. Nonrandom disjunction in *Drosophila*. Annual Review of Genetics 1:71-86.

Ohta T, Kojima K-I. 1968. Survival Probabilities of New Inversions in Large Populations. Biometrics 24(3):501-516.

Olvera O, Powell JR, de la Rosa ME, Salceda VM, Gaso MI, Guzman J, Anderson WW, Levine L. 1979. Population genetics of Mexican *Drosophila*. Evolution 33(1):381-395.

Popadic A, Anderson WW. 1994. The history of a genetic system. Proceedings of the National Academy of Sciences of the United States of America 91(15):6819-6823.

Popadić A, Anderson WW. 1994. The history of a genetic system. Proceedings of the National Academy of Sciences of the United States of America 91(15):6819-6823.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP et al. . 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Research 15(1):1-18.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406-425.

Schaeffer SW, Aguade M. 2000. Evidence for balancing, directional, and background selection in molecular evolution. In: Singh RS, Krimbas CB, editors. Evolutionary Genetics: From Molecules to Morphology. Cambridge, UK: Cambridge University Press.

Schaeffer SW, Anderson WW. 2005. Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*. Genetics 171(4):1729-1739.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW et al. . 2008. Polytene Chromosomal Maps of 11 *Drosophila* Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics 179(3):1601-1655.

Schaeffer SW, Goetting-Minesky MP, Kovacevic M, Peoples JR, Graybill JL, Miller JM, Kim K, Nelson JG, Anderson WW. 2003. Evolutionary genomics of inversions in *Drosophila pseudoobscura*: Evidence for epistasis. Proceedings of the National Academy of Sciences of the United States of America 100(14):8319-8324.

Sperlich D, Pfreim P. 1986. Chromosomal polymorphism in natural and experimental populations.: Academic Press, London.

Stebbins GL. 1945. Evidence for abnormally slow rates of evolution with particular reference to the higher plants and the genus *Drosophila*. Lloydia (Journal of Natural Products) 8:84-102.

Sturtevant AH, Beadle GW. 1936. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. Genetics 21:554-604.

Sturtevant AH, Dobzhansky T. 1936a. Geographical DistributionI and cytology of "sex ratio" in *Drosophila pseudoobscura* and related species Genetics 21(4):473-490.

Sturtevant AH, Dobzhansky T. 1936b. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. Proceedings of the National Academy of Sciences of the United States of America 22:448-450.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. Molecular Biology and Evolution 12(5):823-833.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Molecular Biology and Evolution 24(8):1596-1599.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America 101(30):11030-11035.

Wallace B. 1966. Chromosomes, giant molecules, and evolution. New York, NY: W.W. Norton & Co.

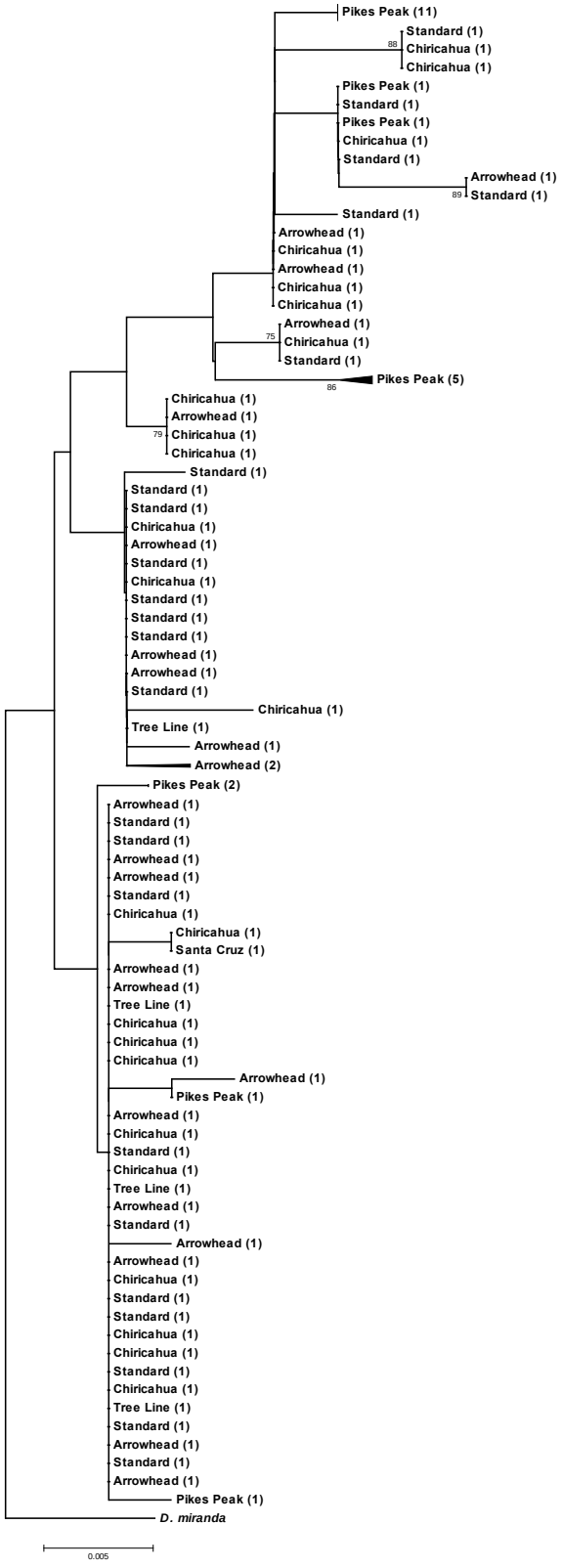
Wallace B. 1968. Topics in population genetics. New York, NY: Norton.

Supplementary Material

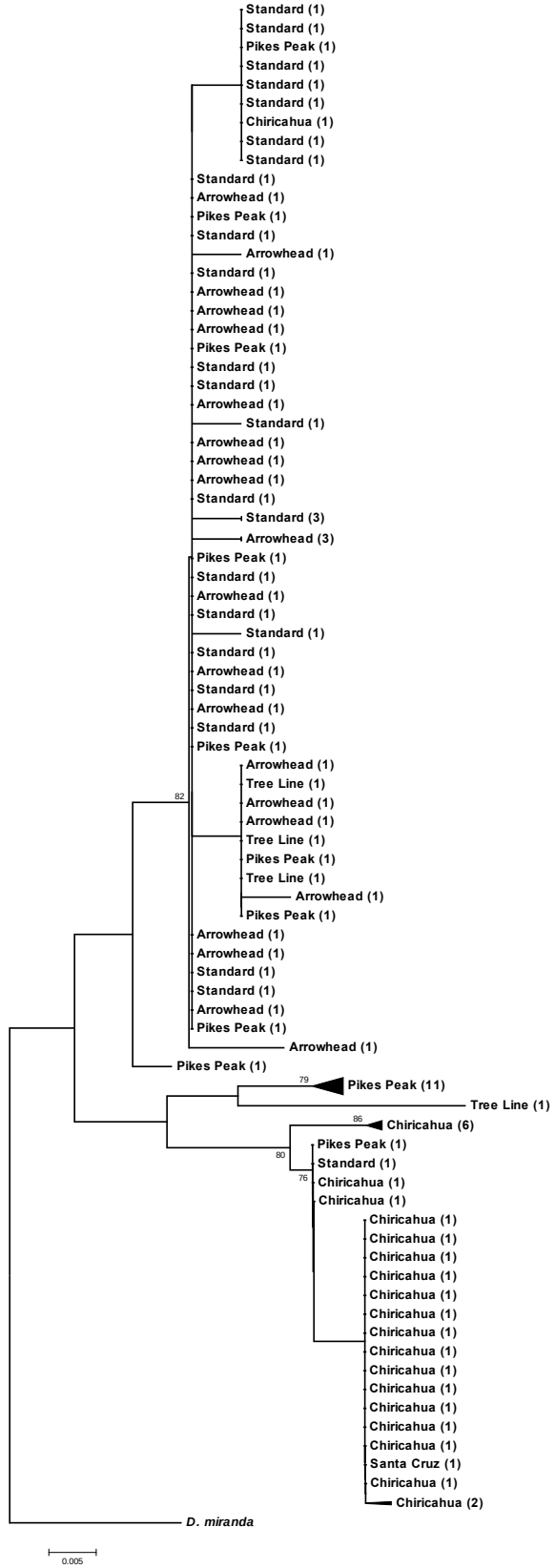
Figure S1. Consensus trees generated by the Neighbor Joining method in MEGA 4.0. (A) pSTPP, (B) pHYSC, (C) pSTAR, (D) pHYST, (E) dSTPP, (F) dSCTL, (G) dSTAR, (H) pSCCH, (I) dHYSC, (J) dSCCH, (K) dHYST. The phylogenetic trees were generated with sequences collected from each breakpoint region. The number of fly strains varies at the different breakpoint regions so each tree may contain a different number of taxa. The taxa names represent the arrangements and the numbers in parentheses represent the number of strains clustering at a particular node. The *Drosophila miranda* sequence serves as the outgroup and is rooted on the midpoint of the dendrogram. Numbers on the branches represent the confidence of each clade which were derived from 1000 bootstrap replicas. Only bootstrap values above 70% are shown in the tree. The trees are listed in the order of which the breakpoint regions occur along the AR chromosome from the centromere to the telomere.

Figure S1

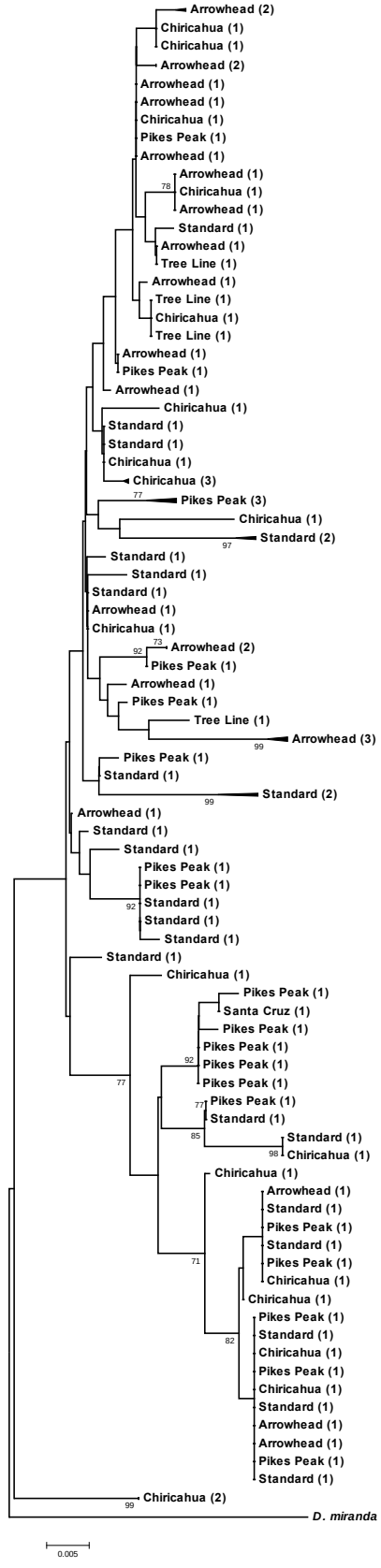
A: pSTPP



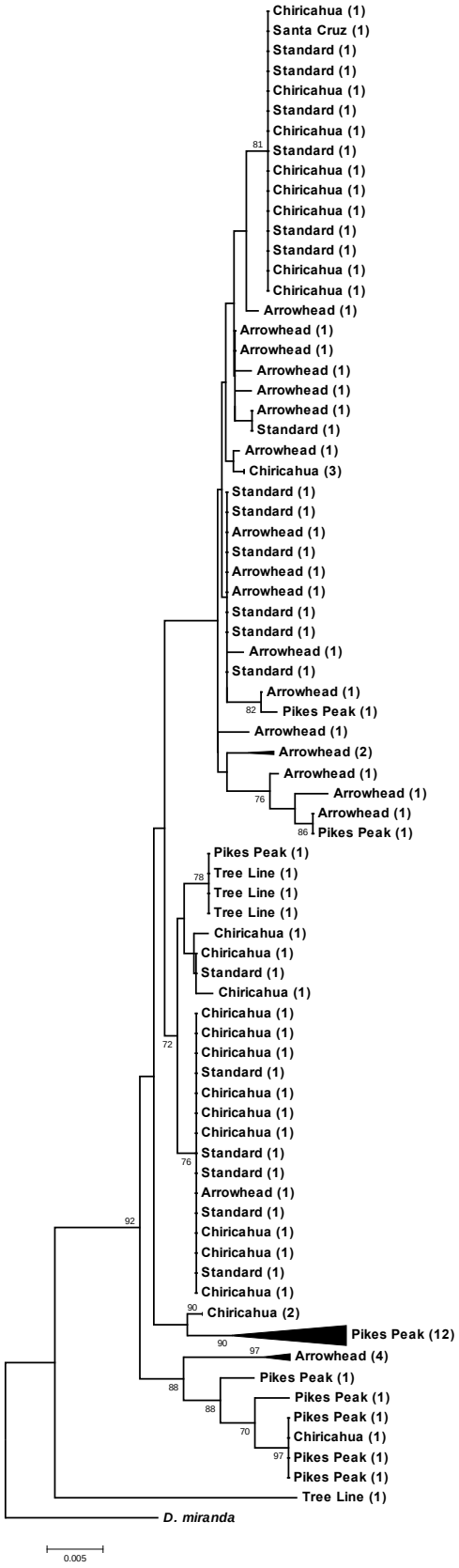
B: pHYSC


Standard (1)
Standard (1)
Pikes Peak (1)
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Standard (1)
Standard (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Standard (3)
Arrowhead (3)
Pikes Peak (1)
Standard (1)
Arrowhead (1)
Standard (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Standard (1)
Pikes Peak (1)
Arrowhead (1)
Tree Line (1)
Arrowhead (1)
Arrowhead (1)
Tree Line (1)
Pikes Peak (1)
Tree Line (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (11)
Tree Line (1)
Chiricahua (6)
Pikes Peak (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Santa Cruz (1)
Chiricahua (1)
Chiricahua (2)
D. miranda
82
79
86
80
76
0.005

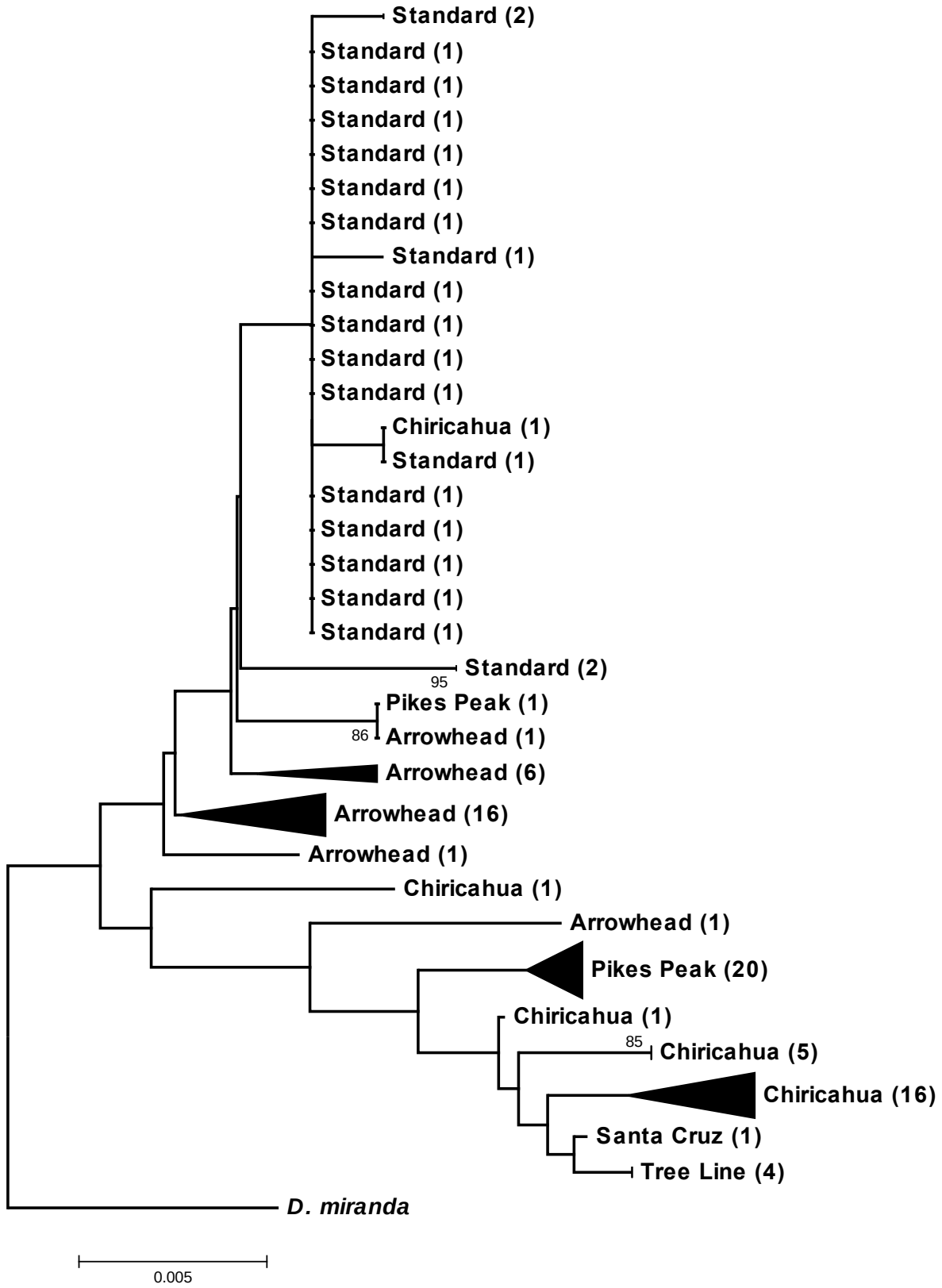
C: pSTAR


Arrowhead (2)
Chiricahua (1)
Chiricahua (1)
Arrowhead (2)
Arrowhead (1)
Arrowhead (1)
Chiricahua (1)
Pikes Peak (1)
Arrowhead (1)
78
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Tree Line (1)
Arrowhead (1)
Tree Line (1)
Chiricahua (1)
Tree Line (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Chiricahua (1)
Standard (1)
Standard (1)
Chiricahua (1)
Chiricahua (3)
77
Pikes Peak (3)
Chiricahua (1)
97
Standard (2)
Standard (1)
Standard (1)
Standard (1)
Arrowhead (1)
Chiricahua (1)
92
73
Arrowhead (2)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
Tree Line (1)
99
Arrowhead (3)
Pikes Peak (1)
Standard (1)
99
Standard (2)
Arrowhead (1)
Standard (1)
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
92
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Pikes Peak (1)
Santa Cruz (1)
Pikes Peak (1)
Pikes Peak (1)
92
Pikes Peak (1)
Pikes Peak (1)
77
77
Pikes Peak (1)
Standard (1)
85
Standard (1)
98
Chiricahua (1)
Chiricahua (1)
Arrowhead (1)
Standard (1)
Pikes Peak (1)
Standard (1)
71
Pikes Peak (1)
Chiricahua (1)
Chiricahua (1)
Pikes Peak (1)
Standard (1)
82
Chiricahua (1)
Pikes Peak (1)
Chiricahua (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Standard (1)
Chiricahua (2)
99
D. miranda
0.005

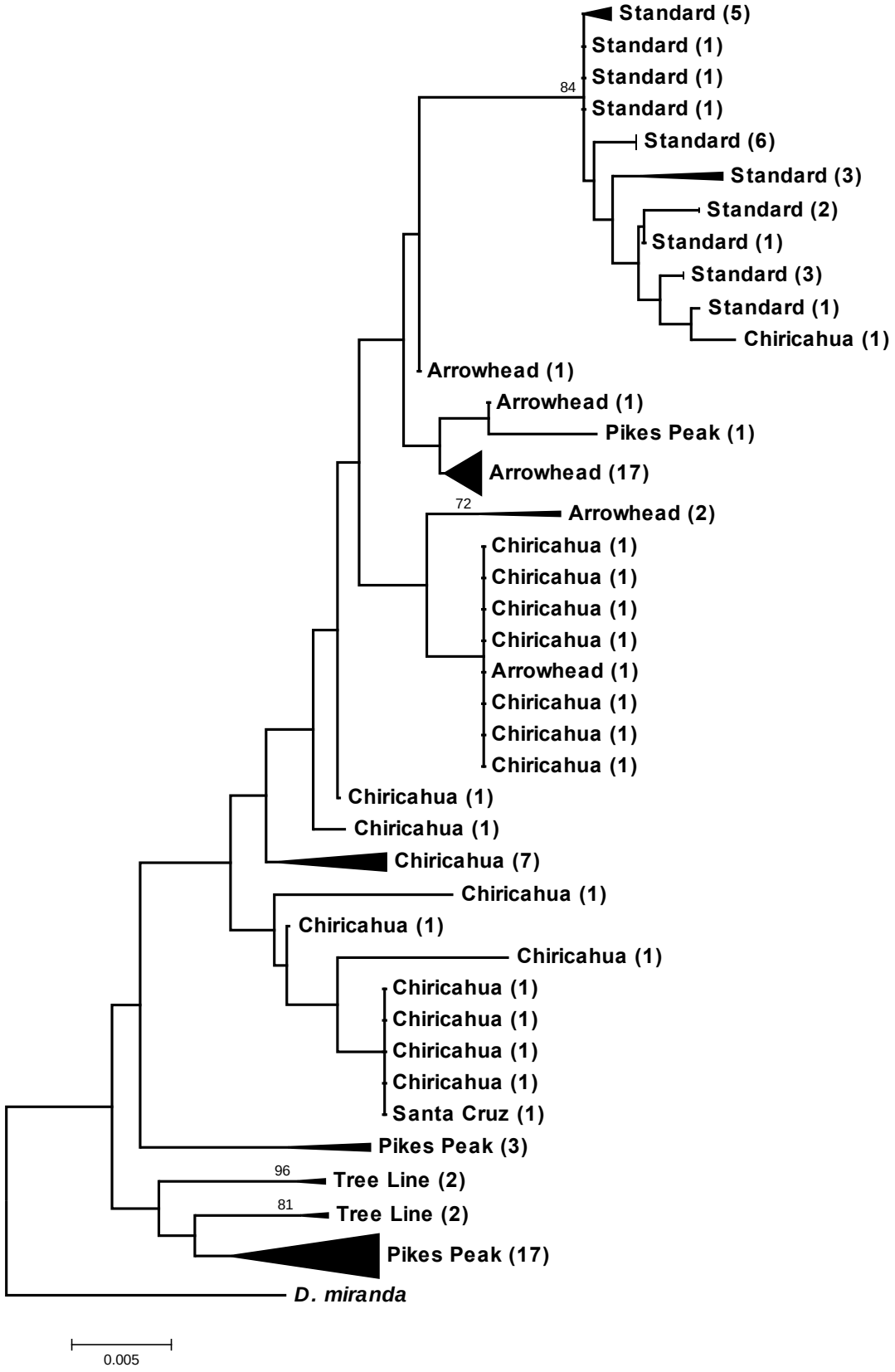
D: pHYST


Chiricahua (1)
Santa Cruz (1)
Standard (1)
Standard (1)
Chiricahua (1)
Standard (1)
Chiricahua (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Standard (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Chiricahua (3)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Arrowhead (2)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Tree Line (1)
Tree Line (1)
Tree Line (1)
Chiricahua (1)
Chiricahua (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Standard (1)
Chiricahua (1)
Chiricahua (2)
Pikes Peak (12)
Arrowhead (4)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Chiricahua (1)
Pikes Peak (1)
Pikes Peak (1)
Tree Line (1)
D. miranda
0.005

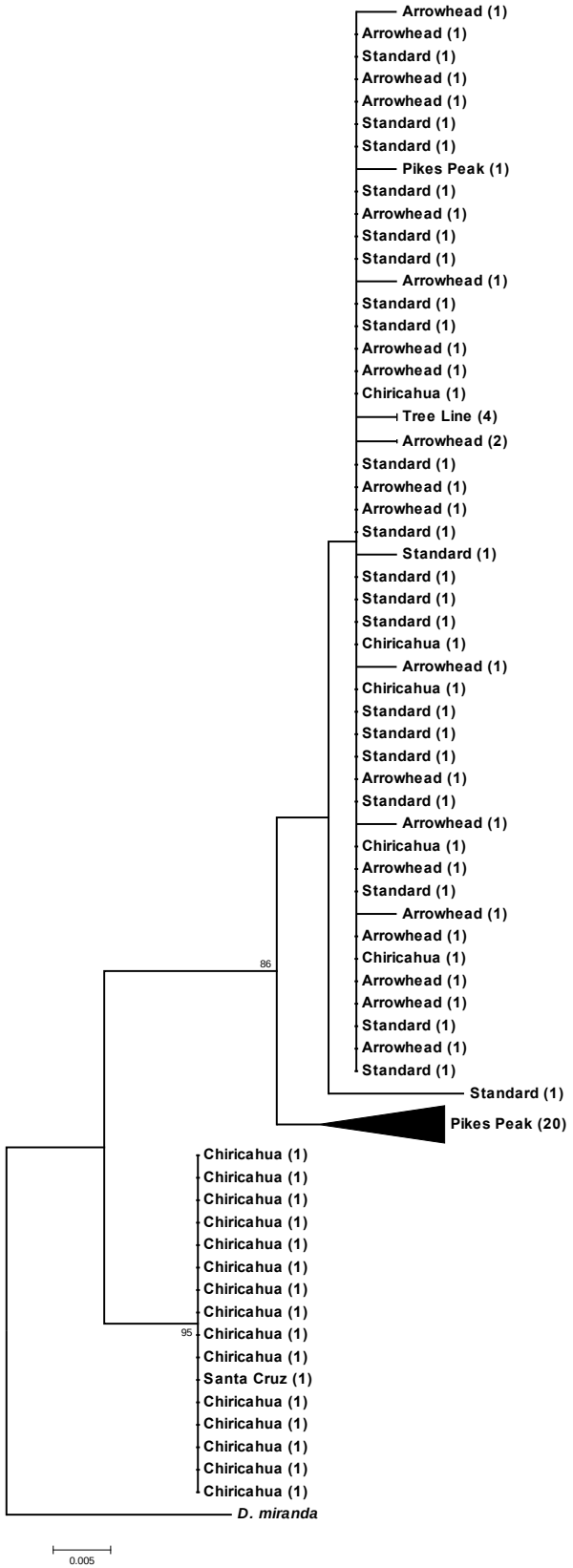
E: dSTPP


Standard (2)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (2)
95
Pikes Peak (1)
86
Arrowhead (1)
Arrowhead (6)
Arrowhead (16)
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Pikes Peak (20)
Chiricahua (1)
85
Chiricahua (5)
Chiricahua (16)
Santa Cruz (1)
Tree Line (4)
D. miranda
0.005

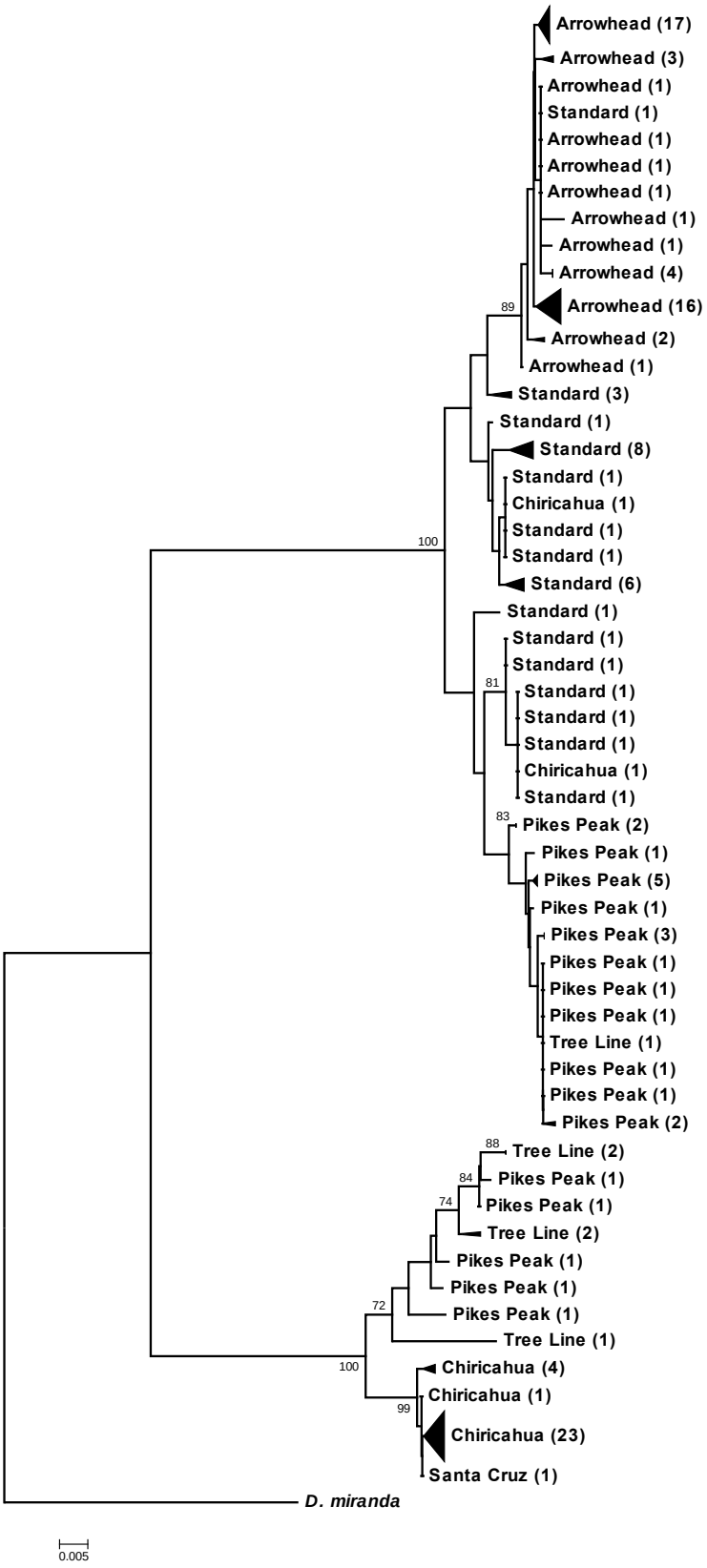
F: dSCTL


Standard (5)
Standard (1)
Standard (1)
84
Standard (1)
Standard (6)
Standard (3)
Standard (2)
Standard (1)
Standard (3)
Standard (1)
Chiricahua (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (17)
72
Arrowhead (2)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Arrowhead (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (7)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Santa Cruz (1)
Pikes Peak (3)
96
Tree Line (2)
81
Tree Line (2)
Pikes Peak (17)
D. miranda
0.005

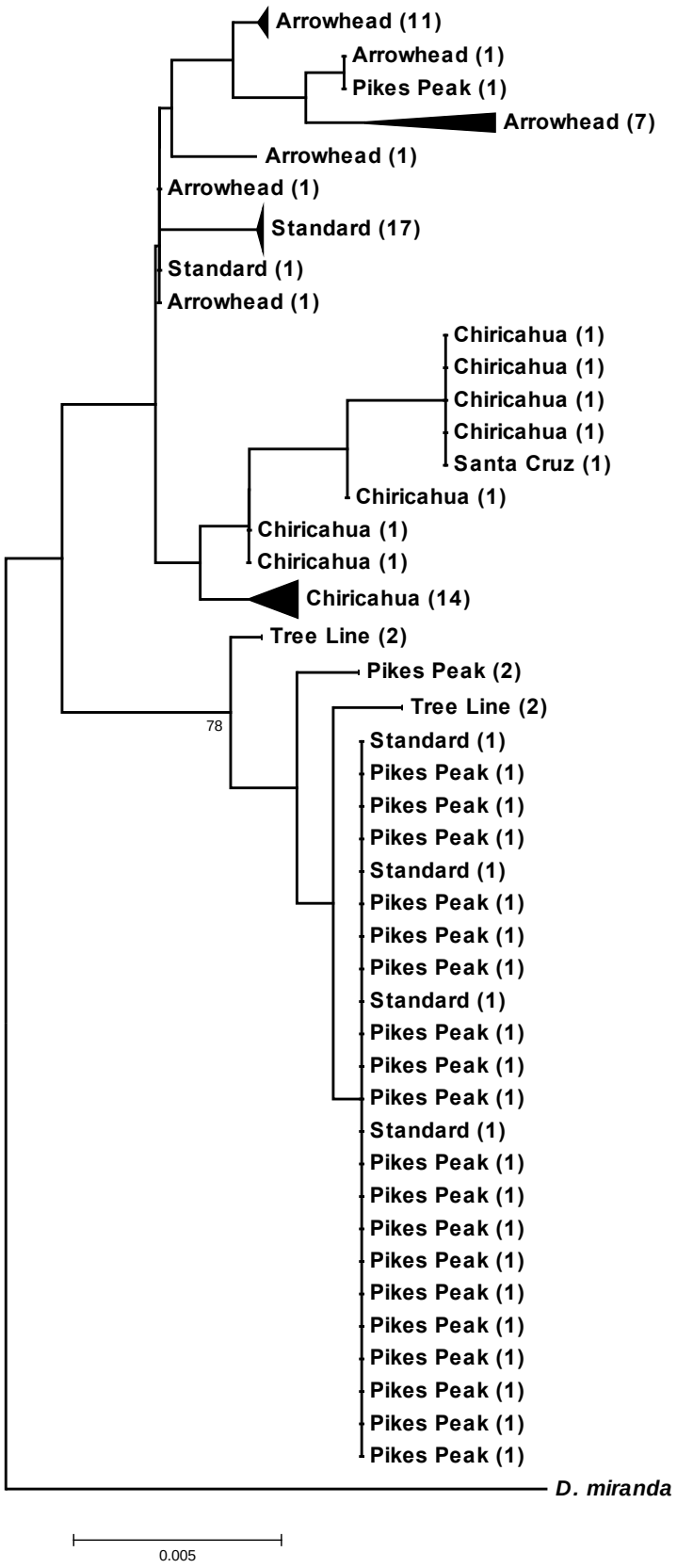
G: pSCCH


Arrowhead (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Standard (1)
Pikes Peak (1)
Standard (1)
Arrowhead (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Chiricahua (1)
Tree Line (4)
Arrowhead (2)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Arrowhead (1)
Chiricahua (1)
Standard (1)
Standard (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Standard (1)
Standard (1)
Pikes Peak (20)
86
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
95
Chiricahua (1)
Chiricahua (1)
Santa Cruz (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
D. miranda
0.005

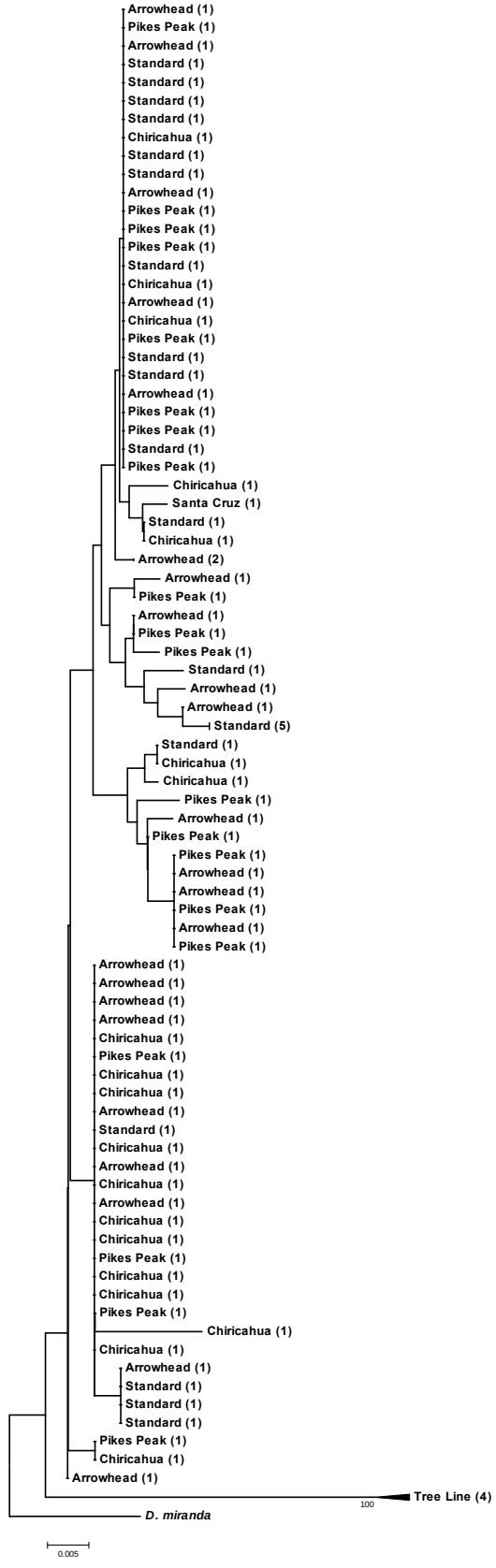
H: dSTAR


Arrowhead (17)
Arrowhead (3)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (4)
89
Arrowhead (16)
Arrowhead (2)
Arrowhead (1)
Standard (3)
Standard (1)
Standard (8)
Standard (1)
Chiricahua (1)
100
Standard (1)
Standard (1)
Standard (6)
Standard (1)
Standard (1)
Standard (1)
81
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Standard (1)
83
Pikes Peak (2)
Pikes Peak (1)
Pikes Peak (5)
Pikes Peak (1)
Pikes Peak (3)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Tree Line (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (2)
88
Tree Line (2)
84
Pikes Peak (1)
74
Pikes Peak (1)
Tree Line (2)
Pikes Peak (1)
Pikes Peak (1)
72
Pikes Peak (1)
Tree Line (1)
100
Chiricahua (4)
Chiricahua (1)
99
Chiricahua (23)
Santa Cruz (1)
D. miranda
0.005

I: dSCCH


Arrowhead (11)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (7)
Arrowhead (1)
Arrowhead (1)
Standard (17)
Standard (1)
Arrowhead (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Santa Cruz (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (14)
Tree Line (2)
Pikes Peak (2)
Tree Line (2)
78
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
D. miranda
0.005

J: dHYSC


Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Chiricahua (1)
Standard (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Chiricahua (1)
Arrowhead (1)
Chiricahua (1)
Pikes Peak (1)
Standard (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Chiricahua (1)
Santa Cruz (1)
Standard (1)
Chiricahua (1)
Arrowhead (2)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Arrowhead (1)
Arrowhead (1)
Standard (5)
Standard (1)
Chiricahua (1)
Chiricahua (1)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Chiricahua (1)
Pikes Peak (1)
Chiricahua (1)
Chiricahua (1)
Arrowhead (1)
Standard (1)
Chiricahua (1)
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Chiricahua (1)
Chiricahua (1)
Pikes Peak (1)
Chiricahua (1)
Chiricahua (1)
Pikes Peak (1)
Chiricahua (1)
Chiricahua (1)
Arrowhead (1)
Standard (1)
Standard (1)
Standard (1)
Pikes Peak (1)
Chiricahua (1)
Arrowhead (1)
Tree Line (4)
100
D. miranda
0.005

K: dHYST

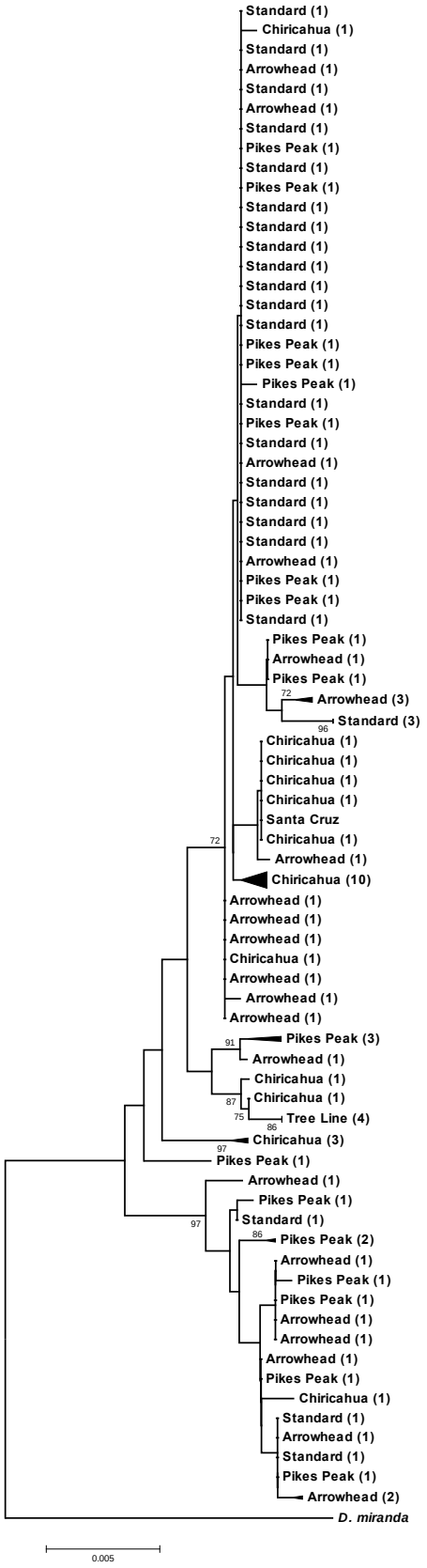
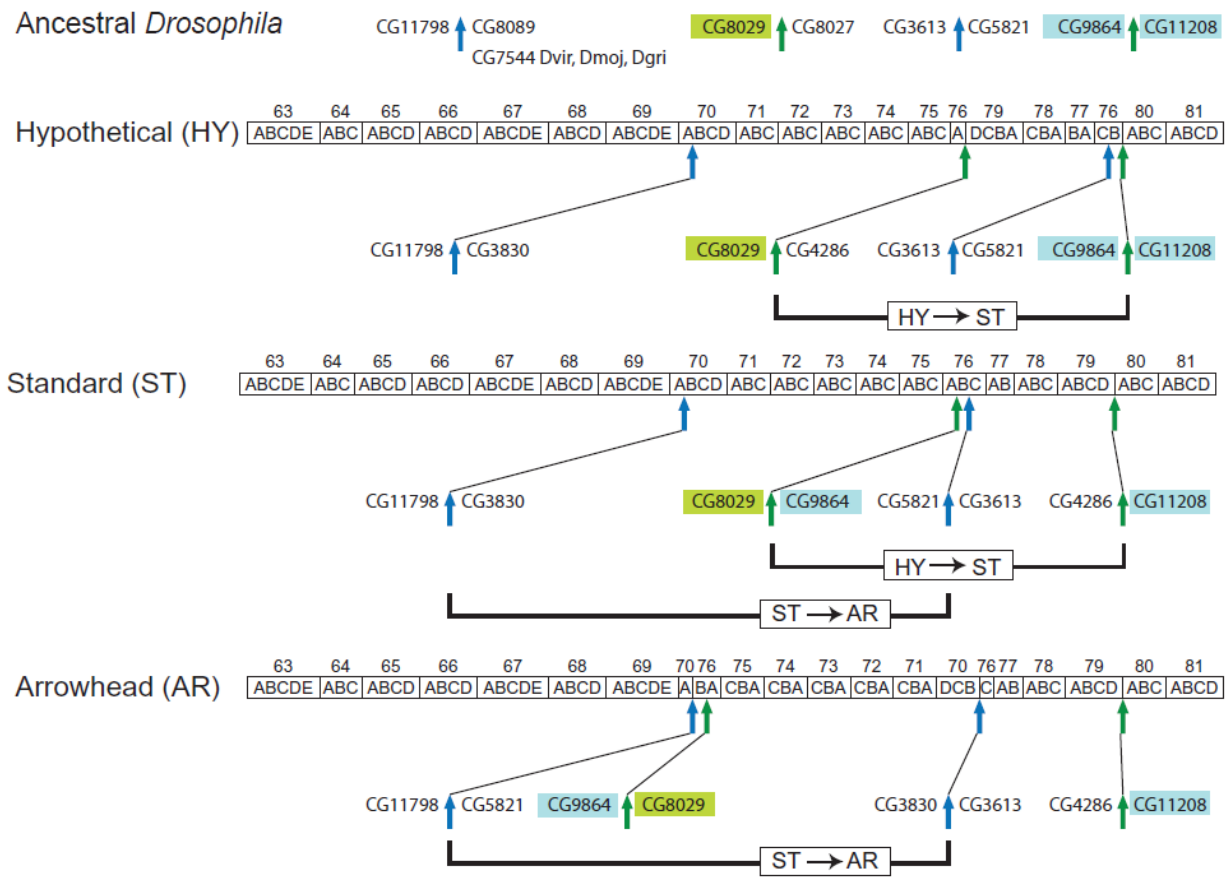

Standard (1)
Chiricahua (1)
Standard (1)
Arrowhead (1)
Standard (1)
Arrowhead (1)
Standard (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Pikes Peak (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Standard (1)
Arrowhead (1)
Standard (1)
Standard (1)
Standard (1)
Standard (1)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Standard (1)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
72
Arrowhead (3)
96
Standard (3)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Chiricahua (1)
Santa Cruz
Chiricahua (1)
72
Arrowhead (1)
Chiricahua (10)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Chiricahua (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
91
Pikes Peak (3)
Arrowhead (1)
Chiricahua (1)
87
Chiricahua (1)
75
Tree Line (4)
86
97
Chiricahua (3)
Pikes Peak (1)
Arrowhead (1)
Pikes Peak (1)
97
Standard (1)
86
Pikes Peak (2)
Arrowhead (1)
Pikes Peak (1)
Pikes Peak (1)
Arrowhead (1)
Arrowhead (1)
Arrowhead (1)
Pikes Peak (1)
Chiricahua (1)
Standard (1)
Arrowhead (1)
Standard (1)
Pikes Peak (1)
Arrowhead (2)
D. miranda
0.005

Figure S2. Linkage Chain Analysis. **A.** Gene adjacency information for the Arrowhead, Standard, and Hypothetical arrangements as well as the Ancestral *Drosophila* species (*D. grimshawi*, *D. mojavensis*, *D. virilis*, *D. willistoni*, *D. ananassae*, *D. yakuba*, *D. erecta*, and *D. melanogaster*). The breakpoints used to convert the different arrangements into each other are indicated along with the changes in gene adjacency. **B.** The linkage chains for the Hypothetical to Standard breakpoints in *D. pseudoobscura* and each of the nine *Drosophila* species. The colors are presented to show how the adjacencies change among species. The light green and light blue boxes highlight the genes that are adjacent in all species except *D. pseudoobscura* where a species specific inversion has occurred. Three inversions are necessary on the *D. pseudoobscura* lineage to go from the ancestral adjacencies to that in the Hypothetical arrangement.

Figure S2

A


Ancestral *Drosophila*
CG11798 CG8089
CG7544 Dvir, Dmoj, Dgri
CG8029 CG8027
CG3613 CG5821
CG9864 CG11208
Hypothetical (HY)
63 64 65 66 67 68 69 70 71 72 73 74 75 76 79 78 77 76 80 81
ABCDE ABC ABCD ABCD ABCDE ABCD ABCDE ABCD ABC ABC ABC ABC ABC A DCBA CBA BA CB ABC ABCD
CG11798 CG3830
CG8029 CG4286
CG3613 CG5821
CG9864 CG11208
HY → ST
Standard (ST)
63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
ABCDE ABC ABCD ABCD ABCDE ABCD ABCDE ABCD ABC ABC ABC ABC ABC ABC AB ABC ABCD ABC ABCD
CG11798 CG3830
CG8029 CG9864
CG5821 CG3613
CG4286 CG11208
HY → ST
ST → AR
Arrowhead (AR)
63 64 65 66 67 68 69 70 76 75 74 73 72 71 70 76 77 78 79 80 81
ABCDE ABC ABCD ABCD ABCDE ABCD ABCDE A BA CBA CBA CBA CBA CBA DCB C AB ABC ABCD ABC ABCD
CG11798 CG5821
CG9864 CG8029
CG3830 CG3613
CG4286 CG11208
ST → AR

B


Dvir
CG9864 CG11208
CG8027 CG8029
CG5473 CG9450
CG6406 CG4984
CG4927 CG9696
CG30394 CG8332
CG4286 CG4643
CG30122 CG6131
Dmoj
CG9864 CG11208
CG8027 CG8029
CG5473 CG9450
CG6406 CG4984
CG4927 CG9696
CG30394 CG8332
CG4286 CG4643
CG30122 CG6131
Dgri
CG9864 CG11208
CG8027 CG8029
CG5473 CG9450
CG6406 CG4984
CG4927 CG9696
CG30394 CG8332
CG4286 CG4643
CG30122 CG6131
Dwil
CG9864 CG11208
CG8027 CG8029
CG5473 CG30122
CG6406 CG4984
CG4927 CG8332
CG30394 CG9696
CG4286 CG9450
Dpse
CG4286 CG11208 dHYST
CG9864 CG8029 pHYST
CG5473 CG9450
CG30122 CG4984
CG4927 CG6406
CG30394 CG8332
CG8027 CG9696
Dana
CG9864 CG11208
CG8027 CG8029
CG5473 CG9450
CG6406 CG4984
CG4927 CG4527
CG30394 CG9696
CG4286 CG30122
CG3416 CG8332
Dyak
CG9864 CG11208
CG8027 CG8029
CG5473 CG30122
CG6406 CG4984
CG4927 CG8332
CG30394 CG9696
CG4286 CG9450
Dere
CG9864 CG11208
CG8027 CG8029
CG5473 CG30122
CG6406 CG4984
CG4927 CG8332
CG30394 CG9696
CG4286 CG9450
Dmel
CG9864 CG11208
CG8027 CG8029
CG5473 CG30122
CG6406 CG4984
CG4927 CG8332
CG30394 CG9696
CG4286 CG9450

CHAPTER 3

MOLECULAR POPULATION GENETICS OF INVERSION BREAKPOINT REGIONS IN *DROSOPHILA PSEUDOOBSCURA*

Research Article

To be submitted to Genetics

Andre G. Wallace , Donald Detweiler, and Stephen W. Schaeffer

Department of Biology

The Pennsylvania State University

Abstract

The presence of paracentric inversions in populations can have a profound effect on the pattern and organization of nucleotide variability. Regions near inversion breakpoints are expected to have higher levels of diversity because of reduced gene exchange while regions in central regions away from breakpoints are predicted to have lower levels of nucleotide diversity. The third chromosome of *Drosophila pseudoobscura* is a model system to test these predictions because over 30 different gene arrangements have been detected in populations. These gene arrangements resulted from a series of overlapping paracentric inversions and their age has been estimated from nucleotide sequence data. Ten of these gene arrangements are frequent and widely dispersed throughout North America. This study examined nucleotide diversity of 11 genetic markers near six pairs of inversion breakpoints to test hypotheses about the molecular evolutionary genomics of different chromosomal arrangements in *D. pseudoobscura*. Approximately 100 third isochromosomal lines of *D. pseudoobscura* collected from four populations in the Southwestern United States were sequenced for each marker. To determine the impact of the marker position (on the chromosome) on nucleotide variability, nucleotide heterozygosity was estimated at each breakpoint region based on the neutral mutation parameter ($4N_e\mu$), π and θ. Nucleotide heterozygosity was different for each breakpoint region and the level of heterozygosity within each arrangement was different for every marker. Overall, markers closer to the distal segment of the chromosome had higher levels of nucleotide heterozygosity than markers within the proximal segment of the chromosome. In addition, our results rejected the hypothesis that newer inversions will have lower levels of nucleotide variability near the breakpoints of derived arrangements than near the breakpoints of older inversions. These observations were similar even after sequences with gene conversion events were removed. We

did observe low levels of nucleotide diversity at regions close to the breakpoint of the new PP inversion than within the inverted segment. Intralocus and interlocus linkage disequilibrium (LD) was estimated to assess whether the inversion polymorphism leads to an elevation of LD and to test whether sites in breakpoint pairs are in LD. High levels of LD were observed within all 11 regions and high levels of LD were observed between most breakpoint pairs. The central region of the chromosome had the highest levels of LD compared to the proximal and distal regions because this is the region that experiences the highest level of recombination suppression.

Introduction

Chromosomal inversions have a long history in population genetics and evolutionary research that extends over several decades. Spatial and temporal patterns, cyclical seasonal fluctuations, meiotic drive and phylogeny are some of the areas of study that utilized inversions (DOBZHANSKY and STURTEVANT 1938) . In addition, inversions have been employed to look for evidence of selection. For example, Wright and Dobzhansky (1946) employed population cage experiments to show that natural selection is a driving force behind seasonal cycling frequencies of *Drosophila pseudoobscura* inversion polymorphisms. The ability to identify inversions with simple cytological analysis proved to be very important in these early studies. The ability of inversions to suppress recombination made them candidates to explore selection (STURTEVANT and BEADLE 1936) and such studies are still being pursued today (HOFFMANN *et al.* 2004; STEFANSSON *et al.* 2005).

Recombination plays an important role in homogenizing nucleotide variability between chromosomes. New chromosomal inversion mutations are a potent isolating mechanism between homologous chromosomes because crossing over is suppressed in gene arrangement heterozygotes. The newly inverted chromosome will capture a unique set of alleles based on the chromosome that initially was inverted. Recent theoretical studies have developed models to help understand how the presence of paracentric inversions can alter patterns of nucleotide variability in *Drosophila* (NAVARRO *et al.* 2000; NAVARRO *et al.* 1997). Navarro et al. (2000) explored the effects of chromosomal inversion polymorphisms on levels and patterns of nucleotide variability. The studies were performed using the Poisson model and the Counting model. In the Poisson model, it is assumed that recombination events are independent while the Counting model assumes inference among recombination events. Based on their models, they

predicted that newer inversions will have lower variability at inversion breakpoints than within the inverted segment while older inversions will have greater variability at the breakpoints than within the inverted segment.

The reduction of recombination by inversions has been linked to the production of inviable gametes from crossing over. However, the production of inviable gametes from crossover is not always the final result since double crossing over can increase the production of viable gametes during meiosis. Although crossing over is reduced because recombinant gametes are inviable, genetic exchange can take place in the form of gene conversion where small tracts of DNA are exchanged between chromosomes. Though not typical, this nonreciprocal transfer of genetic information can prove to be more influential in recombination than crossing over in inverted chromosomes (NAVARRO *et al.* 1997). The *rosy* locus found within inverted segments of the 3rd chromosome of *Drosophila melanogaster* was shown to have a higher rate of exchange among heterokarytypes initiated by gene conversion rather than crossing over (CHOVNICK 1973; HILLIKER *et al.* 1994). Navarro *et al.* (1997) used two models of crossing over to explain these observations, the Poisson model that assumes independence of cross over events versus the Counting model that assumes that there can be interference between cross over events. The Counting Model provided a better explanation for these observations by suggesting a one order of magnitude greater rate of recombination caused by gene conversion versus crossing over within inverted regions (NAVARRO *et al.* 1997).

Navarro *et al.* (1997) showed that genetic exchange varies among the proximal, central and distal regions of inverted chromosomes when a Counting model is assumed. Proximal regions closest to the centromere have lower levels of exchange than distal regions of the

chromosome nearest to the telomere. This predicts that proximal regions will be more divergent among different gene arrangements than distal regions.

The size of an inversion is also suggested as a major factor that determines whether crossing over or gene conversion is the driving force of genetic exchange between chromosomes. Gene conversion is expected to have a constant impact on genetic exchange between inverted chromosomes, but crossing over can only homogenize diversity if two exchanges occur within the inverted segment. Therefore, crossing over will have a greater impact on genetic exchange as inversions get larger (BETRAN *et al.* 1997; NAVARRO *et al.* 1997). Although gene conversion occurs at a constant rate across inversions of all sizes, the amount of DNA exchanged is only 200 to 300 base pairs (BETRAN *et al.* 1997; SCHAEFFER and ANDERSON 2005). This leads to the general prediction that regions closest to breakpoints are likely to diverge because genetic exchange is lowest near the breakpoints (NAVARRO *et al.* 2000; NAVARRO *et al.* 1997). The exact mechanisms governing the effect of gene conversion on recombination remains unclear.

The reduction of recombination is also expected to lead to significant levels of linkage disequilibrium (LD) or nonrandom associations among nucleotide sites. Because new mutations occur on a single chromosome, they generate nonrandom associations with variation on the chromosome. Recombination will shuffle variation on the chromosome breaking up the newly formed nonrandom associations. Since inversions prevent recombination from shuffling nucleotide diversity, nonrandom associations should be maintained. Linkage disequilibrium is proposed as a method to map human genetic disease genes. The presence of inversions can lead to high levels of linkage disequilibrium without being associated to disease genes. This study examines the structure of LD in the presence of paracentric inversions.

Drosophila species are ideal model systems to examine the predictions of Navarro et al. (1997; 2000). The gene content is conserved among *Drosophila* species across the genus (Muller, 1940), but the order of genes has been shuffled through the accumulation of fixed inversion mutations. As might be expected, populations of most *Drosophila* species are polymorphic for paracentric inversions providing the raw material for lineage specific gene orders (Sperlich and Pfreim 1986). For example, *D. pseudoobscura* populations have over 30 different gene arrangements on its third chromosome that were created by a series of overlapping paracentric inversions (DOBZHANSKY and EPLING 1944; POWELL 1992). Inversions in *D. pseudoobscura* provide an excellent system to estimate the levels and patterns of nucleotide variability in *Drosophila* gene arrangements and their breakpoints. In addition to levels and patterns of variability within gene arrangements, information may also be obtained to explain nucleotide variability between different arrangements within particular species.

The availability of the *Drosophila pseudoobscura* genomic sequence data have enabled us to examine patterns and levels of nucleotide variability associated with regions in and around inverted chromosomes (RICHARDS *et al.* 2005). In this study, strains of *D. pseudoobscura* from several different inversion types (chromosomal arrangements) were sequenced. Regions near inversion breakpoints along the third chromosome of *D. pseudoobscura* were sequenced in ancestral and derived gene arrangements to test predictions about breakpoint evolution (NAVARRO *et al.* 2000; NAVARRO *et al.* 1997). First, estimates of nucleotide diversity were examined at each breakpoint region to see if older inversions had higher levels of genetic variation at the breakpoints. Second, levels of variation in proximal regions, in the inverted regions, and in distal segments were estimated to test whether proximal regions are more divergent than distal regions as predicted by Navarro et al (2000). Third, we derived two

estimates of the neutral mutation parameter ($4N_e\mu$), π and θ, among all regions to test for departures from the neutral theory of molecular evolution (KIMURA 1986). Finally, we estimated intralocus and interlocus linkage disequilibrium (LD) to determine if proximal regions have higher levels of LD than distal regions because of the polarity of genetic exchanges between these two segments of the chromosome.

Materials and Methods

Fly strain and DNA extraction. Genomic DNA was prepared from 100 isochromosomal lines of *Drosophila pseudoobscura* using a single-fly DNA extraction protocol (Gloor and Engels 1992). These lines were collected in 1998 from four different locations in the southwestern United States by S. W. Schaeffer and W. W. Anderson (University of Georgia). The localities include: Davis Mountain, TX; James Reserve, CA; Mount Saint Helena, CA; and Kaibab National Forest, AZ (Schaeffer et al. 2003).

PCR Primer Design, Amplification and Nucleotide Sequencing. A total of 18 loci were examined in this study. Seven loci are at some distance from inversion breakpoints while 11 regions were in close proximity to the breakpoints of six inversions. We used primers for eight loci previously sequenced by Schaeffer *et al.* (2003), engrailed (*en*), exuperantia 1 (*exu 1*), Myocyte enhancing factor 2 (*Mef 2*), even skipped (*eve*), amylase 1 (*Amy 1*), vestigial (*vg*), F6, and Ecdysone Receptor (*EcR*). The vestigial locus was used as the marker for the distal Standard to Arrowhead breakpoint since it is in close proximity to the actual breakpoint (RICHARDS *et al.* 2005). The other 10 markers were near the breakpoints of six gene arrangements of *D. pseudoobscura* that are frequent and widely distributed in this species (POWELL 1992). The gene arrangements examined were Pikes Peak (PP), Santa Cruz (SC), Tree Line (TL), Standard (ST), Hypothetical (HY) and Chiricahua (CH) (DOBZHANSKY and EPLING 1944) (**Figure 1**). The approximate locations of the six pairs of breakpoints were inferred from the genome sequence of the third chromosome of *D. pseudoobscura* and the correlation of the nucleotide map with the polytene map (SCHAEFFER *et al.* 2008). PCR primers were designed to amplify the 10 remaining breakpoint regions. Because the Santa Cruz and Tree Line proximal breakpoint regions appear to be coincident on the cytological map, only one primer pair was designed to interrogate sequence

evolution at both inversion breakpoints. We used the comparison of the *D. persimilis* and *D. pseudoobscura* genome sequences of the third chromosome as a guide for regions that were likely to be less constrained at the nucleotide level (Noor et al. 2007). Breakpoint regions were labeled either "p" or "d" to represent the proximal or distal breakpoints, the two-letter code of the ancestral inversion, and the two-letter representation for the derived inversion (Table 1). The PCR fragments amplified were between 400 and 1000 base pairs. The breakpoint regions sequenced included non-coding sequences except the pHYST locus which had less than 50% coding sequence. Total PCR reaction volume was 50µl. To verify that the amplification product was the predicted length (bp), 5µl of the completed PCR reaction was run on 2% agarose gel. Once the fragment size was verified, the completed PCR reaction was treated with ExoSAP-IT (USB, Affymetrix Corporation) to remove contaminants such as unused dNTPs and primers remaining in the PCR product. Fragments were then sequenced at the Penn State University Nucleic Acid Facility (University Park, PA) using an ABI Hitachi 3730XL DNA Analyzer. Each breakpoint region was sequenced in both forward and reverse directions. The forward and reverse sequences were assembled and conflicts between the two reads were resolved using the SEQMANII program (DNASTAR, Madison, WI).

Nucleotide Sequence Alignment. Nucleotide sequences for each breakpoint region were aligned using the MEGALIGN program (DNASTAR, Madison, WI). Alignments were performed manually and inspected visually to insure that indels were scored consistently among the different sequences.

Nucleotide Polymorphism and Divergence at *D. pseudoobscura* Breakpoint Regions. After fragments were sequenced and aligned, alignment files were exported to DnaSP for population genomic analyses (Librado and Rozas 2009). DnaSP was used to measure nucleotide

variability within all sequences at each locus. We tested the predictions of Navarro *et al.* (2000) of how breakpoint position influences levels of nucleotide variability within and among four gene arrangements (AR, PP, CH, and ST) with estimates of nucleotide heterozygosity based on the number of segregating sites (θ) (Watterson 1975) and the number of pairwise differences (π) (TAJIMA 1983; WATTERSON 1975). Estimates of nucleotide diversity used the Jukes and Cantor correction (Jukes and Cantor, 1969; (LYNCH and CREASE 1990). In addition, we computed the variance of θ for free and no recombination (Tajima 1983).

Ancestral and Derived Arrangements. The Hypothetical, Tree Line, and Santa Cruz arrangements have each been suggested as the ancestral arrangement of all the other third chromosome arrangements (Bartolome and Charlesworth 2006a; Popadic and Anderson 1994; Powell 1992). We assume that the Hypothetical arrangement is the ancestral arrangement based on several lines of evidence (Wallace and Schaeffer, unpublished). First, analysis of breakpoints from genome comparisons failed to identify either the Santa Cruz or Tree Line breakpoints. If the Tree Line or Santa Cruz rearrangement is the ancestor, then we should observe interspecific breakpoints that are located near their inferred cytological location on the *D. pseudoobscura* map (SCHAEFFER *et al.* 2008); Wallace and Schaeffer, unpublished). No such interspecific breakpoints exist for either the Santa Cruz or the Tree Line arrangements, but do exist for the predicted positions of the Hypothetical arrangement. Second, phylogenetic data shows that the Hypothetical arrangement serves as a link between the Santa Cruz and Standard arrangements. Therefore, we polarize the inversion events on the *D. pseudoobscura* third chromosome phylogeny based on the Hypothetical chromosome being the common ancestor of all other arrangements. The estimated ages of the gene arrangements are: 1.12 million years for HY, 1.06

million years for SC, 0.91 million years for TL, 0.95 million years for ST, 0.57 million years for AR, 0.95 million years for PP, and 0.58 million years for CH.

DNA Divergence of derived *D. pseudoobscura* arrangements from ancestral arrangements. DnaSP was utilized to estimate nucleotide variability of derived *D. pseudoobscura* arrangements from ancestral arrangements. We estimated nucleotide diversity (π) which utilizes the average number of nucleotide differences (k) per site. The data collected was partitioned based on the proposed cytogenetic phylogeny of the *D. pseudoobscura* arrangements studied (Schaeffer et al. 2003). Each breakpoint region was sequenced in all of the isochromosomal strains, but we chose to partition our data based on whether the breakpoints were involved in the formation of a particular gene arrangement. For instance, the pSTPP and dSTPP breakpoints were used to form the Pikes Peak arrangement. To examine predictions of the Navarro *et al.* model, we asked whether nucleotide variation at the pSTPP and dSTPP breakpoints within the derived Pikes Peak arrangement is less than the variability in the ancestral population comprised of all other arrangements. Nucleotide diversity was computed for the derived and ancestral arrangements as well as for the combination of all arrangements. These values were obtained using the Jukes and Cantor method ((Nei 1987)). These data are represented in **Table 3** and **Figure 3**. When comparing the Arrowhead inversion against all other inversions, the Pikes Peak sequences were eliminated from the analysis because both Pikes Peak and Arrowhead have the same immediate ancestor (Standard). Likewise, the Arrowhead sequences were removed from the Pikes Peak analysis.

Nucleotide variability was computed across individual inversions to determine if diversity was lower at the breakpoints than within an inversion (**Table 3 and Figure 3**). Using **Figure 1** as a guide, regions sampled for this study were ordered based on their occurrence in the

chromosome of each inversion. Divergence was estimated for regions outside the proximal and distal breakpoints, close to inversion breakpoints, and within the inverted segment. The overall nucleotide diversity at each locus for individual arrangements is represented in **Figure 3**.

Tests of an Equilibrium Neutral Model. The Tajima *D* statistic was used to test the neutral theory of molecular evolution (TAJIMA 1989). We estimated values for Tajima's *D* across all sequences at each locus and also within individual gene arrangements. The Hudson Kreitman, and Aguade (1987) test was used to test for departures from a neutral model. The probabilities of more extreme values were estimated within each arrangement and for all arrangements at each locus, using coalescent simulations.

Linkage Disequilibrium Analysis. Fisher's exact test (Sokal and Rohlf, 1981) was used to test pairs of variable sites within and between third chromosomal loci for significant LD. We concatenated the aligned sequences from the 11 breakpoint regions and 7 gene regions of 78 strains to test for intra- and inter-locus LD. Only comparisons of sites capable of generating a significant result with Fisher's exact test were performed (LEWONTIN 1995). We used the Q-value approach with a false discovery rate of 1% to overcome the multiple comparison problem (STOREY 2002).

Results

Nucleotide Polymorphism and Divergence at *D. pseudoobscura* Breakpoint Regions.

The Navarro *et al.* (2000) model makes predictions about the levels of heterozygosity expected near derived and ancestral breakpoint regions of inversions. Nucleotide diversity within individual inversions is likely to be reduced near breakpoints because the initial inversion mutation occurs in a single individual eliminating variation in the newly formed arrangement. Here, we estimated nucleotide heterozygosity at each region to determine if nucleotide diversity is impacted by the position of the breakpoint on the chromosome. Nucleotide heterozygosity was estimated for all sequences at each locus (all). We also estimated the distribution of nucleotide heterozygosity within individual gene arrangement types at each locus. The *D. pseudoobscura* breakpoint regions show a range of nucleotide diversity levels. Measures of nucleotide heterozygosity at breakpoint regions are shown in **Table 1** and **Figure 2**.

Two different measures of nucleotide heterozygosity were estimated, π (Nei, 1987) and θ (WATTERSON 1975). Both π and θ are estimates of the neutral mutation parameter $4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide (TAJIMA 1983). The nucleotide diversity estimates across all arrangements in the 11 loci varied from a low of 0.010 in the pSTPP and dHYST regions to a high of 0.039 in the dSTAR region. Nucleotide diversity estimates (π) varied within arrangements from 0.003 to 0.017 in AR, from 0.002 to 0.030 in PP, from 0.004 to 0.021 in CH, and from 0.002 to 0.022 in ST (**Table 1**). At each locus, nucleotide diversity across all arrangements was greater than nucleotide diversity within individual arrangements, with the exception of the pHYSC/pSCTL and the pHYST regions which both had greater levels of diversity within the PP arrangement than in all inversions combined. Nucleotide

diversity within individual arrangements varied from locus to locus. For instance, θ within the PP group of sequences was 0.026 at the pHYSC locus and 0.010 at the pSTPP locus.

The proximal and distal breakpoints might be expected to have similar levels of diversity because both ends of the inversion will be the slowest to recover from the variation reduction from the original mutation event. However, due to the lack of involvement of the distal regions in fertility, diversity levels will likely be greater at the distal regions than at the proximal regions. This can be explained by the polarity inversions create in the chromosome arm on which they occur which can be correlated to the production of non-viable gametes within the proximal regions (CARSON 1953; NAVARRO *et al.* 1997). Nucleotide diversity (θ) across all arrangements at the proximal and distal breakpoints of each inversion mutation was found to be greater at the distal breakpoint region than at the proximal region. The SCCH and HYST breakpoint regions were exceptions because θ was greater at pSCCH (0.023) than at dSCCH (0.010) and greater at pHYST (0.017) than at dHYST (0.012). The two STPP and the two HYST breakpoint regions possessed the lowest levels of nucleotide diversity in comparison to any other pair of breakpoint regions. The dSCCH region had a significant deficiency in nucleotide diversity when compared to the other regions. At each region, nucleotide diversity levels varied among the different arrangements. For example, at the pHYSC region, π was 0.030 within the PP group and 0.004 within the AR group of sequences. No significant differences in nucleotide diversity were observed between individual arrangements within the dSCCH and dHYSC regions. There was an excess in heterozygosity estimates within PP at the pHYSC and dSTAR regions. There was also excess diversity within PP and CH within the dSCTL and pSCCH regions. Finally, excess nucleotide diversity was observed within the AR and CH arrangements at the dSTPP region.

Tests for Departures from the Neutral Theory of Molecular Evolution. Tajima's (1989) *D* tests whether the difference between two estimates of the neutral mutation parameter ($4N_e\mu$), π and θ, are significantly different from zero. Tajima's *D* will be zero under a neutral model, will be significantly greater than zero in a population experiencing balancing selection, and will be significantly negative under purifying selection or population expansion (INNAN and STEPHAN 2000). There were negative Tajima's D values for 66% of loci within gene arrangements. However, only the dSTPP locus within the PP arrangement had a significant negative value. Eight of the 11 regions studied, had negative values for Tajima *D*, which is consistent with previous observations for loci in *D. pseudoobscura* (HAMBLIN and AQUADRO 1999; SCHAEFFER 2002; SCHAEFFER *et al.* 2003).

DNA Divergence of derived *D. pseudoobscura* arrangements from ancestral arrangements. We examined the pattern of nucleotide diversity outside of and within the inverted regions of the third chromosome arrangements to determine the pattern of variation in the proximal regions closest to the centromere, in the inverted regions, and the distal segment closest to the telomere. The theoretical work of Navarro *et al.* (1997) predicts that nucleotide divergence will be greater in the proximal than the distal regions. Within the inverted regions, nucleotide variability will differ depending on its position on the chromosome since recombination rates are affected by the positions of markers on the chromosome. Using sequence information from the 11 loci sampled, we estimated the nucleotide diversity between derived and ancestral populations of the different *D. pseudoobscura* inversions. These data are shown in **Table 2** and **Figure 3**.

In all inversions, nucleotide diversity was higher within the ancestral arrangements than in the derived arrangement except at the dHYST region of AR, the pSTPP, pHYST and dHYST

regions of PP, and the dSCTL and pSCCH loci of CH. Some loci showed equal levels of diversity between both ancestral and derived arrangements such as the pSTAR locus within the CH arrangement and the pHYSC locus within the PP arrangement. At the ends of the chromosome, nucleotide diversity was reduced in both derived and ancestral arrangements within all inversions. The AR inversion had only one locus (dHYST) that showed higher nucleotide diversity in the derived arrangement than in the ancestral arrangement. The estimates of nucleotide diversity within the ancestral arrangements ranged from 0.007 to 0.046 in AR, from 0.008 to 0.037 in PP, from 0.006 to 0.033 in CH and from 0.011 t0 o.042 in ST. In the derived arrangements, nucleotide diversity ranged from 0.004 to 0.018 in AR, from 0.002 to 0.027 in PP, from 0.004 to 0.022 in CH and from 0.001 to 0.022 in ST.

Navarro et al. (2000) employed a coalescent approach and showed that low levels of nucleotide variability are expected near breakpoints of new inversions and higher levels of variability within the central regions of the inverted segment. In older inversions that have reached equilibrium, the variability at the breakpoints becomes higher than within the inverted segment. We tested the Navarro *et al.* (2000) predictions using the nucleotide variability estimates from the 11 inversion breakpoint regions examined. The location of the breakpoints on the chromosome varies in the different inversion types. **Figure 1** shows the orientation of the sampled breakpoint regions in seven different gene arrangements. It was essential to map the sequenced breakpoint regions on each chromosomal so that we know each marker's chromosomal context with respect to the proximal, inverted, and distal regions. **Figure 3** shows the nucleotide diversity estimates along four *D. pseudoobscura* chromosomal arrangements (AR, PP, ST, and CH). The patterns of nucleotide diversity shown by Navarro et al (2000) were not consistently observed here. Nucleotide diversity at the breakpoint regions of the AR arrangement

(a new inversion) was greater than nucleotide diversity within the central regions of the inverted segment (**Figure 3**A). However, within the PP inversion, which is also considered a new inversion, nucleotide diversity was reduced at regions closer to the proximal and distal breakpoints than at loci within the inverted segment (**Figure 3B**). Therefore, the PP arrangement satisfied the predictions of Navarro et al (2000). The prediction that nucleotide variability would be greater at the breakpoint regions of older inversions was not completely supported by our data. In **Figure 3**C and 3D, nucleotide diversity fluctuated between the regions closer to the breakpoints and regions within the inverted segment. For example, in the CH arrangement, π was 0.004 at the distal breakpoint region and 0.022 at the pSTAR locus which is within the inverted segment.

Linkage Disequilibrium among Loci. To further assess the effects of inversions on rates of recombination, we estimated patterns of linkage disequilibrium (LD) which is defined as non-random association of alleles. Patterns of LD can be affected by levels of recombination, the size and structure of the population being studied, selection, chromosomal location, the age of the alleles, and other factors. More specifically, significant LD can be the result of low levels of recombination or strong positive selection. Pairs of segregating sites across all gene arrangements were tested for significant LD. A total of 51,905 of the possible 226,248 pairwise comparisons of the 644 segregating sites were capable of rejecting the null hypothesis of no association with Fisher's exact test. Intralocus and interlocus LD for the 18 genetic markers were estimated as the percentage of comparisons that were in significant LD (**Figure 4**). Among all the different regions, the pSCCH region had the greatest levels of intralocus LD. Also, when LD at the pSCCH region was compared to other breakpoint regions, strong LD (>40%) was observed in the pSCTL/pHYSC, dSTAR and dSCCH regions which suggested strong interlocus LD. On

the other hand, the dHYST and dHYSC regions had strong intralocus LD, but little to no evidence of interlocus LD. The pSTAR region showed some amount of intralocus LD but no evidence of LD when compared to other regions. The pHYSC and dSTAR markers also had intralocus LD values greater than 50%. All the breakpoint regions showed some evidence on intra- and inter-locus LD while not all the non-breakpoint regions showed evidence in LD. The *EcR* locus showed no significant evidence of intra- and inter-locus LD. Overall, the breakpoint regions showed higher levels of intralocus LD than any non-breakpoint regions, except for the *Mef2* and *Amy1* loci which showed LD levels greater than some breakpoint regions. Overall, the EcR locus had the least amount of intralocus LD and the pSCCH breakpoint region the most intralocus LD. There was strong to moderate evidence of interlocus LD among all regions except the pSTAR and EcR regions.

There were no clear intralocus LD patterns observed between proximal and distal breakpoint regions. One out of six distal breakpoint regions (dSTAR) had greater than fifty percent intralocus LD and two (pSCCH and pHYSC/pSCTL) out of five proximal breakpoint regions possessed more than fifty percent intralocus LD. Although only three markers showed strong intralocus LD (>50%), all breakpoint regions and non-breakpoint regions had some evidence of intralocus LD with the exception of the *EcR* locus. Interlocus LD appeared to be lower for the markers closer to the ends of the chromosome than for markers in the center of the chromosome. The dHYST and EcR markers are closest to the distal end of the chromosome and showed low levels of interlocus LD while the pSTPP and en markers closest to the proximal end of the chromosome showed low interlocus LD. All other markers positioned on other areas of the chromosome have higher levels of interlocus LD with the exception of the pSTAR region that showed no interlocus LD.

LD was estimated between nucleotide sites and the four different gene arrangements (AR, PP, ST and CH). A total of 748 of the possible 3,220 pairwise comparisons of the 644 segregating sites were capable of rejecting the null hypothesis of no association between the segregating site and the inversion type with Fisher's exact test. It was shown that the PP arrangement had the highest levels of LD and the AR arrangement had the lowest levels of LD. The EcR locus showed no evidence of LD in any of the gene arrangements. The pSTPP marker only showed evidence of LD with the PP arrangement. The markers at the extreme end of the chromosome showed no evidence of LD with the arrangements with the exception of the pSTPP breakpoint region in the PP arrangement. The PP arrangement had the longest continuous segment of LD along the chromosome.

Gene Conversion. The predictions of low variability at new inversion breakpoints by Navarro et al (2000), involved the estimation of variability while assuming constant gene conversion. Gene conversion and crossing over are the two main forces driving recombination and therefore, we may have included gene conversion events in our original analysis. To better compare our findings with the predictions of Navarro et al (2000), we determined if there was evidence for gene conversion tracts using the method of Betran *et al.* (1997) by comparing all sets of chromosomal arrangements at all the sampled regions (AR vs ST, AR vs PP, etc.). All 11 regions showed evidence of gene conversion among arrangements. The number of sequences containing gene conversion tracts ranged from 3 at the dHYSC region to 10 at the dSCTL breakpoint region.

Discussion

Nucleotide Polymorphism and Divergence at *D. pseudoobscura* Breakpoint Regions.

Using the Counting model, Navarro *et al.* (1997) showed that within breakpoint pairs, recombination rates are greater in the distal segment than in the proximal segment (NAVARRO *et al.* 1997). Therefore, the nucleotide heterozygosity levels at breakpoint regions shown in **Figure 2** satisfy this prediction. With the exception of two pairs of breakpoint regions (SCCH and HYST) θ was higher at the distal segment than at the proximal segment. Previous studies have linked this particular pattern to the production of inviable gametes during crossover. The production of inviable gametes occurs when crossing over occurs in the proximal segment and within the inverted segment while crossing over in the distal segment has no impact on fertility (CARSON 1953; NAVARRO *et al.* 1997). Overall, nucleotide heterozygosity was consistently low at the proximal end and the distal end of the chromosome. This is consistent with earlier findings (NAVARRO *et al.* 2000; NAVARRO *et al.* 1997) explained by the low gene flux rates in these regions. According to Navarro et al (2000), in areas where flux rates are low, the effect of a partial sweep is greater and variability is reduced within the new inversion. This means that the approach towards equilibrium will be slower.

The higher levels of nucleotide heterozygosity within the PP arrangement at the pHYSC/pSCTL and the pHYST (**Figure 2**) breakpoint regions suggests that the PP arrangement is contributing more to the overall diversity within these breakpoint regions than any other breakpoint region. This may be explained by overall size of the PP inversion. The PP inversion is significantly larger than any of the other inversions studied here (**Figure 1**). Because the PP inversion covers ~75% of the chromosome, there are more chances for this inversion to share sequence information with other arrangements. The heterozygosity within the PP arrangement

was greater than the overall diversity at these two breakpoints (nucleotide diversity among all arrangements at the breakpoint). The PP is shown to be a newer inversion and new inversions may not have been in existence long enough to share genetic information with other gene arrangements of *D. pseudoobscura* in natural populations (NAVARRO *et al.* 2000). It is also possible that the PP arrangement is not widely or evenly distributed throughout the different geographical locations and this would provide lower chances of genetic exchange with other arrangements. Most of the PP arrangement sequenced in this study was collected from the Davis Mountain in Texas. A dataset containing equal numbers of arrangements from the different localities may resolve this concern.

Breakpoint Regions Fail to Reject Selective Neutrality. Evidence has been previously provided to support the occurrence of selection in *Drosophila pseudoobscura* populations (Dobzhansky 1943; Wright and Dobzhansky 1946). To determine if the regions we sequenced show any evidence of selection we applied the Tajima's D statistical test to our data set. Tajima's D test is able to determine if selection is occurring in a DNA sequence and is also able to distinguish between a sequence evolving under a neutral process rather than a non-neutral process. The negative Tajima's D values shown for majority of the regions sampled here support the recent population expansion of *D. pseudoobscura,* signifying positive selection.

DNA Divergence of derived *D. pseudoobscura* arrangements from ancestral arrangements. According to the predictions of Navarro et al (2000), patterns of nucleotide variability along the chromosome of gene arrangements depend on the age of the inversion. This prediction is based on theory that shows that as new inversions increase in frequency, they can eliminate variability in a large segment of the chromosome by means of a possible selective sweep. The resulting effect is low variability at the breakpoints of new inversions and higher

variability within the inverted segment. However, older inversions that have reached equilibrium will show greater variability at the breakpoints than within the inverted segment. We failed to provide strong support for these predictions in this study. One new inversion (PP) satisfied the prediction of Navarro et al (2000) of low variability at the breakpoint regions than within the inverted segment. One reason our results may not agree is the possibility of sequences being moved between inversions by the process of gene conversion affecting nucleotide variability estimates. In the Navarro *et al.* (1997) study, constant gene conversion was assumed. Therefore, we repeated our analysis eliminating the sequences that possessed gene conversion events. The results remained consistent with our initial findings (Supplemental **Figure S1**). Though unlikely, another possibility is that none of the gene arrangements studied here have reached equilibrium. Several studies have generated phylogenies for these *D. pseudoobscura* arrangements showing the order in which they occur (Wallace and Schaeffer, unpublished). However, these inversions may not have existed long enough to have achieved equilibrium. According to Navarro et al (2000), it takes at least 10^7 generations for an inversion to achieve mutation-drift-flux equilibrium. The Tree Line (TL) arrangement is considered one of the oldest *D. pseudoobscura* arrangements (AQUADRO *et al.* 1991). Once enough TL flies are collected and sequenced, it would be interesting to see if nucleotide variability patterns satisfy the predictions made by Navarro *et al.* (2000). Based on the age of this arrangement, we would expect to see greater heterozygosity levels at the breakpoint regions of the inverted segment than within the inverted segment. For example, the pattern that was observed for the ST to AR inversion (**Figure 3A**). The heterozygosity estimates at the breakpoint markers (pSTAR and dSTAR) were higher than estimates obtained for any of the markers between those two regions. We also failed to show the patterns predicted to occur when comparing nucleotide variability between the proximal and

distal breakpoints of gene arrangements (NAVARRO *et al.* 1997). The expected pattern is greater levels of nucleotide diversity within the distal regions than within the proximal regions. Within all four gene arrangements studied, nucleotide variability was greater at the proximal breakpoint region of the inversion than at the distal region. For example, in the CH arrangement the pSCCH region possessed more than twice the nucleotide variability of the dSCCH region (**Figure 3**).

Linkage disequilibrium. The results of the LD analysis within each arrangement provide a link between estimates of LD and the length of inversions. The PP arrangement had significant levels of LD distributed over a large portion of the chromosome (**Figure 4**). In fact, the pSTPP marker only showed evidence of LD in the PP arrangement and it is close to the proximal breakpoint region of the PP chromosome. This is concordant with the theory that newer inversions have low rates of recombination near the inversion breakpoints and low rates of recombination are indicated by significant LD. The breakpoint regions within the other gene arrangements also show significant LD, except for the AR arrangement. Overall, inversion breakpoint regions tend to provide more evidence of LD than non-breakpoint regions. In addition to estimating rates of nucleotide diversity, inversion breakpoints may provide information regarding rates of recombination by estimates of linkage disequilibrium.

The central region of the chromosome harbors the highest levels of LD. The central region of the chromosome is expected to be inverted in most gene arrangement heterozygotes. Thus, the central regions would be expected to experience the strongest amount of recombination suppression compared to either the proximal or distal regions. The reduced recombination leads to strong non-random associations between sites.

One surprising observation was that the two breakpoints for the Standard to Arrowhead inversion are not in LD with each other, despite being captured by the inversion at the same time.

All of the other inversion breakpoints show sites in strong LD with each other. In addition, the pSTAR breakpoint region does not show strong associations with other loci on the chromosome. These data suggest that variation at pSTAR is being shuffled with respect to the nucleotide diversity at all other loci. One might expect low levels of LD if this locus had few segregating sites all with low frequency, but nucleotide heterozygosity in this region is between 0.02 and 0.025, not particularly low values among the loci sampled. Further investigation of this region is necessary to understand its odd pattern of LD.

References

Anderson WW, Arnold J, Baldwin DG, Beckenbach AT, Brown CJ, Williams GO, Bryant S, Coyne J, Harshman LG, Heed WB et al. . 1991. Four decades of inversion polymorphism in *Drosophila pseudoobscura*. Proceedings of the National Academy of Sciences of the United States of America 88(22):10367-10371.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. Human Molecular Genetics 18(14):2555-2566.

Aquadro CF, Begun DJ, Kindahl EC. 1994. Selection, recombination, and DNA polymorphism in *Drosophila*. Non-neutral evolution. Theories and molecular data.:46-56.

Aquadro CF, Weaver AL, Schaeffer SW, Anderson WW. 1991. Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proceedings of the National Academy of Sciences of the United States of America 88:305-309.

Babcock CS, Anderson WW. 1996. Molecular evolution of the Sex-Ratio inversion complex in *Drosophila pseudoobscura*: analysis of the Esterase-5 gene region. Molecular Biology and Evolution 13(2):297-308.

Bartolome C, Charlesworth B. 2006a. Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. Genetics 174(4):2033-2044.

Bartolome C, Charlesworth B. 2006b. Rates and Patterns of Chromosomal Evolution in *Drosophila pseudoobscura* and *D. miranda*. Genetics 173(2):779-791.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356(6369):519-520.

Betran E, Rozas J, Navarro A, Barbadilla A. 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146(1):89-99.

Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. 2008. Chromosomal Rearrangement Inferred From Comparisons of 12 *Drosophila* Genomes. Genetics 179(3):1657-1680.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons From Human, Mouse, and Rat Genomes. Genome Research 14(4):507-516.

Brehm A, Krimbas CB. 1991. Inversion polymorphism in *Drosophila obscura*. Journal of Heredity 82:110-117.

Carson HL. 1953. The effects of inversions on crossing over in *Drosophila robusta*. Genetics 38(2):168-186.

Charlesworth B, Charlesworth D. 1973. A study of linkage disequilibrium in populations of *Drosophila melanogaster*. Genetics 73(2):351-359.

Chen J-M, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP. 2010. Genomic rearrangements in inherited disease and cancer. Seminars in Cancer Biology 20(4):222-233.

Chovnick A. 1973. Gene conversion and transfer of genetic information within the inverted region of inversion heterozygotes. Genetics 75:123-131.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. 2002. A Polytene Chromosome Analysis of the *Anopheles gambiae* Species Complex. Science 298(5597):1415-1418.

Dobzhansky T. 1943. Genetics of natural populations. Genetics 28:162-186.

Dobzhansky T. 1944. Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*. Carnegie Institute of Washington Publication 554:47-144.

Dobzhansky T. 1948. Genetics of natural populations. Genetics 33:588-602.

Dobzhansky T. 1950. The genetics of natural populations. Genetics 35:288-302.

Dobzhansky T, Epling C. 1944. Taxonomy, geographic distribution, and ecology of *Drosophila pseudoobscura* and its relatives. Carnegie Institute of Washington Publication 554:1-46.

Dobzhansky T, Queal ML. 1938. Genetics of natural populations. II. genic variation in populations of *Drosophila pseudoobscura* inhabiting isolated mountain ranges Genetics 23(5):463-484.

Dobzhansky T, Sturtevant AH. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics 23:28-64.

Emmens CW. 1937. Salivary gland cytology of roughest[3] inversion and reinversion, and roughest[2]. Journal of Genetics 34:191-202.

Engels WR, Preston CR. 1984. Formation of chromosome rearrangements by P factors in *Drosophila*. Genetics 107:657-678.

Ester A, Francesca V, Josep E, Joan B. 2006. Genetic reproductive risk in inversion carriers. Fertility and sterility 85(3):661-666.

Gloor G, Engels W. 1992. Single-fly DNA preps for PCR. *Drosophila* Information Service 71:148-149.

Green MM. 1980. Transposable elements in *Drosophila* and other Diptera. Annual Review of Genetics 14:109-120.

Gruneberg H. 1936. A case of complete reversion of a chromosomal rearrangement in *Drosophila melanogaster*. Nature 138:508.

Hamblin MT, Aquadro CF. 1999. DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics 153(2):859-869.

Hasson E, Eanes WF. 1996. Contrasting histories of three gene regions associated with In(3L)Payne of *Drosophila melanogaster*. Genetics 144(4):1565-1575.

Hickey DA, Bally-Cuif L, Abukashawa S, Payant V, Benkel BF. 1991. Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 88(5):1611-1615.

Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. 1994. Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics 137(4):1019-1026.

Hoffmann AA, Sgrò CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation. Trends in Ecology & Evolution 19(9):482-488.

Hudson RR, Kaplan NL. 1995. Deleterious Background Selection With Recombination. Genetics 141(4):1605-1617.

Innan H, Stephan W. 2000. The Coalescent in an Exponentially Growing Metapopulation and Its Application to *Arabidopsis thaliana*. Genetics 155(4):2015-2019.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics 123(4):887-99.

Kaufmann BP. 1942. Reversion from roughest to wild type in *Drosophila melanogaster*. Genetics 27:537-549.

Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: A syndrome of aberrant traits including mutation, sterility, and male recombination. Genetics 86(4):813-833.

Kim Y, Stephan W. 2000. Joint Effects of Genetic Hitchhiking and Background Selection on Neutral Variation. Genetics 155(3):1415-1427.

Kimura M. 1986. DNA and the Neutral Theory. Philosophical Transactions of the Royal Society of London 312(1154):343-354.

King RC, and Stansfield, W.D. 1985. A Dictionary of Genetics. Oxford University Press, New York.

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. Genetics 173(1):419-434.

Krimbas CB, Powell JR. 1992. *Drosophila* Inversion Polymorphism. Boca Raton, FL: CRC Press. p. 1-52.

Kumar S, Nei M, Dudley J, Tamura K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9(4):299-306.

Lande R. 1984. The expected fixation rate of chromosomal inversions. Evolution 38(4):743-752.

Lemeunier F, Aulard S. 1992. Inversion polymorphism in *Drosophila melanogaster*.339-405.

Lewontin RC. 1995. The Detection of Linkage Disequilibrium in Molecular Sequence Data. Genetics 140(1):377-388.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451-1452.

Lynch M, Crease T. 1990. The analysis of population survey data on DNA sequence variation. Mol Biol Evol 7(4):377-394.

Moriyama EN, Powell JR. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Molecular Biology and Evolution 13(1):261-277.

Morrow D. 1970. A Cytological Analysis of the Position of Tree Line in the Phylogeny of Gene Arrangements of the Third Chromosome of *Drosophila pseudoobscura* [MS. Thesis]. [Ithaca, NY]: Cornell University.

Nachman MW. 1997. Patterns of DNA Variability at X-Linked Loci in *Mus domesticus*. Genetics 147(3):1303-1316.

Nachman MW, Bauer VL, Crowell SL, Aquadro CF. 1998. DNA Variability and Recombination Rates at X-Linked Loci in Humans. Genetics 150(3):1133-1141.

Navarro A, Barbadilla A, Ruiz A. 2000. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. Genetics 155(2):685-98.

Navarro A, Betran E, Barbadilla A, Ruiz A. 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146(2):695-709.

Nei M. 1987. Molecular Evolutionary Genetics. New York, NY: Columbia University Press.

Nei M, Kojima KI, Schaffer HE. 1967. Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57:741-750.

Noor MAF, Garfield DA, Schaeffer SW, Machado CA. 2007. Divergence Between the *Drosophila pseudoobscura* and *D. persimilis* Genome Sequences in Relation to Chromosomal Inversions. Genetics 177(3):1417-1428.

Novitski E. 1951. Non-random disjunction in *Drosophila*. Genetics 36(3):267-280.

Novitski E. 1967. Nonrandom disjunction in *Drosophila*. Annual Review of Genetics 1:71-86.

Ohta T, Kojima K-I. 1968. Survival Probabilities of New Inversions in Large Populations. Biometrics 24(3):501-516.

Olvera O, Powell JR, de la Rosa ME, Salceda VM, Gaso MI, Guzman J, Anderson WW, Levine L. 1979. Population genetics of Mexican *Drosophila*. Evolution 33(1):381-395.

Painter TS. 1933. A new method for the study of chromosome rearrangements and the plotting of chromosome maps. Science 78:585-586.

Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America 100(13):7672-7677.

Popadic A, Anderson WW. 1994. The history of a genetic system. Proceedings of the National Academy of Sciences of the United States of America 91(15):6819-6823.

Popadić A, Anderson WW. 1995. Evidence for gene conversion in the amylase multigene family of *Drosophila pseudoobscura*. Molecular Biology and Evolution 12(4):564-572.

Powell JP. 1992 Inversion polymorphism in *Drosophila pseudoobscura* and *Drosophila persimilis*. p. 73-126.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP et al. . 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Research 15(1):1-18.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406-425.

Schaeffer SW. 2002. Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. Genetical Research 80(3):163-175.

Schaeffer SW, Aguade M. 2000. Evidence for balancing, directional, and background selection in molecular evolution. In: Singh RS, Krimbas CB, editors. Evolutionary Genetics: From Molecules to Morphology. Cambridge, UK: Cambridge University Press.

Schaeffer SW, Anderson WW. 2005. Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*. Genetics 171(4):1729-1739.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW et al. . 2008. Polytene Chromosomal Maps of 11 *Drosophila* Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics 179(3):1601-1655.

Schaeffer SW, Goetting-Minesky MP, Kovacevic M, Peoples JR, Graybill JL, Miller JM, Kim K, Nelson JG, Anderson WW. 2003. Evolutionary genomics of inversions in *Drosophila pseudoobscura*: Evidence for epistasis. Proceedings of the National Academy of Sciences of the United States of America 100(14):8319-8324.

Schaeffer SW, Miller EL. 1993. Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics 135(2):541-552.

Sperlich D, Pfreim P. 1986b. Chromosomal polymorphism in natural and experimental populations.: Academic Press, London.

Stebbins GL. 1945. Evidence for abnormally slow rates of evolution with particular reference to the higher plants and the genus *Drosophila*. Lloydia (Journal of Natural Products) 8:84-102.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG et al. . 2005. A common inversion under selection in Europeans. Nat Genet 37(2):129-137.

Storey JD. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(3):479.

Sturtevant AH. 1917. Genetic factors affecting the strength of linkage in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 3:555-558.

Sturtevant AH. 1926. A crossover reducer in *Drosophila melanogaster* due to inversion of a section of the third chromsome. Biologisches Zentralblatt 46:697-702.

Sturtevant AH, Beadle GW. 1936. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. Genetics 21:554-604.

Sturtevant AH, Dobzhansky T. 1936a. Geographical DistributionI and cytology of "sex ratio" in *Drosophila pseudoobscura* and related species Genetics 21(4):473-490.

Sturtevant AH, Dobzhansky T. 1936b. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. Proceedings of the National Academy of Sciences of the United States of America 22:448-450.

Tajima F. 1983. Evolutionary Relationship of DNA sequences in finite populations. Genetics 105(2):437-460.

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

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. Molecular Biology and Evolution 12(5):823-833.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Molecular Biology and Evolution 24(8):1596-1599.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America 101(30):11030-11035.

von Grotthuss M, Ashburner M, Ranz JM. 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus *Drosophila*. Genome Research 20(8):1084-1096.

Wallace B. 1966. Chromosomes, giant molecules, and evolution. New York, NY: W.W. Norton & Co.

Wallace B. 1968. Topics in population genetics. New York, NY: Norton.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7(2):256-276.

Wright S, Dobzhansky T. 1946. Genetics of Natural Populations. XII. Experimental reproduction of some of the Changes Caused by Natural Selection in Certain Populations of *Drosophila pseudoobscura* Genetics 31(2):125-156.

Table 1.
Nucleotide Polymorphism and Divergence at *D. pseudoobscura* Breakpoint Regions.

Breakpoint Region		*n*	*S*	$\pi \pm Sd$	$\theta \pm Sd$	*Tajima's D*
pSTPP	All	93	26	0.010 ± 0.001	0.014 ± 0.004	-0.87786
	AR	22	12	0.008 ± 0.001	0.009 ± 0.004	-0.61015
	PP	22	13	0.010 ± 0.002	0.010 ± 0.004	0.08691
	CH	24	12	0.009 ± 0.001	0.009 ± 0.004	-0.06156
	ST	22	11	0.009 ± 0.001	0.009 ± 0.004	-0.02011
pHYSC/pSCTL	All	91	25	0.026 ± 0.002	0.025 ± 0.008	0.09589
	AR	21	9	0.004 ± 0.001	0.007 ± 0.003	-1.11966
	PP	17	30	0.030 ± 0.003	0.026 ± 0.010	0.6452
	CH	25	14	0.011 ± 0.004	0.018 ± 0.007	-1.29905
	ST	25	11	0.007 ± 0.003	0.015 ± 0.006	-1.78821
pSTAR	All	93	81	0.022 ± 0.001	0.036 ± 0.010	-1.31943
	AR	23	37	0.017 ± 0.003	0.022 ± 0.008	-0.76617
	PP	21	31	0.019 ± 0.001	0.019 ± 0.007	0.00329
	CH	22	42	0.021 ± 0.002	0.025 ± 0.009	-0.61106
	ST	24	45	0.022 ± 0.002	0.027 ± 0.010	-0.75675
pHYST	All	74	59	0.017 ± 0.001	0.017 ± 0.005	-0.11853
	AR	21	34	0.009 ± 0.002	0.012 ± 0.004	-1.09202
	PP	20	48	0.023 ± 0.002	0.019 ± 0.007	0.84446
	CH	26	25	0.008 ± 0.001	0.008 ± 0.003	-0.06224
	ST	7	12	0.007 ± 0.001	0.006 ± 0.003	0.484
dSTPP	All	91	42	0.015 ± 0.0001	0.016 ± 0.004	-0.12442
	AR	23	24	0.008 ± 0.001	0.012 ± 0.005	-1.33946
	PP	20	14	0.003 ± 0.002	0.007 ± 0.003	-1.82978
	CH	24	24	0.009 ± 0.002	0.011 ± 0.004	-0.6391
	ST	20	7	0.002 ± 0.001	0.003 ± 0.002	-1.37596
dSCTL	All	92	45	0.020 ± 0.001	0.024 ± 0.007	-0.507
	AR	22	12	0.007 ± 0.002	0.008 ± 0.004	-0.67466
	PP	20	24	0.014 ± 0.003	0.017 ± 0.007	-0.697
	CH	24	20	0.014 ± 0.002	0.014 ± 0.005	-0.03747
	ST	23	10	0.006 ± 0.001	0.007 ± 0.003	-0.51888
pSCCH	All	86	38	0.023 ± 0.002	0.023 ± 0.007	-0.09399
	AR	21	8	0.003 ± 0.001	0.006 ± 0.003	-1.47353

	PP	21	14	0.012 ± 0.001	0.011 ± 0.004	0.41961
	CH	19	18	0.019 ± 0.005	0.014 ± 0.006	1.32963
	ST	22	7	0.003 ± 0.001	0.005 ± 0.002	-1.40374
dSTAR	All	104	39	0.039 ± 0.002	0.037 ± 0.010	0.33375
	AR	42	18	0.010 ± 0.001	0.011 ± 0.004	-0.2848
	PP	21	21	0.023 ± 0.005	0.023 ± 0.009	-0.04854
	CH	42	18	0.010 ± 0.001	0.011 ± 0.004	-0.2848
	ST	16	26	0.016 ± 0.002	0.016 ± 0.006	0.08609
dSCCH	All	82	18	0.011 ± 0.0001	0.010 ± 0.003	0.12893
	AR	20	8	0.006 ± 0.001	0.006 ± 0.003	-0.3024
	PP	22	6	0.002 ± 0.001	0.005 ± 0.002	-1.76226
	CH	17	5	0.004 ± 0.001	0.004 ± 0.002	0.14213
	ST	19	6	0.005 ± 0.001	0.005 ± 0.002	0.37985
dHYSC	All	83	40	0.015 ± 0.002	0.024 ± 0.007	-1.23262
	AR	23	13	0.012 ± 0.001	0.011 ± 0.004	0.55
	PP	19	13	0.011 ± 0.002	0.011 ± 0.005	-0.21404
	CH	19	14	0.011 ± 0.002	0.012 ± 0.005	-0.37342
	ST	19	12	0.013 ± 0.002	0.010 ± 0.005	0.89802
dHYST	All	92	50	0.010 ± 0.001	0.012 ± 0.003	-0.4989
	AR	23	29	0.011 ± 0.001	0.009 ± 0.003	0.81332
	PP	20	34	0.013 ± 0.001	0.011 ± 0.004	0.91343
	CH	24	33	0.007 ± 0.002	0.010 ± 0.004	-1.02524
	ST	22	21	0.007 ± 0.002	0.006 ± 0.002	0.10175

n, sample size; *S,* number of segregating sites; $\pi \pm$ sd, nucleotide diversity per site based on pairwise differences with its standard deviation; $\theta \pm$ sd, nucleotide diversity per site based on the number of segregating sites with its standard deviation assuming no recombination (Tajima 1983).

Table 2

DNA Divergence of derived *D. pseudoobscura* arrangements from ancestral arrangements

		pSTPP	pHYSC	pSTAR	pHYST	dSTPP	dSCTL	pSCCH	dSTAR	dSCCH	dHYSC	dHYST
AR	n	22	21	23	21	23	22	21	42	20	23	23
	k	2.719	0.695	7.708	6.81	4.000	2.567	0.981	3.314	2.053	4.075	8.893
	π	0.008	0.004	0.018	0.009	0.008	0.007	0.003	0.010	0.006	0.012	0.011
ST, CH, SC, TL	n	48	52	48	33	41	50	44	35	40	41	48
	k	3.053	4.957	9.703	6.223	7.568	6.665	8.124	15.089	2.991	5.905	6.291
	π	0.009	0.026	0.022	0.008	0.014	0.018	0.025	0.046	0.009	0.018	0.007
All	n	70	73	71	54	64	72	65	77	60	64	71
	k	2.935	4.494	9.285	6.941	7.456	6.667	6.641	15.18	3.214	5.214	7.455
	π	0.008	0.023	0.021	0.009	0.014	0.018	0.020	0.045	0.009	0.016	0.009

		pSTPP	dSCTL	pSCCH	pSTAR	pHYSC/ pSCTL	dSTPP	pHYST	dSTAR	dSCCH	dHYSC	dHYST
PP	n	22	20	21	21	17	20	20	21	22	19	20

	k	3.654	5.105	4.067	8.624	5.147	1.974	15.879	5.762	0.71	3.503	10.205
	π	0.01	0.014	0.013	0.02	0.027	0.004	0.023	0.024	0.002	0.011	0.012
ST, CH, SC, TL	n	49	50	44	49	53	47	33	35	40	41	48
	k	3.105	5.882	7.642	10.005	4.988	9.241	5.879	8.839	2.991	5.905	6.332
	π	0.009	0.016	0.024	0.020	0.027	0.017	0.008	0.037	0.009	0.018	0.008
All	n	71	70	65	70	70	67	53	56	62	60	68
	k	4.566	10.374	10.867	9.631	6.865	10.176	12.62	9.593	3.551	5.282	9.114
	π	0.011	0.021	0.025	0.022	0.030	0.016	0.018	0.038	0.010	0.016	0.009

		pSTPP	pHYSC/ pSCTL	pSTAR	pSCCH	dSCTL	dSTPP	pHYST	dSTAR	dSCCH	dHYSC	dHYST
ST	n	22	25	24	22	23	20	7	16	19	19	22
	k	3.000	1.360	9.688	0.455	1.889	1.147	5.048	3.350	1.918	4.281	5.082
	π	0.009	0.007	0.022	0.001	0.005	0.002	0.007	0.015	0.005	0.013	0.006
AR,PP,CH,SC,TL	n	70	66	69	64	69	70	67	86	63	64	70
	k	3.801	5.726	9.367	9.011	7.133	7.904	12.222	8.897	3.989	5.052	9.126

	π	0.011	0.031	0.022	0.029	0.020	0.015	0.018	0.042	0.011	0.002	0.011
All	n	70	66	69	64	69	70	74	102	82	83	70
	k	3.801	5.726	9.367	9.011	7.133	7.904	11.671	8.657	3.775	4.920	9.126
	π	0.011	0.031	0.022	0.029	0.020	0.015	0.017	0.039	0.011	0.015	0.011

		pSTPP	dHYSC	pHYST	dSTPP	dSCTL	pSCCH	pHYSC/ pSCTL	pSTAR	dSCCH	dSTAR	dHYST
CH	n	24	19	26	24	24	19	25	22	17	16	24
	k	3.156	3.602	6.071	4.917	4.446	6.550	2.160	9.632	1.544	3.142	6.366
	π	0.009	0.011	0.009	0.009	0.012	0.020	0.011	0.022	0.004	0.015	0.008
AR,PP,ST,SC,TL	n	68	64	48	66	66	65	65	70	65	86	67
	k	3.773	5.041	13.893	6.641	7.637	4.113	3.138	9.476	3.58	7.094	8.305
	π	0.011	0.016	0.020	0.013	0.021	0.012	0.017	0.022	0.010	0.033	0.010
All	n	92	83	74	90	90	84	90	92	82	102	91
	k	3.507	4.92	11.671	10.221	7.366	7.543	5.093	9.529	3.775	8.657	8.218
	π	0.010	0.015	0.017	0.015	0.020	0.021	0.026	0.022	0.011	0.039	0.010

Breakpoint regions are ordered horizontally based on their order on the chromosome of a specific inversion (**Figure 1**). n, number of sequences analyzed; k, average number of pairwise differences (Tajima 1983); π, nucleotide diversity per site calculated using k; "All", the combination of the two populations for each category.

Figure 1. Representation of the order of breakpoints on the chromosomes of *Drosophila pseudoobscura* gene arrangements. The arrows show the locations of the inversion breakpoints on the chromosome of different gene arrangements. The breakpoints appear in pairs and are color-coded according to the inversion phylogeny depicted on the right side of the figure. For example, red represents the PP arrangement.

Figure 1

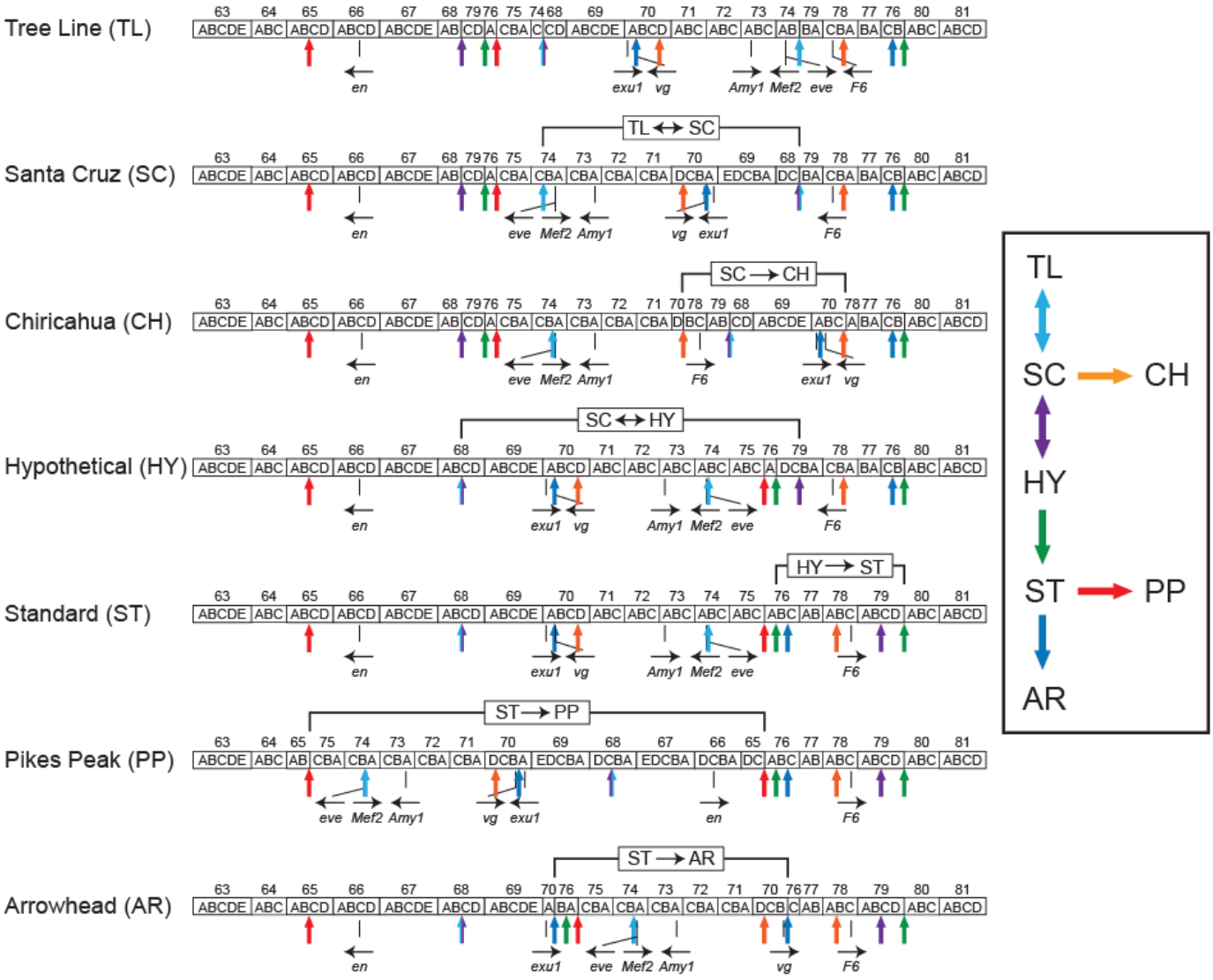
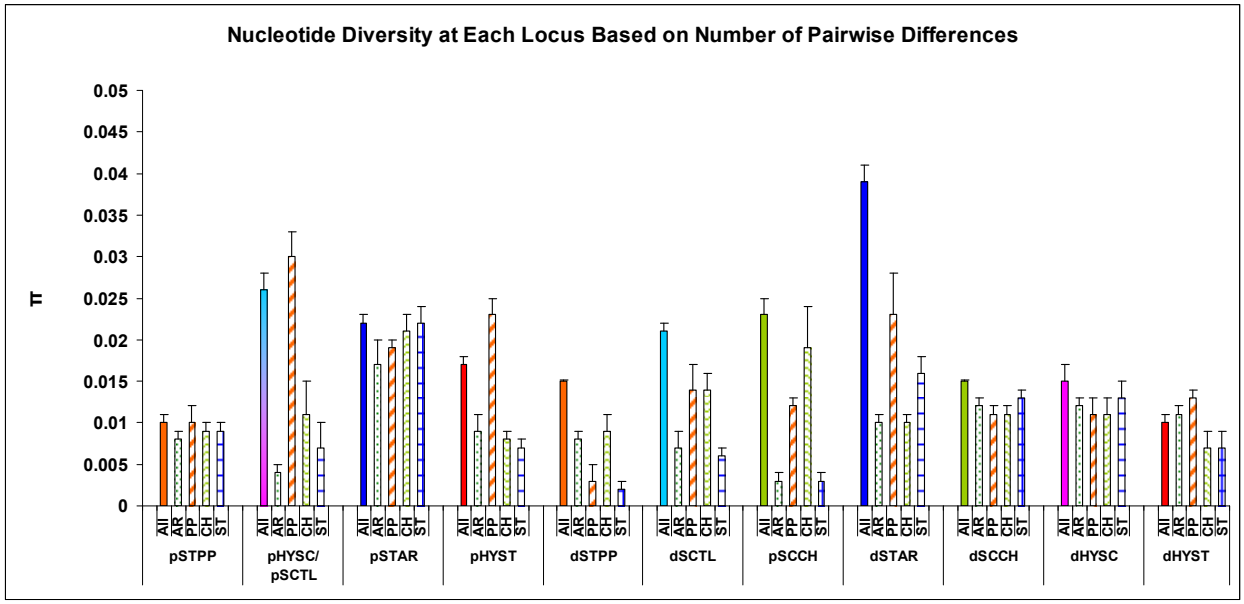

Tree Line (TL)
Santa Cruz (SC)
Chiricahua (CH)
Hypothetical (HY)
Standard (ST)
Pikes Peak (PP)
Arrowhead (AR)
TL ↔ SC
SC → CH
SC ↔ HY
HY → ST
ST → PP
ST → AR
en
exu1
vg
Amy1
Mef2
eve
F6
TL
SC
CH
HY
ST
PP
AR

Figure 2. Nucleotide heterozygosity and divergence in 11 *D. pseudoobscura* Breakpoints Regions. "All" represents the estimates for all samples at a specific loci. Nucleotide heterozygosity estimates within individual arrangements (AR, PP, CH and ST) are also represented here. Values are shown for two measures of nucleotide heterozygosity. (A) represents π and (B) represents θ. Error bars are based on standard deviation values. Each pair of breakpoint regions is represented by the same color.

Figure 2

A


Nucleotide Diversity at Each Locus Based on Number of Pairwise Differences
π
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
All AR PP CH ST
pSTPP
pHYSC/ pSCTL
pSTAR
pHYST
dSTPP
dSCTL
pSCCH
dSTAR
dSCCH
dHYSC
dHYST

B

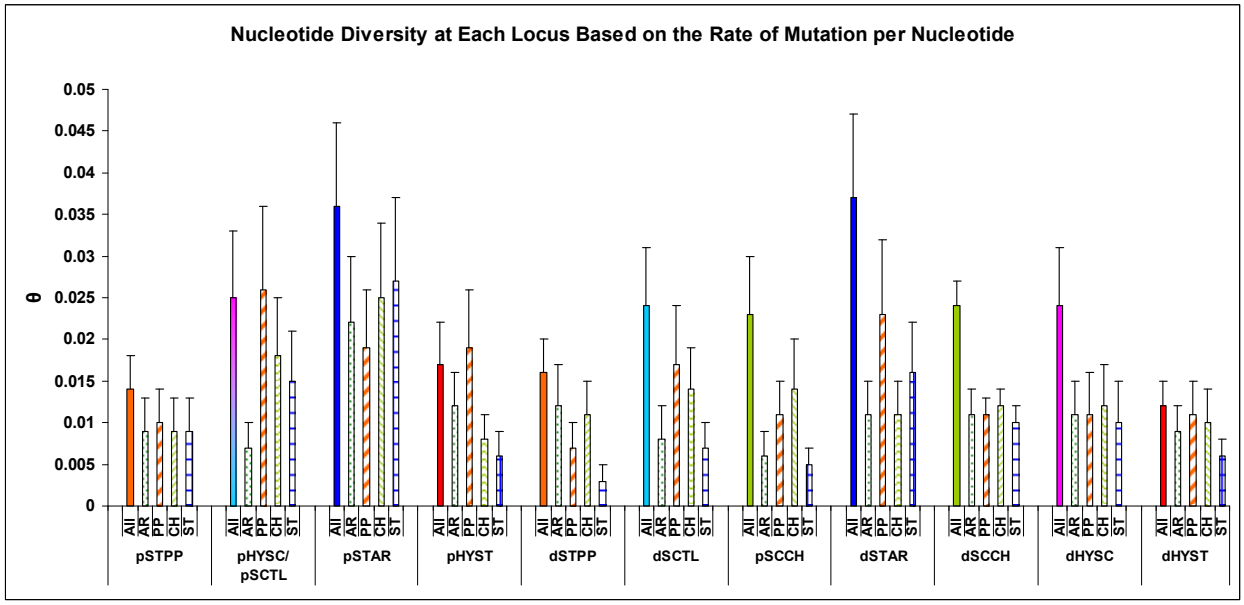
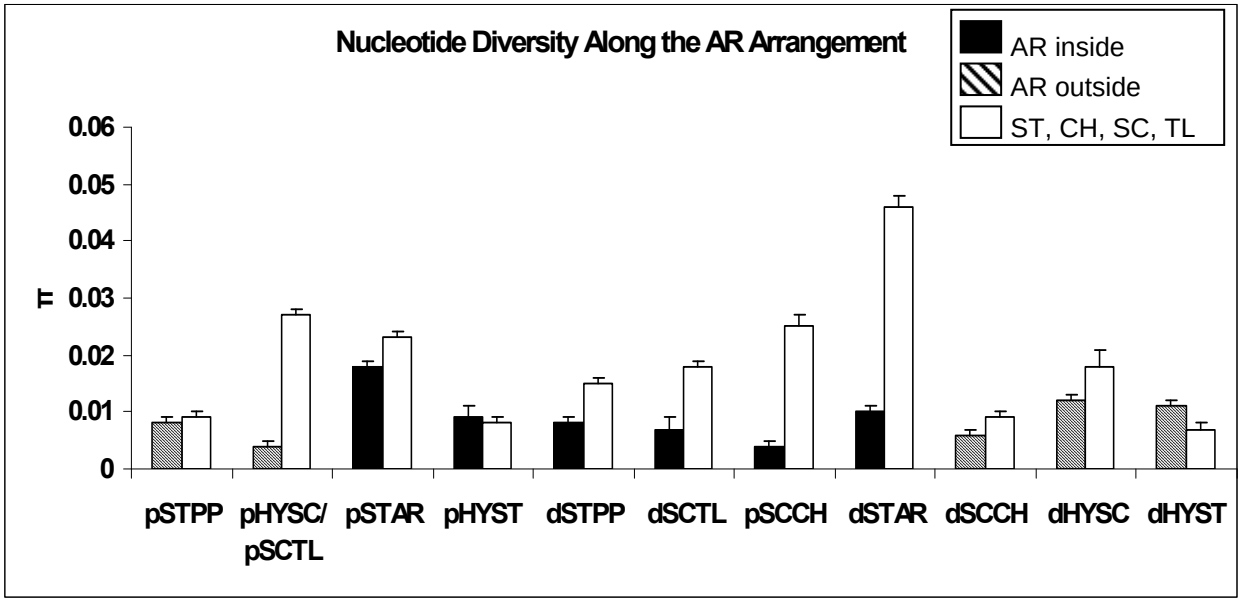

Nucleotide Diversity at Each Locus Based on the Rate of Mutation per Nucleotide
θ
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
All AR PP CH ST
pSTPP
pHYSC/ pSCTL
pSTAR
pHYST
dSTPP
dSCTL
pSCCH
dSTAR
dSCCH
dHYSC
dHYST

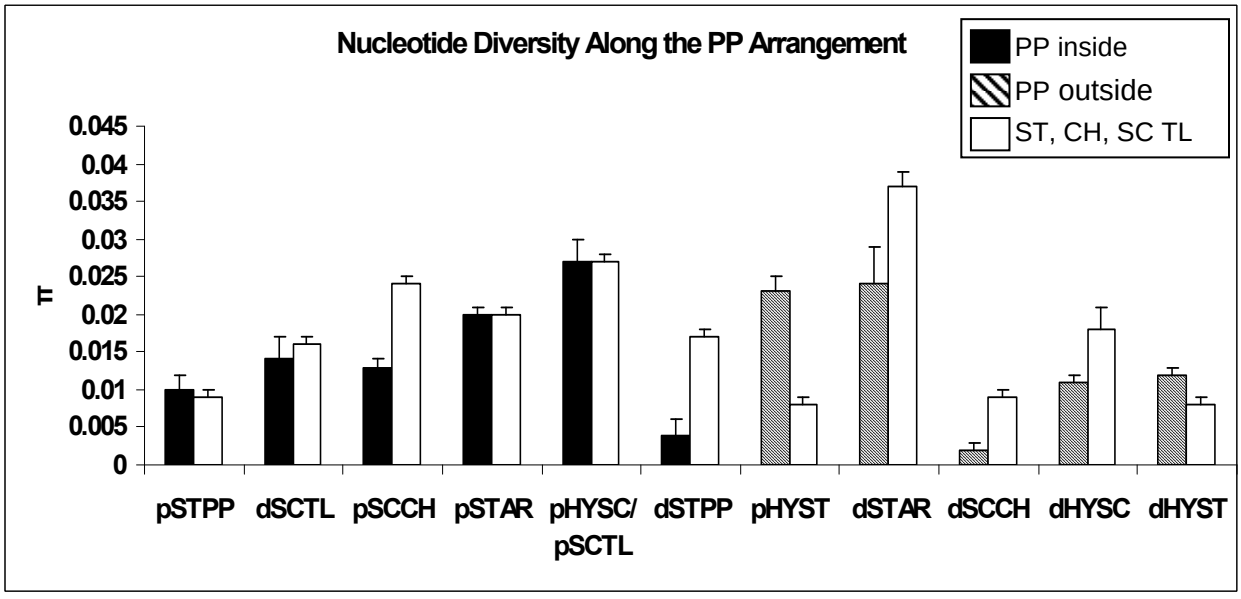
Figure 3. Nucleotide variability along the chromosome of four *D. pseudoobscura* inversions. The graph shows the nucleotide diversity along different arrangements: (A) Arrowhead, (B) Pikes Peak, (C) Standard, and (D) Chiricahua. The loci are organized based on their order on the chromosomes of the four inversions. Loci of the derived arrangements are represented by black columns when they are within the inverted segment and by striped columns outside the inverted segment. White columns represent loci for the ancestral arrangements. Error bars are standard deviations of π obtained using the Jukes and Cantor correction (Jukes and Cantor, 1969; (LYNCH and CREASE 1990).

Figure 3

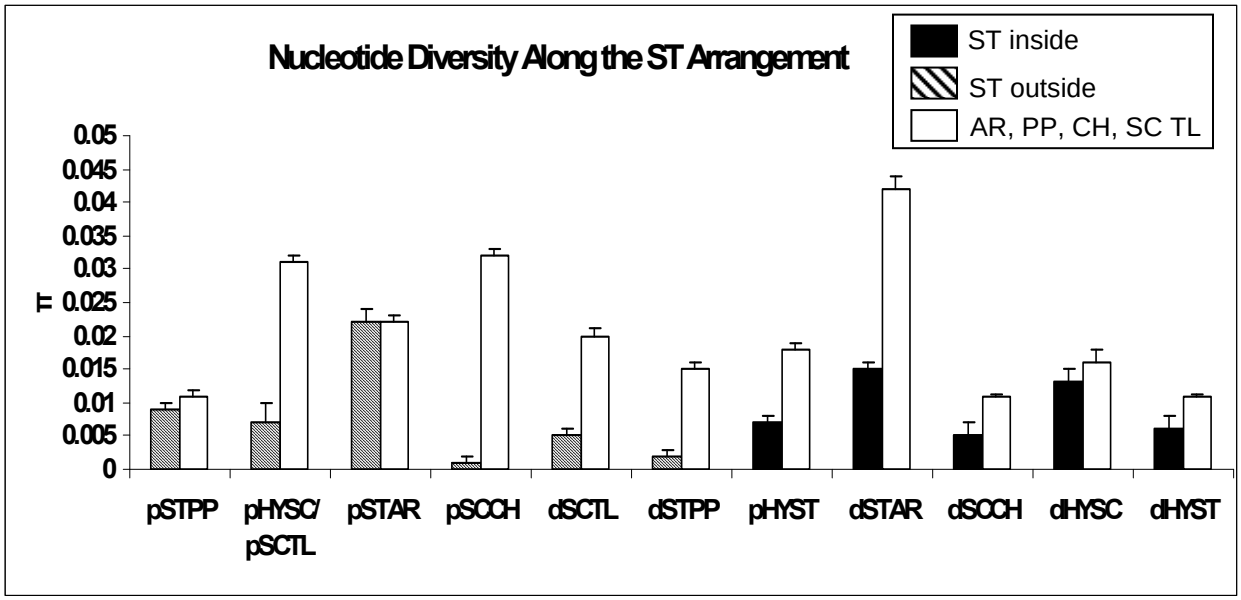
A


Nucleotide Diversity Along the AR Arrangement
AR inside
AR outside
ST, CH, SC, TL
π
0.06
0.05
0.04
0.03
0.02
0.01
0
pSTPP
pHYSC/ pSCTL
pSTAR
pHYST
dSTPP
dSCTL
pSCCH
dSTAR
dSCCH
dHYSC
dHYST

B


Nucleotide Diversity Along the PP Arrangement
PP inside
PP outside
ST, CH, SC TL
π
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
pSTPP
dSCTL
pSCCH
pSTAR
pHYSC/ pSCTL
dSTPP
pHYST
dSTAR
dSCCH
dHYSC
dHYST

C


Nucleotide Diversity Along the ST Arrangement
ST inside
ST outside
AR, PP, CH, SC TL
π
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
pSTPP
pHYSC/ pSCTL
pSTAR
pSCCH
dSCTL
dSTPP
pHYST
dSTAR
dSCCH
dHYSC
dHYST

D

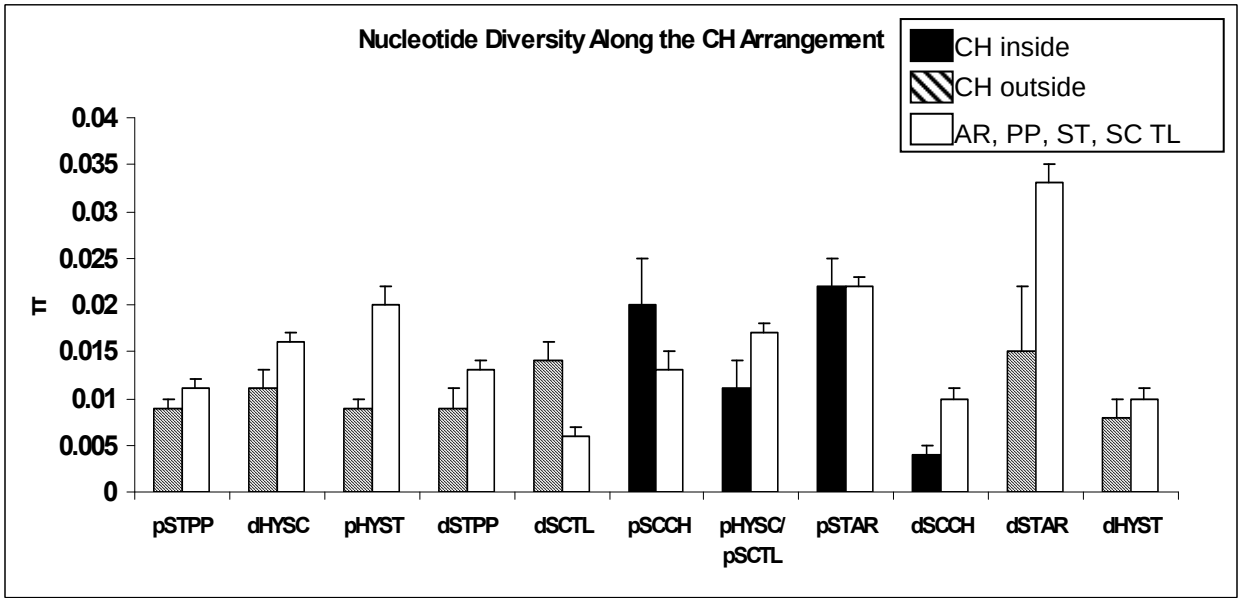
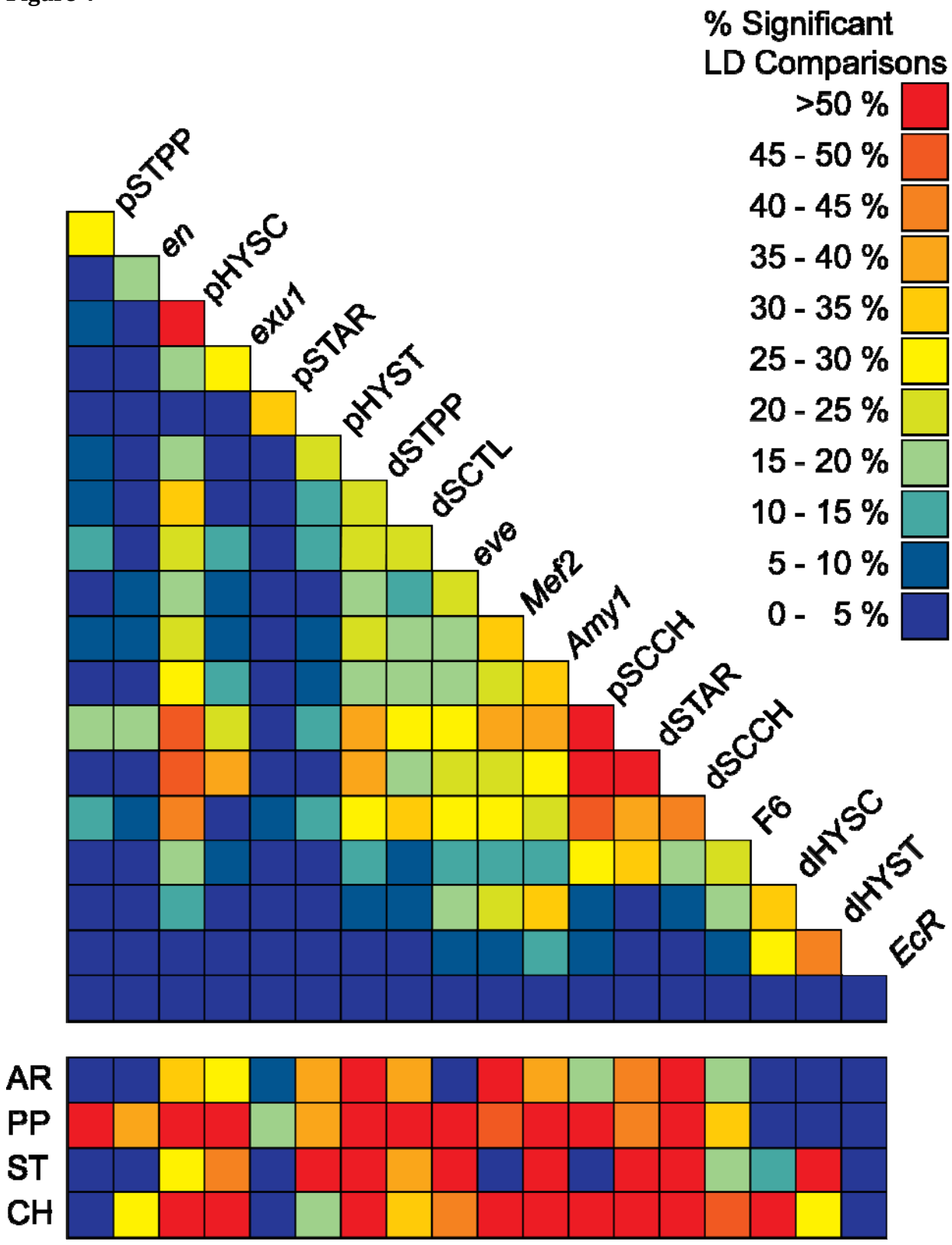

Nucleotide Diversity Along the CH Arrangement
CH inside
CH outside
AR, PP, ST, SC TL
π
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
pSTPP
dHYSC
pHYST
dSTPP
dSCTL
pSCCH
pHYSC/ pSCTL
pSTAR
dSCCH
dSTAR
dHYST

Figure 4. Pairwise linkage disequilibrium plot for within and between loci across the third chromosome of *D. pseudoobscura*. Levels of LD are represented as percentage of comparisons whether within or between loci that were significant with a probability of 0.05 and a false discovery rate of 1 % (STOREY 2002).

Figure 4


% Significant
LD Comparisons
>50 %
45 - 50 %
40 - 45 %
35 - 40 %
30 - 35 %
25 - 30 %
20 - 25 %
15 - 20 %
10 - 15 %
5 - 10 %
0 - 5 %
pSTPP
en
pHYSC
exu1
pSTAR
pHYST
dSTPP
dSCTL
eve
Mef2
Amy1
pSCCH
dSTAR
dSCCH
F6
dHYSC
dHYST
EcR
AR
PP
ST
CH

Supplemental Data

Table S1. Primer Sequences Used for PCR Amplification

Primer Name	Primer Sequence	Coordinate Interval	Length (bp)	Cytological Position
pSTPP_f	GAT ACC ACT CGG CAA GCA GAA G	2,292,730…2,293,129	354	64C
pSTPP_r	CGC CTC AGT TAA TTA GCC CAC AAA			
pHYSC_f	TGG TGT TGA GTA TCT GCC GTG GTT	6,432,192…6,432,612	376	68C
pHYSC_r	CTG CTG CCG CTG CTC CTA TCA			
pSTAR_f	CCT GAT ACC CAC GGA GTC TTC	8,900,335…8,900,844	468	76B
pSTAR_r	TCG CTA CAG GGA TCA GGT TTT			
pHYST_f	CTT ATT CCC GCC TCT TGT GTA GC	9,140,888..9,141,693	806	76B
pHYST_r	GAC GGC CCT CAG ACG ATA GTT G			
dSTPP_f	ATC GGT ACA ACA GCC AGG GAC AAC	9,832,232…9,832,852	573	75B
dSTPP_r	ACT TCG TGG GAT CGC TGG CAT AAT			
dSCTL_f	ATG GCG ATG GAG TCC TCT GTC TAT	10,830,454...10,830,904	404	74B
dSCTL_r	ACT GGC GCC ATG TCT CTG TCT CG			
pSCCH_f	AAC CGG CAT ACA CCC TCA TTC	14,259,585…14,260,006	377	70C
pSCCH_r	GTT GCG CAT TAT TTA TTC CCT GTA			
dSCCH_f	TCC GGA GAT CGC AAA ACT GTC G	15426271…15426670	379	77B
dSCCH_r	TAT GCG CTG CTT CTG ATG CTT GAT			
dHYSC_f	GAG CCC GGG CCA GGT CCA T	17,444,472…17,444,876	362	79C
dHYSC_r	TAT CGT GCG TTG TGC GTA ATC AGC			
dHYST_f	ACA AGA TCC GGG GTA TTA	17,705,213..17,706,152	898	79D/80A
dHYST_r	CTG TTC CGG GTA GAT GTA TTC GTA		898	

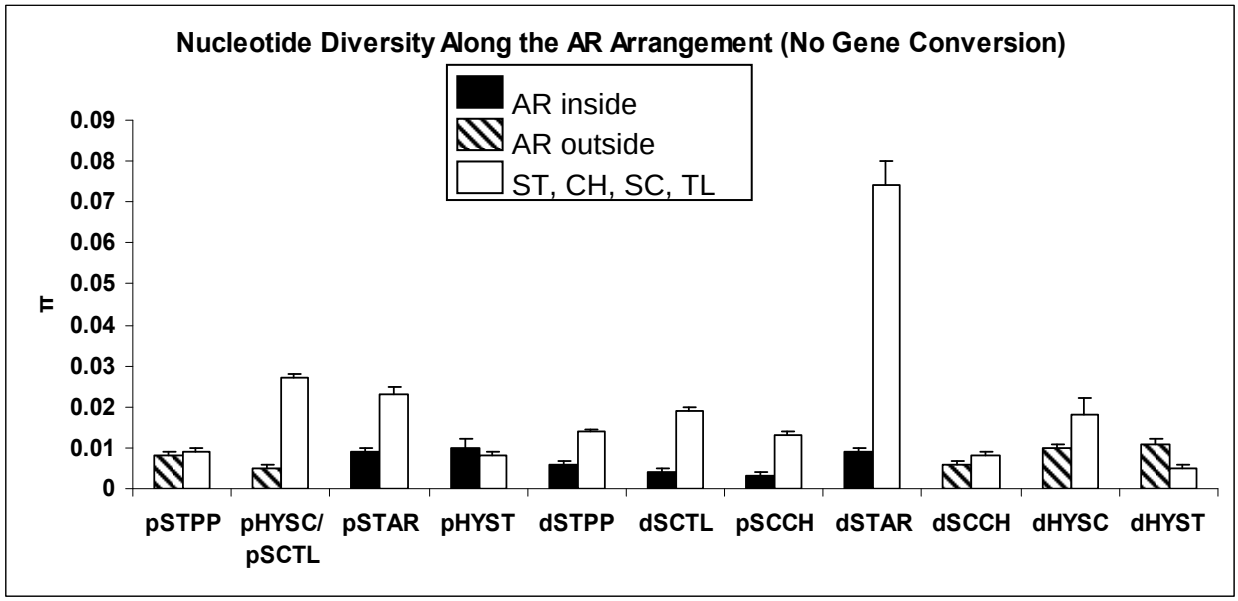
The abbreviated name represents the location (p, proximal or d, distal) of the breakpoint on the chromosome, and the last four letters represent the two arrangements involved in the inversion. The first two letters are the ancestral arrangement and the last two letters are the derived arrangement. For example, the STPP notation is for the breakpoints that converted the ancestral

Standard arrangement into the derived Pikes Peak arrangement. Primer names are the abbreviated region names with the addition of "f" and "r" for forward or reverse. The coordinates are the location of the genetic marker in the genome strain (Richards *et al.* 2005), which carries the Arrowhead arrangement.

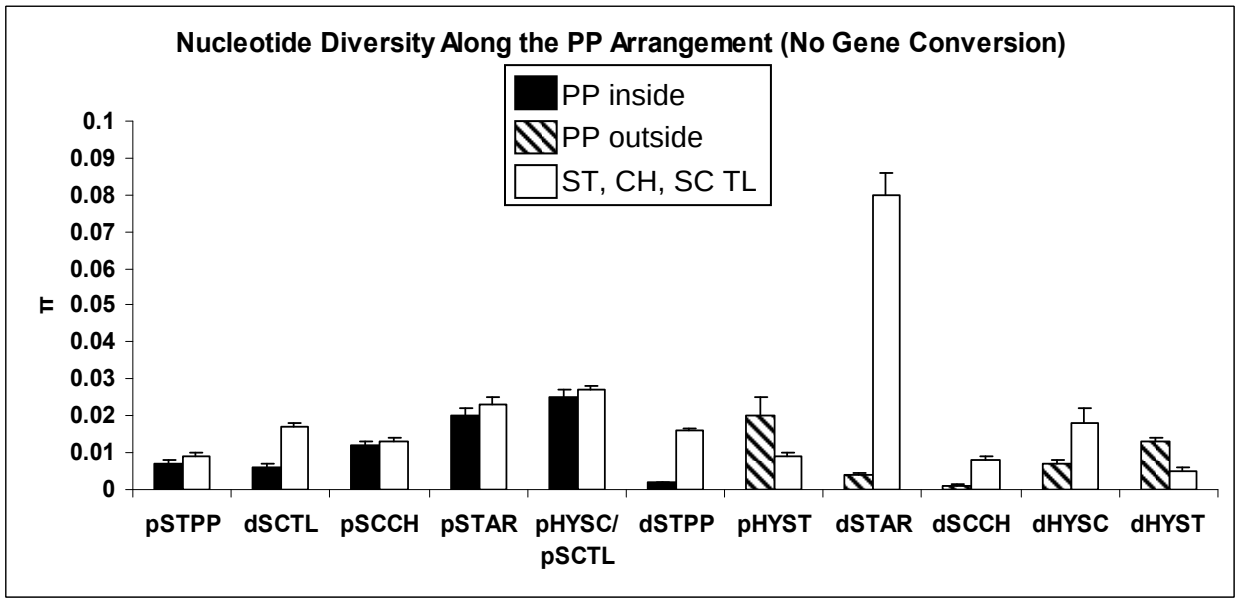
Figure S1. Nucleotide variability along the chromosome of four *D. pseudoobscura* inversions. The graph shows the nucleotide diversity along the (A) Arrowhead, (B) Pikes Peak, (C) Standard, and (D) Chiricahua arrangements. The loci are organized based on their order on the chromosomes of the four inversions. Loci of the derived arrangements are represented by black columns when they are within the inverted segment and by striped columns outside the inverted segment. White columns represent loci for the ancestral arrangements. Error bars are standard deviations of π obtained using the Jukes and Cantor correction (Jukes and Cantor, 1969; (LYNCH and CREASE 1990). This analysis is similar to that performed in **Figure 3** above, except that all sequences involved in gene conversion events were excluded from the analysis.

Figure S1.

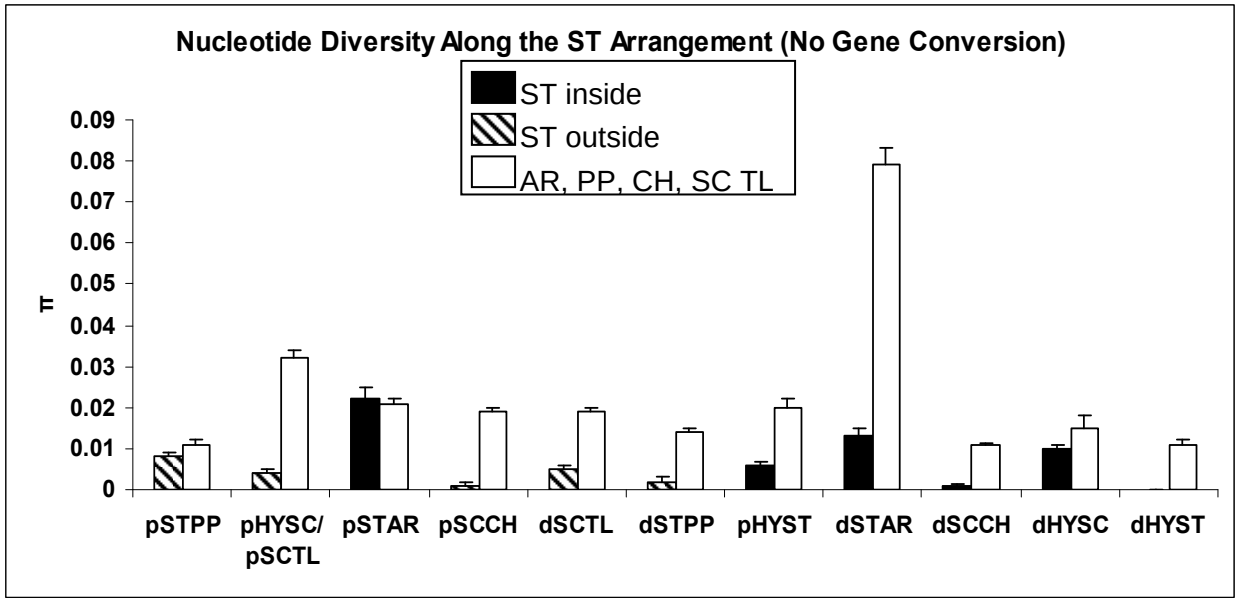
A


Nucleotide Diversity Along the AR Arrangement (No Gene Conversion)
AR inside
AR outside
ST, CH, SC, TL
π
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
pSTPP
pHYSC/ pSCTL
pSTAR
pHYST
dSTPP
dSCTL
pSCCH
dSTAR
dSCCH
dHYSC
dHYST

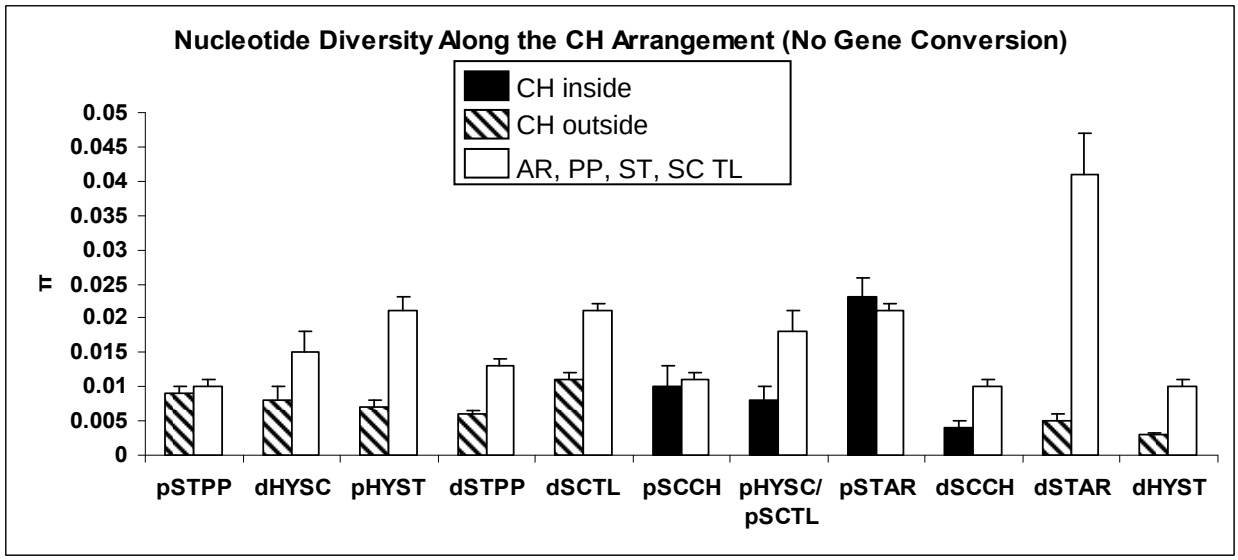
B


Nucleotide Diversity Along the PP Arrangement (No Gene Conversion)
PP inside
PP outside
ST, CH, SC TL
π
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
pSTPP
dSCTL
pSCCH
pSTAR
pHYSC/ pSCTL
dSTPP
pHYST
dSTAR
dSCCH
dHYSC
dHYST

C


Nucleotide Diversity Along the ST Arrangement (No Gene Conversion)
ST inside
ST outside
AR, PP, CH, SC TL
π
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
pSTPP
pHYSC/ pSCTL
pSTAR
pSCCH
dSCTL
dSTPP
pHYST
dSTAR
dSCCH
dHYSC
dHYST

D


Nucleotide Diversity Along the CH Arrangement (No Gene Conversion)
CH inside
CH outside
AR, PP, ST, SC TL
π
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
pSTPP
dHYSC
pHYST
dSTPP
dSCTL
pSCCH
pHYSC/ pSCTL
pSTAR
dSCCH
dSTAR
dHYST

CHAPTER 4

CONCLUDING REMARKS

Concluding Remarks

The experiments in this thesis addressed two major components of the *Drosophila pseudoobscura* inversion polymorphism: evolutionary history of the gene arrangements and the population genetics of nucleotide diversity in a suppressed recombination environment.

In the first set of experiments, we investigated the evolutionary history of *D. pseudoobscura* inversion polymorphism addressing questions regarding its origin, phylogeny and ancestry. Recombination can obscure a true phylogeny by causing unrelated individuals to appear more similar than they actually are. Therefore, regions that experience low levels of genetic exchange are ideal candidates for inferring evolutionary relationships. We used breakpoints of inversion events to infer the history of the gene arrangements in *D. pseudoobscura* because recombination is less likely to occur there than in the central parts of the inversions. In phylogenetic trees produced with individual breakpoint regions, no monophyly of the arrangements was observed except for the breakpoint regions found in the central region of the *D. pseudoobscura* third chromosome. This monophyletic pattern in the central region of the chromosome was validated when markers were concatenated according to their position on the chromosome and phylogenetic trees were produced. If estimates of nucleotide diversity within these central loci are lower than nucleotide diversity at the proximal or distal regions, this result could be combined with the phylogenetic results to conclude that these regions (central) are segments with low recombination rates. Using a concatenated data set of breakpoint and non-breakpoint regions we were able to provide support for a unique origin of *D. pseudoobscura* arrangements. The PP arrangement shares genetic information with other gene arrangements by gene conversion. Our molecular genetic phylogeny strongly agrees with the proposed cytological phylogeny of the *D. pseudoobscura* arrangements because when we concatenated our complete

data set, we obtained a tree with strongly supported clades that is concordant with this phylogeny. From our results, we were not able to determine the exact ancestral arrangement but we were able to eliminate the previously suggested ST arrangement as ancestral. Our data suggests that either the HY or the SC arrangement is the ancestral *D. pseudoobscura* arrangement.

The second set of experiments utilized molecular and genomic approaches to investigate how nucleotide diversity within and between *D. pseudoobscura* inversions was altered by differences in levels of genetic exchange across the chromosome. Based on the estimates of nucleotide heterozygosity, we were able to conclude that proximal breakpoint regions have lower recombination rates than distal breakpoint regions. This conclusion is consistent with the theoretical predictions made by Navarro *et al.* (1997 &2000). Whether it is the size or the age of the PP inversion, we cannot say for sure, but the PP inversion shares more genetic information with all other inversions than any single arrangement. This observation is consistent with what is depicted in the phylogeny. Our data also supports a recent population expansion of *D. pseudoobscura* and evidence of positive selection acting on this species indicated by the negative Tajima's D values obtained for the regions studied. We were not able to provide support for the predictions made by Navarro *et al.* (2000) of low recombination rates near the breakpoints of new inversions and high recombination rates within the center of the inverted segment. We proposed that gene conversion may affect our results since this process could move genetic information across the arrangements. However, our analysis was repeated with the elimination of the sequences involved in gene conversion events but we still were not able to provide evidence to support the previous predictions obtained from computer simulations (NAVARRO *et al.* 2000). Though our analyses produced strong data, our studies may have benefited from more sequences

from the SC and TL arrangements. The HY, SC and TL are proposed to be three of the oldest arrangements and so they would provide deep insight as to how the age of an inversion factors into determining patterns of nucleotide variability. Santa Cruz and Tree Line are rare in the southwestern United States cline. All experiments in this field could benefit from sequence data obtained from the Hypothetical arrangement but unfortunately this strain has still not been collected from the wild.

VITA

Andre G. Wallace

Education:

- Ph.D. Biology (**2010**)
  The Pennsylvania State University (Penn State), University Park, PA, 16802.
- B.S. Biology (**2005**)
  Medgar Evers College, The City University of New York, Brooklyn, NY, 11225.

Training and Research Experience:

- Graduate Assistant, The Pennsylvania State University (**2005-2010**)
- NIH Post-Baccalaureate fellow, Emory University, Atlanta, GA (**1/05-7/05)**
- The New York City Louis Stokes Alliance Research Program, CUNY, NY (**2003-2004**)
- NIH Special Training, National Institutes of Health, Bethesda, MD (**2003**)

Laboratory Skills/Techniques:

- Cell Culture • Calcium Imaging • Flow Cytometry • Molecular Cloning • Protein-protein interaction assays • Protein localization and visualization • Fluorescent Imaging • Genomic Sequencing and Analysis

Teaching Experience:

- Penn State Teaching Assistant for Molecular Biology of the Cell (**2008 and 2009**)

Manuscripts Being Prepared for Publication:

- Evolutionary Genomics of *Drosophila pseudoobscura* Breakpoints. (**Submitted to Molecular Biology and Evolution**)
- Population Genomics of *Drosophila pseudoobscura* Gene Arrangements. (**To be submitted to Genetics**)

Honors and Awards:

- Alfred P. Sloan Fellow (**2007-Present**)
- Penn State Braddock Scholarship for Exceptional Science Students (**2005-2007**)
- Eberly College of Science Bunton Waller Fellowship (**2005-2006**)

Professional and Community Service:

- Judged poster competition at Penn State's Graduate Research Exhibition. (**2009**)
- Served as judge for oral and poster presentations at the Pennsylvania Junior Academy of Science (PJAS) research competition. (**2008-2010**)
- Volunteered for Penn State's Biology Department recruitment weekend. (**2006-2009**)
- Served on the selection committee for PJAS college admission scholarships. **(2007)**

Activities and Affiliations:

- Abstract accepted for poster presentation at the annual ASCB meeting (**December 2010**)
- Poster presentation at the Drosophila Research Conference (**2010**)
- Poster presentation at Penn State's Graduate Research Exhibition **(2010)**
- Oral presentation during Penn State's Sloan Brown Bag series (**2010**)
- Member of the American Society for Cell Biology (**current**)
- Member of the Genetics Society of America (**current**)